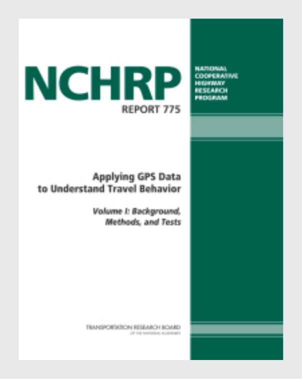


NCHRP REPORT 775

NATIONAL COOPERATIVE HIGHWAY RESEARCH PROGRAM

Applying GPS Data to Understand Travel Behavior

Volume I: Background, Methods, and Tests

TRANSPORTATION RESEARCH BOARD OF THE NATIONAL ACADEMIES

Applying GPS Data to Understand Travel Behavior, Volume I: Background, Methods, and Tests

DETAILS

216 pages | 8.5 x 11 | PAPERBACK
ISBN 978-0-309-28402-8 | DOI 10.17226/22370

CONTRIBUTORS

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 I: Background, Methods, and Tests

Jean Wolf
WESTAT | GEOSTATS SERVICES
Atlanta, GA

William Bachman
WESTAT | GEOSTATS SERVICES
Atlanta, GA

Marcelo Simas 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

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*

FOREWORD

By Nanda Srinivasan
Staff Officer
Transportation Research Board

This report provides guidelines on the use of multiple sources of GPS data to understand travel behavior and activity. The guidelines are intended to provide a jump-start for processing GPS data for travel behavior purposes and provide key information elements that practitioners should consider when using GPS data. The report will be of interest to transportation planners, travel modelers, and travel survey practitioners.

With the high costs associated with primary data collection, methods to improve the use and accessibility of newer sources of data such as Global Positioning System (GPS) data can benefit many transportation practitioners. GPS data can have multiple uses beyond traditional applications such as estimates of speed and travel times. GPS-related data that have been collected from automatic vehicle location systems, from highway sensors, as supplemental information to traditional travel surveys, and via passive technologies [e.g., Bluetooth, radio frequency identification (RFID), and smartphones] have shown promise for additional planning purposes. Some challenges to increased use of GPS data include addressing data bias; balancing precision, coverage, and confidentiality; resolving institutional issues such as data ownership; and addressing the complexity of combining these data with other sources to discern behavioral relationships. While it has been generally accepted that GPS data have a wide variety of uses, research was needed to assist in their use by transportation planners, travel modelers, and travel survey practitioners.

The research under NCHRP Project 8-89 was performed by Jean Wolf, William Bachman, and Marcelo Simas Oliveira of Westat | GeoStats Services, Atlanta, Georgia, in association with Joshua Auld and Kouros Mohammadian from University of Illinois, Chicago, Peter Vovsha from Parsons Brinckerhoff, Inc., and Johanna Zmud from RAND Corporation. Information was gathered via literature review and from interviews with practitioners, data providers, and researchers. The next stage of research explored a number of analytical approaches for extracting information from traces of GPS data. Only a few of those methods could be easily translated into clear and defensible methods (or standards) for processing GPS travel trace data. The research team selected the most promising and valuable analytical procedures for testing and evaluation within the scope of this research effort and applied these methods using datasets available from several GPS-enhanced travel surveys conducted within the past decade.

The report is structured in two volumes. Volume 1 presents the methods used and results of tests conducted. Volume II translates the results of the tests conducted into guidelines for planners and researchers to implement these procedures.

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.

PREFACE

NCHRP Report 775: Applying GPS Data to Understand Travel Behavior spans two volumes. Volume I includes the detailed literature review and industry assessment, an identification of best methods, a plan to perform some tests of these methods, and the results of these tests. Volume II is a guidance document that highlights key information for practitioners who are interested in using Global Positioning System (GPS) data for evaluating travel behavior.

Since the inception of the GPS in the mid-1990s, potential applications of GPS data in transportation planning have been explored broadly and extensively, and the specific use of GPS data for better understanding of travel behavior has held a special interest for transportation modelers and planners. Applications of these highly accurate location and spatial data have been implemented in a wide range of transportation uses, including asset management and tracking, congestion management, and household travel surveys.

However, the full potential of the GPS and similar location tracking technologies for providing travel behavior details is still unknown. Hardware, software, and processing techniques continue to evolve and will likely do so for the foreseeable future. For practitioners and researchers, specific questions still remain regarding the best methods for processing and using these data as a source for travel behavior analyses. Additionally, the availability of new data products that are based on archived consumer product trace data has greatly increased awareness of the need for guidance about these products so that practitioners can develop fiscally responsible and theoretically sound data collection programs. NCHRP Project 8-89 addressed these challenges by documenting the state of practice in 2012 and exploring techniques to extract travel details inherent in simple GPS-based trace data sets.

There were two major challenges encountered and addressed while conducting this research. First, the increasing availability and promise of consumer product trace data has made practitioners anxious to evaluate these new population-based data sources as a replacement for conducting stratified travel surveys as well as other origin–destination studies. While certainly appealing, these data sources are processed and aggregated by private companies that maintain proprietary data management methods and are unwilling to share techniques and algorithms. Furthermore, these private firms are bound to protect the privacy of their data sources; consequently, they are unwilling or unable to release data at the individual data-source level. To address this challenge, the NCHRP Project 8-89 research team interviewed the primary data providers (who provided marketing materials), reviewed published studies that attempted to evaluate these data products, and then summarized key findings. Volume II contains descriptions of standard data product structures and formats and suggests key information elements that users should consider when evaluating these data products.

The second major challenge was limiting the full range of analytic options to include for detailed analysis and documentation of methods. While the research community has explored a number of analytical approaches for extracting information from traces of GPS data, there are few examples that can be easily translated into clear and defensible methods (or standards) for processing GPS travel trace data. The research team selected the most promising and valuable analytical procedures for testing and evaluation within the scope of this research effort. The methods used and results of these tests are described in Volume I. Volume II translates the results of these tests into guidelines for planners and researchers wishing to implement these procedures.

The research team believes that the two volumes that make up this report will provide insight and instruction to the transportation community with respect to past and present uses of GPS data (from a range of sources) for travel behavior analysis, as well as sound guidance on processing GPS data to better understand travel behavior.

CHAPTER 1

Literature Review and Industry Assessment

Transportation professionals have been enamored with the potential uses of Global Positioning System (GPS) data ever since GPS became fully operational in 1995. Early GPS-enhanced household travel surveys, such as the 1996 FHWA Lexington Pilot Study and the 1997 Austin Household Travel Survey, have led the way in evaluating GPS use in travel surveys (Battelle Memorial Institute 1997; Murakami and Wagner 1999; Casas and Arce 1999). These initial studies were hindered by the U.S. government's intentional degradation of GPS's positional accuracy (known as selective availability). Selective availability was eliminated in early 2000, thereby accelerating the rapid development and implementation of a wide range of commercial, consumer-oriented, location-based services (LBS) and supporting GPS devices.

Over the past 11 years, more than 25 household travel surveys (HTSs) conducted within the U.S. have used GPS augments to help assess the level, breadth, and magnitude of travel underreporting or misreporting by the large diary-based reporting sample. And, with each survey, GPS sample sizes have steadily increased, with some of the most recent surveys involving the deployment of GPS data loggers to thousands of households, either with large subsamples (e.g., New York City, Atlanta, and California) or with the entire surveyed population (e.g., Cincinnati, Cleveland, and Jerusalem). Over this same time frame, consumer-based GPS products, such as stand-alone or in-dash personal navigation devices (PNDs), GPS-enabled smartphones, and fleet tracking systems [e.g., automatic vehicle location (AVL) systems] have led to the creation of large-scale GPS data sets that can be mined or translated into detailed travel behavior information. In addition, other fixed-location approaches to tracking personal travel, such as those supported by Bluetooth, radio frequency identification (RFID), and mobile phone tower technologies, offer alternative methods for providing some level of travel behavior information.

The combination of these large-scale GPS travel survey data collection events, the increasing availability of large consumer-based GPS data sets, and ongoing studies evaluating the use and benefits of fixed-location sensors have led to many discussions within the transportation community about the roles, advantages, and disadvantages of various GPS data sources for transportation planning and modeling, as well as for other travel behavior research initiatives. Given the need for more data to support a wide range of transportation planning and modeling activities, combined with ongoing budgetary constraints, the time has come to clearly and objectively evaluate the multiple sources of GPS data that could be leveraged and used for transportation planning beyond the traditional application area of travel time and speed studies.

Overview of Literature Review and Industry Interview Process

The increasing availability of travel data collected from location-aware technology, such GPS devices, combined with the availability of open application programming interfaces (APIs) and open-source software (OSS), has peaked interest in the application of GPS data for use in travel forecasting, planning analysis, and transportation system management. Frequently, however, the initial attraction by public agencies to these detailed travel data has met with roadblocks related to cost, challenges with integration into existing modeling paradigms, concerns about data privacy, sample bias, and data management difficulties. GPS and other tracking technologies can provide a depth of insight into travel behavior and activity patterns that exceeds traditional modeling data needs (such as trip rates and travel times) and that complements standard system performance metrics (e.g., average speed and congestion identification). Realization of these potential benefits will require an objective assessment of these various data sources along with guidance to assist transportation data users in decision making and data management.

In 2011, the Transportation Research Board initiated a study to evaluate these GPS data sources and to provide guidance on the use of these sources by transportation planners, travel modelers, and travel survey practitioners; this study is NCHRP Project 8-89, "Applying GPS Data to Understand Travel Behavior." This chapter reports on a broad literature review conducted on GPS data sources, actual and potential uses of these data in the field of transportation, standards for GPS data collection and storage, and concerns about the various sources with respect to coverage, bias, accuracy, and privacy.

To supplement this literature review, comprehensive questionnaires were sent to industry experts in the areas of travel surveys, travel behavior research, travel demand modeling, and traffic data provision. The responses from these questionnaires provided additional direction to the literature review process and have been summarized at the end of this chapter to provide both state-of-the-practice and state-of-the-art confirmation of industry uses and plans for GPS data. Hereafter, information gathered from these industry responses are referred to as the 2012 Industry Survey.

For the purpose of this report, it is important to clarify the scope of this research initiative. The prioritized GPS data sources evaluated were:

1. GPS data loggers and GPS-equipped smartphones deployed to households recruited within a household travel survey;
2. Passive GPS or cell phone data collected from devices purchased by consumers, such as mobile phones and PNDs;
3. Other GPS and location-based data sources that have been used for understanding various aspects of travel behavior (i.e., transit surveys, transit AVL data, private fleet tracking systems, and probe vehicle studies).

In addition, other fixed-location sensors, such as mobile phone towers, RFID readers, Bluetooth sensors, and Wi-Fi sensors, that have known locations and can detect when relevant devices pass by can also provide useful information about transportation system performance as well as travel behavior. Although not GPS technology as defined in the previous list, these technologies are also discussed in this chapter.

GPS-Based Travel Behavior Data Collection and Uses

The following subsections discuss the use of GPS technology to enhance HTSs, provide notable examples of GPS-augmented HTSs, present ways in which GPS data have been used to improve travel demand models (TDMs), and conclude by identifying other travel behavior study types that have benefited from GPS technology.

Overview of GPS-Enhanced Travel Surveys

This section presents an overview of the use of GPS in travel surveys and its evolution over the past two decades. It concludes with a discussion on the emerging use of smartphones as GPS data collection alternatives for these surveys.

As data needs for developing TDMs have increased and survey participation rates have generally fallen over the past several decades, more sophisticated methods of data collection have been developed by the travel survey community in an effort to address these problems. There was a shift first from traditional travel diaries to activity diaries accompanied with major advances in survey techniques that were generally driven by increases in computing power, portability, and availability along with decreases in cost.

The evolution of travel survey methods has continued with the introduction of GPS-enhanced travel survey techniques. The use of GPS data collection has been found to have many advantages over traditional survey methods. First, GPS-enhanced surveys provide a more accurate and detailed account of the spatial and temporal aspects of personal travel than what survey respondents are able to recall and report, and GPS data sets have been used to correct significant trip underreporting errors associated with pen-and-paper or phone-based activity surveys (Battelle Memorial Institute 1997; Wolf, Bricka, et al. 2004). GPS-enhanced surveys should have less respondent burden for capturing travel details by leveraging passive GPS data collection while collecting more information and more accurate information. In addition, by further reducing respondent burden through the use of automated activity type, location, timing, and travel mode identification routines, GPS-based prompted-recall surveys allow for more complex questions to be asked.

The latest generation of GPS-based surveys includes GPS-only studies, in which basic household information is collected first and then GPS data loggers are used by study participants, with software algorithms and models used to generate all necessary details of travel. Finally, the combination of more accurate spatial–temporal data along with reduced respondent burden allows for multiday data collection, which in turn enables more in-depth aspects of travel behavior to be studied, including variability in travel patterns, route choice, activity location selection, and mode selection. Furthermore, multiday data collection can support reductions in required sample sizes, thereby offsetting some, if not all, of the additional costs inherent in GPS-enhanced and GPS-based travel surveys (Stopher, Kockelman, et al. 2008).

GPS-Based Subsamples for Travel and Activity Surveys

The use of GPS data in activity and travel surveys is a relatively new practice, made possible through improvements in the technology itself and the demand for more accurate travel data. Initially, GPS data collection was used mostly to provide corrections for trip rates obtained from traditional household travel surveys or to demonstrate the feasibility of doing so. These studies tended to be conducted in conjunction with traditional diary-based household travel surveys. GPS-enhanced surveys of this type primarily used passive GPS data collection systems, where the GPS traces were collected and analyzed without any input from the participants, with a few studies using more active or interactive systems that employed a combination of technologies such as an onboard computer or handheld device combined with a GPS receiver to gain additional input from the participants (Battelle Memorial Institute 1997, Doherty and Oh 2012, Guensler and Wolf 1999). Some of these studies compared the GPS-identified trips with the diary report trips from the same GPS subsample as a means to correct larger, traditional activity diary samples. This method is also referred to as the dual-method approach because it requires the GPS subsample to use both GPS devices and diaries, which results in increased burden to these participants.

Several examples of GPS-enhanced surveys that have used the dual-method approach are statewide surveys in California (NuStats 2002) and Ohio (Pierce et al. 2003), and regional studies in Austin (Casas and Arce 1999), Laredo (Forrest and Pearson 2005), Kansas City (NuStats 2004), Seattle (Cambridge Systematics 2007), Chicago [Chicago Metropolitan Agency for Planning (CMAP) 2008], and Denver (NuStats 2010). Similar dual-method studies have recently been completed in California (performed by NuStats and GeoStats) as well as in regional surveys conducted for Philadelphia and Los Angeles by Abt SRBI (2012 Industry Survey). The primary intent of the GPS component in each of these studies was to develop trip rate correction factors. Figure 1-1 shows GPS data collected during the California HTS pilot study conducted in 2011.

Additional analysis was performed using GPS data collected in the 2001 California statewide (Wolf, Oliveira, and Thompson 2003; Zmud and Wolf 2003), Ohio statewide (Pierce et al. 2003), Kansas City (Wolf, Bricka, et al. 2004; Bricka and Bhat 2007), and Denver (Bachman et al. 2012) surveys, among others, to gain insight into the underreporting phenomenon. Most recently, a study by Bricka and Murakami (2012) used a combined GPS and diary sample from Indianapolis to not only evaluate trip underreporting in traditional surveys, but also to test potential trip reporting errors with the use of GPS-only samples.

These survey and research efforts have led to a large body of knowledge about trip underreporting in household travel surveys as well as the methods for identifying and correcting this problem. The use of a GPS subsample within a larger traditional travel survey for correction factors continues to be an important way for this technology to support travel demand modeling needs. In the 2012 Industry Survey of market research firms that specialize in travel surveys (conducted as part of this research effort), most respondents reported either recently using or continuing to use GPS samples to correct self-reported trip rates (2012 Industry Survey).

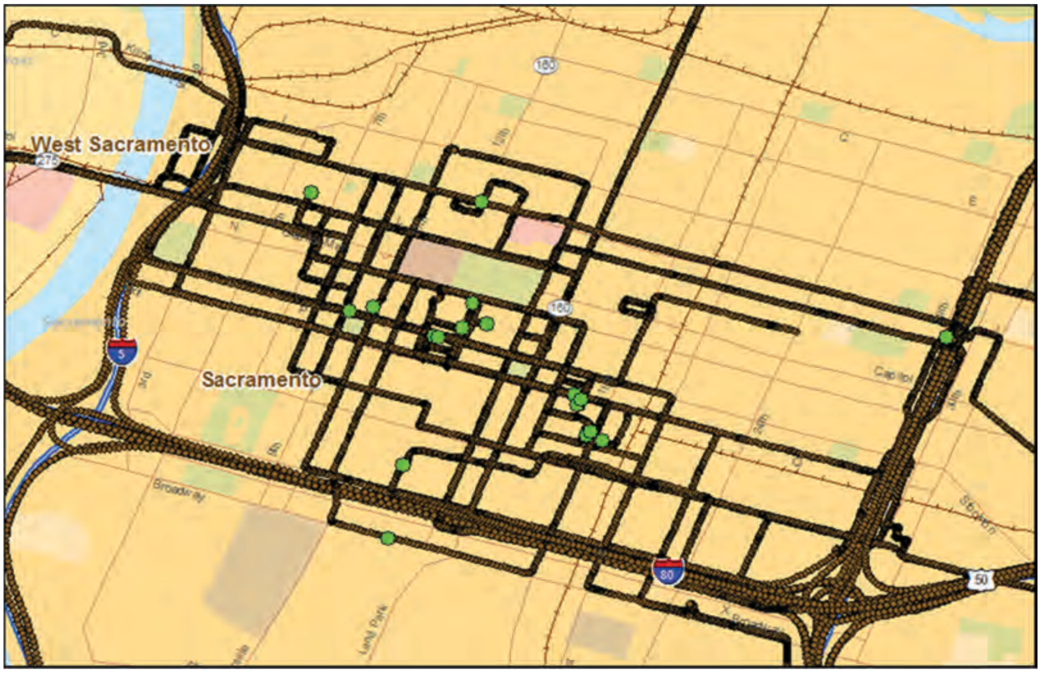


Figure 1-1. Example of GPS data collected during 2011 California HTS pilot study.

A Move Toward the Replacement of Travel Surveys with GPS

Although the first significant use of GPS in travel surveys was to measure and correct for trip underreporting, it has long been thought that GPS-based surveys could someday completely replace the travel reporting component of household travel surveys (Wolf 2000). The expectation has been that a completely GPS-based survey would significantly lower the respondent burden while increasing the quality and quantity of information captured, specifically in the automatic collection of trips and their attributes, including trip start and end times, activity locations and durations, and route choices (Wolf 2000; Murakami, Morris, and Arce 2003). Accurate travel reporting has traditionally been a challenge for survey respondents due to limitations in memory recall, the tendency to filter out what is considered by the participant as either unimportant (i.e., ATM visit or convenience store stop) or confidential, and the inherent complexities of trip reporting methods.

Furthermore, there has been interest from travel demand modelers to extend the reporting period of traditional travel surveys beyond a single day to better measure the variability in day-to-day travel. A few travel surveys conducted outside of the United States have done this; for example, the Mobidrive survey conducted in Germany collected travel information for 6 weeks (Axhausen et al. 2002). However, in the United States there have been few travel surveys that have attempted to collect even 2 days of travel data due to a significant decline in participation rates and trip rates attributed to higher respondent burden (Chicago Metropolitan Agency for Planning 2008; Bricka 2008). Consequently, reducing respondent burden is critical to recruiting and retaining a good, representative sample of the targeted population—and even more so if multiday travel information is desired. It is worth noting that most of the recent GPS-enhanced travel surveys conducted in the United States have collected multiday GPS data ranging from 2 to 7 consecutive days.

During the industry interviews conducted for this project, a leading researcher from the Institute of Transportation and Logistics Studies (ITLS) touched on many of these aspects in his industry survey response, "Accuracy [of GPS] is clearly far greater than in diaries. People are notoriously bad at estimating the times at which they travel, how long they travel, and certainly how far they travel. . . . A huge advantage is the ability to collect multiday data as well as the accuracy and coverage already described. We believe that personal passive GPS loggers reduce respondent burden substantially" (2012 Industry Survey).

With the relative ease and accuracy of collecting travel data through GPS logging established by early studies, subsequent research has looked into using processing techniques on the collected GPS data to completely replace the traditional travel and activity diaries and associated retrieval methods. Efforts in this area have been conducted along two main lines: (1) processing the GPS data into basic trips and attributes and then having the participants confirm, complete, and/or correct these data through a GPS-based prompted-recall interview, and (2) using a GPS-based prompted-recall subsample to calibrate models that are then used to impute details on completely passive GPS data collected by the majority of the sample without further input from survey participants. The following sections discuss these two approaches in more detail.

GPS-Based Prompted-Recall Travel Surveys

The use of passive GPS logging coupled with a follow-up survey that is based on the trips identified within the GPS data is usually referred to as a GPS-based prompted-recall (PR) survey. This is because the GPS data are used to reconstruct the activity-travel pattern of the respondent, with the detected trips and trip attributes then presented to the participant, who is prompted for further responses. This mode combines the automated data processing routines of purely passive surveys with respondent verification of the auto-generated data and, usually, the collection of additional data that may be difficult to extract from GPS traces alone (i.e., trip purpose, vehicle occupancy, parking cost, etc.). Several different types of prompted-recall surveys have been conducted, including both vehicle-based and person-based, that use various strategies for prompting the individual recall of travel patterns. The primary advantages of this survey mode are the collection of detailed information about aspects of travel and activity participation that cannot be automatically deduced (Auld et al. 2009) and the reduced respondent burden during actual travel, which is limited to carrying the device, something that most respondents do not seem bothered by (Lawson, Chen, and Gong 2010).

An early example of a GPS-based prompted-recall survey was implemented by Bachu, Dudala, and Kothuri (2001), who used vehicle-based GPS data to track a sample of 10 households over several days. The results of this study showed that the survey participants could recall the details of trips identified in the GPS traces several days after initial data collection with little loss of recall ability. A small pilot study was also conducted by Stopher, Bullock, and Horst (2002) using prompted-recall survey methods with a similar process of auto-identifying activity-travel episodes with manual adjustment. In this study the travel patterns were shown in maps and in a sequential tabular format, with unknown attributes (such as purpose, travel companions, and costs incurred) left blank for the respondents to fill in. This survey also had the respondents correct the generated travel patterns. A similar mail-out PR follow-up study was conducted for a portion of the Kansas City GPS subsample, in which respondents were prompted to fill in details of GPS-identified trips that were not recorded in the standard diary the respondents filled out or mentioned by the household during the travel reporting interview (NuStats 2004).

As mentioned previously, most of the early prompted-recall studies involved creating maps or other displays, then mailing these to the respondents for completion, which could involve significant delays and, therefore, a potential degradation of recall ability. More recently, GPS-based prompted-recall surveys have been implemented using web-based data collection platforms. The use of web-based prompted recall allows much more detailed information regarding travel behavior to be collected. Yasuo Asakura of Kobe University states, "The combination of GPS [and] web has made [it] possible to obtain whole travel behavior data that were not observed only by the GPS," (2012 Industry Survey). Examples include the collection of detailed travel planning behavior (Auld et al. 2009) and activity rescheduling strategies (Clark and Doherty 2010), among others. Studies by Marca (2002), Stopher and Collins (2005), Lee-Gosselin, Doherty, and Papinski (2006), Li and Shalaby (2008), and Auld et al. (2009) were performed using PR surveys over the Internet, and web-based computer-assisted self-interviews (CASIs). A computer-assisted telephone interview (CATI) GPS prompted-recall component was added to the recent household travel survey for the New York Metropolitan Transportation Council, which also used a web-based CASI PR component (Chiao et al. 2011; Wilhelm, Wolf, and Oliviera 2012).

In each of these PR studies, members of recruited household wore GPS data loggers for one or more days, and the data were later transferred to a central server for processing, either by direct uploading of the data removed from the device after the survey was complete, as in the surveys by Stopher and Collins (2005); Auld et al. (2009); Wilhelm, Wolf, and Oliviera (2012); and Oliveira et al. (2011), or through continuous wireless communication as in Lee-Gosselin, Doherty, and Papinski (2006). Regardless of the data transfer method, the collected raw points were then processed to identify the activities and trips, and the recall survey was built upon the identified activity-travel episodes.

Another variation of a GPS-based prompted-recall survey was implemented in Jerusalem in 2010–2011, where the regional planning agency used an internal team to conduct a 100% GPS-based travel survey. They used laptops to administer face-to-face interviews using commercial off-the-shelf (COTS) computer-assisted personal interview (CAPI) software that was integrated with a custom GPS prompted-recall tool developed by GeoStats. This approach was used to carry out both the initial recruitment and subsequent GPS-based prompted-recall interviews (Oliveira et al. 2011). This survey collected detailed GPS-based prompted-recall travel data from more than 8,800 households located within the Jerusalem region and was remarkable in its aggressiveness with respect to technology adoption by the planning agency and acceptance by diverse population groups within the region. Furthermore, the survey platform made it possible to conduct PR interviews immediately following GPS data downloads to the laptops in the participants' homes, with no opportunity for interviewer preprocessing or cleaning of the GPS trip data prior to participant review.

The Ohio Approach to GPS-Based Household Travel Surveys

Since 2009, the Ohio Department of Transportation has initiated two large-scale GPS-based household travel surveys within the state, which advanced the current state of practice in GPS-based travel surveys. The first survey, derived from the work previously implemented by Stopher and Collins (2005), was conducted in the Cincinnati area where a completely GPS-based household travel survey was performed. The travel survey included over 2,000 households, of whom 601 completed a 1-day prompted-recall study (Stopher et al. 2012). The study found that GPS-only travel surveys were feasible, although the authors state that prompted-recall data had limited usefulness because "it was also quite clear from the results of the prompted recall that it does not provide 'ground truth,' because people still misunderstand what is required and misremember what they did" (Stopher et al. 2012).

The second GPS-based travel survey in Ohio was conducted in Cleveland by GeoStats (2012 Industry Survey). In this study, more than 4,000 households provided travel information using GPS data loggers, with approximately 1,300 of these households participating in a GPS-based prompted-recall interview using CASI or CATI survey methods. The purpose of the prompted-recall sample was to assist in the calibration and validation of the algorithms and models developed for imputing critical trip attributes such as travel mode, companions, and trip purpose for the remaining GPS-only sample. Another 453 households composed entirely of persons over the age of 75 reported their travel using travel logs (which seems simpler and more appropriate for this demographic group), yielding a final overall sample size of 4,545 households.

Smartphone Use in Household Travel Surveys

According to the Pew Research Center (Smith 2012), 46% of adults in the United States own a smartphone, with almost three-quarters (74%) of them getting real-time location-based information on their smartphones. These statistics are particularly impressive given that the two most popular smartphone platforms did not exist until the middle of the first decade of the 2000s.

In addition to being capable of running custom software applications (commonly referred to as "apps"), most smartphones integrate multiple technologies, such as keyboard and voice inputs, GPS, accelerometers, gyroscopes, and cameras, most of which are applicable to conducting travel surveys. Modern platforms make the use of these imbedded technologies available to software developers, with GPS-based location-referencing services becoming one of the most popular features in smartphones. For example, there are several free apps available for the most popular platforms that allow users to record GPS locations as frequently as one point per second.

The opportunity to leverage smartphones is appealing to travel behavior researchers, and these devices are quickly becoming another method in the travel survey toolbox for collecting GPS-based travel data. According to Murakami and Bricka (2012), the possibility of using devices owned by participants can address common implementation challenges in GPS-based travel surveys by: (a) eliminating the need to ship out and retrieve GPS loggers, (b) shortening the time between travel date collection and data review, and (c) reducing costs associated with equipment loss. When combined, the growing market penetration and technical capabilities of smartphones makes them an attractive medium for conducting travel surveys. This has sparked a growing interest in the travel survey community, with several pilot studies having been conducted over the past few years (Bricka and Murakami 2012).

Smartphones have the ability to be used in travel surveys as either active or passive data collection devices. In the active scenario, respondents would use the phone app to respond to survey questions before (i.e., recruitment questions) and during their travel day, either by confirming stops or explicitly starting and stopping the recording of GPS traces. Early deployments of smartphones to collect GPS data within travel surveys have used this approach, with notable examples being the TRAC-IT research project (Center for Urban Transportation Research, University of South Florida 2012) and the PTV Pacelogger app (Bricka and Murakami 2012).

Passive use of smartphone technology requires participants to download and initialize the app and identify themselves within the household persons roster; from that point on, all recording takes place automatically in the background, with the app detecting when the monitoring period ended and transmitting the captured data for processing. This passive data collection scenario can also be complemented by a PR interview completed in the same app or via the web. Relevant examples of this approach are the Quantifiable Traveler app developed by UC Berkley and the Future Mobility Survey conducted in Singapore (Murakami and Bricka 2012).

However, there are still a few technological and methodological challenges to overcome before smartphone solutions can become dominant in household travel surveys; these challenges include:

1. Market fragmentation,
2. Power management,
3. Data plans and associated costs, and
4. Self-selection and capture mode biases.

The first issue is that there are several active smartphone platforms in the United States, each with multiple versions in active use and varying levels of API and technology support. Table 1-1 shows the breakdown of the top five smartphone platforms in the United States and includes the number of active different versions for each.

This fragmented reality makes it difficult and costly to develop apps for multiple platforms and operating system (OS) versions to support the majority of participants with smartphones. For example, the iPhone 3G running iOS versions older than 4 does not have the ability to log GPS data in the background while the phone is performing other activities, such as running a different app or during a call. There are also significant API differences between platforms that make it challenging to offer the same features and logic consistently across all apps offered to all participants.

The second issue that needs to be addressed is that of power management when logging trace data, be it from GPS or other sources such as Wi-Fi. A viable travel survey app has to allow a participant to make normal use of his or her smartphone while capturing the required data. Incremental improvements in battery technology, central processor units (CPUs), and GPS chipsets have alleviated this limitation, but it is still the case that continuously logging GPS data will rapidly deplete a smartphone battery. Researchers have dealt with this limitation through various strategies such as providing external power sources (Doherty and Oh 2012), limiting the use of GPS and relying mostly on Wi-Fi and cell tower data for positioning combined with algorithms for identifying when to start and stop logging (see Quantifiable Traveler

Table 1-1. Top U.S. smartphone platforms.

Platform	February 2012	Number of Major OS Versions*
Android (Google)	50.1%	8
iOS (Apple)	30.2%	6
Blackberry (RIM)	13.4%	4
Windows (Microsoft)	3.9%	3
Symbian (Nokia)	1.5%	4
Others	0.9%	N/A

*Only includes versions released since 2007.
Source: comScore MobiLens from http://www.comscore.com/Press_Events/Press_Releases/2012/4/comScore_Reports_February_2012_U.S._Mobile_Subscriber_Market_Share, accessed on 08/31/2012.

and Future Mobility Survey in Chapter 26 of the *Travel Survey Manual Update* by Murakami and Bricka and also Battelle Memorial Institute 2012), and providing direct control over the logging to participants (Center for Urban Transportation Research, University of South Florida 2012).

The third issue is related to the need to transmit and download data from the app and the fact that participants may have limitations on their data plans. Even after applying data compression, high-resolution GPS traces can get fairly large, and transmitting them back for processing could have considerable cost impacts on an unknowing participant. This can be alleviated by providing materials that explain the expected data transfer demands of the application up front, applying trace simplification algorithms such as SQUISH (Muckell et al. 2011), providing incentives that will offset data transmission costs, or only transmitting minimal information back to a central location (Center for Urban Transportation Research, University of South Florida 2012).

The fourth challenge has to do with the fact that travel surveys are typically conducted at the household level and that not all adult members will have a compatible smartphone. This means that passive GPS data loggers will still need to be shipped to households even if there is a smartphone owner/user in the household. Of course, households without any smartphone will require one or more passive GPS loggers as well. These mixed GPS methods could be confusing for a survey household.

Finally, as seen with other data collection methods and technologies, there are multiple biases (e.g., age, gender, income, and ethnicity) related to smartphones. To mitigate these biases, it is important to provide alternative means of participating and to ensure that the data collected, regardless of survey mode, is properly integrated into the overall survey platform and framework. Developing a comprehensive system to support and integrate multiple survey modes across and within households is not a trivial task, and the costs to develop, maintain, and update this system and all components, as well as to provide technical support to participants, will be incurred on an ongoing basis.

Despite all of these challenges, it should be noted that the widespread availability and use of smartphones are relatively recent phenomena, and the technology as well as its uses are likely to continue changing and evolving at a fast pace over the next several years.

Examples of GPS-Enhanced Household Travel Surveys

This section describes the data requirements of three recent household travel surveys: the 2012–2013 California Household Travel Survey (CHTS), the 2011 Atlanta Regional Travel Survey, and the 2012–2013 Northeast Ohio Regional Travel Survey (covering the greater Cleveland region).

The relevant travel demand model's data requirements have a large influence on the data elements collected in a household travel survey. Simple four-step TDMs may require basic household, person, vehicle, and trip-level information, whereas advanced activity-based models require more precise details about household, person, and vehicle characteristics, as well as expanded information about actual travel behavior, travel options, and costs.

These three surveys have been selected as representative examples of an expansive survey design (whose requirements were driven by many state agency and regional agency stakeholders), a typical survey design (in which the requirements were driven by a regional metropolitan planning organization (MPO) that has both a four-step TDM and an activity-based TDM), and a GPS-/technology-driven design (that was intentionally defined to require the minimum data elements needed to support the current four-step regional TDM), respectively. The purpose of this comparison is to show the range of current data requirements in household travel surveys to be considered when evaluating the reduced respondent burden associated with GPS-based travel surveys as well as when evaluating other data sources to replace travel surveys.

The California Statewide Household Travel Survey—An Expansive Survey

The California Department of Transportation (Caltrans) sponsors the decennial CHTS, the most recent of which began survey data collection in February 2012 and was completed in January 2013. This statewide survey was designed to support the statewide travel demand model. Additionally, an attempt was made to accommodate regional travel demand models by including representatives from the MPOs and councils of governments from across the state in the planning and design process. Other California state agencies, such as the California Energy Commission, were also active participants in the design of the survey to meet their own agency data needs. By trying to accommodate the data needs of this wide range of users, the CHTS was a significantly longer and more comprehensive survey than typical household travel surveys. The design of the survey also included a long-distance trip diary in addition to the regular single-day travel diary; the purpose of the long-distance trip diary was to collect information about inter-regional travel within the state that is not captured in a typical 1-day survey. The final sample size for the full survey was approximately 42,500 households, and 5,717 of these households also participated in the GPS component. The survey used a dual-method approach (participants receive both diaries and GPS devices) with three GPS subsamples: diary and wearable GPS, diary and vehicle GPS, and diary and vehicle GPS supplemented with an onboard diagnostic (OBD) device (or engine sensor).

The 2011 Atlanta Regional Travel Survey—A Typical Travel Survey

The Atlanta Regional Commission (ARC) conducted its most recent regional travel survey in 2011; this survey had a targeted sample size of 10,000 households with a subset of 1,000 GPS households (PTV NuStats 2011). Recruitment methods offered to participants were telephone (CATI) or web (CASI) interviews, and retrieval methods included CATI, CASI, and diary mail back with data entry into the web-based retrieval system. The purpose of the 10% GPS subsample was to collect detailed information about all trips made to estimate levels of trip underreporting that could be applied to the larger, non-GPS sample. Consequently, the dual GPS and diary method was implemented. A split design was also recommended, with the objective being to obtain 667 complete households with in-vehicle GPS data and the remaining 333 complete households with wearable GPS data. The GPS devices were used for 7 days by the vehicle sample and 3 days by the wearable sample, with the first day coinciding with the assigned diary/travel day.

This split technology design allowed for the collection of 7 days of highly accurate vehicle-based data with minimal respondent burden while focusing the use of the wearable GPS device to those households that reported some incidence of transit use for a work or school commute. Households selected for the wearable GPS component were deployed for 3 days, with all household members between the ages of 16 and 75 receiving GPS equipment. A $25 incentive per instrumented vehicle or person was offered to all recruited GPS households for successfully reporting travel data, using all GPS devices provided, and for returning all devices. The final data sets for the survey contained 10,278 completed households and 1,061 completed GPS households.

The Northeast Ohio Regional Travel Survey—A State-of-the-Art, GPS-Only Survey

The Northeast Ohio Regional Travel Survey, covering the Cleveland metropolitan area, was one of only three large-scale travel surveys to use GPS for nearly 100% of the participating households. (The other two were a smaller 2,583 household study conducted in Cincinnati and an 8,800 household GPS-based travel survey conducted in Jerusalem.) The Cleveland survey collected detailed socio-demographic and travel data from 4,545 households, including a 30% subset who participated in a GPS-based prompted-recall interview designed to confirm trip details via CATI or CASI survey methods. Trip details for the remaining GPS-only sample were imputed based on land use data, geocoded addresses, GPS data characteristics, and information collected during the recruitment interview.

This means that the majority of households in the study completed a recruitment interview, wore a GPS device for 3 or 4 days, completed a record of usage, and sent the device(s) back, thereby concluding their participation. The smaller percentage of the sample used GPS to record their travel while recording a few basic details of the trips made on their assigned travel day to reference during their retrieval interview.

Given the use of GPS as a primary means of data collection and software algorithms for imputing travel details, the survey sponsors (the Northeast Ohio Areawide Coordinating Agency and the Ohio Department of Transportation) agreed that they would also try to minimize the number of data elements required in both the recruitment and retrieval interviews so that only essential variables needed for model development or support were required. Consequently, this survey represents a minimalistic approach to HTS data collection.

Comparison of Data Requirements

Table 1-2 and Table 1-3 provide summaries of the counts of variables for each of the three surveys discussed in this section. The summary includes the count of variables in each of the typical tables provided in the final data set. The two rightmost columns show the difference between the variables collected for the two sample types in Cleveland, with a notable difference in the second table illustrating the reduced burden for the GPS-only participants. As mentioned previously, the number of variables in these TDM data sets is also a reflection of what might be required when trying to use or reuse existing data sets for travel survey purposes. Appendix A contains the complete listing of all variables delivered by table (household,

Table 1-2. Number of delivered variables.

Description	California Statewide	Atlanta	Cleveland PR	Cleveland GPS Only
Household variables	50	38	32	32
Person variables	104	92	93	93
Vehicle variables	29	15	7	7
Location/place/trip/activity variables	53	54	43	43
Long-distance travel	51	0	0	0
Totals	**287**	**199**	**175**	**175**

Table 1-3. Number of questions asked of participants.

Description	California Statewide	Atlanta	Cleveland PR	Cleveland GPS Only
Household variables	25	15	8	8
Person variables	97	85	83	83
Vehicle variables	27	13	5	5
Location/place/trip/activity variables	45	36	34	0
Long-distance travel	47	0	0	0
Totals	**241**	**149**	**130**	**96**

person, vehicle, location/place/trip, and long-distance travel) and by survey.

Use of GPS Travel Data in the Development of Transportation Models

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, Inc., 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. More specifically, GPS-based travel/activity surveys have been implemented with the expressed intent of improving the quality of behavioral data needed for trip, tour, and activity-based models. Data from GPS and consumer technologies are also emerging as a source for identifying baseline network operating conditions and for validating model outputs.

Over the last several years, the primary incentive for regions to invest in a GPS-enhanced travel survey component has been the identification of trip rate correction factors that adjust model trip rates based on unreported travel measured in the GPS subsample for the larger diary-based samples. While GPS survey data have been used for other investigative analyses during model development, the systematic use of these data sets for other purposes is just now beginning to grow. Perhaps most intriguing is the use of passive GPS data collected by survey participants to replace traditional diary-based reporting methods (as discussed in the previous section).

Over the past decade, GPS data have been applied in transportation planning model development to:

- Generate trip rate correction factors,
- Identify activity schedules,
- Explore activity interactions within a household and within larger social networks,
- Identify activity locations,
- Identify route choice and mode choice preferences,
- Explore variability and pattern formation in activity-travel patterns,
- Identify baseline network roads and conditions,
- Evaluate bike/pedestrian travel behavior,
- Validate travel demand models, and
- Identify trip purpose and activity type.

While uses such as trip rate correction factors and the identification of baseline transportation network conditions have been applied in several regional model development efforts to date, other uses are still emerging and are found only in research studies. Table 1-4 lists some of the known uses of GPS data that have been applied in practice.

Table 1-4. Uses of GPS data in transportation model development.

Use of GPS Data	Applied in Practice
Trip rate correction factors	Atlanta, California, St. Louis, Kansas City, Washington, D.C., Chicago, Massachusetts, New York City
Activity schedule development	Jerusalem, New York, Cincinnati
Activity interaction analysis	Jerusalem, Cincinnati, New York
Activity/trip end geocoding	Cincinnati, Jerusalem, Cleveland
Baseline network development	Many (GPS probe vehicle data, consumer data)
Route and/or mode choice analysis	Jerusalem, Zurich, Seattle, San Francisco, Portland
Model calibration/validation	Many (GPS probe vehicle data, consumer data)
Bike/pedestrian models	San Francisco, Monterey Bay
Full travel diary replacement	Cincinnati, Jerusalem, Cleveland

Trip Rate Correction Factors

One of the main attractions of using GPS devices to identify travel behavior is that the data set is observed (passive) instead of reported (active). The basic limitations of reported travel behavior through diaries have been recognized for many years (Stopher and Greaves 2007; Casas and Arce 1999) as respondents frequently forget some trips and interpret the definition of a "trip" differently. When GPS became viable for use as part of a household travel survey, one of the first analyses was the comparison of GPS trips to diary trips. The differences between the observed and reported trips and the need to account for these differences in demand models have justified the inclusion of GPS subsamples in many large-scale travel surveys over the last decade (Wolf, Loechl, et al. 2003; Bradley, Wolf, and Bricka 2005). The extent of underreporting varies by region and demographic profile (Wolf, Loechl, et al. 2003; Bricka and Bhat 2007). A review of five recently completed surveys conducted for Denver, Atlanta, Nashville, Massachusetts, and California revealed overall underreporting levels ranging from 11% to 25%; however, these percentages should neither be interpreted nor applied broadly—additional analyses are needed to generate appropriate, targeted correction factors based on specific trip, tour, and socio-demographic characteristics.

Generating trip rate correction factors can be accomplished for surveys in which households report travel in addition to GPS travel data, and also for surveys in which some households report travel using diaries only and others use GPS with prompted recall only. (The recent 2010–2011 New York City regional travel survey used this latter approach.) The techniques for either situation are similar in that correction factors are generated for subsets of travel. For trip-based models the corrections should be for specific trip types (i.e., home-based work), and for tour-based models the corrections should be for specific tours (i.e., school tours for children).

Activity Schedules and Interactions

Activity-based models (ABMs) tend to require data on the full activity-travel pattern of individuals and such hard-to-collect information as planning times and flexibility measures. ABMs operate with disaggregate individual daily patterns and schedules. From this point of view, it is essential to collect a full-day list of person trips and activities with no gaps, overlaps, or inconsistencies. If one of the trips or activities of the person is missing, miscoded, or underreported, this essentially makes the entire person-day unusable for some of the ABM components. Underreporting in aggregate four-step models can be somewhat improved by applying trip rate correction factors derived, for example, from a 10%–15% subsample of GPS-assisted households (Wolf, Bricka, et al. 2004). This approach is less useful for ABMs since it is impossible to restore the individual disaggregate details that include not only the number of trips by purpose but also their sequence and timing based on a small subsample. This becomes especially important as scheduling models begin to account for intra-household interactions among household members, which require consistent schedules for all household members. Consequently, missing or underreported data for just one household member can invalidate the data for the entire household, increasing the prevalence of unusable data.

One advanced surveying approach that fully addresses this issue and minimizes the underreporting biases is a 100% GPS-assisted prompted-recall method (Oliveira et al. 2011). The recent comparisons between GPS and non-GPS subsamples of the Jerusalem HTS have shown that, all else being equal, a GPS subsample provides trip rates that are 50%–70% higher and tour rates that are 10%–20% higher than diary-only households, with rates varying based on trip/tour purpose. The most frequently underreported travel components are short trips, nonmotorized trips, and intermediate stops on commuting tours. While these trips might not contribute significantly to regional vehicle miles traveled (VMT), they are important for understanding and modeling travel behavior and other (longer) trips. For example, the presence of intermediate stops on commuting trips (like dropping off a child at school on the way to work or visiting a gym or shopping mall on the way from work) can be a major reason for a person's resistance to switch to transit. A data collection method that systematically simplifies tours may result in an overly optimistic mode choice model that would overpredict the number of transit users for a new service.

Such GPS data collection efforts as the types described previously will be even more vital to emerging travel demand modeling paradigms. These include advanced activity-based models that focus more on the dynamic behavioral aspects of the traveler, such as in models by Habib and Miller (2008), Nijland et al. (2011), Auld and Mohammadian (2012), and others. These models all have a focus on day-to-day dynamics and choice behavior that cannot be observed in standard travel diary data. This is especially true for several large-scale, next-generation travel demand models in the process of development, including ADAPTS (Auld et al. 2009), SimAGENT (Goulias et al. 2012), and SIMTRAVEL (Pendyala, Konduri, and Chiu 2012), which could greatly benefit from new GPS-based data collection techniques.

Many of these models rely on somewhat esoteric concepts, such as choice set formation, activity time-space constraints, and scheduling flexibility, which individuals rely on when forming activity-travel patterns but are likely to have a fairly limited ability to recall in a survey setting. For example, a concept such as flexibility (e.g., spatial flexibility) can be factored into models when attempting to formulate realistic choice sets from which survey participants can choose. However, the individual will often have difficulty articulating the actual constraints underlying a location decision, as was found in the UTRACS survey (Frignani et al. 2010). Rather than rely on individuals to recall their location choice decision-making process, long-term observations can be used to directly observe the variability of location decisions for activities to formulate a more accurate measure for use in model development. This can similarly be done for other factors such as timing flexibility, and route choice variability.

Activity and Trip End Locations

Another issue that has plagued HTSs and subsequent travel model development for many years is geocoding of locations (trip origins and destinations) and ensuring proper trip arrival and departure times. For any travel model, whether four step or ABM, a trip record with unknown or incorrect destination location(s) is unusable for most sub-models. Rounding and other mistakes 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 CATI- and GPS-collected travel time data and how they compare with modeled estimates based on the same origin–destination (OD) pairs were analyzed by GeoStats (Wolf, Oliveira, and Thompson 2003) for the three regions in California that participated in the GPS component of the 2001 statewide travel survey. This analysis revealed that CATI tends to significantly overestimate travel time 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.

Baseline Transportation Network Conditions

All transportation models require some sort of baseline network and measurement of operating conditions. Travel demand model networks typically require estimates of free flow and congested speeds. Activity and tour models require these same speeds but at a more refined temporal scale. Simulation models and dynamic traffic assignment (DTA) models need detailed speed and condition data for the specific model scope. Many model developers have used GPS-based probe vehicles to collect these data over the last 15 years. Probe-vehicle data are typically collected using a sampling plan to ensure that the roads of concern are collected at designated times of day. GPS tracking has also been used to monitor network performance measures such as travel times, speeds, and delay (Quiroga 2004; Hackney, Marchal, and Axhausen 2005).

More recently, consumer data that is collected by private companies is being used to identify baseline conditions. Several private data companies have roadway speeds archived over multiple years and can generate custom queries. The original data come from a number of different technologies, including personal navigation devices, commercial vehicle GPS, and smartphones. These technologies are discussed in more detail in a later section.

GPS data that have been collected as part of a household travel survey can also be used to identify new model links that are needed in the network. Since travel demand model networks do not typically include all roads, some collector and local roads that are heavily traveled may not be represented. The GPS travel data can be used to identify these frequently traveled links.

Another related research area is measuring travel behavior changes that result from changes in network conditions. Much work has been completed in Australia measuring travel behavior changes in response to the TravelSmart policy (Stopher, Fitzgerald, and Biddle 2006; Stopher, Swann, and Fitzgerald 2007). These studies use either 1-week or 4-week GPS panels, repeated over a period of years, to extract some basic travel behavior measures, such as the vehicle kilometers traveled and number of trips. As the GPS data allows for a much more accurate method of determining these values and the data are easier to collect, it has proved useful in measuring travel behavior changes.

Route Choice Analysis

Route choice decisions are very difficult for survey respondents to reproduce using any reporting method. This has led to a lack of useful data on route selection behavior outside of simulated experiments. However, once GPS technology started being used in travel surveys, it was realized that the route selection behavior of the travelers would also be captured (Jan, Horowitz, and Peng 2000; Li, Guensler, and Ogle 2005; Papinski, Scott, and Dougherty 2008). In the research of Jan, Horowitz, and Peng, data from the Lexington study were used to form general observations about route selection behavior, comparing variations in path selection and deviations from assumed shortest paths (Jan, Horowitz, and Peng 2000). Georgia Tech researchers used GPS to observe variations in the chosen morning commute route within the Commute Atlanta project (Li, Guensler, and Ogle 2005). Papinski, Scott, and Dougherty compared route pre-planning to actual morning commute routes and made observations about how routes are planned (Papinski, Scott, and Dougherty 2008).

The recent Traffic Choices Study sponsored by FHWA and conducted in Seattle, WA evaluated the before and after choices (including route) when different corridors were tolled (Puget Sound Regional Council 2008). This type of survey can be combined with any travel demand policy experiment. (Imposing differentiated tolls was the essence of the Seattle study.) In this case, the individual response to the policy can be analyzed through car use. An important additional benefit of this type of study that has not been fully utilized yet is the ability to calculate travel condition measures like travel time reliability in addition to average travel time and cost. Travel time reliability has been widely recognized as a very important characteristic of highway service quality. However, this characteristic has never been included in travel models [and travel behavior analysis in general except for stated preference (SP) studies] because of the absence of network data on travel time variation at the trip origin–destination level. There is a wealth of data on travel time distribution at the highway segment level provided by sensor-based data collection systems (loops, cameras, RFID, etc.). However, a GPS-assisted traffic choices study is a unique way to track travel times and conditions for entire trips (for example, commuting to and from work) implemented by the same individual over a substantial period of time.

The San Francisco County Transportation Authority developed a route choice model based on GPS travel data collected as part of the CycleTracks data collection effort (Hood, Sall, and Charlton 2011). GPS data collected on smartphones using the CycleTracks application were analyzed to identify activities, mode transfers, and network paths. The results were used to create a multinomial logit model to reveal route and condition preferences. A smaller-scale but similar effort was conducted in Portland in 1999 (Broach, Gliebe, and Dill 2009), and recent efforts in Austin, TX and Monterey, CA also used the same approach to route choice modeling.

Model Calibration and Validation

Travel model calibration and validation are conducted by comparing travel time forecasts from the baseline model with some sort of ground-truth data. The ground-truth data in this situation can come from trip origin–destination travel times and traffic volume counts at major screenlines. (Screenlines refer to locations around an urban area where traffic is funneled into a few crossing points. Screenlines typically occur at river crossings, rail crossings, or border crossings.) More recently, GPS origin–destination travel times have been used from probe vehicles, GPS-based travel surveys, or consumer product origin–destination data sets. Trip ends are geocoded and assigned to a traffic analysis zone (TAZ). Trip travel times between TAZs are recorded based on their start times. The resulting table of TAZ-to-TAZ trips is aggregated and averaged into time-of-day bins based on the model design needs. The results can then be used in comparison with a similar output table generated by the model.

In addition to GPS-based household travel survey data, the same information can be tabulated from reported information in travel diaries. A more recent development by a cell phone data provider allows modelers to purchase zone-to-zone travel times for large percentages (30%–70%) of the population where zones are based on census geometry. While still needing some independent evaluation, the approach is palatable to the modeling industry and fits the needs of independent and comprehensive validation data sets.

Additional Modeling Needs

Visitor/Tourist Travel Behavior. There are very few visitor models applied in practice. Most that have been applied are greatly simplified and aggregate in nature. However, for some major cities like New York, visitors represent a very significant travel component. In advanced ABMs, a disaggregate approach for modeling visitors has been considered that is based on the same principle of micro-simulation of individual behavior as the core model for residents. Hotels provide the basis for structural synthesis of the population of visitors. To support such a model, a sample of daily travel diaries of hotel customers similar to individual questionnaires of HTSs can be collected. However, daily travel of visitors might be quite intensive in terms of number of trips and chains of trips (especially for nonbusiness purposes). Further, visitors and tourists are less familiar with the area, and it would be more difficult to retrieve trip end locations with them by address in a conventional non-GPS setting. Tourists and visitors may be even more reluctant to participate in a comprehensive, long survey compared to residents. It is possible that GPS-assisted methods with prompted recall would be attractive for this type of survey. A simplifying aspect of this type of survey is that the sampled unit is a person rather than an entire household (when compared to a household travel survey).

One approach would be to integrate an airport survey with the hotel visitors' survey. Visitors can be recruited as they arrive at the airport and be equipped with a GPS device for the duration of their stay, with travel information retrieved using a prompted-recall method when they return to the airport for their departing flight. It should be noted that visitors' trips to and from airports themselves represent an important travel market with unique characteristics such as very high willingness to pay for travel time savings and reliability.

Taxi Vehicle Activity Patterns. Taxi as a travel mode has never been fully represented in travel models. In many regional travel models, taxi activity is entirely missing from the forecast. In other models, it is represented in a simplified way in terms of actual availability for the trip and associated travel time (wait and ride) and cost. However, in some major cities like New York, taxis represent over 30% of motorized trips. Making the taxi share more accurate in the mode choice model is not the only task to adequately represent taxies in travel models. Another important issue that has never been fully addressed in operational models is a proper representation of taxi vehicle movements that is much more complicated than for private automobiles. Taxi vehicle movement represents a complicated daily chain of trips with and without passengers that do not directly relate to passenger tours. For advanced travel models, one can envision a special new sub-model being developed for converting passenger taxi trips into vehicle trip chains. This sub-model will require a new data source that could be envisioned as a multiday GPS-assisted survey of taxis (with GPS devices installed on taxis). Some recent and ongoing research has been completed in modeling taxi behavior and taxi demand assisted by GPS data collection. One such a study was conducted by Liu, Andris, and Ratti (2010), who used a large database of taxi driver GPS traces to analyze how taxi drivers evolve their travel patterns to increase revenue.

Commercial Delivery Vehicle Activity Patterns. Similar to the additional data recommended for taxi movements, one could consider a GPS-assisted multiday survey for trucks, deliveries, and other non-passenger vehicles contributing to daily circulation in major cities. Behavior of commercial vehicles is very different from passenger travel behavior and, in general, has been less explored and understood. A delivery truck might have 10 to 20 chained trips per day that are very difficult to retrieve reliably in a conventional survey setting. GPS-assisted technology is the only way to retrieve actual truck movements as a basis for a more advanced freight delivery model. A truck GPS study of a major grocery chain in the Chicago region was recently completed by the University of Illinois – Chicago (Mohammadian et al. 2013).

Emissions Modeling. Household travel survey GPS data can be used in emissions modeling to develop and evaluate driving profiles and to generate link-level speed estimates. Emissions models require speed data to generate accurate estimates of vehicle emissions. The driving profiles provide the fraction of time respondents spend at different speed bins. Both Mobile 6.2 and the new MOVES model need this information to generate emissions estimates. Many regions also use a link-based emissions assessment that is based on the average speed of a road network link and the road volume. GPS data can provide the average speeds as well as their variability, thereby giving planners more options for improving emissions estimates. GPS data also provide measurement of time between engine starts. This value is used in estimating the intensity of cold start emission events.

Local road travel is often unknown. A properly designed GPS component can help analysts identify the fraction of VMT and vehicle hours traveled (VHT), two common emissions modeling metrics, that occurs on links outside of the model network. These data can increase the accuracy of emissions estimates, particularly local travel, which may have a lot of stop-and-go activity.

GPS data from household travel surveys have been used to support secondary research into emissions and greenhouse gas formation by the National Renewable Energy Lab in Colorado and the U.S. EPA (Gonder et al. 2007). Emissions researchers are relying less on typical driving cycles for emission rate estimates and instead, with the proliferation of GPS and other research, relying more on actual driving patterns from the general population. This interest is expected to continue with an increased focus in greenhouse gas emissions and tighter emissions standards. Involvement of these researchers in the formation of a possible GPS supplement to a household travel survey could result in additional financial support and increased benefits from the study.

In fact, the California Energy Commission and the California Air Resources Board (CARB) spearheaded a GPS augment to the 2012–2013 CHTS California Statewide Travel Survey in which 1,200 households received both GPS devices and OBD engine sensors to install in up to three household vehicles for 7 days. These data streams were processed, delivered, and used to estimate fuel consumption and vehicle emissions as required by recently passed state laws that require greenhouse gas emissions monitoring.

An early use of GPS in 1999 by CARB evaluated heavy-duty truck activity (Battelle Memorial Institute 1999). This study, and many other special-purpose studies, has evaluated detailed trace data from instrumented vehicles to improve on the understanding of vehicle activity.

Other Types of GPS-Based Travel Behavior Studies

Physical Activity/Health Research

Given that many regions are interested in smarter and more transit-oriented development, many are including a physical activity component to the planned travel surveys. The 2001 SMARTRAQ survey in Atlanta included a component similar to this that was funded by the Centers for Disease Control. Health researchers have been very active in transportation planning in recent years due to the clear impact of travel behavior on physical activity (and obesity). People who become car-dependent not only have an impact on traffic congestion but may also have limited physical activity. Wearable GPS devices combined with activity monitors (or accelerometers) have been deployed for multiday periods in many studies to quantify the levels of physical activity and travel for different populations of interest. These studies have included before and after mobility and travel evaluations of the cane training rehabilitation program offered for visually impaired veterans, an analysis of travel patterns and environmental exposures of children with asthma, and an examination of the level of physical activity on and off trails by trail users across the state of Massachusetts (Wolf and Lee 2009; Wolf and Trost 2009; Troped et al. 2008). GPS and accelerometer data can be used to answer key questions such as where does most physical activity occur (at home, a non-home location, or as a by-product of travel), how much physical activity occurs on a daily basis, and at what intensity does this activity occur.

Mark Freedman of Westat stated, "We are excited about the potential opportunities to combine the transportation and health sectors in joint data collection efforts where the use of travel diaries, GPS, and accelerometer data can be combined" (2012 Industry Survey). Such a study was recently conducted by Thompson and Kayak (2011) using GPS combined with accelerometer data to quantify individual daily activity levels. A similar study by Lee et al. (2012) used GPS traces from travel surveys to assess the amount of physical activity respondents engage in when choosing active transportation modes (i.e., walking, biking, etc.). Such studies are likely to become more and more important as public health challenges in the developed world, such as obesity and cardiovascular disease, grow more prevalent (Thompson and Kayak 2011; Doherty and Oh 2012).

Inclusion of a physical activity component in a household travel survey can serve health research needs, as well as provide data for transportation modeling and regional planning needs. The relationship between physical activity and transportation planning is clear, and this knowledge is useful in the development of strategies that promote nonmotorized travel. Beyond this, understanding the relationship between physical activity and travel mode must factor into the design of livable communities and other built environment planning exercises. The most recent Nashville regional travel survey was designed with these joint goals in mind. This survey was branded as the Nashville Transportation and Health Study and included a 10% subsample of households who used GPS devices and accelerometers and also completed an extensive health survey in addition to the household travel survey.

Road Pricing and User Fees

An ongoing study being conducted in Minnesota has looked at the feasibility of road pricing strategies using a GPS-enabled device to assess user fees, either through onboard calculation or communication with a centralized management center (Pierce et al. 2011). The study has looked at the suitability of GPS in terms of its accuracy in being able to distinguish various road segments, privacy issues, and various architectures for collecting the data and assessing the fees. Samsung smartphones were provided to recruited participants, who installed the devices in their vehicles for the duration of the study. Some of the findings are likely applicable to more general GPS survey efforts, such as the trade-offs between thick and thin client data processing strategies (i.e., processing and aggregating data within the device instead of at a centralized location), which could help mitigate some of the privacy issues in other surveys as well (Pierce et al. 2011).

The aforementioned Traffic Choices Study sponsored by FHWA and conducted in Seattle, WA evaluated the before and after behavior when different corridors were tolled at different times of the day (Puget Sound Regional Council 2008). Findings suggested that participants made small-scale adjustments that could, as a whole, have an impact on traffic congestion. Further, the study suggested that open road tolling is technically feasible but will require a more robust business model to achieve the tolling program's goals.

Transit Passenger Surveying

To comply with Title VI requirements as well as to improve service planning, transit agencies conduct surveys at varying frequencies to gather data regarding travel patterns (which may include origin–destination information), ridership demographics, and customer satisfaction information. The data are a valuable tool for effective transit planning and travel demand modeling. Traditional surveys collect data on the sequence of the current one-way trip from origin to boarding and then alighting to destination. These surveys also generally ask about payment method, fare subsidies available to the customer, and options for alternative modes, as well as demographic details about the rider, including age, race, income, and frequency of transit use. A 2005 survey of 52 transit agencies found that larger agencies conduct five or more onboard and intercept surveys annually, while smaller agencies conduct surveys every 1 to 3 years (Schaller 2005). Transit agencies are also required to have data collected by a recent onboard survey (within 5 years) to apply to the Federal Transit Administration's New Starts program.

Prior to the rise of GPS and mobile technology, the latter of which allowed for real-time geocoding, the most common approach to the collection of these data was to have a survey administrator hand out pencil and paper questionnaires on selected bus routes or trains. An advantage of the pencil and paper method is the relative ease of handing out a questionnaire to a large proportion of the riders of a given route. A disadvantage to this approach is the possibility of a failure by the participant to understand the objective (e.g., participants will report a round trip instead of the requested one-way trip). Additionally, the completeness, precision, or accuracy of reported origin, boarding, alighting, or destination locations are not always known and not easily verifiable. While this approach is still widely used today, innovative methods are being used to replace or supplement this traditional approach.

One such method used to collect and audit responses to transit surveys uses GPS-enabled personal digital assistants (Oliveira and Casas 2010). However, this approach does not fully address concerns about the accuracy of self-reported data. A method that addresses the concerns about the accuracy of the data, while also leveraging the capabilities of real-time geocoding and GPS data, is the use of tablet devices to conduct face-to-face personal interviews (Atlanta Regional Commission 2012). The use of this combination allows for more accurate, complete, and representative responses and also provides the interviewer with details about the transit system and study area that would otherwise be unavailable. Tablets with cellular connectivity allow the survey managers to adjust goals in real time. While there may be concerns about the costs of the tablets or the labor-intensive nature of face-to-face personal interviews, there is also evidence to suggest that the costs become fairly comparable when calculated using completed, usable surveys.

Automated passenger counter (APC) and AVL systems are used frequently as a means to measure level of service (LOS), manage dispatching and scheduling, and provide feedback to drivers about schedule adherence (Furth et al. 2006). While it is possible that these technologies could be integrated to provide high-quality and accurate passenger movement data, it is not a prevalent option at this time. Challenges to this possibility include the fact that not all vehicles in a fleet are outfitted with APC and AVL devices and that their integration is not always a straightforward task. Other applications of merged AVL/APC data include the estimation of dwell times.

Standards, Guidelines, and Common Practices for Travel Demand Model Data Collection

Household travel surveys constitute one of the most important sources of disaggregate travel behavior data for TDMs. They were initially conducted in the United States in the 1950s, and, during most of the decades since then, little has been done "to standardize the processes or to institute consistent practices of acceptable quality of reliability" (Stopher, Alsnih, et al. 2008).

The U.S. Department of Transportation requires that transportation management areas (TMAs), defined as urbanized areas with populations exceeding 200,000, go through a certification review every 4 years per 23 CFR § 450.334(b) (U.S. Department of Transportation 2012). During this review, a TMA may be required to update the data used in its TDM, especially if the travel survey data used were collected 15 or more years previously (Murakami and Bricka 2012). In fact, for most MPOs, conducting a travel survey constitutes one of the largest routine expenditures made from its planning budgets (Stopher, Alsnih, et al. 2008).

Evolution of Household Travel Survey Standardization

The 1990s saw an increase in the demands placed on TDMs and travel surveys as a consequence of changes put in place by the 1990 Clean Air Act amendments, the 1991 Intermodal Surface Transportation Efficiency Act (ISTEA), and other earlier legislation (Tierney et al. 1996). At the same time, travel surveys went through several technology evolutions throughout the 1980s and 1990s with the introduction of computerized interviewing systems, centralized call-centers, and online address validation and geocoding technology.

Guidance on recommended processes and practices in travel surveys was needed to better meet these added demands by improving quality, reliability, and transferability of the collected data. These initial efforts were crystalized in the FHWA *Travel Survey Manual* (Tierney et al. 1996). Although this document did not prescribe standards, it did provide significant guidance by compiling material from previous guideline documents as well as technical papers into a single comprehensive source. The 580-page report identified the various types of travel surveys and covered several important subjects, such as management and quality control, precision and accuracy, geocoding, and emerging trends identified at the time (including stated response and longitudinal surveys). In addition, the report dedicated individual chapters to the main types of travel surveys identified by FHWA, which were household travel surveys, vehicle intercept and external station surveys, transit onboard surveys, commercial vehicle surveys, workplace and establishment surveys, visitor surveys, and parking surveys.

A subsequent push for standardization came about in the early 2000s with the publication of the *NCHRP Report 571: Standardized Procedures for Travel Surveys* (Stopher, Alsnih, et al. 2008). This report was jointly created by a team of travel survey experts from around the world and focused on identifying aspects of personal travel surveys that could be standardized. It also provided recommendations on how to implement the drafted standards, identified areas for future research, and included templates for requests for proposals

(RFPs). The main aspects for which standards were developed and presented in this report were:

- Design of survey instruments,
- Design of data collection procedures,
- Pilot surveys and pre-tests,
- Survey implementation,
- Data coding and geocoding,
- Data analysis and expansion, and
- Assessment of survey quality.

Several of these standards were incorporated into the design of the 2009 National Household Travel Survey (NHTS). More recently, members and friends of the Transportation Research Board's Travel Survey Methods Committee (ABJ40) have started maintaining an online version of the travel survey manual (http://www.travelsurveymanual.org). Initial content for the website came primarily from *NCHRP Report 571* (Stopher, Alsnih, et al. 2008) and FHWA's 1996 *Travel Survey Manual* (Tierney et al. 1996). Since its initial release, the content has been updated and expanded by the professional community.

The remaining parts of this section summarize the main aspects of these latest standards with regard to how they affect the collection of travel survey data for supporting GPS data processing and augmentation, present a review of the basic elements of a successful GPS-enhanced travel survey, and present three examples of recent travel surveys that illustrate different design approaches.

Relevant Travel Survey Standards Guidance from Other U.S. Federal Agencies

Other federal agencies that provide guidance relevant to travel surveys are the Bureau of Transportation Statistics (BTS), the Office of Management and Budget (OMB), and the U.S. Census Bureau. After the passage of ISTEA in 1991, the BTS was created in 1992 for the purpose of administering "data collection, analysis, and reporting and to ensure the most cost-effective use of transportation-monitoring resources." The stated mission of the BTS is "to create, manage, and share transportation statistical knowledge with public and private transportation communities and the nation" (U.S. Bureau of Transportation Statistics 2005). To that end, the BTS has released a statistical standards manual that provides general guidelines about the planning and design of all types of surveys run by a government agency and includes guidelines for the actual collection of data, the processing of data, and subsequent analysis, dissemination, and evaluations of data quality (U.S. Bureau of Transportation Statistics 2005).

The OMB oversees the Office of Information and Regulatory Affairs, which hosts a document developed to provide guidance on the development and deployment of statistical surveys (U.S. Office of Management and Budget 2006). The document is similar in form and content to the standards and guidelines found in the BTS document and contains the following caveat:

> The standards and guidelines are not intended to substitute for the extensive existing literature on statistical and survey theory, methods, and operations. When undertaking a survey, an agency should engage knowledgeable and experienced survey practitioners to effectively achieve the goals of the standards. Persons involved should have knowledge and experience in survey sampling theory, survey design and methodology, field operations, data analysis, and dissemination as well as technological aspects of surveys.

As with the BTS, the OMB has provided a firm general framework for survey design and deployment, and it has stressed the need to conform to sound, proven statistical methods when conducting surveys but resists providing any more concrete requirements, thereby leaving discretion to the survey design process.

The U.S. Census Bureau conducts a decennial census, an economic and government census every 5 years, and the annual American Community Survey (ACS). It is common practice to use ACS question wording and choice lists in travel surveys. The bins used for questions about income, for example, are often identical to the bins found in income questions posed by the Census Bureau. Questions about race/ethnicity, gender, occupation, age, and other socio-demographic attributes may also be based on census structure, wording, or choices. These intentional design consistencies between surveys allow for comparisons at appropriate person/household and geographic aggregation levels, which are an important means of checking and controlling the sample for biases and representativeness during the survey effort.

Guidelines in GPS Data Collection and Basic Processing for Travel Surveys

Although U.S. federal agencies do not provide formal standards or guidance on how to use GPS technology in travel surveys, there exists a considerable body of literature on the topic. There are also standards that are related to how GPS data are collected, archived, and shared, which are applicable to household travel surveys. This section discusses these two topics. The next section of this chapter covers the literature review results in the area of GPS data processing and imputation methods.

Guidance from Technical Literature

Wolf proposed guidance in the form of simple steps that can be used to convert a stream of raw GPS points into trips (Wolf 2000). The process included the filtering of GPS points based on quality indicators and zero speed followed by the computation of dwell time at each point. It was reported that a dwell time of 120 s worked well for identifying most trips. Stopher, Jiang, and Fitzgerald provided additional guidance on how to prepare raw GPS point data for processing along with suggested thresholds for GPS data-quality indicators, more detailed procedures for filtering non-movement, and suggestions for how to detect and handle cold starts and signal dropouts (Stopher, Jiang, and Fitzgerald 2005). More specifically, the team proposed that GPS points with zero speed and that show movements of less than 15 m be removed during filtering (Stopher, Jiang, and Fitzgerald 2005).

Tsui and Shalaby (2006) suggested that points with fewer than three satellites in view and with horizontal dilution of precision (HDOP, a measure of the quality of a GPS coordinate solution, where smaller numbers indicate better data) values above 5 be automatically removed, as well as points with zero directional heading and speed values. The resulting points are then reviewed for positional jumps, which tend to occur in urban canyon areas. Schüssler and Axhausen (2009b) used altitude in lieu of traditional dilution of precision quality indicators in cases where the collected data do not contain them by removing points that reported unrealistic altitude readings. Schüssler and Axhausen combined this filtering with the examination of points with near-zero speed for a minimum of 120 s to detect trip ends; activity locations were subsequently identified by looking at "bundles of GPS points" consisting of at least 15 points in sequence (Schüssler and Axhausen, 2009b). Alvarez-Garcia et al. (2010) used a minimum distance threshold of 30 m between points to filter raw data before identifying stops, while Lawson, Chen, and Gong (2010) employed 50-m buffers to detect trip ends by searching for records outside a point buffer within a 120-second time window.

With regard to device selection, information can be obtained in reports from previous projects as well as from findings in research papers. However, given the rapid evolution and constant change in the consumer market for GPS data loggers, it may be necessary to conduct independent evaluations such as the ones contained in Lawson et al. (2008) and Anderson et al. (2009).

Related GPS Data Standards Applicable to Travel Surveys

The National Marine Electronics Association (NMEA) developed one of the earliest standards for encoding data from GPS receivers, as well as for use with other navigational sensors. This standard, which is named NMEA 0183, is supported by most GPS receivers and has undergone several updates since being released in the 1990s. The NMEA 0183 standard defines a set of text messages that can be sent over a serial protocol; parsing these messages yields information that can be used for navigation or logged for later processing.

The GPS exchange format (GPX) uses XML (extensible markup language) schema for the lightweight encoding of waypoints, routes, and tracks (Foster 2004). This format is very popular in consumer-grade devices and is supported by several websites and web services. Another XML-like standard that has been used for exchanging GPS data is the Keyhole Markup Language (KML), which can be used by both Google Earth and Google Maps to display GPS data on an image of a globe and map respectively (Google 2012). Since its initial adoption, KML has also become an Open Geospatial Consortium (OGC) standard (Open Geospatial Consortium 2008). GeoJSON is a geospatial data interchange format based on JavaScript object notation (JSON) and is commonly used in web development (Butler et al. 2008). The focus of these standards is on mapping and navigation applications, making them less than suitable to travel survey and transportation planning applications.

More recent advances in both the capabilities and usage of web and mobile (i.e., smartphone and tablet devices) mapping APIs have generated a few de facto standards that the information technology industry follows when processing GPS data, such as:

- Point coordinates are stored in decimal degrees using the World Geodetic System (WGS) 84 datum,
- Point date and time information are provided in coordinated universal time (UTC),
- ISO 8601 or NMEA 0183 formats are used when encoding date and time information into text, and
- Speed is indicated in meters per second.

Imputation and Data Fusion of Travel Behavior Details

Travel behavior researchers have recognized that detailed travel data from GPS traces combined with other trip details and geographic information has the potential to provide travel behavior details without participant reporting. For example, trip ends derived from passively collected GPS data can logically be combined with geographically referenced land use data to estimate trip purpose (i.e., home-based shopping trips). Several other travel details can be similarly estimated when combined with other measured or reported data elements. These processes of imputation and data fusion are attractive to travel behavior researchers because data generation moves away from participant-reported variables and more toward measured variables, which are deemed more reliable. Further, there is growing interest in using GPS data generated through consumer devices and apps that are archived and resold by private companies. This type of data does not come with additional travel details other than the GPS trip trace and is limited in usefulness unless additional behavioral details can be imputed. The following sections present how these tasks have been accomplished by researchers and practitioners.

Trip End Identification

Before trip identification can occur, it is often necessary to perform basic formatting of the logged data to have it in a consistent format that is convenient for processing. This includes converting date and time information (which is often provided in UTC) to local date and time, as well as performing unit conversions (i.e., some devices may report speeds in atypical units such as knots). It is also important when performing these initial steps to be aware of the logging rules used to configure the GPS loggers prior to data collection. These rules define when and how often a new GPS position is to be logged. For example, knowing the expected logging interval helps in detecting signal dropouts. An initial data quality step is often undertaken prior to any analysis to remove data points identified by the device as having too few satellite connections or low precision due to poor satellite configurations.

Next, the remaining points are processed to remove further erroneous observations to produce reasonable traces. The traditional method involves the filtering of GPS points moving at very slow speed (e.g., less than 1 mph) followed by the computation of time intervals between subsequent points; these time intervals represent time over which the logger did not move. Whenever a gap of 120 s or more is found, a new trip end is placed (Wolf, Guensler, and Bachman 2001). This method has also been extended by looking at distance covered between points, spatial buffers, and heading changes (Stopher, Jiang, and Fitzgerald 2005; Lawson, Chen, and Gong 2010). Other studies have used various clustering algorithms such as *k*-means clustering (Ashbrook and Starner 2003), spatial density analysis (Flamm, Jemelin, and Kaufmann 2007), and land-use–constrained spatial buffering (Auld et al. 2009). Once the initial set of trip ends is identified, it is necessary to perform additional processing to deal with the potential presence of cold starts and signal dropouts.

Cold start events occur when the GPS receiver is either powered down or has not acquired satellite signals for an extended period of time (i.e., more than a few hours). Under these circumstances, a receiver may take several minutes to restart acquiring and reporting GPS positions, which results in the start portion (or the entirety) of the trip not being captured. Smartphones and other connected devices can shorten this time through the use of assisted GPS technologies that use the cell-data network to download updated satellite orbit information and time offsets, which can be used to speed up receiver positional reacquisition. Cold start events can be detected by comparing the end and start locations of adjoining trips and searching for distance gaps. The trip start location and path can then be corrected using information from the previous trip as well as geographic information system (GIS) data.

Signal dropouts occur when the receiver no longer has a lock on the minimum number of satellites needed to compute a positional fix (i.e., three for a two-dimensional solution and four for computing a three-dimensional solution). Typical causes of signal dropouts are urban canyons and overhead blockages such as bridges, tunnels, and tree foliage. Signal dropouts can be detected by computing the speed between the starting and ending points of a data gap (defined as a gap consisting of two points that are separated by a time interval longer than a multiplier of the logging epoch but shorter than the defined trip end criterion) and comparing it with a minimum movement speed. If the computed average speed is greater than the minimum movement speed, the dropout can be ignored. If the computed average speed is less than the minimum movement speed, then it should be inspected for possible short stops like those that occur when picking up and dropping off passengers.

These methods are usually combined with analyst follow-up and inspection procedures to ensure that consistent and accurate results are produced. These are especially important when reviewing and tagging short stops that last less than the usual trip end delay criteria of 120 s (Steer Davies Gleave and GeoStats 2003).

Determining Basic Trip Details

Once the end points of trips are identified, it becomes a simple computational problem to derive basic trip attributes. For example, one can define P as the ordered set of n GPS points (p_i) belonging to a trip, D as a function that returns the distance between any two points p_i and p_{i+1}, and T as the function that returns the amount of time between any two points. Using these, one can identify the basic trip attributes using the formulae that appear in Table 1-5.

Additional filtering and post-processing of the GPS points may be conducted before these formulae are applied. For

Table 1-5. Basic trip attributes from GPS.

Name	Formula
Origin location and time	$p1$
Destination location and time	pn
Trip duration	$T(pn, p1)$
Trip distance	$\sum_{i=1}^{n-1} \mathrm{D}(\mathrm{p_i}, \mathrm{p_{i+1}})$
Trip path	P

example, the points' positions could be smoothed to remove outliers before computing distance or generating a final trip path (Li, Guensler, and Ogle 2005).

Travel Mode Detection and Processing

Once the data are segmented using basic trip ends, it is necessary to detect mode transitions that may have occurred within a GPS trip. The resulting sub-trips are often referred to as trip or mode segments or elemental trips. Once these are identified, travel modes are assigned.

Detecting Mode Transitions

Work presented by de Jong and Mensonides proposed an approach whereby trips were segmented into single-mode stages based on the assumption that a short period of zero speed was necessary for each mode change (de Jong and Mensonides 2003). The travel mode of the stages was then determined by leveraging the speed characteristics and the proximity to public transport stops and routes. In addition, the proposed logic tested whether the generated mode sequence was reasonable; for example, the logic would not allow direct transitions from bus to auto without an intermediate walk stage. The logic uses the fact that the walk mode has consistently low speeds and accelerations.

Tsui and Shalaby presented an integrated system to process person-based GPS data for travel surveys (Tsui and Shalaby 2006). This system (GPS-GIS) included two versions; version one included modules for performing data filtering, identifying trip ends, and detecting mode transitions within trips and mode identification, while version two used link matching of the GPS data to a GIS representation of the transportation network to support further GIS-based processing. The mode transition identification module in version one of the GPS-GIS system segmented each trip into single-mode stages 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 by 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 until 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.

Travel Mode Identification

Different approaches have been used to associate a travel mode with a sequence of GPS points. Most mode identification algorithms rely on central tendency measures of instantaneous GPS point speeds, such as mean, median, mode, as well as indicators of speed variability, such as acceleration and deceleration rates, standard deviation values, and maximum speed and acceleration values for different high percentiles (e.g., 85th and 95th), as well as measures of positional quality. The main methods used to perform mode identification can be classified into three groups: rule-based, probabilistic, and artificial intelligence. Artificial intelligence methods include both fuzzy logic and neural network applications.

Stopher, Clifford, and Zhang proposed a hierarchical, rule-based process in which candidate modes are tested against the GPS data in the following predetermined sequence: walk, bicycle, off-street network transit, bus, and auto (Stopher, Clifford, and Zhang 2007). In all cases, the 85th percentile values for speed, acceleration, and deceleration are used to rule out a travel mode candidate. The authors point out that using the 85th percentile values for the decision variables has the benefit of dealing with outliers typically found in person-based GPS logs (i.e., the few points that stray from the captured trajectory). Oliveira et al. (2011) used pre-computed values for average, maximum, and standard deviation of mode speeds to select the most likely candidate. The first step in the process was to compute average and standard deviation values of the trip segments' point speeds. The algorithm selected the travel mode that most closely matched the values for average and standard deviation of point speeds while having its 95th percentile speed lower than or equal to the mode's maximum speed.

Oliveira et al. (2006) demonstrated how a probabilistic modeling approach could be used to identify travel mode using data obtained by fusing GPS points with personal accelerometers. The developed multinomial logit model was shown to accurately identify modes for 75% of the validation cases. It was also found that using the accelerometer data improved the identification accuracy for nonmotorized travel modes. Moiseeva, Jessurun, and Timmermans (2010) also used a probabilistic approach, this one called the Bayesian belief network, to identify travel modes achieving a high accuracy rate of 92%.

Fuzzy logic has its roots in machine control and is applicable to imprecise situations, which cannot be defined with crisp true-and-false rules. For example, trips can be defined in terms of slow, medium, or fast using median travel speed. Most people typically do not reason that there is an exact cutoff between someone traveling slowly and fast. Fuzzy logic allows you to describe a speed value in terms of how slow and how fast it is at the same time using continuous values from 0 to 1 (Lawson, Chen, and Gong 2010). Tsui and Shalaby (2006) and Schüssler and Axhausen (2008) employed a fuzzy logic basic approach to define whether a mode candidate was suitable to a set of GPS points; in both cases an ordered set of most-likely candidates was associated with each trip segment to be classified. The fuzzy variables used in the membership functions were average speed of GPS points, 95th percentile maximum speed of GPS records, positive median acceleration of GPS records, and data quality of GPS records (based on HDOP). Tsui and Shalaby performed additional discriminatory analysis on the initial set from the fuzzy logic functions utilizing the results from a map-matching procedure, further refining the resultant choice set (Tsui and Shalaby 2006).

Neural networks are able to generalize conclusions for data without the need to define relationships or rules in an a priori manner. A neural network can learn the subtle differences between car, bus, and walking trips and, therefore, automatically detect the mode of transportation for a new, previously unseen trip (Gonzalez et al. 2008). Byon, Abdulhai, and Shalaby (2007) demonstrated how neural networks could be used to automatically detect the mode of transportation. Their research used data collected using a laptop connected to GPS receivers; attributes used in their neural network included instantaneous acceleration, speed, and HDOP. The impact of different logging frequencies was analyzed, with the authors finding that mode detection performance was close to 80% with sampling intervals as long as 3 min (Byon, Abdulhai, and Shalaby 2007). Shorter reporting intervals produced better results.

Gonzalez et al. (2008) took the application of neural networks to model identification further by examining how well it could work on data collected using mobile phones equipped with assisted GPS technology. The authors also showed how the method could be applied to a reduced version of the input data, which were called "critical points." These were defined as the minimum set of points necessary to reconstruct a participant's path (Gonzalez et al. 2008).

Lawson, Chen, and Gong (2010) conducted a controlled experiment where the performance of three mode selection methods was compared. The first applied a rule-based algorithm similar to the one proposed in Stopher, Clifford, and Zhang (2007), the second implemented the neural network method previously described by Gonzalez et al. (2008), and the third was the approach documented in Schüssler and Axhausen (2008). The overall result was that the neural network approach produced the best results, with a success rate of 84% (Lawson, Chen, and Gong 2010).

Research on improving instantaneous and post-processed travel mode identification techniques is ongoing, and improvements in the reported success rates stated in this section are quite possible. Processing enhancements such as including trip chain logic, evaluating related geospatial data, and expanding to traveler-specific behaviors across multiple days may improve trip mode identification; however, research results in these areas were not formally published at the time the literature review was performed for this NCHRP study.

Route Identification

One of the post-processing steps implemented with GPS data sets is to relate their point locations with spatial data sets representing the transportation network in a process known as map matching (MM). When applied to GPS data sets, it allows the identification of the routes taken on the network.

MM processes can be performed on the fly (i.e., in a navigation device) or as a post-processing step to previously collected GPS trajectory data sets. The first application has been covered extensively in the literature and is mostly concerned with accurate predictions of where a GPS receiver is along a link and does not necessarily need to run faster than real time. This discussion focuses on the latter application, which is frequently applied in the context of household travel surveys (Doherty et al. 2000), performance measurement studies (Marchal, Hackney, and Axhausen 2005), and the analysis of utility vehicle behavior (Blazquez and Vonderohe 2005). Some additional applications of the results of a post-processing MM exercise of GPS point trajectories to a transportation network are to associate speed information to network links, identify locations where congestion occurs, generate data sets that help to understand route choice, and associate network data elements with the GPS data.

Initial map-matching algorithms used with GPS data would identify match candidates either by finding the closest line feature to the point or locate the line feature with the shape point closest to the GPS point. Greenfeld proposed an algorithm that created lines between subsequent GPS points and used similarity comparisons to decide which network link best matched each GPS point pair (Greenfeld 2002). A weighting scheme was used to balance the degree of parallelism between the GPS line segment and the link, the shortest distance to the link, and the size of the intersecting angle between the GPS-derived line and the street arc. White, Bernstein, and Kornhauser (2000) suggested a similar algorithm that used differences in heading to rule out matches.

Blazques and Vonderohe (2005) developed a rule-based MM algorithm that made use of shortest-path computations and turn restriction information to verify matches as the route was built. Although a high success rate was observed while applying this approach, the authors felt that it could be further improved, particularly when poor quality data were used, at the intersection of divided highways and where false negatives failed to snap (Blazquez, Ponce, and Miranda 2010).

The improved algorithm added dynamic resizing of the match distance tolerance, filtering of snaps based on the vehicle azimuth (or heading), and a revised set of matching rules.

Quddus et al. (2003) proposed an algorithm for use in a real-time navigation application that fused GPS points with data from a dead-reckoning (DR) device using a Kalman filter. It made use of weights to compute a score for each match. The score took into account the distance between the GPS points and the network arc, as well as the deviation between the point's heading (computed by the DR device) and the bearing of the link, and the position of the point relative to the link. Byon, Abdulhai, and Shalaby (2007) also proposed a MM routine to be used in real-time monitoring applications using data from GPS-enabled mobile phones.

Quddus, Noland, and Ochieng (2006) proposed an algorithm that used fuzzy logic to deal with the errors and uncertainties present when matching GPS points to a road network and that also used data from a DR device. It used multiple inputs to evaluate if a GPS point should be matched to a link, but placed emphasis on heading differences and the perpendicular distance to links. Fuzzy inference systems were used to combine the multiple input variables to generate likelihoods of links being appropriate matches in the different stages of the algorithm, namely the selection of the first route link, the determination of whether a point snapped to the current link, and the selection of links to be added to the route.

The MM process proposed in Velaga, Quddus, and Bristow (2011) built on the approach proposed by Quddus et al. (2003) by adding an additional step that optimizes the weights used in the scoring scheme using a genetic algorithm optimization technique. The authors also proposed the use of different weights based on the operating environment where the GPS data were collected (e.g., urban, suburban, and rural).

A common characteristic of these and most existing MM algorithms is the concept of match modes, with most of them containing an initialization mode, which identifies the first route link; a link snapping stage, which computes projections, or snaps, of the GPS points on the current link; and a new link search mode, which identifies the next link to be added to the route. The algorithm proposed by Dalumpines and Scott (2011) did not follow this pattern; it instead relied on network topology and used the GPS trace to construct a series of gates that limited the road network connectivity. This way, a shortest-path solution connecting the start of the trace to its end was likely to match the actual route taken. This process assumed that a complete and up-to-date road network, including information on turning restrictions, was available. It is not likely to be able to handle incomplete road networks, such as the ones typically used by travel demand models (i.e., missing local and collector type roads).

Most MM algorithms in the literature reconstruct the original route using a linear process with a single link at a time added. Once each decision to extend the route with a link is made and accepted, this link will remain on the final selected route. MM procedures based on the multiple hypothesis technique (MHT) address this limitation by maintaining several alternative paths and eventually selecting the best one at the end of the process.

Pyo, Shin, and Sung (2001) demonstrated how MHT can be used in a navigation setting, which included integration with a DR device. However, due to its origins in navigation, the proposed logic focused on the accuracy of the point projections on the network instead of the final vehicle path. Marchal, Hackney, and Axhausen (2005) applied MHT to the problem of post-processing MM, with a focus on the operational performance. This meant that only the two-dimensional GPS point coordinates along with the line features representing the network were used as inputs. Furthermore, the scoring of point snaps and routes was mostly based on the distance between the GPS points and the network links. One of the findings of this application of MHT was that reasonable results could be obtained despite these simplifications. The authors also looked at the impact of keeping different numbers of candidate routes during the course of a route derivation; it was found that keeping 30 candidates produced quality results at reasonable computational speeds. This algorithm was later extended by Schüssler and Axhausen (2009a) to be used within the context of trips with the added ability to fill in gaps in the identified routes.

Trip Purpose Identification

Identifying trip purpose or activity at destination from GPS-derived trips remains a difficult problem to solve. This is because it inherently requires a multi-factorial approach where extensive data sources are combined with the basic trip attributes derived directly from the underlying trace data.

Wolf, Guensler, and Bachman (2001) demonstrated how trip end locations could be matched to GIS data to derive basic trip purpose classifications. At the time of the research it was noted that the trip purpose determination step worked well for approximately 78% of the test cases.

Schönfelder and Samaga (2003) used a multistage hierarchical matching procedure to infer trip purposes. This process involved the calculation of clusters of trip ends, the identification of trips whose purposes could be trivially deduced, and determining relationships between trip purposes and the socio-demographics of the respondents, as well as time of day when the activity occurred. Since the authors did not have information on the true trip purposes, the distribution of the inferred purposes was compared to that from a regional household travel survey. The results indicated differences in a number of purposes, including private business, work and work-related, shopping, and leisure activities.

Bohte and Maat (2009) also used GIS data (land use and points of interest) to pre-assign trip purposes to participants in a GPS PR survey. The developed survey system applied corrections to the derived purposes based on responses from participants. This rather simple trip identification approach was able to correctly predict purpose for 43% of the collected trips (Bohte and Maat 2009). The ability to accumulate these corrections across multiple participants was later added and pilot tested by Moiseeva, Jessurun, and Timmermans (2010).

Stopher, Clifford, and Zhang refined this method with the addition of several improvements, including the use of frequently visited locations (e.g., home, work, and school) and activity duration (Stopher, Clifford, and Zhang 2007). More recently, improvements have been made to this core method by adding the concept of tours to the data processing and review (Shen and Stopher 2012). The latest version of this method was applied to the GPS data collected in the 2009–2010 Greater Cincinnati regional travel survey, in which every member in the household over the age of 12 was asked to carry a passive GPS device for 3 days. A subset of the households also completed a PR interview where the GPS-derived trips were presented and additional attributes, including trip purpose, were added. The 4,133 GPS trips from the PR data set were used to evaluate the accuracy of the method. The authors found that the accuracy of the base method, without the tour-based corrections, was approximately 59%. The added validation and corrections were able to increase the accuracy to 67%. The authors also compared the derived trip purpose distribution with that generated using data collected as part of the 2009 NHTS and found that the two were not significantly different (Shen and Stopher 2012).

Griffin and Huang (2005) used decision trees, built using the C4.5 algorithm, to identify trip purposes for GPS activity locations. The authors reported a high accuracy in the determined purposes, but the focus of the paper was on the clustering approach used to determine trip end locations from the GPS data stream. McGowen and McNally (2006) demonstrated the use of classification trees and discriminant analysis to identify the most likely trip purpose. The developed models were based on personal, household, and trip attributes. A total of 22 different variables were employed, with the source of the data being the 2000–2001 CHTS. Only out-of-home activities were considered in the developed models, and these accounted for 40% of destinations. The reduced set of purposes contained 26 types, which were also aggregated into five major activity categories. The developed classification trees and discriminant models performed very similarly, with average accuracies in the 73% to 74% range for the major activities and 62% to 63% in the 26 disaggregated activities (McGowen and McNally 2006).

Chen et al. (2010) combined the approaches proposed in Schönfelder and Samaga (2003) and Stopher, Alsnih, et al. (2008) with probabilistic multinomial logit models (MNL). The approach consists of two steps: (1) clustering trip ends into origins and destinations and (2) identifying trip purposes. The second step is performed through two sub-steps; the first one assigns trip purposes deterministically, while the second applies MNL models to calculate the probability of four trip purposes: work/school related, personal business, shopping, and social recreation. Two MNL models were developed, one for home-based trips and another for non–home-based trips. The factors taken in as independent variables in these models fell into three different types: time of day, history dependence, and land use characteristics. The authors reported 67% and 78% match rates for home-based trips and non–home-based trips, respectively (Chen et al. 2010).

Cell Phones, Personal Navigation Devices, Smartphones, and the Emerging Role of Consumer Technologies in Travel Behavior Research

Cell phones, PNDs, smartphones, and other consumer technologies are commonly used across most demographics and are now viable sources of travel data. Unlike documented research studies and applied public-sector data collection, data and methods for collecting and using GPS capabilities found in consumer devices and applications are closely guarded by the private companies that are developing commercial products. Further, consumer products and applications change rapidly, and systems evaluated one day may be irrelevant the next. For this section of the report, attention will be focused on documented studies using consumer products and the most relevant but generalized capabilities of these products for extracting travel behavior details. First, active data collection strategies are presented, followed by the more prevalent passive data collection approach. This section concludes with a sample of current offerings from private data providers.

Research institutions and private companies have developed capabilities to use consumer devices to generate travel behavior data in both an active and a passive manner. Active data collection refers to the use of common personal consumer products to administer a survey or seek direct user feedback. Passive data collection refers to the use of the archived travel data generated by consumer technologies and consumer applications.

From a transportation planning or research perspective, the technology and data access methods are less important than the characteristics of the resulting data set. However, researchers and practitioners must be aware of the limitations of new technology solutions and new consumer data products, particularly before they are applied in a forecasting model. All solutions, regardless of type, require three key considerations for evaluating a consumer data product's ability to accurately identify travel behavior:

1. Identifying data bias—Data from consumer products can be biased because it is the user's choice to purchase and operate the product. Some demographic groups have been slower to use consumer products and, therefore, may not be represented in the data samples. It should also be noted that just because a device may be more prevalent in a certain demographic does not mean that the data has the same bias. For example, travel speed data from a smartphone may not be biased because the user must follow traffic laws like everyone else, thereby negating some of the demographic bias.
2. Verifying data quality—Location and attribute accuracy is important for understanding potential error ranges in subsequent analyses. Furthermore, the consumer product market is very competitive, and independent validation of quality is important until data standards are implemented that provide some level of quality assurance.
3. Ensuring privacy protection—Given that these data sets come from private consumers, special attention to privacy protection must be maintained. Planners and researchers must be assured that they are legally protected and that the original consumers have provided consent. Existing and proposed legislation at the state and federal level is focused on this issue.

Initial questions that arise with consumer products address market penetration. For example, how many people use the data-generating product, and is the user group demographically biased? The Pew Research Center provides market statistics about consumer product usage and technology penetration into American society (Pew Research Center 2012). These statistics are useful in determining any high-level bias that might be intrinsic in the use of consumer products. The following statistics were noted in April of 2012:

- One in five people do not use the Internet. The non-users are biased toward senior citizens, Spanish speakers, those with less than a high school education, and those with low income.
- Those that do use the Internet use it very frequently.
- Eighty-eight percent of American adults have a cell phone.
- More demographic groups are using smartphones as their main Internet access source (including minorities and low-income households that have been traditionally low-level Internet users).
- The use of cell phones is not biased by race.
- Forty-six percent of American adults have a smartphone, and the market share is growing. In less than 1 year, between May 2011 and February 2012, estimated smartphone ownership increased by 11%. The forecast is for this trend to continue.
- Smartphone use is not biased by race, but is biased toward those with a high income and those who are well-educated, urban/suburban, and under age 50.

As transportation planners and researchers identify new methods to access travel data, statistics such as these can provide some insight into survey design and data usage limitations. The use of cell phones and smartphones across different races appears to be unbiased, which is encouraging given that travel data from minorities have been difficult to access using traditional or GPS-based methods (Bricka et al. 2009). It should also be noted that known bias in device usage can be accounted for in survey design. In fact, some of the bias in device usage may aid in capturing individuals that do not respond well to traditional methods.

Active Data Collection from Consumer Devices

Interest in the active use of smartphones for surveying has increased in the last several years as travel behavior researchers have tried to find ways to reduce respondent burden and the equipment costs (and deployment costs) for GPS-based travel surveys. Given that smartphones are typically equipped with GPS (plus accelerometry and motion sensing capabilities) and carried by users almost everywhere, the attraction to tap into these devices for location and human activity information as well is logical.

Before smartphones were widely available, active electronic data collection was possible using personal digital assistant (PDA) devices configured with GPS receivers and other external devices. One early example of such an instrumentation package was the Electronic Travel Diary solution developed for the physical activity sub-survey of the 2000 SMARTRAQ travel survey done in the Atlanta region. The package included a Palm III device and a wearable GPS data logger; custom programming was done to encode the main portions of the travel survey into the Palm device, which was later fused with the GPS data in a post-processing step (Wolf et al. 2000).

The TRAC-IT application is one of the first examples of using smartphones to actively collect travel behavior data. TRAC-IT was developed in 2007 for use in a Florida DOT research project with the objective being to "better understand and pattern household travel behavior for the purpose of educating, promoting, and encouraging households to use alternatives to driving alone." The application required users to actively start and stop trip capture, followed by the input of information such as place type, purpose, mode of transportation, and travel companion counts. The phone captured and transmitted GPS trace information while the trip was active. Using the data collected, the application provided feedback to participants in the form of personalized suggestions on how to save time and money by streamlining travel behavior (Center for Urban Transportation Research, University of South Florida, 2012).

Another good example of active data collection using smartphones is the CycleTracks app, which was originally developed by the San Francisco County Transportation Authority to document bike routes with the purpose of using them to build the model network support for bicycles. The application design requires users to actively start and stop logging GPS data and to associate comment information with the collected trace. The original CycleTracks source code was made available publicly using an open-source license. In 2011, NuStats modified this code base to create a proof-of-concept app called PTV Pacelogger. This effort was done in close collaboration with Portland Metro for use in a pilot survey in Portland within the context of Oregon Household Activity Survey (Bricka and Murakami 2012).

Smartphones have also been used to collect vehicle-based GPS data as part of the Minnesota DOT Mileage Based User Fee demonstration project. In this project, smartphones and supporting equipment were provided to 500 volunteer participants, primarily in Minnesota's Wright County, with the intent of keeping the phone permanently installed in the vehicles for the duration of the study. Participants were asked to use the equipment for 6 months and agreed to pay a mileage-based fee accumulated during the testing period. The collected GPS data were used in the smartphones to estimate charges based on roadway type, geography, and other factors and were not transmitted to the central processing location (Battelle Memorial Institute 2012).

Passive Data Collection from Consumer Products

Transportation planners and decision makers are in a dynamic age of transportation data availability. The rapid market penetration of location-enabled consumer technologies is providing new, nontraditional sources of travel data regarding persons, vehicles, and transportation networks. Mobile phones and personal navigation devices can generate massive amounts of archival data regarding personal travel, and this information is increasingly used by transportation professionals in planning and research. Initial consideration is generally met with enthusiasm regarding the potential of the data to identify detailed travel behavior but is soon tempered with questions regarding bias, data quality, value, and integration into existing models and planning procedures. Further, planners and researchers are facing these issues in a rapidly evolving consumer marketplace where formal guidance is limited and new product offerings are common.

There is a wide range of active research initiatives into social behavior patterns, as evidenced by the digital breadcrumbs created by everyday technologies. The transportation community has been slower than several other disciplines in finding applied roles for consumer data to support traditional transportation planning. Most of the consumer data products currently used in transportation are targeted at supporting traveler information systems and marketed directly to consumers. Within the past few years, consumer data marketed for public agency consumption has found footing and is now poised to play a role in improving understanding of travel behavior and supporting future transportation planning decisions.

These sources of passive travel data are the result of standard technologies that are used in everyday life (i.e., communication, navigation) but also generate archival traces of information. While the list of potential consumer products includes familiar devices such as mobile phones, smartphones, and PNDs, there are other, lesser-known devices that can also be used as sources of travel data. Given that products and product updates come and go with regularity, one way of categorizing each solution may be to base the categories on the fundamental location technology and the data collection method.

Most consumer data that are available to public agencies today are generated for a purpose other than supporting travel behavior analyses. Almost all original products were designed to support real-time traffic data offerings where sensor, GPS, and/or cell phone data were collected in real time and translated into link-level speeds. As this market took hold and as data archives grew, other uses for the data arose. These archival data sets can be extensive with respect to the amount of data collected, the duration of the data collection period, and the geographic coverage. Generally, these data are converted into a data product (e.g., NavTeq traffic patterns data) or a data query service (e.g., such as that offered by TomTom) that are, in turn, offered for purchase. More recently, these products have been adapted to public-sector planning applications promising details of population movement or facility-based performance.

Table 1-6 lists potential technologies, example providers, market focuses, and potential travel behavior values.

From the public-sector perspective, there are a number of attractive advantages and complicating disadvantages when considering these data products.

Advantages of these data sets include that they contain:

- A massive amount of data,
- Data that are both comprehensive and continuous (i.e., many days of data),
- Data that represent real-world experiences/conditions, and
- Raw data that allow for a wide variety of delivered product solutions custom made for certain studies (e.g., speed, origin–destination, delay, and turning movement studies).

Table 1-6. Emerging consumer data sources.

Technology	Example Provider	Primary Market Focus	Potential Public-Sector Interest
In-vehicle navigation device	TomTom	Navigation and real-time traffic information	Transportation system performance, repetitive travel patterns
In-vehicle service	OnStar	Location-based services	Origin–destination data, parking, transportation system performance
Mobile phone tower-to-tower handoffs	AirSage	Traffic data, population movement data	Origin–destination data, population movement, long-distance travel times
Smartphone application	INRIX Traffic	Real-time and predictive traffic information	Transportation system performance, origin–destination data, trip-making patterns

Some disadvantages of these data sets are that:

- They have privacy issues with releasing detailed travel data,
- They are unproven in applied travel behavior models,
- They have a potential demographic bias in source data (i.e., actual demographics are unknown or proprietary),
- They have uncontrolled equipment usage (users can activate or deactivate device),
- They are unable to automatically identify short trip ends with low-resolution data,
- Current products are heavily aggregated and are typically averaged at segment levels,
- Black-box processing (i.e., proprietary algorithms) makes data quality uncertain, and
- There is a dynamic market space—data providers and products might not be around for long, may change content, or may shift focus.

Passive location data are obtained from the user in different ways. Some data are obtained from remote sensors, other data may be directly uploaded from the consumer device, and some data may be volunteered from users. The primary methods of gaining access to passive location data from consumer products are:

- Cellular activity detection by cell towers,
- User-installed application with automatic data upload of location data,
- Imbedded application within product not under user control,
- User-initiated data upload or sharing of location data, and
- Signal detection by sensors (Bluetooth, Wi-Fi, etc.).

From a traveler behavior standpoint, the method of data upload or access is less important than the various data characteristics that define its quality and usefulness, including logging logic, point resolution/frequency, post-processing algorithms, coverage, bias, quality, and privacy protection. One method of validation that can be implemented in these early travel behavior products is to compare results to other travel surveys like the NHTS. Comparing average trip length distributions for certain trip types (home-based work, home-based school, and non–home-based) may be a simple way to evaluate a product's potential.

Early Proof-of-Concept Studies for Passive Data Capture

The earliest use of data from consumer products relied primarily on cell phones and occurred as part of academic studies for traveler information systems. A research paper by Qiu and Cheng provides a good summary of the early history of cell phone use for traveler information systems (Qiu and Cheng 2007). As early as 1996, researchers were exploring the concept of using cell phone signals to find the location of a phone and to estimate its travel speed. These studies were precursors to efforts in using the same technology to measure travel behavior.

Researchers at Kobe University developed methods for extracting travel details from mobile phone data (Asakura and Hato 2004). They developed techniques and algorithms for extracting movement information based on archived cell phone activity data. Their procedures were successfully tested on 100 participants. They also recognized that additional data besides time and location are needed for use in travel behavior applications.

Researchers in South Africa tracked cell phone data for 83 participants over a 2-day period where time and position were updated every 5 minutes (Krygsman and Schmitz 2005). Findings showed that it was possible to use cell phone tracking techniques to generate time and location information for activities. They also noted that substantial effort was required for establishing additional details needed to support the data needs of transport models.

Research papers available at the SENSEable City Laboratory at the Massachusetts Institute of Technology (MIT) include a wide range of concept papers on consumer product data analysis. The earliest papers focused on issues surrounding location-based services and wireless hotspots. Their research expanded with time and now includes a number of papers and findings that provide insight and direction into the future transportation planning roles of consumer data. In 2006, a paper by Ratti et al. identified a vision for using cell phone data for understanding people's behavior (Ratti et al. 2006). The vision was demonstrated with data from Milan, Italy, and showed the possibility of true travel monitoring for both typical and atypical events. Ratti et al. (2007) followed up this research with a study of cell phone movement data in Graz, Austria. In that study, cell phone users volunteered to have their locations tracked for a 24-hour period to show how users' travel could be monitored. These early efforts did not fully address the mass processing and imputation needs but rather focused on the idea that behavior could be identified based on the passive data. Since 2007, the SENSEable City Lab at MIT has had available several additional papers focusing on the technological challenges and opportunities for using consumer data to identify actionable travel behavior information.

In all of these early studies, the focus was on proof of concept with limited sample sizes. All studies mentioned concerns over privacy as a significant hurdle to both the future application and the viability of publicly sponsored research efforts. Most defended against the concept of bias due to the market penetration of cell phones.

Public Agency Applications

The first documented large-scale public agency use of archived consumer product data for estimating travel behavior was in Israel in 2009 (Gur et al. 2009). In this study, data from ITIS Traffic Services Ltd. was acquired by the Israel Department of Transport and contained archived data from 10,000 mobile phones for 16 one-week periods. The project was conducted as part of an effort in building a countrywide travel demand model. The data were used to identify interzonal trips to and from home and other non–home-based trips. While the effort was successful in supporting large-scale movements of the population, it was suggested that the data should be used in conjunction with more traditional survey methods to support the specific modeling requirements of city and regional planning.

In 2011, AirSage was contracted by the Capital Area Metropolitan Planning Organization (CAMPO) in Raleigh, NC to conduct an origin–destination study (AirSage Inc. 2011). Data were collected over a 60-day period from more than 600,000 Sprint phones. Data were tabulated to generate origin–destination matrices for TAZs to support CAMPO modeling efforts. A similar project was conducted for the South Alabama Regional Planning Commission in 2011 (Mobile MPO 2011). Final results from these projects have not yet been published. CAMPO also used AirSage data as part of a speed study for their region and found the results comparable to those obtained using GPS probe data. Some concerns over validation were mentioned.

In 2011, MyGistics conducted an evaluation of time-varying travel demand for a major interchange in Roseville, CA using cell phone data from AirSage (Ma et al. 2012). Zonal trips for the study area were tabulated from cell phone data to estimate the travel demand for peak periods. The results showed that the time-varying demand estimation is possible using passive sources.

In 2012, the Virginia DOT studied the origin–destination patterns of travelers passing along the joint section of I-95 and I-64 (Business Wire 2012). TomTom origin–destination data were used to identify zone-to-zone travel patterns and were then used as inputs into a micro-simulation model of the corridor. The results of this study have not yet been published.

Table 1-7 shows additional project characteristics for known applications of consumer data for travel behavior analysis.

Sample of Current Offerings from Private Companies

Several private companies that sell transportation data generated from consumer devices (and other sources) were contacted regarding their data offerings and were sent custom industry-specific questionnaires. This section briefly summarizes the information provided by each company that responded to these questionnaires. It is worth noting that most responses were provided in the form of marketing materials. The following sections summarize the information gathered from each provider.

AirSage. AirSage has agreements with two of the top three major mobile phone carriers that allow AirSage to gather information regarding the location of mobile phones in real time, which, in turn, enables them to generate real-time traffic data for traveler information systems and to generate archived cell phone movement information to support origin–destination studies. They have a refined and patent-protected process [known as Wireless Signal Extraction (WiSE)] for triangulated signals based on signal strength at cell towers and can classify data points as transient or stationary. Their coverage area includes the continental United

Table 1-7. Summary of efforts applying consumer data for travel behavior analysis.

Study	Data Source	Retrieval Method	Bias Treatment	Noted Data Quality Issues	Stated Privacy Protection
Ministry of Transport and Road Safety, Israel, 2008 (Gur et al. 2009)	Cell phone	Passive cellular activity detection	Overnight locations compared to aggregated demographics Conducted separate CATI survey of population regarding cell phone use	Trip end location limitations, processed data not exactly matching model needs	Legal review by agency lawyers, removal of private information from raw data
Capital Area MPO, Raleigh NC, 2011 (AirSage Inc. 2011)	Cell phone	Passive cellular activity detection	Overnight locations compared to aggregated demographics	Very short trips excluded, some merging of TAZs needed in dense areas	Removal of private information, final products only include aggregated results by TAZ
South Alabama RPC, 2011 (Mobile MPO 2011)	Cell phone	Passive cellular activity detection	Unreported	Unreported	Unreported
Virginia DOT, 2011 (Business Wire 2012)	PND	User-initiated app with auto data upload	Unreported	Unreported	Covered under license agreement
MyGistics 2011 (Ma et al. 2012)	Cell phone	Passive cellular activity detection	Not explicitly discussed, but mentioned "data cleaning"	Unreported	Unreported

States and Hawaii. Most of AirSage's data products are based on all cell phones covered by the carrier agreements that provide these triangulated data points. Other data products are based on location data collected from opt-in devices that allow more detailed tracking (called FastCache). AirSage estimates home and other activity locations based on time-of-day, day-of-week, and location patterns over many weeks. Demographic distributions can then be applied to the estimated home locations.

Bias: AirSage reports that their coverage includes 70% of cell phones in the United States.

Data Quality: Origin–destination information is aggregated to the U.S. Census block level or higher. An internal data validation process is conducted within WiSE.

Privacy Protection: No information about private consumers is extracted to any data products. Exact trip end information is not provided for their origin–destination data; locations are aggregated to the U.S. Census block level. FastCache users have license agreements in place that allow sharing of more detailed information.

INRIX, Inc. INRIX gathers real-time travel information from a wide range of original sources, including GPS/AVL data from fleet vehicles, personal vehicles, roadside sensor-based systems, cell phone data, smartphone applications, RFID, and other sources. They have products that have been generated specifically for public agencies that are based on archived travel condition data. The majority of their solutions are designed to support operations, system planning and measurement, and system optimizations. For travel behavior and model support, archived INRIX speed data have been used for developing baseline speeds by times of day.

Bias: No information provided other than as indicated by their listed sources of information. Generally, roadway speed and delay information is not susceptible to the same impact of demographic bias as trip information.

Data Quality: Historical data products provide statistical distributions to reflect variability. Quality is related to sample size that may vary based on traffic volumes and functional classification. Multiple data sources contribute to data validation.

Privacy Protection: INRIX does not reveal private information from sources; only aggregate data at a road segment or route level are provided.

Nokia/NAVTEQ. Nokia gathers real-time travel information from a wide range of original sources, including GPS/

AVL data from fleet vehicles, personal navigation devices, cell phones, roadside sensors, and other sources. For public agencies, Nokia provides Traffic Patterns and Traffic Analytics that aggregate traffic speeds as needed to support performance evaluation or planning efforts.

Bias: Approximately half of the data sources are private, consumer-based sources, and the other half are fleet vehicle-based sources. Bias is limited because they only provide road segment speeds.

Data Quality: Data quality procedures are applied to all data before they are released into a real-time or archived data product. Multiple data sources contribute to data validation. Quality is related to sample size, which may vary based on traffic volumes and functional classification.

Privacy Protection: No private information is released in any product. Private information is not collected, and unique IDs are periodically assigned in the data processing to ensure anonymity. Nokia private consumer data usage is authorized with direct user acceptance of a data sharing agreement at the activation of a Nokia device.

TomTom. TomTom gathers GPS data from personal navigation devices, and GPS/AVL from fleet vehicles, smartphones, cell phones, and third-party data. TomTom data products serve real-time information in support of traveler information systems and archived historical data for transportation planning and operations.

Bias: The prevalence of consumer devices has indicated a potential demographic bias typical of cell phone market penetration rates. Bias is limited for road segment and route speed information.

Data Quality: Smartphone GPS data are controlled for mode bias by limiting data usage to when the smartphone is docked in a car holder. Multiple data sources contribute to data validation. Quality is related to sample size, which may vary based on traffic volumes and functional classification.

Privacy Protection: Data from consumer products are provided according to data usage agreements with the consumer. Private information is given the highest regard, and no private information is released in any data products. Random IDs are generated at regular intervals in the internal processing of data.

TrafficCast. TrafficCast gathers GPS/AVL data from fleet vehicles and from mobile and fixed-sensor–based sources to generate estimates of real-time traffic information (provided as Dynaflow). TrafficCast also provides Bluetooth sensor systems and data delivery for interested agencies (provided as BlueTOAD). Travel-behavior–related data products include archival data for estimating baseline speeds and OD data at setup using Bluetooth sensors for specific study areas.

Bias: There is potential bias in traffic data given the heavy reliance on fleet vehicles. Bluetooth sensor data are biased toward drivers that are more likely to have Bluetooth-enabled devices. In both of these situations, traffic conditions mitigate the impact of final product bias.

Data Quality: Quality metrics are used internally to identify potential sensor or data-quality issues. Quality is related to sample size, which may vary based on traffic volumes and functional classification.

Privacy Protection: No private information is gathered from any of the information sources.

Privacy Protection in Consumer Data Sets

As noted previously, a wide range of information can be gathered about travelers moving on transportation networks through various data collection methods. As the capabilities of sensor-based systems and consumer products advance, more personally identifiable information is accessible and, therefore, comes with an increased risk of exposure of private information. In the last couple of years, it was revealed that most phones, GPS devices, and smartphone applications have stated and unstated capability to archive and upload information about usage, including location data. The hardware and software companies controlling these elements protect their data access with license agreements that users must agree to prior to operation. These companies are also very sensitive to the concerns of their customers and generally avoid risking their primary markets. Therefore, while the idea of selling consumer data has some attraction, there is some hesitation to providing any sort of private data to a government agency.

At the time of this writing, there were three proposed legislative acts intended to control government access to private data collected from consumer products: the Geolocation Privacy and Surveillance Act of 2011 in the U.S. Congress, the Wireless Surveillance Act of 2012 in the House of Representatives, and the California Location Privacy Act (which passed with bipartisan support in August of 2012).

The Geolocation Privacy and Surveillance Act is a bipartisan bill that provides "a legal framework designed to give government agencies, commercial entities, and private citizens clear guidelines for when and how geolocation information can be accessed and used." The primary focus of this act is to prevent law enforcement from tracking individuals without a search warrant except in cases of emergency. More importantly, the bill also addresses private companies' use of consumer data and limits their use unless there is explicit consent by the individual. Passage of this bill may have some impact on the availability of some consumer data sets, particularly those with blanket or third-party data access agreements.

The Wireless Surveillance Act of 2012 is designed to limit access to electronic communication. This act, less important to travel behavior research, addresses email and phone communications. The recently passed California Location Privacy Act of 2012 makes it mandatory for law enforcement agencies to obtain a search warrant before gathering any GPS or location tracking data from a personal cell phone or other device. While most of the attention on privacy legislation is in law enforcement, there are definite concerns in all three (and in older efforts) regarding data access by any public agency.

One of the first transportation planning studies to address the issue of data privacy was the Connected Vehicle Road User Fee Test (Pierce et al. 2011). The researchers were confronted with public concern about the tracking capability of the technology to be used. The issue and the project's focus on privacy protection identified that the concerns were valid and that other transportation planning studies should have thorough procedures for checking legality and providing the utmost security for private information.

The types of private data that can be accessed vary by device.

Table 1-8 details the type of information that can be collected by detector type. There are several categories of agreements between users, service providers, and data collectors when data are actively or passively collected. In some cases there is no agreement at all, as in video surveillance typically purposed for traffic and incident management.

GPS and Cell Phones. Once selective availability, which refers to the degradation of GPS signals, was eliminated in 2000, the accuracy of GPS devices improved from 30 m to 100 m in 2000 to 5 m to 10 m by 2003 (Zmud and Wolf 2003). The information that can be collected from GPS devices has such precision that much information can be derived from the time-stamped position data. Another move by the government that affected location capabilities by cell phone rather than GPS technology alone was the development of E-911, where the Federal Communication Commission required all cell phone manufacturers to have built-in location-detecting technology that would allow wireless network operators to provide latitude and longitude information on callers within 300 m to support emergency response. This requirement was part of the Wireless Communications and Public Safety Act of 1999 (Wolf 2000; Karim 2004).

As a result of the E-911 mandate, cell phones are now equipped with a number of location-sensing technologies. The location can be determined by GPS, Wi-Fi, Bluetooth, and their interaction with the wireless network. Data collected using the first three detection types can be controlled by the owner by altering the settings on the phone. Data collection via the wireless network interaction, however, is done so that anytime a phone is turned on and interacts with the wireless network in some way, whether a tower-to-tower handoff or the initiation of a phone call or text, the location can be determined through triangulation. This location detection is done passively and has been processed by private companies for real-time traffic and population movement applications.

Data from smartphone applications is an area of location detection that is advancing swiftly. People agree to certain terms to use the services that require location detection and/or sharing. Many location-based services such as navigational apps (e.g., Google Maps) and social networking applications (e.g., foursquare) are very popular, and the location data generated by these apps can be archived and resold according to licensing agreement terms.

RFID, Bluetooth, and Wi-Fi. RFID systems can detect the location of travelers when their assigned RFID tag is within range of a receiver. Generally, this allows the provider of the RFID tag to look up private account information for the person subscribed to the unique RFID. This capability can be employed for applying and enforcing road use-based fees.

Bluetooth readers are much less obtrusive and function by identifying the media access control (MAC) address as

Table 1-8. Individual information accessed from consumer products and detectors.

Location detector	Individual information
Video surveillance	Vehicle – location, vehicle type, speed, time, occupancy Pedestrian/transit – location, time, activity, company License tag ID – can be linked to vehicle registration data
Bluetooth	Location, time
GPS device (includes PND)	Location, speed, route, frequent locations, time, acceleration
RFID	Location, speed, time
Cell phone	Location, speed, time
Smartphone application	Location, user information, context
Transit smart cards	Origin, destination, frequented stations/stops, and times

Bluetooth-equipped devices pass by a sensor. By pairing the MAC address with observations downstream, the travel time and speed information can be generated. A similar approach is possible for devices emitting Wi-Fi signals.

When transit fares are collected from passengers via electronic fare media, records of passengers' travel times (i.e., when their transit trips started and ended), their transit trip durations (i.e., how much time it took to travel within the transit system), and routes used are automatically captured. Furthermore, at the transit vehicle level, boarding counts (as well as alighting counts for some transit systems) are also stored. With this information, transit agencies know when, where, and how their passengers use the system; in turn, the transit agencies are able to monitor and improve their services. Electronic fare media can also have additional information associated with the user, such as school, employment, other organization, age, credit card details, and home address.

Data Processing Techniques to Reduce Privacy Exposure

There are ways to process the data to reduce the ability to trace it back to a specific household or person. Almost all processing of data includes a step that renders the data anonymous. However, even though the data are anonymous, it is still possible to use accurate location information within the data to identify a person's home location, work location, and other frequently visited places. A recent research project conducted by Elango and Guensler explored two traceability-reducing post-processing techniques for GPS-collected data prior to distribution (Elango and Guensler 2011). Both techniques involve creating polygons around home locations and trimming the location data from within that polygon. By keeping the authentication and filtering process in the communication servers separate from the analyzing processes of the traffic servers, the privacy protection of users can be ensured while still maintaining data integrity (Hoh et al. 2006).

License Agreements for Secondary Use

For many users and service providers, it is important that information collected cannot be traced back to the individual user unless explicitly authorized by the user. To protect the information that can be obtained from cell phone usage, Congress passed the Telecommunications Act of 1996, which contained Section 222 that requires telecommunication customer approval before customer proprietary network information is distributed to third parties (Karim 2004).

The following user agreement excerpts were found on different manufacturer or service provider websites regarding the use or reuse of consumer data on April 27, 2012:

> "We collect information about your use of our products and services. Information such as call records, websites visited, wireless location, application and feature usage, network traffic data, service options you choose, mobile and device number, and other similar information may be used for billing purposes, to deliver and maintain products and services, or to help you with service-related issues or questions."

> "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."

> "This type of information may be aggregated or anonymized for business and marketing uses by us or by third parties."

> "We may also draw upon this Personal Information in order to adapt the Services of our community to your needs, to research the effectiveness of our network and Services, and to develop new tools for the community."

> "We receive and store any information you enter on our Service or provide to us in any other way. . . . We automatically receive your location when you use the Service."

> "We may automatically collect location information from your mobile device, but such information will not be directly linked to a specific person. Your location data will only be provided to us in accordance with Terms governing your app, and will then be aggregated with other data."

> "These companies, often called ad servers or ad networks, may place and access cookies on your device to collect information about your visit on our websites."

> "We may put together your current city with GPS and other location information we have about you to, for example, tell you and your friends about people or events nearby, or offer deals to you that you might be interested in. We may also put together data about you to serve you ads that might be more relevant to you."

These excerpts exemplify the variety of ways companies inform users of how their personal data may be accessed, used, and shared with others. However, consumers are often unaware of these terms, do not know where to find these terms, or do not understand the implications of possibly ambiguous terms such as "to develop new tools for the community" or "to deliver and maintain products and services." The ambiguity in terms combined with the variability of policies across license agreements adds to the inconsistency and future unreliability of data sources from consumer products.

Fixed-Location Sensors

Application of Fixed-Location Sensors to Transportation Data Collection

Fixed-location sensors are devices that are positioned along a transportation system and have a short-range detection capability. Historically, license plate surveys and video capture have been used to support OD and travel time studies when information was needed for the modeling of specific transportation facilities or areas. The introduction of RFID and Bluetooth sensor technology allowed the same types of studies to be conducted with reduced labor cost, increased accuracy, and potentially larger sample sizes (due to increased study durations). With enough sensors along a transportation system, travel patterns can be identified and developed into OD matrices that can be used to support travel demand models. Since Bluetooth sensors only capture observed travel times between fixed locations without socioeconomic or demographic information regarding drivers, the data provided are limited to short-term modeling and generally are only applied in specific situations. RFID sensors, on the other hand, are typically used for toll tags or transit smartcards that individuals carry for their travel needs and hence have a connection to a specific user. This connection allows the potential to use other data regarding that person that may be part of a customer database. Further, the customer information also provides an ability to contact that person for follow-up surveys, granted that previous consent was provided.

In addition to these current technologies, there is a significant amount of research regarding connected vehicles. The U.S. DOT is sponsoring research, known as the Connected Vehicle Initiative and the Smart Roadside Initiative, that will allow data transmissions between vehicles [vehicle to vehicle (V2V)] and from vehicle to infrastructure (VTI), where infrastructure includes roadside control systems and sensors. Conceivably, these data could also be used for travel behavior analysis when they become available. Data streams from a fully implemented system would capture vehicle trips. Like most of the vehicle sensing technologies, travel from nonmotorized modes would not be captured. Regardless, these initiatives have the potential to be powerful data collection sources for anonymous vehicle trajectory data.

Bluetooth Technology and Data Collection

The Bluetooth protocol is widely used for exchanging data over short distances from fixed and mobile devices. Bluetooth technology has been used in transportation data collection since 2007. Bluetooth sensors can be fixed along roadways, nonmotorized transportation facilities, onboard transit vehicles, or a number of other pathways. The sensors or stationary receivers can detect the presence of vehicles or individuals with Bluetooth-enabled devices (when the device is in discoverable mode) such as in-vehicle navigational devices, mobile phones, and wireless headsets. The receiver does not collect any information other than the unique MAC address of the device and the time of the observation. As a person or vehicle travels along a network link with multiple sensors, its signal is detected by each sensor generating a trip path, trip time of day, and a trip travel time. Not all individuals or vehicles have active Bluetooth transmitters. Estimates of detection percentage range from 3% to 8% in the multiple studies reviewed over the last few years (Lee, Agnello, and Chen 2011; Voigt 2011; Bullock, Haseman, and Wasson 2010). Since the sensors operate independently, they can record data for long periods to capture enough data to support analysis needs.

There have been applications of Bluetooth devices around the country to support transportation initiatives. Most of these applications focus on performance evaluation and traveler information systems. A few, however, have used Bluetooth technology to understand travel behavior and to feed modeling efforts. A recent Virginia DOT study compared OD data from traditional video data collection and Bluetooth (Lee, Agnello, and Chen 2011). Both approaches were used to identify travel patterns of vehicles entering and leaving the Richmond, VA area (external–external, external–internal, and internal–external counts). The Bluetooth capture rates varied from between 3.73% and 5.82%, while the video captured between 52% and 88% of vehicles. A comparison of trip tables showed significant differences and led the researchers to believe that the video capture was more reliable since the Bluetooth sensors have smaller capture rates and much more signal noise in the data. Another application of Bluetooth sensors for identifying external travel was conducted by the Texas Transportation Institute in Houston (Voigt 2011). The study explored a broad deployment of sensors along different road classes to support a traveler information application. It found capture rates for Bluetooth sensors as high as 20% and determined that capture rates were increasing over time.

A study by Bullock, Haseman, and Wasson was conducted over a 12-week period in 2009 along the I-65 corridor in Indiana and collected 1.4 million travel times (Bullock, Haseman, and Wasson 2010). Portable Bluetooth devices were placed along the main line of the Interstate with semi-permanent ones placed along diversion routes. The possible uses for the data included determination of travel delay times, driver diversion rates, and work zone mobility performance. Bluetooth was also used in a Florida DOT study to identify traffic pattern changes related to a new interchange in Jacksonville (Carpenter, Fowler, and Adler 2012). Fourteen sensors were deployed for 1 week to identify OD matrices for the primary access points in the study area. The resulting OD travel time matrix was also used to validate the final model.

One of the most interesting future travel behavior applications of Bluetooth is its potential role in the Connected Vehicle Initiative research program. The Connected Vehicle Initiative, originally envisioned to improve safety and reduce congestion, is a large research initiative involving intelligent transportation systems (ITSs) that allow vehicle-to-vehicle and vehicle-to-infrastructure communications. Bluetooth is one of the communication technologies being considered. If implemented on a wide scale, sensors could detect vehicles and gather travel path information.

Another interesting study explored the use of multiple Bluetooth sensors to identify in-home activities. Schenk et al. (2011) explored the combination of collecting data from both Bluetooth devices and smartphones to build a complete spatial and temporal "lifespace" pattern. Each of the participants carried a smartphone, which collected GPS data and transmitted a signal to a Bluetooth receiver installed in their homes. For two of the participants, 30 days of data were recorded, and for the remaining participants only 21 days of data were recorded due to battery limitations. While the study was successful in collecting detailed activity data, the need for installing multiple fixed sensors prohibits wide-scale deployment for large-scale behavioral studies. If, however, existing consumer devices can capture and record the same signals, then this approach could become more feasible.

In essence, Bluetooth will continue to play a role in providing OD data for modeling. However, it is likely that this technology will be limited to small-scale or corridor implementations where models are needed to evaluate changes in geometric or operation conditions. Broader policy and behavioral changes will likely not depend on this technology in the near future.

RFID Technology and Data Collection

RFID is a technology that is applicable to a variety of fields for the purpose of tracking vehicles, people, and goods wirelessly. For transportation planning, RFID allows organizations and agencies to passively collect data about the tag-equipped users of a roadway or transit facility. Similar to Bluetooth, the user must pass within a certain distance of an RFID sensor for detection. The primary difference between RFID and Bluetooth is that the RFID technology is deployed in transportation infrastructure as a means to identify users of a particular service such as toll collection, parking, and transit fare capture. This implies that additional information such as home address can be linked to any RFID-derived travel data. This added capability provides for a deeper understanding of the socio-demographics of the travelers that pass a sensor and affords the opportunity for follow-up contact. Transit and toll agencies have used this capability to track system performance, to measure demand, and for user satisfaction surveys.

Another implementation of fixed-location sensors is the automated fare collection (AFC) system. AFCs are being used in a variety of transit systems such as MARTA, the NY Metro, and the Chicago Transit Authority. Research into using these systems as a potential replacement for traditional OD surveys is currently under way (Munizaga, Devillaine, and Amaya 2012; Chakirov and Erath 2012). Generally, AFC cards require that a passenger tap a fare box at a train station or on a transit vehicle to debit stored value from the card. Some systems require that the passenger tap out to exit a transit station. Most of these systems are the exclusive means for transit passengers to use for fare payment and thus are used by nearly 100% of riders. The location of the fare boxes can be geocoded, allowing for the boarding and alighting data of passengers to be accurately collected or recorded. In cases where the AFC is integrated with an AVL, bus and light rail transactions may also be geocoded with a reasonable degree of accuracy. Registration of cards is an option but is not compulsory. Such registration does not require the passenger's home, work, or school address to be provided. However, the lack of these attributes does not necessarily mean that data about origins and destinations cannot be deduced. With the addition of land use data, there is a reasonable means for determining whether a trip is originating in an area likely to be the home, work, or school location.

Managing Large Data Sets

The advent of high-frequency GPS logging applications for transportation in the early 2000s brought new data management challenges to transportation researchers and practitioners. This type of data has the potential of being quite large. For example, a person traveling for an average of 90 min a day for a week will generate 37,800 GPS points at a 1-s logging resolution if points are only collected during travel. If one then tries to log the same amount of travel from 1,000 persons, there would be 37.8 million points. Translated into disk space requirements, this number of points would require 2 to 3 gigabytes (GB), depending on the number of attributes stored per point, the resolution, and the level of indexing. However, if no filtering of non-travel points is applied, these numbers would be increased 16-fold.

When compared to today's large disk drives, with capacities measured in hundreds of gigabytes, storage requirements in the 2-GB to 3-GB range (or even 32 GB to 48 GB if all points are logged) may not seem like much, but one needs to realize that the real challenge is to be able to effectively retrieve, clean, process, visualize, and attach attributes to groups of records. The current state-of-the-practice approach to solve this problem is to use server-based relational databases (Wolf, Schönfelder, et al. 2004; Oliveira et al. 2011) to store the GPS data in context. The availability of open-source server-based relational databases such as MySQL and PostgreSQL has made this an affordable and popular solution.

One approach to make these large data sets more manageable is to apply compression and filtering schemes at import time. For example, when the objective is to measure travel, one can simply filter out speeds below a minimum threshold. Other strategies for reducing the number of records stored (and therefore memory/storage requirements) include sampling the data at lower frequencies (e.g., record only a point every 10 s instead of every second) and applying data simplification or compression procedures such as SQUISH (Muckell et al. 2011). Unfortunately, these latter approaches result in loss of information, which may not be a problem for immediate uses of the collected data but may limit future secondary applications.

Best practices should be followed when using a relational database to store large data sets. These include using appropriate field data types to keep record sizes small, ensuring that tables have primary keys defined, and applying normalization to schema. Expected data storage and processing needs should be used to guide the selection and configuration of a relational database platform, including the hardware on which it will run. From a hardware standpoint, it is important to use servers that have as much random access memory (RAM) as the budget can afford, to use servers with a redundant array of inexpensive disks for both performance and failure protection, and to have good backup and restore procedures in place.

Despite being capable of storing, managing, and processing very large data sets, relational databases have architectural limitations that make them unsuitable for keeping data sets whose size is measured in terabytes. This magnitude of data is becoming increasingly common with the advent of continuously monitored data sources such as those data sets generated from permanently instrumented vehicles. For example, the American Trucking Research Institute records approximately 4 billion position data points from commercial trucks annually (Bernardin et al. 2012). A research project in Singapore reported on a comparable data set with over 4 billion GPS observations coming from approximately 15,000 Singapore taxicabs (Koh, Nguyen, and Woodard 2010). This specific data set occupied over 300 GB of disk space and was loaded in a PostgreSQL database.

The current trend in household travel surveys toward data collection methods that are primarily based on GPS methods or sources is likely to increase the amount of data that needs to be managed (Giaimo et al. 2010; Oliveira et al. 2011). Smartphone applications that allow participants to collect GPS and other sensor-based data (such as accelerometer data) within existing travel surveys are also likely to reduce the data acquisition costs associated with the deployment of specialized devices while contributing to the creation of significantly larger large data sets (Bricka and Murakami 2012).

A common practice used when managing and processing larger data sets is segmenting (also known as partitioning or sharding) the accumulated data into smaller units (Nemala 2009). Partitioning the data in this manner allows traditional database software to find and process records quickly by loading much of the data into RAM. Data partitioned in this manner can also be placed on different servers, which allows the resulting system to scale out to meet demand. Unfortunately, when data are segmented in this way, additional steps are required to conduct analysis over the entire data set. This added management overhead increases as the size of the data set in question grows and more segmentation is needed.

To deal with these issues, Google developed distributed storage and processing systems using commodity (i.e., inexpensive and numerous) computer clusters running Linux and capable of handling hardware failures through the use of redundancy when partitioning the data on the cluster (Chang et al. 2006). Adding computers to the cluster would automatically increase processing capacity and reduce total run time (also referred to as scaling out).

These new technologies make it possible to run very large data crunching jobs in a reliable and efficient manner, allowing software engineers to focus on the algorithms and processing logic instead of the management overhead associated with these types of tasks. The research papers Google published on this approach inspired other companies and individuals to develop software implementing the same approach. Some of these efforts, in turn, became open-source projects, and a community grew around them.

As a whole, these new technologies have been referred to as "big data." The storage solutions that drive these new data management solutions are collectively called "noSQL" and are commonly combined with a programming approach called MapReduce. These new data storage technologies focused on simpler data structures (typically consisting of key value pairs) as opposed to the more complex representations used in relational data modeling. This is explained by the fact that most of the initial applications consisted of processing document-based data such as web pages and web server logs. Table 1-9 shows examples of popular open-source big-data technologies.

Finally, the recent emergence of cloud computing has made it possible to rent large computer clusters on demand at reasonable costs. The combination of this new availability and big-data technologies has made it possible for small organizations to tackle sizable data management and processing jobs. The flexibility inherent in cloud-based solutions also allows computing resources to be added and removed as needed.

For example, StreetLightData is a start-up company which is developing technology for processing massive amounts of GPS data collected using smartphones. The processed data will then be used to develop site selection and planning data products. According to a recent interview given by its chief technology officer and founder, Paul Friedman, StreetLightData is using a Hadoop-based big-data cloud solution from Claudera to batch process initial processing done using several servers running PostgreSQL (i.e., the initial database is segmented). StreetLightData's website is http://www.streetlightdata.com/.

Table 1-9. Popular open-source big-data technologies.

Name	Description
Hadoop	Apache Hadoop is an open-source software framework for data-intensive distributed applications and was originally created by Doug Cutting to support his work on Nutch, an open-source web search engine. It is currently one of the most popular frameworks for distributed processing.
Cassandra	Apache Cassandra is an open-source distributed database management system developed by Facebook to power its Inbox Search feature. Facebook abandoned Cassandra in favor of HBase in 2010, but Cassandra is still used by a number of companies, including Netflix, which uses Cassandra as the back-end database for its streaming services.
HBase	Apache HBase is an open-source, non-relational columnar distributed database designed to run on top of Hadoop. It provides fault-tolerant storage and quick access to large quantities of sparse data. HBase is one of a multitude of NoSQL data stores that have become available in the past several years.
MongoDB	Created by the founders of DoubleClick, MongoDB is another popular open-source NoSQL data store.
CouchDB	Apache CouchDB is still another open-source NoSQL database. It uses JSON to store data, JavaScript as its query language, and MapReduce and HTTP for an API.

Source: http://www.networkworld.com/slideshow/51090

Survey of Industry Experts

To assess data needs, current capabilities, and future directions of both transportation data providers and users, customized questionnaires were sent to industry experts who worked for companies or organizations that (1) collect or analyze GPS data for travel behavior research, (2) use GPS data for travel behavior and activity modeling, or (3) sell consumer travel or traffic data. Each person selected to complete a questionnaire is considered an expert in his or her respective discipline/category based on the research team's knowledge, available publications, and conference presentations. Representatives from different firms and research organizations were selected within each industry category to give adequate coverage and to minimize bias.

Table 1-10 provides a complete list of all questionnaire respondents, with their company affiliation, organization or firm affiliation, and industry group. The questionnaires

Table 1-10. List of experts by industry.

Industry	Organization or Firm
Travel Survey Practitioners	Abt SRBI Battelle Memorial Institute ETC Institute GeoStats NuStats Resource Systems Group (RSG) University of Sydney Westat
Travel Behavior Researchers	Argonne National Laboratory/UIC Delft University of Technology ETH Zürich FHWA / USDOT IFSTTAR (French Institute for Science & Technology of Transport, Development & Networks) Texas Transportation Institute Tokyo Institute of Technology University at Albany University of Tokyo

Table 1-10. (Continued).

Industry	Organization or Firm
Transportation Planners and Modelers	Atlanta Regional Commission Cambridge Systematics, Inc. Chicago Metropolitan Agency for Planning Chicago Transit Authority Jerusalem Transport Masterplan Mark Bradley Research and Consulting Metropolitan Council (Minneapolis & St. Paul) Ohio DOT Parsons Brinckerhoff, Inc. San Francisco County Transportation Authority Texas Transportation Institute
Traffic Data Providers	AirSage INRIX, Inc. Nokia (NAVTEQ) TomTom, Inc. TrafficCast International

sent to these industry experts were designed to collect both current and future plans for GPS data use or provision, and were customized for each industry (see Appendix B for each questionnaire). As responses were received, follow-up contact was made as needed for clarifications or to collect reference material mentioned in the response. These references were then reviewed and included if relevant in the literature review synthesis.

Summary tables have been created that contain the main themes and responses received for each question, by industry, along with a few key, representative quotes. The responses appear in Table 1-11, Table 1-12, and Table 1-13, respectively.

It should be noted that the traffic data providers did not respond to the questionnaires directly; instead, many of them simply copied marketing information into the questionnaire itself. Consequently, there is no summary table of responses for this last industry category. Instead the marketing information offered by the traffic data providers has been integrated within the relevant literature review sections of this report and their complete individual responses are provided in Appendix C.

Table 1-11. Summary table of travel survey consultant responses.

Question	Summary	Relevant Quotes
1. Current use of technologies	• All use passive vehicle or personal/wearable GPS loggers • Surveying purposes include: – Correction of trip diaries (still the predominant use) – As supplemental data to trip diaries (for more accurate timing/location info) • As sole source of travel information with either passive diary creation or used to generate prompted-recall questionnaires for nonspatial-temporal information • Other uses include: – Driving behavior analysis in response to policy intervention at either the aggregate (i.e., road pricing – Minnesota) or individual level (TravelSmart – ITLS) – Health and physical activity surveys, either to match location to physical activity (2012 Nashville Transportation and Health Study) or to correct self-reports – Vehicle emissions and fuel consumption studies [GPS used in tandem with engine sensors (i.e., California Energy Commission interest in 2012 CHTS)] – Real-time vehicle information studies (i.e., bus tracking) – Survey administration – selection of intercept sites, track surveyors, etc.	• "We have used GPS in a range of travel behavior surveys over the last decade including pilot tests for conventional household travel surveys, evaluation of travel behavior change interventions, and in-vehicle driving behavior studies."

(continued on next page)

Table 1-11. (Continued).

Question	Summary	Relevant Quotes
2. GPS plans for near/long term	The near-term plans for GPS data collection mostly involve three main thrusts: • Enhanced processing and imputation algorithms utilized in GPS data processing • Sensor fusion or use in tandem efforts [i.e., adding accelerometers or OBD sensors (ITLS, GeoStats, Westat)] • Transitioning to smartphone data collection, either as a survey application for respondents who have smartphones or possibly as replacements for GPS loggers	• "We are ... investigating [the] potential of smartphones, which I think are the future in this space." • "We foresee more GPS-only designs that leverage data processing and imputation algorithms to derive trip details." • "We are actively testing/fielding a new smartphone application."
3. Impact of GPS use on participation rates and sample represent-ativeness	• In general, not a lot of information provided on impacts of GPS data collection on participation or response rates. • Representativeness is not generally found to be an issue, with GPS either having no effect or actually increasing sample representativeness (at least in the ITLS/Sydney case and in the 2010–2011 NYC regional travel survey). Establishing representative samples for household travel surveys is often a requirement and handled by proprietary methods (Abt SRBI). • Anecdotal evidence is mostly given, with wide variance depending on the type of study. Short-term data collection replacing or supplementing diary surveys has generally found either no change (ETC Institute, GeoStats) or an increase in participation or at least compliance (ITLS, Westat). On the other hand, longer-term surveys, which require more from participants, seem to be more challenging to recruit for, as was found in the Battelle road pricing study and was observed by GeoStats.	• "Recruit and retrieval rates seem to be most impacted by the overall level of burden associated with survey participation. Offering more options for participation . . . increase[s] both participation rates and sample representativeness by bringing in different population groups." • "While we can't isolate the impact of the GPS per se on recruitment, anecdotally it had both positive and negative impacts."
4. How do you process GPS data to generate deliverables?	• Mostly proprietary algorithms • Most GPS-enhanced travel surveys reported following some variation of clustering points about stops and segmenting traces into separate trips or trip segments. • Either fully automated or automated with manual review and correction – ITLS algorithms available for review (Stopher et al. 2012) – Other survey purposes require less processing—for example speed limit studies, road pricing impacts (which just get VMT, etc.), and some driving behavior studies. • Simplified methods for vehicle-based data collection (not multimodal, which tends to produce messier data due to continuous power supply and more line-of-view obstructions with GPS satellites)	• "Most travel surveyors/firms have some algorithms to process the retrieved household travel survey GPS data from participants. These processes include determination of origins and destinations, travel paths, travel speeds, and travel modes."
5. GPS-based travel behavior details included in deliverables	• Travel surveys generally provide, at a minimum, the trip segments (by mode if using person-based devices), and calculated origins and destinations for trips (corrected by diary data if a combined survey). • Some provide link-matching details from GIS, which can correct inaccuracies in GPS data and link the trip records to detailed network information. • Others match to TAZ or census tract. • Most also provide speed, distance, and travel time from the GPS data. • For combined sensor surveys (i.e., health studies), GPS data can be combined and enhanced with other sensor data (i.e., accelerometers). • Many now include automated data imputation for modes (most common) and imputed purpose (if the survey did not include diary collection). • Driving behavior data are extracted from GPS (i.e., VMT, start/end times).	• "Our standard GPS data deliverable contains tables with complete details on GPS households, GPS persons or GPS vehicles, GPS trips, GPS trip segments, GPS points, and network links matched to GPS points. For dual-method (diary comparison) studies, we also provide tables with GPS trips matched to reported trips and missed trip analysis results."
6. Methods for privacy protection	Privacy and security enforcement is generally maintained through the following mechanisms: • Institutional controls [such as institutional review boards (IRBs), ethics guides, confidentiality agreements, human subjects training] • Physical security – secured storage, protected databases • De-identification – removing identifiers from travel information, including names and addresses • Data separation This list is similar to solutions in all types of surveys. GPS raises new issues not addressed by the above, such as: • Fuzzification of trip start/end (i.e., add random error, round to zone centroid, etc.) • What to do about pattern information that can be uniquely derived • Concerns about privacy can be mitigated as in thick-client paradigm (i.e., equipment collects and aggregates the data before transmission, actual raw data never sent).	• "One consistent prevailing theme in this research is the concern expressed by the general public regarding privacy, or more specifically how studies like this could result in a reduction in personal privacy. Frequently, these fears are manifested through concerns of Big Brother type statements or how these studies would enable the government to track the movement of individual citizens." • "We are concerned about the release of raw data and would welcome some easily processed method of hiding the specific locations in the data for the purpose of releasing the data."

Table 1-11. (Continued).

Question	Summary	Relevant Quotes
7. GPS coverage and accuracy compared to other methods	• Clear improvements in collecting spatial–temporal data [i.e., location within 5 m at every second (provided the device is charged, taken, and used correctly)] – Charging and carrying along are primarily issues with personal loggers. – Accuracy and coverage issues remain due to cold start delays/satellite acquisition times, sky blockage, and urban canyon issues. – Some issues may be addressed with improved technology (i.e., differential GPS, secondary sensors) • Studies consistently show that GPS-collected travel surveys are more accurate than traditional diary surveys in terms of trip reporting. • Technology continues to improve. • Positional accuracy is now to a level that link identification, even between parallel roads, is possible with a high degree of accuracy.	• "Accuracy is clearly far greater than in diaries. People are notoriously bad at estimating the times at which they travel, how long they travel, and certainly how far they travel." • "Each generation of GPS units is better than the previous generation. For example, the delays in getting signals have been almost eliminated."
8. Advantages of GPS	• Advantages largely relate to the previously identified improvements in – Accuracy of trips, – Detail of trips, and – Reduction in respondent burden. • Additionally, the greater level of trip detail available allows other survey purposes beyond travel surveys, as mentioned (i.e., road pricing, travel behavior modification interventions, etc.). • GPS data support next-generation travel models by providing more detailed data (e.g., exact routes, fine-grained location data) • It is possible to reduce sample sizes through longer data collection periods, enabled with burden reduction	• "There is less respondent burden for capturing travel details while collecting more information and more accurate information." • "[GPS] is ubiquitous and available using off-the-shelf technology that is widely popular in the U.S. (i.e., a cell phone). No special equipment is needed. The phone almost always accompanies the individual and thus permits capturing both vehicle and non–vehicle-based travel."
9. Limitations/ concerns about GPS	• Major concern exists regarding bias introduced through using GPS (i.e., respondents with privacy concerns regarding the collection method may select themselves out of the survey) • Accuracy of processing algorithms, and time required to manually correct or the extra burden introduced for respondent correction • Battery life • Costs, compounded by potentially high loss rates of equipment • Data loss due to compliance, device failure, or environmental limitations (i.e., weather) • Cost	• "The final GPS data is still only as accurate and reliable as the individuals who have been recruited to carry the GPS devices." • "Perhaps the most salient limitation of GPS is cost. Like most technology, the cost of units is decreasing; however, the cost of acquiring, maintaining, deploying, and retrieving enough units to conduct a large-scale data collection effort is still high."
10. Pricing information	• Data collection firms are reluctant to share cost information. • GPS survey costs can range from slightly higher to much higher than traditional surveys, depending on scope of survey. • GPS device prices have decreased dramatically over the past decade, and use of smartphone apps may further reduce costs. • Low marginal cost of extending data collection period once devices are in place • Can make costs comparable to traditional data collection with longer deployment periods and smaller sample sizes. • Not relevant for surveys that can only be accomplished by GPS (behavior modification, road pricing, etc.)	• "Costs are rapidly becoming comparable to those for equivalent conventional survey procedures."

Table 1-12. Summary table of travel behavior researcher responses.

Question	Summary	Relevant Quotes
1. Current use of technology	• Many have used GPS data collection devices that include in-vehicle, wearable GPS loggers, smartphones, PDAs, or in-tablet PCs. • Others have conducted GPS-based speed studies and employed GPS technology in bicycle studies, parking studies, traffic operations, and transit scheduling/planning to evaluate vehicle drive cycles and related emissions and to obtain route choice behavior. • The ubiquitous presence of GPS and other embedded technologies in smart phones and PDAs has turned those devices into great travel logger tools. • GPS logs have been used to obtain detailed trajectory of travel, travel mode, departure and arrival times, origin and destination, and (even) trip purpose. • One approach is to combine a GPS-enabled mobile phone with a web-based prompted-recall travel survey. • The GPS log data have been further enhanced by the use of other technologies, such as three-dimensional acceleration data (using accelerometers), and air pressure data for estimating detailed travel behavior, such as the microscopic movement for horizontal direction and the movement in a room. • Several recent studies have employed a combination of a GPS data logger and an accelerometer followed by a prompted-recall instrument on the web. • Others have used GPS on an experimental subsample (e.g., 10%) of a larger group participating in traditional travel diary surveys. • GPS devices have provided the opportunity to collect passive data for longer durations and multiple days, thereby allowing for the collection of additional data (e.g., processes data) that were otherwise not presented to avoid survey burden. • Availability of GPS data loggers with flash memory for data storage, as opposed to other devices that require proprietary software, has eased the process of data extraction, avoiding the need to collect and process the GPS data in separate steps.	• "The combination of GPS + web has made it possible to obtain whole travel behavior data that were not observed when only using the GPS." • "We wrote customized software that was loaded into the flash memory, which handled the data cleanup, conversion to trace format, extraction, and uploading transparently from the user perspective. The primary purpose of this was to shorten the recall period to the same day as data collection."
2. GPS plans for near/long term	• The research and practice trends are toward developing apps for smartphones and tablet PCs. • Use of other technologies to collect location data where GPS signal is not available is being considered. • The improved technology would allow focusing on data collection for longer periods and eventually longitudinal GPS data collection where travel behavior could be studied as a function of life-cycle changes. • Further extension of work on automatic processing of the GPS traces is expected, especially for trip purpose imputation, parking search, and mode changes. • Post-processing of GPS data is becoming more important since it makes it possible to analyze and understand massive sets of GPS data by extracting knowledge from raw data while also maintaining its accuracy for planning purposes. • Shifting to GPS-only household travel surveys	• "FHWA has a project looking at alternatives for collecting long-distance travel information. These could potentially include using cell phone tracking (e.g., from a commercial source), Twitter feeds, or Facebook posts from smartphones (which includes location) and combining these data sources with other more traditional sources." • "We are interested in collecting GPS data for use in an ongoing project involving travel demand modeling incorporating ITS strategies."
3. Impact of GPS use on participation rates and sample represent-ativeness	• Most participants agree that the experience was interesting since the data collection burden for a respondent seems to be much less than for the traditional travel and diary surveys, thus allowing for a longer duration of survey. This will result in somewhat higher cooperation rates. • The cost of equipment and limited number of units in hand may affect the sample size and the duration of the study both in terms of total time needed to conduct the survey and duration of data collection from each participant. • Recruitment and sample representativeness might be challenging in GPS + web studies since those without Internet access are sometimes eliminated from the study, resulting in sampling bias. In particular, older populations and those with limited Internet access will not be adequately covered to capture all travel markets. • Therefore, using multiple survey modes and an appropriate method for weighting responses will be necessary. • A study by INRETS suggests that "GPS survey participation is positively correlated with higher education, higher income and, therefore, higher access to cars and greater mobility."	• "The GPS technology generally tends to attract the higher income and higher educated respondents. We do see evidence of higher interest among the younger crowd, and avoidance from the elderly, and have had to adjust sampling plans to ensure an equitable distribution of participation."

Table 1-12. (Continued).

Question	Summary	Relevant Quotes
4. How do you process GPS data?	• Typically there are multiple layers of data processing: – Quality assurance of the data collection – Log generation – Identifying trips by using a stay-and-move identification algorithm – Identifying travel attributes (e.g., route, mode, destination, purpose) using various heuristic, machine learning, data mining, or statistical models. • Mostly use an in-house program to parse the data into trips and an in-house algorithm to impute trip attributes Map-matching algorithms are also developed to identify the route (path) in the network.	
5. GPS-based travel behavior details included in deliverables	• Many aim to obtain data as comparable as possible to those produced by conventional surveys. • GPS-based travel surveys generally result in full traces of the logger movement at up to second-by-second resolution. • The processed trip and activity records include basic activity-travel information for each episode, including origins and destinations for trip segments, departure and arrival times, trip frequencies, chains, and the route in the network. • In some prompted-recall surveys, additional information is collected on what respondents were doing, who they were with, what the activity at the trip ends was, how the activity-travel episodes were planned, the time constraints at trip ends, the payment of fares, and so on. Qualitative data, such as attitude and opinion, can be also obtained in relation to the actual travel behavior. • Most (if not all) trip and stage details include location coordinates (or geocodes). Mostly have geocoded the stage/trip details.	
6. Methods for privacy protection	• Methods for privacy protection in GPS studies are similar to those employed in traditional surveys: – Institutional controls (IRB, ethics guides, confidentiality agreements, human subjects training) – Removing identifiers from travel information, including names and addresses – Physical and digital security, including the use of firewalls, secured storage, and protected databases, as well as data separation where data identifiers are kept separate from travel data – There have been recent efforts by the Department of Energy and the National Renewable Energy Laboratory to establish the Transportation Secure Data Center to improve access to GPS data while maintaining individual confidentiality. • The right to participate and start the survey remains with respondents, and they can turn on/off the equipment whenever they desire to do so. • Synthetic GPS traces and multi-trace compression can be used as a solution to deal with the privacy issue.	• "We are experimenting with synthetic GPS traces that could be developed from actual trace characteristics but still be shared with the public as they won't expose participant behaviors. Multi-trace compression is another approach we are using to 'shelter' the information in individual traces."
7. GPS coverage and accuracy compared to other methods	• The data received from GPS units (both spatial and temporal) are fairly accurate—certainly more accurate than traditional diary surveys—although minor corrections might be necessary to the logs from inaccurate location tagging. • Researchers have found that GPS surveys result in higher trip frequencies. • There are several well-known issues with the accuracy of GPS studies, including: – Signal losses through urban canyons, tunnels, and buildings; – Cold start and loss of signal at the beginning of trip; and – Need to apply an effective and suitable imputation method to fix these issues.	• "We have seen the quality of the GPS traces improve as the GPS technology has improved." • "The location data itself was also fairly accurate with very few corrections made to the logs from inaccurate location tagging. The accuracy as far as respondent-identified to algorithm-identified activity locations were above 95%." • "Recent work has shown that by using Wi-Fi most of the time and GPS only when Wi-Fi is not available, the draw on the battery can be much less. However, this increases the error on route delineation and does not provide sufficient information about travel speed (if someone wants speed)."

(continued on next page)

Table 1-12. (Continued).

Question	Summary	Relevant Quotes
8. Advantages of GPS	• There are several clear advantages to GPS studies as opposed to traditional surveys, including: – Ability to collect all movements, precise times, locations, and routes; – Respondent burden reduction on data collection, with the individuals not needing to remember exact times and locations; longer reporting and ability to collect multiple days of travel to examine variability of travel; – Ability to capture route choice and speed; – Improved data quality: not reliant on self-report; and – Ability to look at activity time/space prisms and use in travel micro-simulation. • The use of GPS with smartphones also seems to allow for richer, more interactive data collection, where the location-based services provided by phone companies, Google maps, etc. can be used to enrich the data set. – Using a respondent's smartphone reduces the cost and time compared to sending and retrieving GPS equipment. – Smartphones could be used for longitudinal studies, attitudinal surveys, daily travel, or long-distance travel diaries. • GPS data collecting supports activity-based models and next-generation travel models with more detailed data.	• "[Use of smartphones] can be an excellent tradeoff – giving participants a significant cash [that is saved by not sending GPS units to them] incentive (toward their monthly cell phone bill) to participate in a travel behavior project."
9. Limitations/ concerns about GPS	• Sampling bias in GPS studies is inevitable. It seems better to develop the analytical methods with the biased samples. • Despite significant improvements, the life of batteries when GPS is constantly in use is still not satisfactory. Several solutions have been suggested to remedy battery life issues: – Collect GPS data less frequently by setting a longer data collection interval (i.e., not every second) – Use GPS along with other technologies (accelerometer) so the unit can go to sleep when the person is not in motion – Use combinations of GPS and Wi-Fi to collect data • While some attributes (e.g., start and end time, speed, duration, and mode) could be detected, additional questionnaires might be needed to capture other travel attributes like trip purpose. • There are signal losses in urban canyons, tunnels, and buildings. • Privacy issues and securing fine-grained GPS records is a point of concern.	
10. Pricing information	• Most researchers have performed small GPS studies, so their cost estimates significantly vary and do not reflect the true cost of a major GPS study. • The survey cost is decreased as the price of devices drops. • Assuming that a GPS mobile phone is given to a respondent for a month and he/she has to carry the mobile phone and update the web diary every day, the direct survey costs consist of shipping costs, communication costs, and monetary incentives for the respondent.	• "The devices were approximately $60, and we gave out a survey incentive of $35 per household. Other costs included the student employees to deliver the devices and provide about 1 hour of training to the respondents on taking the survey and using the device."

Table 1-13. Summary table of transportation modeler responses.

Question	Summary	Relevant Quotes
1. Current roles of GPS	• Common use of GPS subsample data for comparison to traditional diary methods • Common use of GPS floating car data for travel time and delay studies, as well as for defining link-level speeds by time of day • Common use of GPS for performance evaluation of transportation networks • Use of smartphone application for collecting special use data (CycleTracks) • Growing use of GPS as primary source of travel data for model development • Use of AVL data for bus speed models and route planning • Some analysis of GPS travel data in route choice • Some use of GPS data for model calibration and validation	• "We have participated in SFCTA's CycleTracks application to allow cyclists to voluntarily provide us route information for analysis. We are considering expanding this program and developing our own applications."
2. Specific roles of GPS data in model development	• Development of trip correction factors • Estimate trip rate variability from multiday data sets • Provide baseline network speeds • Calculating core travel time and distance statistics for trip purposes and demographics • Use as primary source in revealed preference household survey • Use as primary source for trip, tours, and activity patterns • Estimate bicycle route choice model • Validation of travel times in DTA	• "In a 100% GPS-assisted HTS (like the Jerusalem HTS), the GPS traces of individual person travel are the basis for extracting trips, tours, and activity patterns. In this case, and especially if the prompted-recall method is applied, the GPS data constitute the core component from which all other data items are derived, and not just correction factors. My personal view is that is the best approach."
3. Secondary uses of GPS data	• GPS data commonly being shared outside of agencies for research purposes: – Air quality analysis – Active transportation (bike/walk) – Congestion analysis – NCHRP Project 8-57, SHRP 2 C04, and SHRP 2 L04 • GPS travel survey data used for congestion analysis and bottleneck ID	
4. Plans for future use of GPS data for travel demand modeling	• Plans to use GPS as the primary source of travel behavior data • Find volunteered GPS data instead of conducting large recruiting efforts. Sampling bias is probably as bad as self-selection bias. • Implement 100% GPS-based prompted recall, with multiday surveys (for at least 1 week) • Use GPS for focused subsamples (visitors, taxis, trucks/commercial vehicles) • Use GPS for surveys of visitors, taxis, and commercial vehicles to assist with modeling • Explore data imputation methods for mode and purpose • Use GPS data to generate models of transit customer trip times, access times, and wait times. • Development of route choice models	• "We hope to use GPS trace data (with subsequent follow-up questions) more as the main data source for surveys and models. Hope that there are ways that more and more 'volunteer' GPS data can be used to support modeling, as opposed to (or in addition to) launching expensive surveys that purport to have 'probability-based samples,' but, due to inevitable sampling biases these days, probably aren't much better than self-selected samples (or at least have compensating biases)." • "GPS will remain an integral part of household (HH) travel surveys, providing actual measured, revealed preference data on times, paths, durations, amounts, locations to supplement surveyed responses on motivation, purpose, costs, scheduling, etc."

(continued on next page)

Table 1-13. (Continued).

Question	Summary	Relevant Quotes
5. Other sources of origin–destination data	• Several modelers are exploring the use of cell phone and sensor-based systems for origin–destination travel times. • Several are currently using consumer data for improving baseline network data and congestion analysis. • RFID data can be used for identifying transit trip start and end locations and trip travel times.	• "As it stands now this data can be used only for certain types of analysis (travel times and speeds as well as individual route choice trajectories). This data cannot replace HTS because it is non-behavioral in nature. The main difference between behavioral and non-behavioral data is that behavioral data includes characteristics of the individual (such as age, gender, and income) and characteristics of the associated daily activities and travel (trip purpose, car occupancy, other trips made on the same day, etc.).
6. Recent purchases of travel behavior data	• Bluetooth readers for travel time and special OD studies • TomTom speed data for model validation and baseline speeds • INRIX for real-time speed monitoring is also being used for network speed validation. • TomTom speed data used to identify travel time reliability • Considering AirSage OD data	
7. Plans for short-term modeling	• Multiple efforts for building/integrating DTA models to support activity-based models • Regional DTA tools to evaluate system improvements • DTA to evaluate toll roads and managed lane projects	• "Observed activity patterns would also be useful in validating activity-based models in addition to using for short-term forecasts. Having the data at the activity-pattern level allows one to more easily test mode and time of day shift possibilities, among other things."
8. Key data needs for long-range modeling	• Better baseline network and demographic data • Better LOS skims • Better data on hard-to-reach populations • Identification of long-term trends in changing travel habits • Better survey data – Spatial and temporal accuracy – Oversample some population groups – Complete/accurate spatial traces • Better networks • Better parking location information • Revealed travel data • Traveler stated preferences • Tourist and visitor surveys for major cities • Taxi models • Delivery vehicle models • Ideally, real-time data would feed self-calibration model routines. • Reliability metrics • Data for understanding the inertia effects on shifting patterns over time • Passive data that can have key behavioral information imputed	• "Long-term forecasts require both the revealed travel data that can be obtained from counts, GPS and cell phone/AVL, and also surveyed data, which, in the future, needs to focus more on preferences, scheduling, flexibility, purposes, etc. rather than respondents trying to tell you what they did (which will be measured instead)."
9. Key travel behavior data issues for activity-based models	• Completeness of household interactions and schedule coordination • Complete household travel data (no missing persons or trips) • Geocoded data that can be tied to parcel location • Completeness of individual daily patterns • Intra-person time/space consistency • Inter-person, intra-household consistency • Survey information for an entire week • Observed behavior • Long-term and travel-related decisions	• "Quality is important for all aspects; less good data is better than more bad, focus needs to be more on quality than sample sizes." • "Data fidelity and resolution, finer level grain of detail."

Table 1-13. (Continued).

Question	Summary	Relevant Quotes
10. Benefits of GPS for understanding travel behavior	• Increased accuracy of spatial and temporal travel details • Completeness and minimization of underreporting • Spatial and temporal resolution • Reduced burden on the respondent to report addresses and timing for all trips/locations, thereby significantly speeding up the survey • Attractive high-tech image of the survey, especially if the prompted-recall method applied is integrated with GIS (such as in the case of the Jerusalem HTS, where the recruitment rate was 70%–80%) • Possibility of collecting data non-invasively for multiple days with subsequent automated imputations of travel modes, purposes, and other data	• "A fusion of the traveler's own experience (GPS) with adjacent system conditions (Bluetooth/INRIX) and an effective means for gathering and recording traveler perceptions of conditions and reactions."
11. Disadvantages of GPS for understanding travel behavior	• Sampling issues due to reliance on traditional recruitment methods • Specific errors associated with GPS not as well understood relative to traditional survey methods • Need for auxiliary data in addition to GPS trace (can't just use passive data) • Signal issues in some locations • Respondent burden of special GPS device • Cost • Limited set of tools for processing GPS into trips	• "GPS has a high cost, and charging requirements coupled with device costs create a high respondent burden. The emergence of smartphone technology provides a great opportunity to bring down the costs of data collection and analysis." • "GPS has its own potential data error/quality issues that are not as well understood yet as the types of errors/biases that are inherent in travel diary surveys."

CHAPTER 2

Summary of Best Data Sources and Methods to Test

Introduction

Chapter 1 discussed several methods for processing and deriving information from GPS traces in the context of HTSs, as well as a wide range of applications of GPS data in the development of transportation models. Also, the research identified the need for guidelines in the processing and archiving of GPS-derived travel survey data. In addition, the first chapter covered several emerging mobility data sources and data providers that are currently using these data sources to derive commercial traffic and aggregate transportation data products. The challenges associated with managing ever-increasing archival data sets were recognized, together with a new set of technologies developed to better handle big data.

This chapter presents a multidimensional analysis of the main candidate data sources that can be used in the analytical tests presented in Chapter 3. Based on the findings in this analysis, candidate test data sets are identified. This is followed by a review of the data processing, imputation, and fusion methods that will be used to augment GPS traces with more complete travel details, and, in some cases, with socio-demographic information.

Inventory and Discussion of Available Data Sets

To evaluate the performance of the various data fusion methods proposed later in this chapter, the research team used the Chapter 1 literature review findings along with information about other recently available data sets (as identified by NCHRP Project 8-89 panel members) to identify potential data sources. These sources included:

- GPS data sets collected as part of HTSs, including those collected in Atlanta, Denver, New York City, Chicago, and California;
- Smartphone GPS data collected either as part of an HTS (such as what was collected in Portland) or for another purpose (such as CycleTracks);
- GPS or other location data collected by traffic data vendors;
- GPS data collected in other transport studies [such as value pricing or mileage-based user fee (MBUF) studies]; and
- GPS data collected by personal navigation devices (such as TomTom data, often sold by traffic data vendors).

This inventory process also categorized each potential data source as one of three types: (1) GPS data collected in tandem with HTSs; (2) anonymous GPS bulk traces from instrumented vehicles, mobile phones, and navigation devices; and (3) fixed-location sensor data. The various characteristics of these types of data are presented in Table 2-1. Of these three data types, only the first and second are truly applicable for deriving the type of behavioral information necessary for developing transportation demand models that are based on modeling individual choices. The third data type, fixed-location sensor data, can only be used for model validation (and aggregate calibration in a very limited sense) and estimating base-year transportation network conditions. This is because it is necessary that the source data contain complete tours from sampled persons (rather than unrelated trips) and, as such, provide enough information to explain the factors causing the observed travel choices. These explanatory factors relate to the person and household characteristics such as age, gender, income, and occupation, as well as activity contexts such as the placement of the particular trips in the individual's daily activity chain.

Other elements that are important for model development include accurately capturing intra-household interactions in the form of shared travel and activity information. For example, the necessity of escorting a child to school on the way to work is an important determinant of commuting mode choice. This behavior component cannot be analyzed and understood solely from the trace of the work commute itself.

Table 2-1. Available data set types and characteristics.

	GPS Data from HTS			Bulk Traces			Fixed-Location Sensors	
Characteristics	**Person**	**Vehicles**	**Smartphones**	**Mobile Phones**	**Instrumented Vehicles**	**Navigation Devices**	**Bluetooth**	**RFID**
Sample size	Small	Small	Small	Large	Large	Large	Large	Small
Spatial accuracy	High	High	Depends on hardware and software	Low	High	High	Limited	Limited
Path completeness	Yes	Yes	Depends	No	Yes	Depends on use	No	No
Complete tours?	Yes	Yes	Not certain	Not certain	Yes	Not certain	No	No
Household interactions?	Yes	No	Depends on usage	Depends on usage	No	No	No	No
Person or vehicle?	Person	Vehicle	Person	Person	Vehicle	Vehicle	Both	Both
Socio-demographics	Yes	Yes	Yes	Derived from home location	Derived from home location	Derived from home location	No	No
Expected biases	Controlled by survey design	Controlled by survey design	Controlled by survey design	Age and market penetration of mobile phone service Contains commercial vehicle travel	Unknown Contains commercial vehicle travel	Market penetration of navigation devices and level of usage Contains commercial vehicle travel	Market penetration of Bluetooth equipment Contains commercial vehicle travel	Market penetration of cards and tags equipped with RFID Contains commercial vehicle travel

Without this contextual information foundation, it becomes very challenging to develop the analytical models that provide the foundation for modern TDMs.

The overall trend in travel model development today is to apply individual behavioral models that explain the outcome (i.e., travel by such dimensions as origin, destination, mode, and time of day) by means of explanatory variables through a plausible decision-making process. In this sense, the GPS traces themselves only provide the snapshot of the outcome, albeit with a very high level of accuracy and spatial–temporal resolution. Supporting the GPS data with behavioral explanatory variables is paramount for applying results within a forecasting model.

Based on this assessment, and given the research team's authorized access to GPS data sets collected as part of household travel surveys, initial efforts were focused on the second data type (i.e., bulk traces). The research team approached two traffic data vendors to obtain test data sets. One was selected due to its focus on cell-phone–based products and large market penetration, and the other was selected based on its use of GPS-based solutions. Unfortunately, the efforts to obtain bulk GPS trace data sets from these traffic data vendors did not succeed given end user licensing restrictions that prevent them from sharing high-resolution trace data with third parties.

It is also important to note that traffic data vendors have historically relied on the instrumentation of fleet vehicles as a primary data source and on smartphone apps that only collect data when users are checking traffic conditions. This means that if one of these data sets was made available for this or any similar personal travel behavior study, the results could be biased due to this significant commercial fleet component (especially for driver owned-and-operated vehicles) and, in the case of personal travel, would likely show partial day traces clustered during morning or afternoon commute hours.

Given the restrictions in obtaining bulk GPS traces from traffic data vendors for this study (data sets that will likely not be made available to planning agencies for the same reasons), as well as the potential personal mobility measurement biases that do exist in these traffic data vendor sources, it became obvious that data sets from the first group, GPS-assisted HTSs, were the most appropriate for use in the test experiments that appear in Chapter 3. More specifically, the research team felt that two types of GPS data from travel surveys could be tested: (1) person-based, GPS-assisted HTS

Table 2-2. Available person-based GPS HTS data sets.

Study Name	Number of Households	Number of Instrumented Persons	Number of Trips on First Day
ARC 2010 Person GPS HTS	334	649	3,613
DRCOG 2009 Person HTS	170	332	2,308
MTC 2012 Person HTS*	1,732	3,386	19,839
Total	**2,236**	**4,367**	**25,760**

*Numbers as of October 2012.

data sets collected by stand-alone GPS data loggers deployed as part of the study; and (2) smartphone-collected GPS data collected in tandem with a household travel survey. Data from this type of GPS data source have the most potential for testing various data fusion methods because of the wealth of information associated with the sampled households and persons. In addition, depending on the original study design, the GPS-derived travel may be associated with trips reported by participants, which can provide calibration data for trip-level imputation models.

The research team obtained permission from ARC and the Denver Regional Council of Governments (DRCOG) to use the GPS data collected as part of their recently completed household travel surveys. In addition, the most recent CHTS, which consisted of a year-long data collection effort, included a 100% person-based GPS target sample of 3,100 households collected in the Oakland/San Francisco region for the Metropolitan Transportation Commission (MTC). MTC agreed to make this data set available for this project's methods tests. Table 2-2 presents summary information on these candidate data sets.

With respect to smartphone data sets collected as part of an HTS, the best candidate identified is the PaceLogger data set that was collected by a subsample of households in Portland as part of the recent Oregon Household Activity Survey (OHAS). This data set was collected using a modified version of the original CycleTracks iPhone app and contains data from 308 smartphone users from 256 households. The research team received permission to use this smartphone data from the OHAS subcommittee of the Oregon Modeling Steering Committee (OMSC), and obtained a copy of the data set and documentation to continue its assessment of the data.

As mentioned previously, the research team is aware of other transportation-related GPS data sets, such as the vehicle-based GPS data collected as part of the Puget Sound Regional Council's Value Pricing Study and the vehicle-based smartphone GPS data collected in the recently completed MBUF project conducted in Minnesota. Given the 100% vehicle focus of these studies, however, it was decided that it would be more informative from a comprehensive research perspective to use a person-based data set with multimodal travel patterns. Furthermore, the MBUF data set was not available at the time this research was conducted. Although the CycleTracks data set was also available for use in testing, it was decided that the intended use of this smartphone app for collecting bicycle trips would limit its usefulness in the analysis of multimodal and motorized travel behavior.

Review of Data Fusion Methods

In the context of this study, the term "data fusion" refers to the process in which two or more data sets are integrated to generate a single reliable data source for modeling and other applications. In data fusion, when two or more data sets are to be integrated, the analyst should find data elements that are statistically compatible across data sets (e.g., income, household structure) and that can perform data integration by applying normalization of the common data elements across data sets along with necessary weight adjustments. It should be noted that since data sets are collected in different contexts, significant differences may exist among them. Therefore, various statistical tests are required to reconcile the differences across data sets.

The data fusion approach relies on data mining and pattern recognition tools combined with statistical distribution updating methods to add demographic characteristics to the GPS traces. The general processes involved in data transferability are reviewed in the following and generally relate to the transference of travel characteristics.

Data Fusion Methods

Data fusion deals with the problem of merging different data sets from a variety of sources into a single data set. The approach allows the merging of two or more data sources collected through various surveys or at different aggregation levels. Data sets typically contain missing variables that complement each other in such a way that the resulting data set includes a complete list of consistent variables. Data fusion could be seen as a special type of data imputation where several variables are missing in data sets because they have not been collected to reduce respondent burden during the survey, or where multiple surveys were conducted to obtain different samples where the questions of interest are split in two sets with a common set of socio-demographic variables. There are several classical approaches to data fusion problems that are presented in the literature (Saporta 2002).

Explicit Model-Based Estimation

In this approach, each missing value can be estimated using a simple model such as regression, discrete choice, machine learning, or cross-classification. Estimations are made variable by variable, not taking into account their correlations, and may lead to inconsistent results. The other problem with this approach is homogeneity of estimated values, in which two units will have the same estimates if their independent variables are the same, and hence it will lack heterogeneity in estimates. It appears that an explicit estimation technique is useful when few missing data points need to be estimated; however, it might not be a good approach to apply when large blocks of missing data need to be generated in a data fusion practice.

Imputation with Implicit Models (e.g., Nearest Neighbors)

This approach is similar to a copy-and-paste practice in that a whole vector of variables for record *i* from a source data set is transferred to record *j* of the target data set where records *i* and *j* have close profiles. The closeness of profiles is measured by identifying the nearest neighbors within an appropriate distance. Another commonly used approach in this category is file grafting, which is based on principal component analysis (PCA).

Data Fusion by Maximizing Internal Consistency

The approach is based on multiple correspondence analysis (MCA) or homogeneity analysis. The essential idea is to assign categories to the set to minimize a loss function. MCA of a disjunctive table can be viewed as the minimization problem of a loss function that is, in fact, equivalent to getting maximum eigenvalues for the completed table (Saporta and Co 1999).

Double Imputation Method

This approach is a file grafting technique that combines the explicit and implicit approaches and is also called nonsymmetrical grafting. It is based on the constrained principal component analysis technique that allows imputing the missing information into a target sample, taking into account knowledge of the relationship structure among variables (Piscitelli 2008).

Data Fusion and Transferability

There has been extensive research in recent years on the transferability of travel attributes of individuals from one context to another. Travel attributes like number of trips, distance traveled, and modes used for each individual are critical requirements in any disaggregate travel demand analysis, and data transferability approaches are seen as reliable alternative solutions for smaller communities where data collection is more costly and challenging. "Data transferability" broadly refers to any approach that utilizes data or models from one context to generate data or models for use in another context. This can be used either in a spatial context, such as generating a model or data for a region on the basis of data that is obtained from another region, or in a temporal context, such as forecasting data for a region based on existing data from the same region. Transferring travel data either temporally or spatially is a common practice that is typically performed in an ad-hoc fashion using household-based cross-classification tables. While the focus of much of this work has been on transferring relations between demographics to travel patterns, the methods should be applicable to the converse situation of interest in this study (i.e., inferring demographics from travel patterns).

Data transferability models are basically built upon data mining methods that can explore the data and detect the interdependencies and correlations among variables (Stopher, Greaves, and Bullock 2003; Reuscher, Schmoyer, and Hu 2002). In the literature, various models have been proposed to transfer disaggregate travel attributes using statistical methods. Mahmassani and Sinha (1981) studied spatial transferability of trip frequency for small urban areas in the state of Indiana at three levels: area wide, zonal, and household. They compared cross-classification tables of trip frequencies among urban areas and their distributions for different trip purposes across different socioeconomic groups. Wilmot (1995) used multiple linear regression models to perform a similar analysis. Unlike the regression models that generate continuous results for discrete variables, Zhao (2000) applied discrete choice models to account for more of the behavioral process of trip generation. Ben-Akiva and Bolduc (1987) and Zhang and Mohammadian (2008) used a Bayesian updating approach to improve spatial transferability of travel attributes. Zhang and Mohammadian transferred data from the NHTS to smaller areas in Iowa and New York and showed that using a small, local sample and Bayesian updating can significantly improve the quality of the synthesized data (Zhang and Mohammadian 2008). Long, Lin, and Pu (2009) applied small area estimation models to identify household- and census-tract–level travel characteristics, such as number of work trips for small and midsize metropolitan areas, where few travel samples are available from various data sources.

The dependency between the travel attributes is a challenging issue that has typically been ignored in the transferability of transportation models. For example, the number of recreational trips for an individual in a day might be dependent on the work trips for the individual on that day. This means that modeling the number of daily recreational trips and work trips independently could add estimation bias to the results. Rashidi and Mohammadian (2011) attempted to study disaggregate trip rates for different trip purposes in the transferability context. They presented household travel attribute models using an exhaustive chi-squared automatic interaction detection (CHAID) data mining algorithm to address several limitations concerning complexity of models, limited explanatory variables, and lack of accurate disaggregate models. In a follow-up study, Fasihozaman, Rashidi, and Mohammadian (2013) applied a significantly modified version of the same algorithm in an attempt to explore and discuss a more disaggregate, and policy-sensitive, individual-based data transferability approach. This was achieved by using a broad set of socio-demographic and land use variables. Using the 2009 version of the NHTS, the modeling approach was further enhanced by using a wide range of probability density functions.

Applicability

There are two main problem areas where data fusion techniques could be used to augment GPS data sets. The first involves the association of socio-demographic and household structure information with individual GPS traces, while the second is related to identifying travel and activity characteristics from the GPS spatial and temporal dimensions.

Figure 2-1 depicts a basic understanding of the two major technical tasks. The first problem area can be called demographic estimation, which consists of attaching person and household characteristics to the individual GPS records. This task is relevant only for anonymous, massive data. The second problem area can be handled by behavior-ization of the person or vehicle traces, and begins with the conversion of the individual traces into a sequence of trips. The major steps of behavior-ization include identification of individual trips, trip modes, purposes, and activity types. For both tasks, multiple additional data sources are used.

Demographic Estimation Using GPS Traces

The data fusion approaches discussed in the previous section have generally been shown to transfer at least some travel characteristics from one context to another with some degree of accuracy, and generally seem to offer potential applications for transferring the relation between travel pattern and personal characteristics to anonymous GPS data traces. Therefore, in addition to the test described in the previous section, the data transferability approach was also tested in this study.

The approach to the personalization of GPS followed in this study begins by developing clusters of travel patterns observed in the source data set to be transferred that exhibit similarities in the types of individuals who engage in them. There are many ways to accomplish such clustering that have been pursued in the transferability literature. To narrow the scope of the project, the research team tested the decision tree methods of cluster development using the C4.5 approach. The general effect of the decision tree models is to split the travel pattern observations into pattern clusters with maximal homogeneity of demographic data within each cluster, in a similar manner to that discussed in the previous section.

In this case, the sample data would be from a high-quality, representative data source, such as the NHTS or other household travel surveys from one or more regions. Anything that can serve as a reliable source to link travel patterns and travel characteristics could be used. While the ideal data source would clearly be a locally collected household travel survey, the assumption in this study is that such data are not available in sufficient quantity to feasibly estimate a travel demand model. However, small-scale, local data may be incorporated into the demographic estimation procedure through a data transferability or updating process.

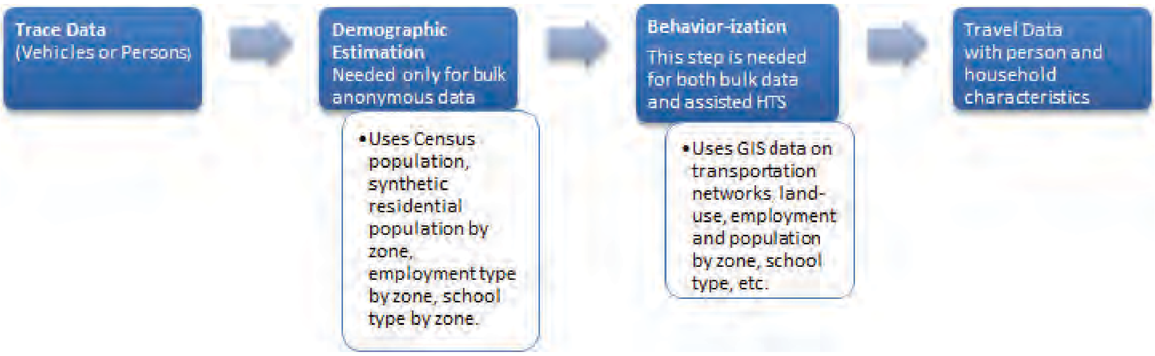

Trace Data
(Vehicles or Persons)
Demographic Estimation
Needed only for bulk anonymous data
•Uses Census population, synthetic residential population by zone, employment type by zone, school type by zone.
Behavior-ization
This step is needed for both bulk data and assisted HTS
•Uses GIS data on transportation networks land-use, employment and population by zone, school type, etc.
Travel Data
with person and household characteristics

Figure 2-1. Data fusion tasks.

The clustering procedure using the source data is followed by an updating procedure that is used to update the dependent variable distributions. In this step, clusters from the transferred sample can be updated with small local samples using, for example, Bayesian updating methods, as in the related work described previously on transferring travel pattern data. A local household travel survey complementary to the anonymous GPS data would be used for this purpose. Alternatively, the procedure can be tested with a household travel survey alone, which would involve both developing and applying the clusters using the same data set, which would allow the updating procedure to be skipped. This would be the case if a household survey with an attached GPS data collection component is to be used.

The result of this process would be a set of clusters (or rules, neurons, models, etc.) that relate travel characteristics to specific sets of demographics, from which demographics for a specific target pattern can be drawn using secondary models, as described previously. These distributions need to conform to known marginal distributions of the target demographic characteristics, which can be derived from census data. The models can also be constrained by joint distributions of demographic variables if these are available from either the census data or from population synthesis. The transferred models will then need to be calibrated to reflect the constraints on the population characteristics.

Identifying Behavior from GPS Traces

The first challenge to overcome when extracting behavior from GPS data is to clean and process it into trips and activities. Performing these types of tasks can take significant effort when processing raw GPS data from emerging sources such as smartphones and wearable (and continuously powered) GPS data loggers. This issue has been tackled in the past using various heuristics that are not necessarily consistent. Based on the literature review findings, the research team proposed that the tests in Chapter 3 focus on the core processing methods necessary to perform basic GPS trip processing, which excludes map matching and route identification. The complete list of methods along with their references is provided in Table 2-3.

The test consists of implementing code or using implementations from the original authors, when available, for each method and using it to process the raw GPS data into trips. In the case of the wearable GPS loggers' HTS data sets, the performance was measured by comparing the outputs with those originally identified, which were reviewed by analysts at GeoStats. However, the research team does not believe that there is much benefit in applying these methods to the smartphone data set from the OHAS study given that it was by definition recorded as separate trips by participants.

Table 2-3. Proposed GPS data cleaning and trip identification methods for testing.

Task	Method Types	Source References	Description
Noise filtering	Complex heuristics	Stopher, Jiang, and Fitzgerald (2005)	Remove zero-speed points and points that show movements of less than 15 m.
		Lawson, Chen, and Gong (2010)	Remove points based on HDOP, number of satellites, zero speed or heading, and presence of jumps.
		Schüssler and Axhausen (2008)	Grouping of points between position jumps combined with an iterative removal process based on segment length. Data are then smoothed using kernel density, and points are removed based on altitude.
Trip identification	Simple dwell time	Wolf, Guensler, and Bachman (2001)	A 120-s dwell time between GPS points is used to identify trip ends.
	Complex heuristics	Schüssler and Axhausen (2008)	Data stream is classified into activity clusters based on position density, with clusters being grouped if they are too close in time, and trips are derived from the clusters.
		Oliveira et al. (2011)	Stream of points is segmented based on dwell time and mode transitions; the resulting trips are then compared against a set of quality parameters (number of jumps, spatial coverage) to determine whether they are real.
Mode transition identification	Heuristics	Tsui and Shalaby (2006) and Schüssler and Axhausen (2008)	Classifies transitions as either EOW, SOW, or EOG points using speed and acceleration thresholds.

Once the traces have been converted into trips, classifier methods can be used to identify behavior. These methods select attribute characteristics from a limited set of choices and can be applied to augment GPS traces with travel mode, trip purpose, and activity information. In this scenario, the additional sources of data that are to be fused with the GPS traces include information about the transportation infrastructure (e.g., proximity to transit facilities and segregated travel modes), land use (e.g., points of interest and parcel and zonal data), common household locations (e.g., home, work, and school), and schedules.

The literature review identified three main groups of methods that could be used to solve this problem: heuristics, probabilistic, and artificial intelligence (AI). Table 2-4 identifies candidate methods from these groups along with potential applications tested. The probabilistic approach mentioned for identifying trip purpose in Table 2-4 consists of developing a multinomial logit (MNL) choice model relating trip purpose to various trip and person attributes. This is a method similar to the one proposed in Chen et al. (2010) and that was applied to the Northeastern Ohio Areawide Coordinating Agency (Cleveland) GPS-based HTS.

Methods based on heuristics may appear at first to be easier to implement since they require little calibration and only a basic understanding of statistical modeling. However, they tend to contain various constants and thresholds that need to be examined and adjusted based on local deployment conditions. Making these adjustments requires expert knowledge on the local conditions as well as the logic behind the algorithms being used. It can also be the case that the logic embedded in the algorithms is tied to local characteristics.

On the other hand, probabilistic and artificial intelligence models require more advanced analytical knowledge (statistical and mathematical) and more extensive calibration. The first aspect of this latter requirement is the need to obtain or

Table 2-4. Proposed travel mode and trip purpose identification methods.

Task	Method Types	Source References	Description
Travel mode identification	Heuristics	Stopher, Clifford, and Zhang (2007)	Series of rules that employ both point speed values and GIS data relationships. More recent variations also employ checks based on tour relationships and acceptable mode sequences.
	Probabilistic (MNL)	Oliveira et al. (2006)	Used multinomial logit model to assign mode based on GPS and accelerometer data
	AI – fuzzy logic	Tsui and Shalaby (2006) and Schüssler and Axhausen (2008)	Membership functions were specified for each travel mode.
	AI – neural networks	Gonzalez et al. (2008)	Trained neural networks using GPS data collected for car, bus, and walking trips. Also examined the performance of the mode identification network while using a subset of the captured points, which the authors defined as “critical points.”
Trip purpose (activity) identification	Decision trees	Griffin and Huang (2005)	Applied the C4.5 algorithm to build a decision tree capable of classifying trip ends into multiple trip purposes
	Probabilistic (MNL)	Chen et al. (2010)	This is an approach GeoStats developed with Parson Brinckerhoff for use in the Cleveland HTS. It uses both rule-based heuristics (for home purposes) and a probabilistic model to compute trip purpose probabilities based on person, household, and trip attributes. The nesting structure is based on natural trip purpose aggregations, from the simpler structure used in traditional four-step modeling to a more detailed one used in an ABM.

collect calibration data that can be used to refine and specify the models for use. If no calibration 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 transferability. It is worth pointing out that the AI fuzzy logic method does not necessarily require a calibration data set but rather a review of the parameters used by membership functions for the various outputs.

As mentioned previously, the need to calibrate models poses challenges to their transferability. This is an aspect that will be evaluated in the next chapter by comparing the characteristics and specifications of the models developed for the different test data sets. Transferability will also be examined by cross-validating models across the different data sets (i.e., calibrating a model with one data set and validating it against another).

The overall performance of the methods was evaluated by comparing their results with the responses reported in the original data. In the case of the HTS data sets, these responses came from the set of GPS trips that were matched to the traditionally reported travel. Only results for trip purpose identification were evaluated on the smartphone OHAS data set since travel mode information was not captured by the data collection application. The number of matches and the characteristics of the mismatches will be explored with the help of tables and charts.

The result of the application of these data imputation and fusion methods was an improved understanding of their performance and shortcomings when applied to person-level GPS trace data. This allowed the research team to make suggestions on the applicability and use of these methods to practitioners.

CHAPTER 3

Methods Evaluation

Introduction

Chapter 3 summarizes the results from the tests conducted on the methods identified in Chapter 2. This demonstration was conducted through two experiments:

- Experiment A: augmenting person-based GPS HTS data with trip details.
- Experiment B: enriching anonymized smartphone GPS data with socioeconomic and demographic information.

Overview of Experiment A

The goal of Experiment A was to evaluate data fusion methods that can be used in the context of GPS-only household travel surveys. The tests consist of implementing code, or using implementations from the original authors when available, for each method and using it to process the raw GPS data into trips. Figure 3-1 shows an overview of the process of turning raw GPS data into processed travel information using the methods tested in Experiment A.

It should be noted that a balanced approach with respect to level of effort allocated was pursued when implementing these methods; in other words, the researchers attempted to perform a comparable amount of calibration and setup across all methods (i.e., to invest a similar level of effort across methods). This ensured that a fair evaluation was performed and also benefited methods that were simpler to calibrate and set up; however, this also means that the full potential of these methods was not necessarily extracted from the tests.

Initial work focused on data preparation and standardization. This work included exporting original GPS data files to comma-separated value (CSV) text files and also locating original raw GPS data files from the ARC and DRCOG surveys.

The resulting CSV input file types generated included:

- Raw GPS points – as collected by the field devices;
- Processed GPS points – filtered to exclude data outside the travel date range and noise; and
- Mode segments – one record per unlinked trip (also referred to as an elemental trip) identified in the processed GPS data; these correspond to individual mode segments in a multimodal trip.

Survey data, which were used to calibrate mode and purpose identification models, were converted from their original database structures to a standardized place-based relational schema. In addition to places, this schema included tables for storing household, person, and location data.

Overview of Experiment B

The purpose of Experiment B was to evaluate methods for attaching person- and household-level information to travel patterns observed in GPS-based household survey data or other sources of GPS trace data. As such, the experiment was designed to be as general as possible, with very few assumptions about the data that would be available in the source data set. So, while the experiment included models derived from household travel survey data, these models should be generally applicable to any source of GPS traces.

Figure 3-2 provides an overview of the process developed for Experiment B. The experiment can be broken down into four stages:

- Stage 0: processing of input trip data from Experiment A into person-travel data records.
- Stage 1: development of the primary demographic clusters.
- Stage 2: selection of optimal person-type clusters based on travel attribute similarity.
- Stage 3: development of person-attribute assignment models for a selected set of demographics.

The outputs of Experiment B include open-source computer code that processes the mode segment data from Experiment A into person-travel records and a set of model files that can be applied to the processed travel records using various open-source modeling packages, including WEKA and BIOGEME.

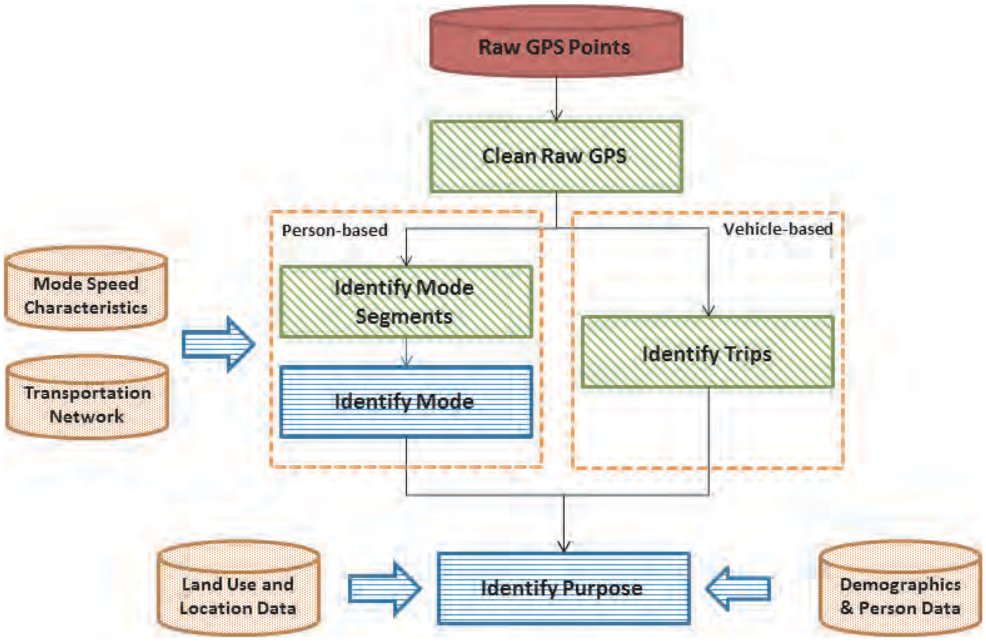


Figure 3-1. Overall sequence of steps covered in Experiment A.

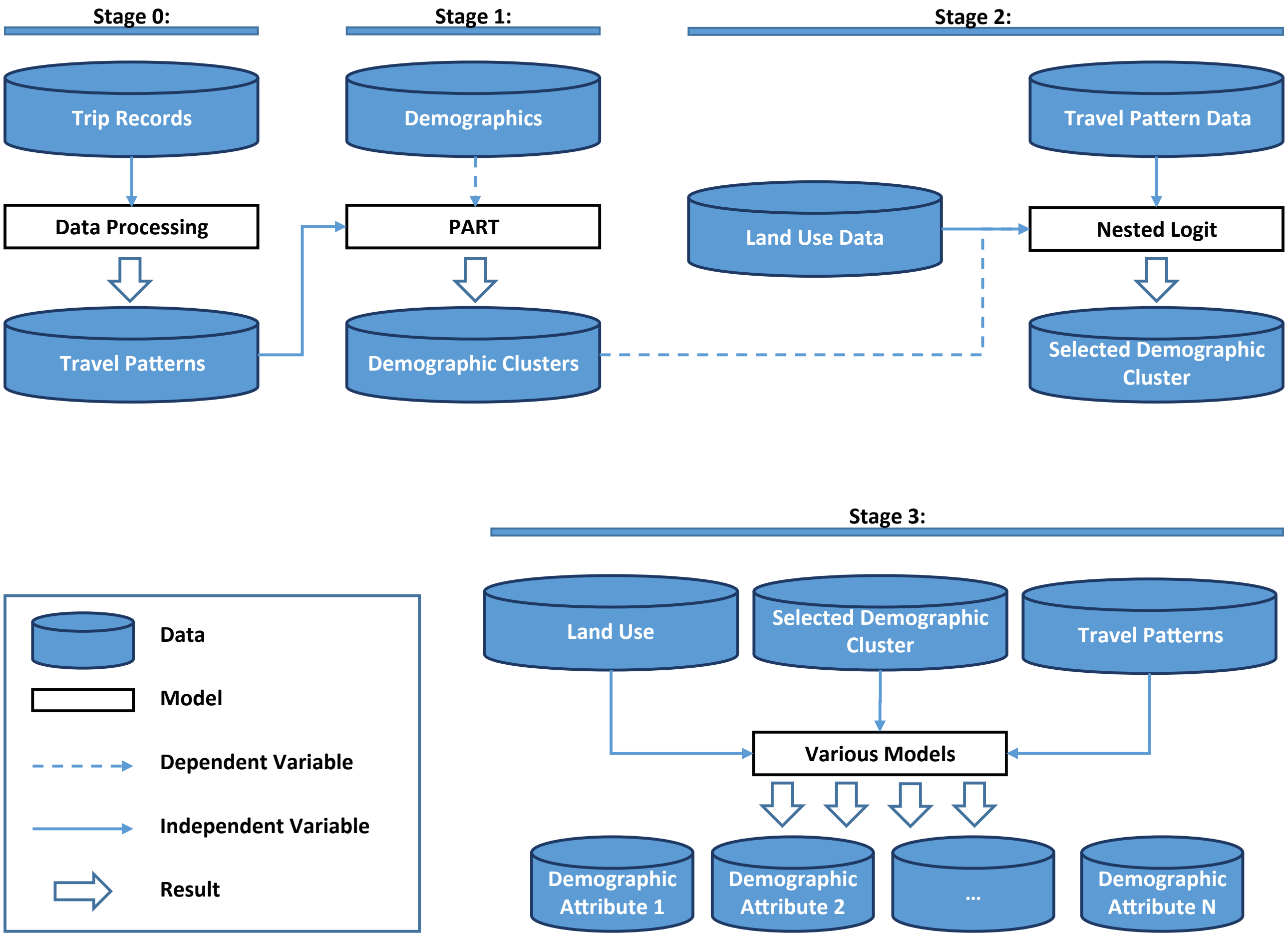


Figure 3-2. Experiment B demographic characterization process.

The remainder of this chapter is organized as follows. First, it introduces the data sets used to carry out the selected experiments along with an outline of the implementation and testing approaches used to evaluate the methods. This is followed by discussions of the two experiments, along with the main findings and results. Additional information on the models implemented in Experiment A and Experiment B is included in Appendix D and Appendix E, respectively. Appendix F contains explanations of the various tools used to conduct the experiments, along with script or code instructions and listings, where applicable.

Reference Data and Software Tools

The research team obtained permission from ARC for using the GPS data collected as part of its recently completed household travel survey. Both person and vehicle GPS data were used to test the GPS processing methods. GPS vehicle data were used to test the noise filtering data because they included values for HDOP and number of satellites, which were not available in the person-based GPS data due to limitations in the wearable GPS logger that was employed. Person-based GPS data were used in all subsequent GPS processing method tests. Both the vehicle-based and person-based GPS data were extensively reviewed as part of the original HTS effort. The original cleaning process included the review of individual travel days by analysts using custom data processing and visualization tools. Because of this original review, the processed data can be treated as a reliable benchmark against which the tested methods' outputs can be evaluated. Three types of reference GPS data sets were created: filtered points, trip ends, and mode segments (corresponding to unlinked trips). A subset of the Atlanta diary data was also used to calibrate trip purpose identification methods; the idea here was that these would simulate the data that are typically collected using GPS prompted-recall methods in GPS-only HTSs.

With respect to smartphone data sets collected as part of an HTS, the research team obtained the PaceLogger data set, which consisted of a subsample of households in Portland as part of the recent OHAS. This data set was collected using a modified version of the original CycleTracks app and contains data from 308 smartphone users within 256 households.

Permission to use and access this data set was obtained from the OHAS subcommittee of the OMSC. In addition to the smartphone GPS points, data from the regular survey were used to test the fit of the models used in the experiments. The data used to train the demographic characterization models were drawn from the 2008 Chicago HTS, which included 2 travel days for about 40% of the respondents (9,736 total observations). The data used for model estimation were limited to the 2-day sample to reduce the confounding effects of intrapersonal, day-to-day variability on the models. This is discussed further in the Experiment B discussion. Table 3-1 provides a summary of the reference data used in Experiment A.

To increase the reproducibility of the tests implemented as part of the research project, a decision was made to use, as much as possible, free and open-source software tools for data processing, modeling, and data analysis. Consequently, the tools (and supporting software) selected to implement the algorithms and models in Experiments A and B were:

- R 3.0 (R Core Team 2013) for heuristics methods and for calling fuzzy logic routines in Java;
- BIOGEME 2.2 and BIOSIM (Bierlaire 2003) for multinomial and nested logit choice modeling;
- WEKA 3 data-mining tool set (Hall et al. 2009) for neural networks, classifier trees, and clustering;
- PostgreSQL; and
- C++ for developing Experiment B data processing scripts.

Table 3-2 shows the assignment of the three programming packages to the experiment tasks.

Table 3-1. Reference data details and sources used in Experiment A.

Study Name	Number of Households	Number of Persons
ARC 2010 Diary	10,278	25,810
ARC 2010 Person GPS HTS	334	649
ARC 2010 Vehicle GPS HTS	727	1,422
OHAS 2009 Portland HTS Diary	4,799	11,133
OHAS 2009 Portland HTS Smartphone	256	307
Chicago 2008 HTS Diary (Total)	10,552	23,808
Chicago 2008 HTS Diary (2-day w/travel)*	2,395	5,125

*Excludes anyone with a travel day on Saturday or Sunday, anyone only responding for 1 day, and anyone who did not travel. This sample was used for the analysis, implementation, and testing approach.

Table 3-2. Programming packages used in the processing of GPS data.

Procedure	Tool Components		
	R	WEKA	BIOGEME
Cleaning raw GPS data	X		
Identifying trips and mode transitions	X		
Identifying travel mode	X	X	X
Identifying trip purpose		X	X
Inferring demographics		X	X

Scripts and code sets created for many of these experiments are provided in Appendix F, along with some basic instructions for use. The data associated with these experiments cannot be made available to the public given privacy concerns for the original participants in the survey efforts under which the GPS data were collected. This is mostly applicable to the original raw GPS coordinate data that were used as input for the data cleaning and trip identification methods.

Experiment A: Basic GPS Data Processing

This section covers the testing of data processing methods that are used to convert raw GPS data into clean points, trips, and mode segments. The application of these methods constitutes the first step necessary for turning raw trace data into transportation behavior information. The methods presented in this section correspond to the boxes with diagonal hashing, as presented in Figure 3-1.

GPS Data Cleaning Methods

GPS data cleaning or noise filtering methods seek to identify points that do not indicate real participant movement and that can hence be removed from the data without loss of travel information. They are typically necessary to improve visualization of the data and also to improve the accuracy of trip identification methods. Furthermore, by removing points that are not indicative of actual movement, they have the added benefit of making the data more manageable.

Original, raw, vehicle-based GPS point files (each one representing all the points collected for an instrumented vehicle) were used as input into the data cleaning methods. The devices used were continuously powered by an internal rechargeable battery, which was also charged by the vehicle's cigarette lighter connector whenever the car was driven. The logging frequency of these devices was set to one point per second, and only points with instantaneous speeds above 1.9 mph (3 km/h) were recorded.

The resulting filtered points from the application of each method were compared against the processed and filtered points in the original GPS data deliverable, also referred to as the reference data, which were reviewed by GeoStats analysts. The null hypothesis here was that each point's final filtered state was to be the same in both data sets. Errors were categorized as belonging to one of the following groups:

- Type I error: Point is not blocked in method when it is blocked in the reference data.
- Type II error: Point is blocked in method when it is not blocked in the reference data.

The noise filtering methods were run against 100 randomly selected raw GPS point vehicle-based data files from the ARC data set. These files contained a total of 2,446,984 GPS points. Each point had a flag appended to it that indicated whether it was considered to be noise by the method. These flags were then used to compare the method's results with the reference, filtered GPS data. Error percentages were calculated by taking the number of errors in each category and dividing by the total number of points. Table 3-3 summarizes the error results of the tested data cleaning methods along with insights obtained while implementing the method.

The first evaluated method (Stopher, Jiang, and Fitzgerald 2005) was originally developed for use in vehicle-based surveys, while the other methods were developed to be applied to person-based GPS data. More specifically, the method by Schüssler and Axhausen (2008) featured additional steps that were added to deal with limitations of the originally used device (i.e., it did not capture instantaneous speeds, HDOP, or number of satellites).

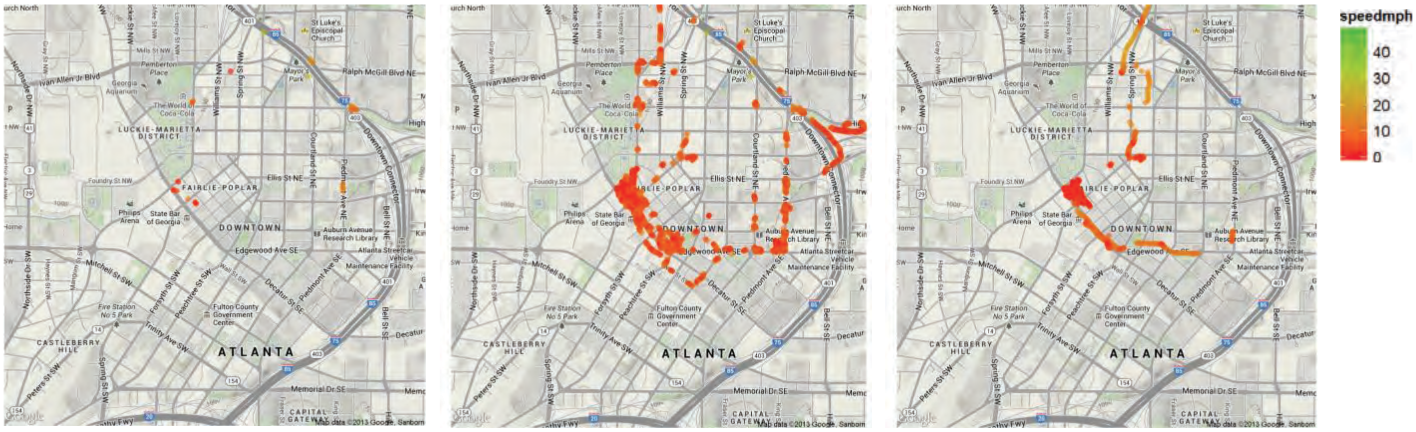
The results indicated that simple rule-based methods based on point quality and speed data available from the GPS device were the most effective. While all three methods had similarly small counts of Type II errors (point was filtered by method, but it was not filtered in the reference data), two of them displayed higher rates of Type I errors (point was not blocked in method, but it was blocked in the reference data).

Table 3-3. Data cleaning methods tested and results from Experiment A.

Source References	Implementation Findings	Type I Error	Type II Error
Stopher, Jiang, and Fitzgerald (2005)	First filtered out all points with fewer than three satellites in view and HDOP equal to or greater than 5. Then removed points that showed no movement (speed equal to zero, less than 15 m of movement, and heading also being zero or unchanged). Point movements were calculated using the great circle distance according to the Vincenty (sphere) method (Vincenty 1975).	0.00%	8.89%
Lawson, Chen, and Gong (2010)	Remove points based on HDOP, number of satellites, zero speed or heading, and presence of jumps. The thresholds for considering points to be of poor quality using HDOP, number of satellites, and speed proposed in the paper were used; these were: HDOP > 5, number of satellites < 3, and speed < 3 m/s (6.7 mph). The paper and its sources did not contain details on how the jump-detection procedure was implemented, and this information could not be obtained. As a result, the stop-flag procedure proposed by Chung and Shalaby (2005), which consisted of discarding points that showed less than 0.00005 decimal degrees of movement, was used.	13.79%	6.90%
Schüssler and Axhausen (2008)	Points are removed if their altitude is not within the study area. They are then smoothed and filtered by speed and acceleration. Since the GPS data used included instantaneous speed data from the actual devices (which are more accurate), the implementation did not calculate speeds as specified in the paper.	14.15%	7.25%

This is to be expected since it is much more likely that the methods will try to err on the side of caution (and thus fail to block some points) than it is that they will block valid points. The Type I error rate for the Stopher, Jiang, and Fitzgerald method was found to be the lowest of the three methods, with a very small number of points being incorrectly flagged as noise.

Figure 3-3 illustrates the results of the three tested methods using a sample of the points from one of the processed files, which contained activity in downtown Atlanta. The maps show points that were classified as noise by the various methods, with the shade of gray varying based on each point's instantaneous speed (converted to miles per hour). The maps indicate that the Stopher, Jiang, and Fitzgerald method only



Stopher, Jiang, and Fitzgerald (2005)

Lawson, Chen, and Gong (2010)

Schüssler and Axhausen (2008)

Figure 3-3. Sample of points from downtown Atlanta identified as noise by tested methods.

identified a small number of points as noise in this area; the maps also illustrate how the Lawson, Chen, and Gong method tended to filter out points around intersections. These were likely points with speeds below the prescribed 6.7 mph (3 m/s) but above the original data collection's minimum speed setting of 1.9 mph. Finally, the maps show that the Schüssler and Axhausen method did block some valid traces, but also captured some of the same noise identified by Lawson, Chen, and Gong.

GPS Data Cleaning Findings

Based on the test results, it is suggested that practitioners select devices (or data sources) that can provide instantaneous speed, as well as HDOP and number of satellites, given the importance that these data elements have in the process of filtering noise out of raw GPS trace data. Another finding from this effort was that further analyst review may be necessary after applying automated filtering to raw GPS data to deal with points that should have been identified as noise but were ignored by the methods (Type I error).

It is also worth noting that the Schüssler and Axhausen method arrived at sound results without relying on point quality indicators like HDOP and number of satellites. However, this method was much more complicated to implement and took significantly longer to process than the other methods. (It was at least 10 times slower.) In the end, the Lawson, Chen, and Gong method clearly performed the best, which shows the importance of having access to the number of satellites, HDOP, and instantaneous speed in the raw GPS data.

Trip Identification Methods

Trip identification methods can take clean GPS points as input and generate a list of trips as output. These methods are not able to detect mode transitions within multimodal trips and are hence more appropriate for use with vehicle-based GPS data. They can also be used to generate trips whose point sequences can be further processed into separate mode segments using mode transition detection methods (reviewed in the next section).

The test consisted of comparing trips identified in the reference data with those generated by the evaluated methods. These methods are typically rule-based, and the criteria for these rules will require a simple calibration effort, which will heavily depend on how the data were collected (i.e., logging rules) and the performance of the GPS device used. As part of this test, the default values found in the original work were used to configure the methods. Input data for this test consisted of noise-filtered points in the original ARC person-based GPS data deliverable. These only included points that were used by delivered trips, which were limited by the original household's assigned travel date range. The devices used to collect the original GPS data were configured to record a point every 3 s, and only points with instantaneous speeds above 1 mph were recorded.

Success of the method was measured by comparing the detected trips against the trips in the reference data set. The null hypothesis in this test was that the same set of trips was identified in both data sets. Trips were deemed to match if their end locations were approximately the same between the reference and processed data sets, where "approximately the same" means the end times of the two data sets were within 15 min and the start and end locations were within 75 m. Based on this, the following errors were computed:

- Type I error: Trip is not found in method but is found in reference data.
- Type II error: Trip is found in method but is not found in reference data.

The trip identification methods were run against 300 randomly selected processed GPS files from the ARC GPS data set, which included 336 individual linked trips. The percentages for Type I errors were calculated by taking the number of trips that did not match the reference data and dividing that by the total number of generated trips. The percentages for Type II errors were calculated by taking the number of reference trips that did not match any generated trips and dividing that by the total number of trips in the reference data. Table 3-4 presents findings from the process of implementing the tested identification methods along with the failure rates of the two detected errors.

Both methods ended up generating fewer trips than what was contained in the reference data set, and this was reflected in the higher rates of Type II errors (trip is found in method but not in reference data). The simple approach proposed by Wolf, Guensler, and Bachman showed a slightly higher Type I (trip is found in reference data but not in tested method) error rate than the one proposed by Schüssler and Axhausen. This result was expected given that the method has no mechanism for detecting short stops (i.e., those which last less than the 120-s threshold). However, this simpler approach ended up with a lower Type II error rate, which indicates that it is less likely to erroneously consider short stops as valid trip ends. The higher Type II error rate found for Schüssler and Axhausen was caused by the fact that the method consistently failed to find trip ends arriving at a smaller number of trips than the method proposed by Wolf, Guensler, and Bachman. An examination of the characteristics of the trips identified by the two tested methods also revealed that the trips identified by the Schüssler and Axhausen method tended to be longer and to have lower average speeds, which is consistent with failing to identify stops. It is possible that this was a side effect of the method's original design, which was customized to a specific data collection device and may not be well suited for processing data that were already filtered for noise points.

Table 3-4. Trip identification methods tested and results from Experiment A.

Source References	Implementation Findings	Type I Error	Type II Error
Wolf, Guensler, and Bachman (2001)	This method consists of calculating point dwell time based on subsequent time steps. Trip ends are then identified based on a minimum delay threshold of 120 s. The method was straightforward to implement.	4.25%	10.27%
Schüssler and Axhausen (2008)	The exact sequence of steps included in the paper was unclear. The test implementation first did a pass over the data to determine the dwell time and activity detection based on point density and time. It also included a short series of points with the activity. Then a second pass was made to determine whether any of the remaining point sequences had a density ratio higher than 2/3. The rest of the data was split into trips. Even though there were several rules applied to the data, the bulk of the detection typically occurred based on the first point density rule. The dwell time activity detection rarely happened, possibly due to the high threshold specified in the paper. There were also only a few instances where the second pass for activity detection found any activities.	3.09%	35.92%

Trip Identification Findings

Based on the results of this test, it is suggested that trips found using automated methods be reviewed for potentially missed short stops. The test also made it clear that a very simple approach like the one proposed by Wolf, Guensler, and Bachman can generate a reasonable first estimate of trip ends. The simplicity of the method also makes it very computationally efficient, thus making it suitable for real-time processing.

Mode Transition Identification Methods

In instances where multimodal travel was captured using GPS, it is necessary to further parse identified trips into mode segments (also referred to as unlinked or elemental trips). Each of these segments features a consistent travel mode. For example, a typical single-leg transit trip will consist of a sequence of three mode segments: walk → bus → walk. The methods presented in this section take as inputs the points belonging to trips and find mode transitions within them.

The null hypothesis in this test was that the mode transition points identified by the method would match the reference data. The test involved passing the filtered GPS points from the reference data (from the person-based GPS subsample of the ARC HTS data set) to the implemented algorithm and then comparing the resulting mode segments against the unlinked trips in the reference data. Unlinked trips were deemed to match if their end locations were within 75 m between the reference and processed data sets and had end times that were within 15 min of each other. The two following errors were detected from this check:

- Type I error: Mode segment end point is not found in method but is found in reference data.
- Type II error: Mode segment end point is found in method but is not found in reference data.

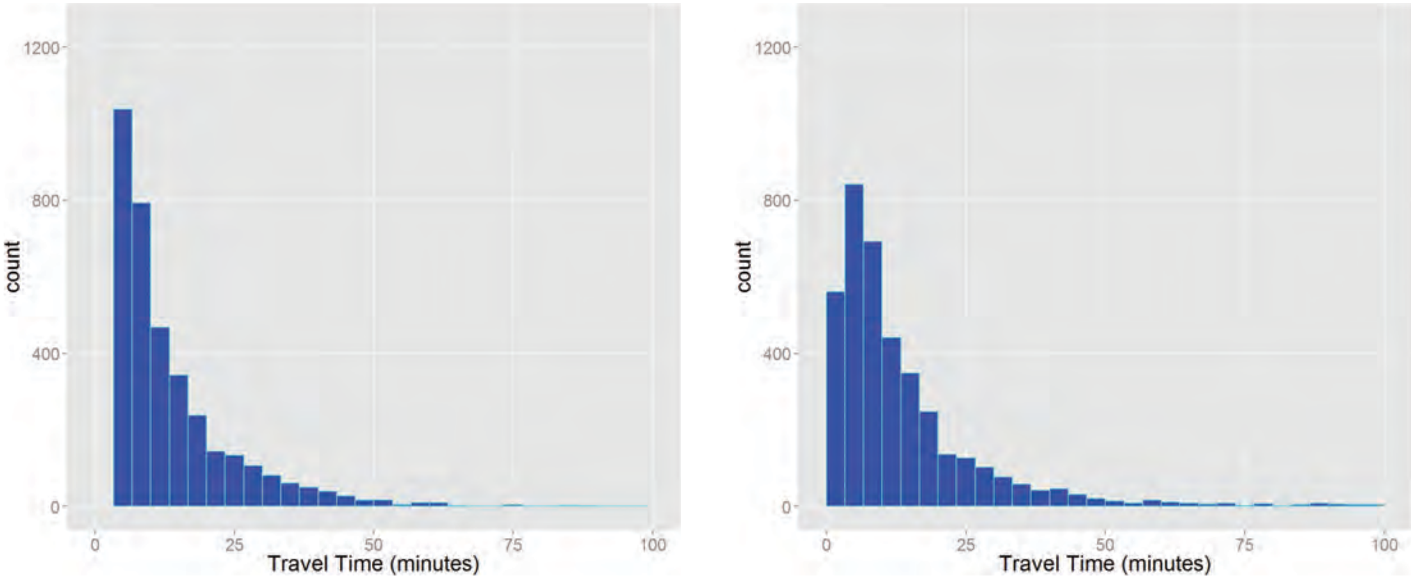
Table 3-5 presents observations based on the exercise of programming the tested mode transition identification method along with the two detected errors. The methods were applied to the filtered GPS points from 300 randomly selected person-based GPS files in the reference data (same set used in the trip identification test) and their corresponding mode segments.

The tests revealed that the first method (Tsui and Shalaby 2006; Schüssler and Axhausen 2008) clearly performed better, with lower Type I (mode transition is not detected by method) and Type II (mode transition end point is found in method but not in reference) error rates. And while the first method showed lower Type I error rates, the second method featured higher Type II error rates.

Examining the distribution of travel times of the identified mode segments in Figure 3-4 reveals that the first method tended to identify shorter mode transitions, likely to be short walk segments. These short segments were attached to longer, and likely motorized, mode segments in the second method. This is consistent with the fact that the second method found many fewer (a 10% decrease) mode segments than what was present in the reference data. On the other

Table 3-5. Mode transition identification method tested and results.

Source References	Implementation Findings	Type I Error	Type II Error
Tsui and Shalaby (2006) and Schüssler and Axhausen (2008)	The papers were unclear about whether to discard an SOW point if another SOW point is detected before an EOW or EOG point is detected. Also unclear about whether to discard an SOW point if an EOG point fails to be an EOW point but could be another SOW. The tested implementation keeps the first SOW in both cases. If no SOW points are detected, the entire file is considered non-walk. The initial implementation tended to keep data together in segments, even though long dwell times were present in its points. To account for this, the implementation added a step that ended a mode segment if a dwell time of at least 120 s was found.	16.58%	9.59%
Oliveira et al. (2011)	This method produces elemental trips (separate trips per mode or unlinked trips), so it was treated as a mode transition identification method. The implemented logic did not create places as per the paper but mode segments instead so that the results could be compared with the other tested method. Since places bound mode segments, this was a straightforward change. An error was found in paper about multiplying the segment's average speed by 1.96 times the point speed standard deviation; it should be adding the result of the second multiplication to the segment's average speed. Finally, since the input data used was already filtered of noise, the original paper's noise filtering logic was disabled for this test.	24.76%	31.96%



Tsui and Shalaby (2006) and Schüssler and Axhausen (2008)

Oliveira et al. (2011)

Figure 3-4. Travel time distribution of identified mode segments.

hand, the first method also identified a higher number (a 3.8% increase) of mode segments than what was reported in the reference data.

Mode Transition Identification Findings

The method originally proposed by Tsui and Shalaby (2006) and later refined by Schüssler and Axhausen (2008) produced better results than the method proposed by Oliveira et al. (2011), with error rates that were 1.5 to 3 times smaller. Furthermore, the first method was also better at capturing short nonmotorized mode segments, which typically occur before and after motorized travel. It should be noted, however, that the first method ended up identifying more mode segments than what was present in the reference data.

Experiment A: Classifier Data Fusion Methods

This test covered the evaluation of classified methods in the context of travel mode and trip purpose identification. Three types of methods were evaluated: heuristic, probabilistic, and AI. The input data for the tests consisted of processing subsets of the unlinked trips present in the reference data set by using the evaluated methods. Some of the methods required calibration (or machine learning), and in these cases the data set was split into calibration and validation groups. The methods presented in this section correspond to the boxes with horizontal hashing presented in Figure 3-1.

Travel Mode Identification Methods

To evaluate the performance of the travel mode identification methods selected from the literature, a set of unlinked GPS trips from the reference data was obtained. The set was constructed so as to have the same number of records per evaluated travel mode. This set was then further divided into calibration and validation records. The first contained 36 records for each evaluated travel mode, while the second included 18 records per mode set. The travel modes used in the tests were walk, bicycle, auto, bus, and heavy rail. Initially, these records had no attributes (other than an identifier), but various attributes were computed from the GPS points that the trips covered, according to the needs of each method.

The null hypothesis for the tests was that the mode selected from the method matched the reference data. Based on this, the following errors were computed:

- Type I error (detected wrongly): Mode classified does not match reference data.
- Type II error (failed to detect): Reference data mode does not match mode classified.

Note that Type I and Type II are paired for this hypothesis (picking a wrong answer implies that you failed to pick the right answer), but it is still interesting to treat them separately to see what modes are frequently overused (Type I) and underused (Type II). Table 3-6 identifies the tested methods and presents findings that derived from the implementation of the methods.

The validation step's goal was to estimate the reliability of the developed process (or model) and to document its performance. Reliability of the calibrated models was assessed by applying them to the validation portion of the data sets and then comparing predicted purposes to actual respondent choices. Classification errors were tabulated as a function of the actual choices selected by respondents, and the distribution of imputed trip purposes was compared to that of the validation data sets.

A confusion matrix, also referred to as a prediction-success table in travel forecasting, was constructed for each method of application. This is a matrix that shows actual choices as rows and modeled outcomes as columns; correct classifications appear on the matrix's diagonal. Within the context of a confusion matrix, Type I errors are the sum of each column without the diagonal value, and likewise the Type II errors are the sum of each row without the diagonal value. It should be noted that the mode-specific error percentages in these matrices are not supposed to add up to 100%; that is because they are generated using different totals (i.e., the horizontal or vertical sum of outcomes for a given mode, not the sum of outcomes across all modes). Type I error rates are computed with respect to column totals, while Type II error rates are calculated using row totals. Table 3-7, Table 3-8, Table 3-9, and Table 3-10 present the confusion matrices for the four tested travel mode identification methods.

Table 3-11 and Table 3-12 summarize the Type I and Type II errors for the tested travel mode identification methods. They show the error rates by mode across rows and tested method across columns.

The heuristics-based method worked best with walk and bicycle trips, but performed poorly with bus trips. It was also not effective at differentiating between auto and bus modes, and failed to classify most bus mode segments as such. Another limitation of this approach is that it may result in some mode segments not being assigned a travel mode—11 in the case of this test. When rail lines travel along the same path as roads (for instance, a metro line that travels between the directions of a highway), this method has difficulty assigning to rail because to qualify for rail, a trip cannot be on the road network. This is the reason why a small buffer of 50 ft was used. The bicycle travel mode was also disqualified if the household reported that they had no bikes, even if the trip matched a biking signature.

Table 3-6. Travel mode identification methods.

Method Types	Source References	Implementation Findings
Heuristics	Stopher, Clifford, and Zhang (2007)	This method needed baseline statistics for each mode, so the 180 training trips were used. The implementation used a 50-ft buffer around road, rail, and bus stops, and to count as being on the road or rail networks, over 50% of the points in the trip had to be within that buffer. For bus mode, both start and end of trip had to be within 50 ft of a bus stop, but the path of the trip was not verified. The 95th percentile speeds in the training data were compared against the 85th percentile speeds in the test data. This was done to account for underestimation of higher speeds in the training samples, which was found in early applications of the method. This method can return N/A results for cases where the mode could not be classified.
Probabilistic	Oliveira et al. (2006)	First, the model in the original paper had to be changed since no accelerometer-based physical activity data were available in the test data set. The first model specification included alternative specific constants and betas for average speed and standard deviation of acceleration. Although this original model's coefficients could be estimated, the resulting model performed very poorly (adjusted rho-square < 0), so a new specification was created that used dummy variables for low, middle, and high speed and acceleration levels by mode in addition to alternative specific constants. This final specification performed better (adjusted rho-square = 0.537), but the small number of observations made it so that the majority of its coefficients did not pass the *t*-test for significance. The final model specification is listed in Appendix D.
Fuzzy logic	Tsui and Shalaby (2006) and Schüssler and Axhausen (2008)	The low, 95th percentile acceleration (m/s^2) category was (0, 0, 0.5, and 0.6) in the papers, but the used data had negative 95th percentile acceleration values (where the person was mostly decelerating for the whole mode), so -9999 was used to cover these cases. The fuzzy logic gives a score of between 0 and 1 to each mode. If there was a tie for greatest value, a value was randomly picked between the winning values. (The papers did not specify a tiebreaker.) The random seed was set to 1 at the beginning of the process so that the same random values were chosen each time.
Neural networks	Gonzalez et al. (2008)	Given that the stopped time is defined as "a certain threshold" in the paper, an assumption was made to consider time spent at 5 mph or less to be stopped time. Since each point was 3 s apart, two successive points must be at or below 5 mph to count 3 s toward the stopped time. Furthermore, HDOP and percent cell-ID fixes were not available. (All points were captured by GPS.) The paper did not specify how to determine stops within a trip; since stops are used to track bus stops, a minimum of 20 s of dwell time was used to identify stops. The test used a learning rate of 0.1 and 300 iterations as given in the final results of the paper, rather than computing which thresholds were best as the paper did. The researchers also did not make use of critical points as suggested in the original paper.

Table 3-7. Heuristic confusion matrix (55/90 = 61% correct, 12% indeterminate).

Ref\Method	Walk	Bicycle	Auto	Bus	Heavy Rail	N/A	Type II
Walk	17	0	1	0	0	0	1 (06%)
Bicycle	0	10	3	0	0	5	8 (44%)
Auto	0	0	17	0	0	1	1 (06%)
Bus	0	0	16	2	0	0	16 (89%)
Heavy rail	1	0	3	0	9	5	9 (50%)
N/A	0	0	0	0	0	0	–
Type I	1 (6%)	0 (0%)	23 (58%)	0 (0%)	0 (0%)	11 (100%)	

Table 3-8. Probabilistic confusion matrix (56/90 = 62% correct).

Ref\Method	Walk	Bicycle	Auto	Bus	Heavy Rail	N/A	Type II
Walk	14	4	0	0	0	–	4 (22%)
Bicycle	2	16	0	0	0	–	2 (11%)
Auto	0	0	9	3	6	–	9 (50%)
Bus	0	1	7	7	3	–	11 (61%)
Heavy rail	1	0	5	2	10	–	8 (44%)
N/A	–	–	–	–	–	–	–
Type I	3 (18%)	5 (24%)	12 (57%)	5 (42%)	9 (47%)	–	

Table 3-9. Fuzzy logic confusion matrix (58/90 = 64% correct).

Ref\Method	Walk	Bicycle	Auto	Bus	Heavy Rail	N/A	Type II
Walk	17	1	0	0	0	–	1 (06%)
Bicycle	1	16	0	1	0	–	2 (11%)
Auto	0	0	18	0	0	–	0 (00%)
Bus	0	0	14	4	0	–	14 (78%)
Heavy rail	1	0	13	1	3	–	15 (83%)
N/A	–	–	–	–	–	–	–
Type I	2 (11%)	1 (6%)	27 (60%)	2 (33%)	0 (0%)	–	

Table 3-10. Neural net confusion matrix (74/90 = 82% correct).

Ref\Method	Walk	Bicycle	Auto	Bus	Heavy Rail	N/A	Type II
Walk	17	0	1	0	0	–	1 (06%)
Bicycle	0	17	1	0	0	–	1 (06%)
Auto	0	0	10	5	3	–	8 (44%)
Bus	0	2	1	14	1	–	4 (22%)
Heavy rail	1	0	0	1	16	–	2 (11%)
N/A	–	–	–	–	–	–	–
Type I	1 (6%)	2 (11%)	3 (23%)	6 (30%)	4 (20%)	–	

Table 3-11. Mode identification Type I error rates by travel mode and method.

Ref\Method	Heuristics	Probabilistic	Fuzzy Logic	Neural Network
Walk	6%	18%	11%	6%
Bicycle	0%	24%	6%	11%
Auto	58%	57%	60%	23%
Bus	0%	42%	33%	30%
Heavy rail	0%	47%	0%	20%
N/A	100%	–	–	–

Table 3-12. Mode identification Type II error rates by travel mode and method.

Ref\Method	Heuristics	Probabilistic	Fuzzy Logic	Neural Network
Walk	6%	22%	6%	6%
Bicycle	44%	11%	11%	6%
Auto	6%	50%	0%	44%
Bus	89%	61%	78%	22%
Heavy rail	50%	44%	83%	11%
N/A	–	–	–	–

The final probabilistic model worked well for nonmotorized modes but had a difficult time discerning between motorized modes, particularly bus and auto. This was likely due to the limited attributes used as independent variables (i.e., all based on speed and acceleration). As with the probabilistic approach, the fuzzy logic method worked best with nonmotorized modes and in making the distinction between them and motorized alternatives. But it also struggled when selecting between motorized alternatives.

The neural network method worked the best across all alternative modes. As with the other methods, its main challenge was differentiating between auto and bus modes, but even in these cases it fared better than all other methods. A key difference here was that the network calibration data included information about time spent at or below 5 mph, which may have helped with differentiating between bus and auto modes.

Interestingly, the neural net was the only method that had a nearly even overuse and underuse bias toward bus mode (the other methods heavily under-selecting it). As observed in most of the papers referenced here, detecting walk and bicycle modes was relatively easy for all methods. The fuzzy logic thresholds were taken as written in the papers, rather than computing them as the paper specified. It is possible that this method would have performed better if the thresholds were calibrated using the 180 training cases that the other methods used.

Mode Identification Findings

Based on these test results, the neural network method should be employed whenever calibration data are available given its lower error rates. If no calibration data are available, then the next best approach should be to use the fuzzy logic method, which performed reasonably well with no calibration. In cases where a calibration data set is not available to train a neural network and another method is selected, it may be necessary to perform additional review of motorized mode selections to properly classify them between competing choices. This additional review may be automated through the use of GIS transit infrastructure data, which helped lower the Type I error rate of the heuristics method for bus and heavy rail modes.

Trip Purpose Identification Methods

The research team tested two modeling techniques for identifying trip purpose: discrete choice modeling, using nested multinomial logit (NMNL) models, and decision trees (Griffin and Huang 2005). Both decision trees and NMNL methods can be calibrated using revealed trip purpose responses from existing HTS data and can then be applied to identify trip purpose for GPS-derived data or GPS-like data (i.e., containing only basic trip attributes). Decision trees have the benefit of graphically organizing the variables that go into trip purpose selection and, in turn, can be used to help direct the development of a model specification for the discrete choice method. For the purpose of this research effort, the team used the WEKA (http://www.cs.waikato.ac.nz/ml/weka/) data mining tool for estimating decision trees using the C4.5 method.

Probabilistic Method for Identifying Trip Purposes

A nested logit model structure was used for this test order to logically group choices according to aggregate purposes such as: at home, at work, nonwork, university or school, airport, and loop trip. At the same time, the participants were classified as belonging to one of eight life-cycle categories listed in Table 3-13.

The BIOGEME modeling tool was used to calibrate these nested logit models. Choice simulations were generated using the associated BIOSIM simulation to perform model validation. Although BIOGEME works in Microsoft Windows, it has known memory management problems in this platform. To avoid these issues, all computations were done using a Linux virtual machine.

Table 3-13. Person life-cycle categories.

ID	Category	Description
1	FT worker	Person is a full-time worker.
2	PT worker	Person works only part-time.
3	University student	Person is 18 years of age or older and a student.
4	Nonworker	Person is 18 years of age or older and does not work or go to university.
5	Retiree	Person is retired.
6	Driving-age school child (16–17 yrs)	Person is between 16 and 17 years of age and goes to high school.
7	Pre–driving-age school child (6–15 yrs)	Person is younger than 16 years of age and older than 5 years of age.
8	Preschool child (<6 yrs)	Person is younger than 6 years of age.

The first step was to categorize each household member into one of the eight categories shown in Table 3-13. If household members qualified for more than one category (e.g., university student and part-time worker), they were then classified into the lowest number for which they qualified in Table 3-13.

Once the data were imported, a series of computed variables were added to the place and person records. These variables and their definitions are identified in Table 3-14.

In addition to the variables listed in Table 3-14, additional dummy variables for specific time-of-day periods as well as their interactions were computed. For all trips in which participants counted themselves as another party member, that passenger was considered a non-household member (so the count of people on the trip stayed the same, but the count of household members on the trip was reduced by one). For determining if a place matched the previous destination, the variable "originisdestination" was computed; places were considered the same if they were within 75 m of each other and had the same name. The two following subsections provide details on some of the challenges encountered when processing the two test data sets selected for this task.

Spatial attributes were attached to the place records using relationships between their destination coordinates and GIS data sets. Table 3-15 identifies these variables, as well as the source spatial data, and also provides a description for how they were set.

It is worth noting that land use data for the Atlanta planning region was sparse in its classification, with only 23 land

Table 3-14. Explanation of computed variables used in the utility equations.

Variable	Description
nonworker	Person is not a worker.
mode	Trip mode
mode_aft	Mode of the trip to the next place
nonauto	Non-auto trip mode
tottr	Total travelers on the trip to current place
tottr_aft	Total travelers on the trip to the next place
actdur	Activity duration (minutes)
nonmand	Trip to a non-mandatory location other than home, usual school location, or usual workplace
transfervariable	Variable indicating possibility of a transfer between two non-auto modes
adultparty	Party of only adult members
childparty	Party of only child members
mixedparty	Party of both adult and child members
someonedropped	A person was dropped at this destination.
someonepicked	A person was picked up at this destination.
dropoffvariable	Variable indicating possibility of drop off
pickupvariable	Variable indicating possibility of pickup
worklocationmatch	Destination location is usual work location, but excluding work from home cases.
schoollocationmatch	Destination location is usual school location, but excluding home schooling cases.
subtourdummy	Set to one if the given trip is a part of a sub-tour (tour starting and ending at the primary destination of the main tour)
simplesubtour	A sub-tour in which only one destination is visited
complexsubtour	A sub-tour in which more than one destination is visited
hhmem	Number of household members, excluding the respondent on this trip
groupgroceryduration	A trip with a household member to a non-mandatory location and taking between 20 and 40 min., indicating possibility of grocery shopping
groupeatoutduration	A trip with a household member to a non-mandatory location and taking between 40 and 60 min., indicating possibility of a typical family eat-out trip
walkmode	Includes walk and wheelchair
bikemode	Includes bike, skates, skateboard, Segway, and scooter
grouprecreationduration	A trip with a household member to a non-mandatory location and taking between 110 and 150 min., indicating possibility of a typical family recreational trip
groupsocialvisitduration	A trip with a household member to a non-mandatory location and with activity duration greater than 150 min., indicating possibility of a typical family social visit
notworklocation	Destination location is not the usual workplace.

Table 3-15. Spatial variables added to the Atlanta Regional Travel Survey data set.

Variable	GIS Data Sets	Source	Description
nearchurch	Church and places of worship	ESRI (2010)	The destination geocode is within 150 m of a feature in the GIS data sets.
nearbigbox	Walmart, Target, and shopping mall locations	ARC (2010)	The destination geocode is within 150 m of a feature in the GIS data sets.
lu_commercial	LandPro	ARC (2010)	The destination geocode is contained by or within 25 m of a commercial land use area.
nearschool	School locations from all surveyed participants	ARC HTS	The destination geocode is within 150 m of a school location.
lu_institutional	LandPro	ARC (2010)	The destination geocode is contained by or within 25 m of an institutional land use area.

use categories available. This low level of resolution made it difficult to differentiate trip purposes for trip ends located at or near multipurpose land uses, such as attempting to differentiate areas in which health care is dispensed from ordinary commercial developments, government buildings, or schools.

Before a model specification was developed, the original list of trip purposes was simplified. This simplification involved collapsing all purposes that took place at home and work to "any other activities at home" and "work doing my job," respectively. The final list of input purposes contained 21 entries. This was done to consolidate purposes that were identified as either being too similar or very difficult to differentiate based on household, traveler, and trip characteristics. Table 3-16 shows the original ARC trip purposes and those that correspond to them in the processed (simplified) data set.

This basic nesting structure developed for estimating trip purpose for the Atlanta HTS data set is shown in Figure 3-5. This diagram includes some simplifications such as not listing home or loop purposes and only including one work purpose that can take place at the work location.

Utility equations for the final 21 trip purposes were defined using the computed variables added to the data set and the interactions between them. These interactions included choice and life-cycle–specific coefficients to all purposes (total of $8 \times 21 = 168$). The utilities included purpose-specific coefficients that captured the impact that certain trip attributes were expected to exert on specific activities (e.g., short activity duration, change in party size, and mixed part for pickup and drop-off events), disaggregate nest-specific coefficients (applied to all purposes under a specific aggregate nest), which were applied to time of day, and person and GIS variables using interactions (e.g., commercial land use for shopping, maintenance, eating out, and discretionary activities). This first model specification contained 295 utility coefficients to be estimated.

During the model's initial successful estimation runs, a large number of the estimated coefficients either failed the null hypothesis *t*-test or ended up with coefficient estimates for which BIOGEME could not estimate *p*-values. Furthermore, two purposes (#12: all other activities at school and #24: attend major sporting event) did not have enough observations to calibrate their coefficients. These two purposes only appeared in 49 and 46 places, respectively.

Two actions were taken to deal with these challenges. First the list of trip purposes was further simplified into 12 choices by combining shopping and maintenance activities [under #15 and #7 respectively; drive through (#7) is considered a maintenance activity for this model, as are vehicle service (#14), household maintenance (#16), health care (#19), and personal business (#20)] and by combining entertainment activities [indoor recreation or outdoor recreation (#23) and attend major sporting event (#24)].

Figure 3-6 shows the final nesting structure. Second, choice, and life-cycle coefficients were simplified so that they could be shared across activities on the second nest level. These changes improved the estimation results, and after three rounds in which coefficients were removed, a final model specification was obtained. The final model specification contained 150 estimated parameters (three of which were nesting coefficients) and had an adjusted R^2 equal to 0.54.

The two strongest positive coefficients in the final model were the ones associated with school and work location matches for school and work purposes (6.99 and 6.35), while the two strongest negative coefficients corresponded to person life-cycle coefficients for retired and nonworker persons and the school purpose (−3.02 and −1.38). Only one of the three nesting coefficient (dis_work) ended up not being significant, with the final estimated value being 1.0, which effectively collapses the nest into two choices at the root level.

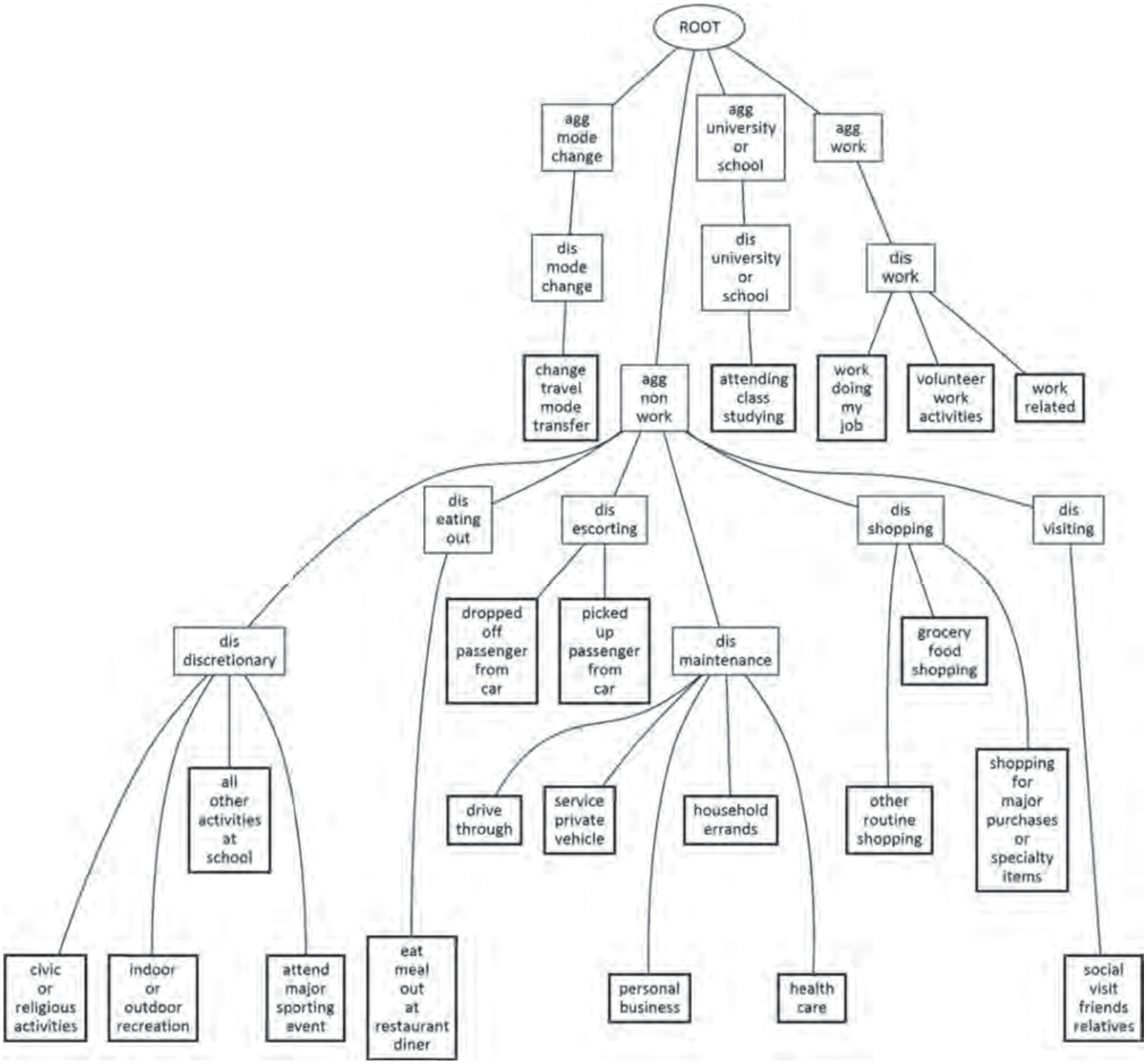
Table 3-16. Atlanta HTS trip purpose simplification.

ARC Trip Purpose		Model Purpose	
Code	Description	Code	Description
1	Working at home (for pay or volunteer)	N/A	
2	Shopping (online, catalog, or by phone)	N/A	
3	Any other activities at home	3	Any other activities at home
4	Change travel mode/transfer	4	Change travel mode/transfer
5	Dropped off passenger	5	Dropped off passenger
6	Picked up passenger	6	Picked up passenger
7	Drive through (ATM, bank, fast-food, etc.)	7	Drive through (ATM, bank, fast-food, etc.)
8	Work/doing my job	8	Work/doing my job
9	Other work-related activities at work	9	Work/doing my job
10	Volunteer work/activities	10	Volunteer work/activities
11	Attending class/studying	11	Attending class/studying
12	All other activities at school (eat lunch, recreational, etc.)	12	All other activities at school (eat lunch, recreational, etc.)
13	Work related (meeting, sales call, delivery)	13	Work related (meeting, sales call, delivery)
14	Service private vehicle (getting gas, oil, lube, repairs)	14	Service private vehicle (getting gas, oil, lube, repairs)
15	Grocery/food shopping	15	Grocery/food shopping
16	Other routine shopping (clothing, convenience store, household maintenance)	16	Other routine shopping (clothing, convenience store, household maintenance)
17	Shopping for major purchases or specialty items	17	Shopping for major purchases or specialty items
18	Household errands (bank, dry cleaning, etc.)	18	Household errands (bank, dry cleaning, etc.)
19	Health care (doctor, dentist, etc.)	19	Health care (doctor, dentist, etc.)
20	Personal business (visit government office, attorney, accountant)	20	Personal business (visit government office, attorney, accountant)
21	Eat meal out at restaurant/diner	21	Eat meal out at restaurant/diner
22	Civic or religious activities	22	Civic or religious activities
23	Indoor recreation or outdoor recreation	23	Indoor recreation or outdoor recreation
24	Attend major sporting event	24	Attend major sporting event
25	Social/visit friends/relatives	25	Social/visit friends/relatives
96	Loop trip	96	Loop trip
97	Other, specify	N/A	

Using BIOSIM, the NMNL purpose model was applied to the validation data set using Monte Carlo simulation; results were output for 10,316 of the 10,512 destinations without a home purpose. The average success rate of three enumeration runs was 60%. When including places with home purposes, which were not modeled, the overall success rate was 77%. A confusion matrix was generated using the mode from 10 enumerations of the validation set and is presented in Table 3-17 (see Table 3-16 for definitions of the trip purpose codes).

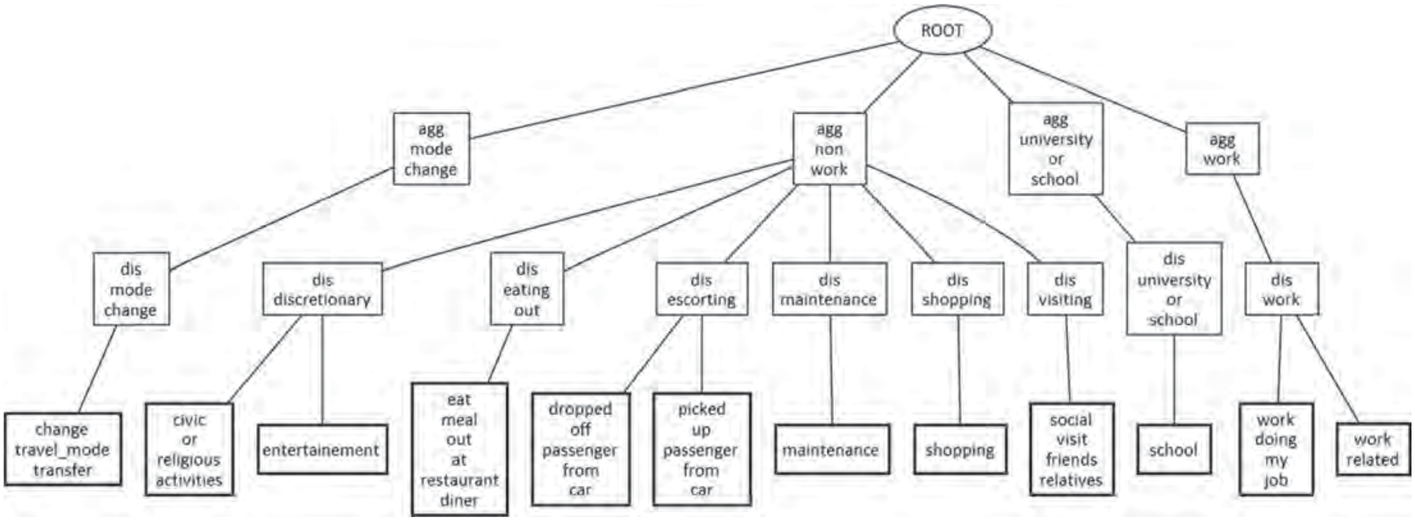
As seen in Table 3-17, the purposes that were incorrectly selected most often (Type I error) were #22 (civic or religious activity) with a 75% Type I error rate and #25 (social visit) with a 68% Type I error rate. These same trip purposes were also the reported purposes that most often showed different model results (Type II error), with the first (#22) showing a match in only 15 out of 173 reported instances (91% Type II error), while the second (#25) was correctly identified 103 out of 630 occurrences (50% Type II error). Another interesting observation is that the largest absolute cells outside the diagonal correspond to the reported-modeled and modeled-reported purpose pairs for activity #7 (maintenance) and #15 (shopping), which indicates that the model cannot easily differentiate between these two purposes.

Figure 3-7 shows the distribution of actual choices according to the modeled purposes; it illustrates the high degree of uncertainty that the model has for discretionary purposes. Match rates for each choice are shown at the top of each choice's bar.



Note: agg = aggregate, dis = disaggregate.

Figure 3-5. Full nested logit trip purpose network structure for ARC model.



Note: agg = aggregate, dis = disaggregate.

Figure 3-6. Final nested logit trip purpose network structure for ARC purpose model.

Table 3-17. Confusion matrix with results of the NMNL purpose model.

Reported / Modeled	*4*	*5*	*6*	*7*	*8*	*11*	*13*	*15*	*21*	*22*	*23*	*25*	*Type II Error*
4	137	49	40	123	–	2	7	34	3	–	12	9	67%
5	1	606	17	30	12	10	3	17	4	1	3	9	15%
6	3	11	521	20	13	3	1	32	3	1	9	2	16%
7	46	22	24	923	3	1	88	503	105	16	62	54	50%
8	2	–	1	11	1,516	–	17	7	1	–	6	7	3%
11	2	–	4	–	3	761	–	1	–	1	2	1	2%
13	31	12	26	186	33	2	161	66	62	4	24	23	74%
15	16	11	5	542	10	1	56	841	119	6	35	26	50%
21	13	5	16	164	2	3	29	289	207	3	38	23	74%
22	5	6	4	40	6	1	22	15	14	15	28	17	91%
23	42	23	20	90	5	6	36	99	70	8	153	45	74%
25	26	34	50	99	3	1	65	50	17	5	65	103	80%
Type I Error	58%	22%	28%	59%	6%	4%	67%	57%	66%	75%	65%	68%	

The purposes that were correctly identified most often were #8 (work/doing my job) and #11 (attending class/studying), which showed match rates of 97% and 98%, respectively, closely followed by #5 and #6 (dropping off/picking up passengers), with match rates close to 85%.

Decision Tree Method for Identifying Trip Purposes

Trip purpose decision trees were created using WEKA's open-source implementation of the C4.5 algorithm, called J48. In addition to the life-cycle Boolean variables listed in Table 3-13, the computed variables identified in Table 3-14, and the spatial variables in Table 3-15, the arrival hour was added to the list of inputs available to the tree-building algorithm. The tree was built using a confidence factor of 0.25 and at least 25 instances per leaf. The confidence factor determines how closely the tree conforms to the training set, and 0.25 is the default in both C4.5 and J48. The 25-instances-per-leaf setting keeps the tree from being overly specific by forcing every leaf to have at least 25 samples. A training sample size of 11,854 was used. The produced tree was able

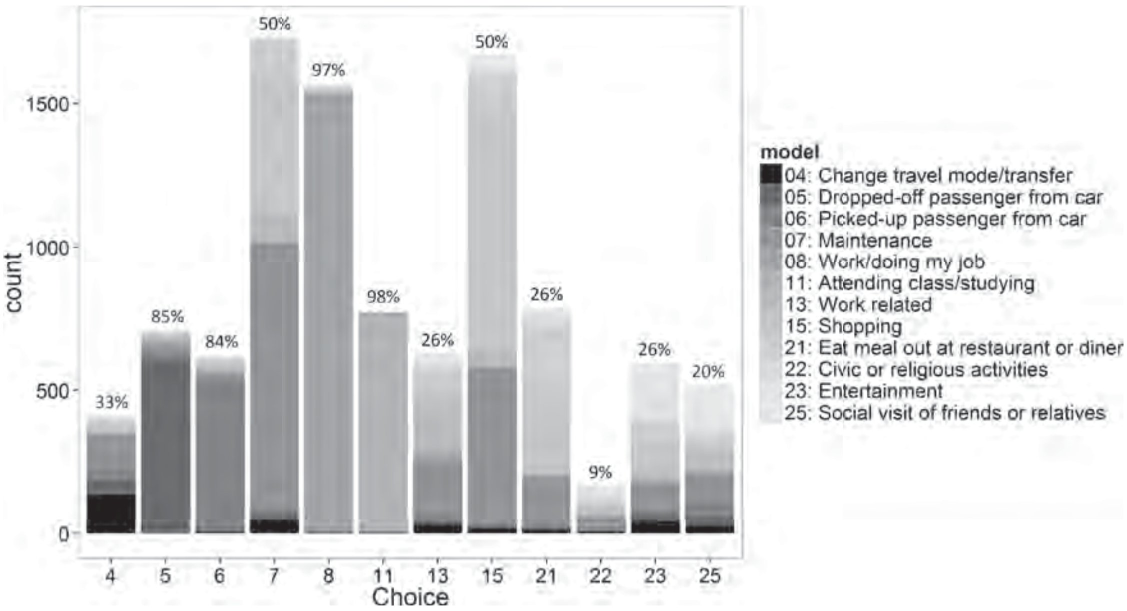


Figure 3-7. Actual choices in the validation data set by simulated choice.

to correctly classify 70% of the training data and 68% of the cross-validation data. The resulting decision tree included 369 decision nodes and is included in Appendix D. The most notable difference between the generated tree and the developed NMNL model was the tree's use of the same variable in paths within the overall the decision process, combined with the occurrence of multiple final decision nodes for the same purpose. These two factors make a complex tree like this one difficult to review and present.

The Atlanta HTS decision tree used the 10,512 trip destinations from the households in the GPS subsample and correctly classified 65% of their activity codes. Counting the places with home purposes, which were excluded from the model, as correctly identified resulted in a final 80% match rate. The purposes that were incorrectly selected most often were #16 (other routine shopping – clothing, convenience store, household maintenance) and #21 (eat meal out at restaurant/diner). Thus, when a mistake was made by the classifier, it was often to select one of these two. Table 3-18 presents the confusion matrix obtained after applying this decision tree to the reported trip purposes.

The reported purposes that were most often incorrectly identified by the classifier were #22 (civic/religious activities) and #23 (entertainment); these were incorrectly identified 66 times out of 104 (63%) and 223 times of 515 (57%), respectively. In other words, these two activities were the hardest to identify correctly and showed wider error ranges by the classifier. Similar to what was revealed in Table 3-17, the largest absolute cells outside the diagonal correspond to the reported-modeled and modeled-reported purpose pairs for activity #7 (maintenance) and #15 (grocery/food shopping); this indicates that the decision tree also cannot easily differentiate between these two purposes.

Figure 3-8 shows the actual trip purpose frequencies according to the selections identified by the decision tree. Match rates are noted as percentages on top of the choice bars.

Trip Purpose Identification Findings

The models developed as part of this research achieved accuracy levels comparable with previous efforts documented by Shen and Stopher (2012) and McGowen and McNally (2006), albeit without the post-validation tour logic added by Shen and Stopher. As expected, mandatory activities (i.e., going to work and school) were easier to identify than discretionary ones. This matches findings reported by Stopher et al. (2012). In fact, accuracies for mandatory purposes were above 95% in both modeling approaches. Escorting activities (i.e., pickup and drop off) were also identifiable to a good degree of accuracy (approximately 85%) using party size and companion information and showed relatively high match rates.

Non-mandatory activities proved harder to identify. This can be attributed to the greater variability displayed by the characteristics of places where these activities occur, particularly with respect to time and space. 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). Better spatial data would include extensive databases with the locations of places associated with discretionary activities; examples of places such as these are restaurants, gas stations, grocery stores, and government buildings. Unfortunately, the availability of public sources of these types of data varies greatly from region to region, so it

Table 3-18. Confusion matrix with deterministic results of decision tree purpose model.

Reported / Modeled	*4*	*5*	*6*	*7*	*8*	*11*	*13*	*15*	*21*	*22*	*23*	*25*	*Type II Error*
4	371	32	15	29	0	1	3	26	1	0	0	2	23%
5	13	609	13	29	4	8	7	4	6	0	10	10	15%
6	21	8	547	16	5	0	5	3	3	0	5	6	12%
7	32	20	31	1,112	15	3	82	417	58	13	48	55	41%
8	6	2	1	8	1,508	0	33	5	2	3	3	7	4%
11	1	0	2	1	3	759	1	1	0	0	3	4	2%
13	16	5	28	153	28	2	264	58	39	5	26	32	60%
15	18	9	12	544	7	4	44	898	88	2	28	20	46%
21	11	3	15	150	1	1	30	239	257	8	64	13	68%
22	1	2	0	30	6	1	7	14	18	38	33	26	78%
23	25	8	10	73	2	10	32	70	59	27	223	73	64%
25	15	23	39	73	1	6	42	47	24	8	72	201	64%
Type I Error	30%	16%	23%	50%	5%	5%	52%	50%	54%	63%	57%	55%	

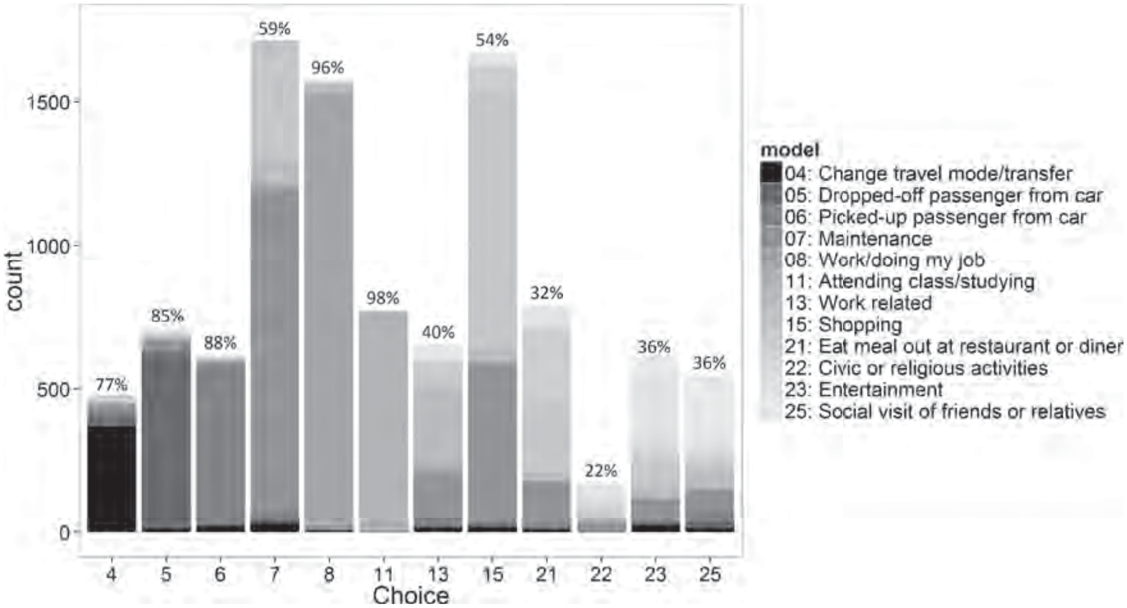


Figure 3-8. Distribution of actual choices in Atlanta HTS validation data set by tree choice.

may be necessary to investigate the availability of commercial data sources.

With regard to the two tested methods, it can be said that it was much faster to generate working models using decision trees than it was using the choice models. This is due to the complexity involved in specifying large models such as these, as well as the long run times needed to estimate model coefficients. For example, the final model specifications took, on average, 5 hours to conclude in BIOGEME, while decision trees could be built and evaluated in WEKA in less than 5 minutes. It should also be noted that the tasks associated with survey data preparation and preprocessing are not trivial, and adequate time should be factored in to project schedules to accommodate the development of automated procedures as well as multiple revision and correction cycles.

Based on this research, the following set of suggestions was developed. These are relevant for efforts that plan to use modeling to identify trip purposes using attributes that can be automatically derived from passively collected trace data (such as GPS data) as well as household and person characteristics.

- Research the availability of detailed land use and point of interest (POI) 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. This helps in the estimation of models and improves their success rates when applied to identify 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. This is especially important for mandatory purposes given that the single most important driving variable for these purposes was found to be a location match with the personal locations. It may be possible to derive work and school locations from data sets such as HAZUS (http://www.fema.gov/hazus).
- Capture frequently visited locations (along with activities conducted at them) as part of the recruitment process (for example, grocery stores, ATMs, and gas stations). The availability of these data will assist in disambiguating choices between competing non-mandatory purposes.

Overall Findings

One of the main findings of Experiment A is that the tested methods for automatically filtering noise out of GPS points, identifying GPS trips, and splitting trips into mode segments generated results that would likely require considerable manual review and cleaning before being deemed usable. This highlights the importance of software tools that can help by increasing the efficiency with which reviews and edits are performed on the processed results. Assuming that these extra steps are taken, these methods can be considered ready for implementation, with some of them having already been implemented in large-scale GPS-based HTS projects.

Regarding travel mode identification methods, it was found that neural networks should be used if calibration data are available. If no calibration data exist, then the next best approach found was the one using fuzzy logic rules. As with the GPS cleaning and processing methods, mode identification results would benefit from additional consistency and logic checks to avoid using unlikely mode sequences, and also would benefit from analyst review.

As for trip purpose identification, the methods evaluated here can be considered the most experimental ones and are likely further from being ready for implementation. However, they did show promising results once the purpose classification was simplified. Both evaluated methods performed well for mandatory and escorting purposes, but had difficulty differentiating between discretionary and maintenance activities. The biggest obstacles for the implementation of these methods are the availability of (1) detailed data on the households and persons within the households, (2) location information for locations most frequently visited by the household members, and (3) detailed land use and POI data.

Experiment B: Demographic Characterization of GPS Traces

Introduction

Experiment B was performed to evaluate various methods for attaching person- and household-level information to travel patterns observed in GPS-based household survey data and other sources of anonymized GPS trace data. For this reason, the experiment was designed to be as general as possible. Consequently, very few assumptions were made about what would be available in the input data sets, and all of the models were estimated using the limited amount of information that can be derived using the methods in Experiment A. The more detailed activity-travel and socio-demographic information available from the HTS data sets was generally ignored or used for validation purposes only. This allows the models developed to be generally applicable to any source of GPS trace data. The methods developed here, then, are appropriate for situations where there is an interest in demographically characterizing mass anonymous GPS traces for further analysis; for example, identifying facility utilization by user classes and developing segmented origin–destination estimates. The methods developed here are not intended to replace the HTS for purposes of travel demand model estimation. Since the process is in fact a simplified, inverted travel demand model, using the model results to estimate or calibrate further models would likely introduce substantial error.

Some assumptions regarding the input data were necessary for model estimation and should hold for any data set to which the model is applied. These include:

- The GPS traces can be uniquely linked to one person.
- The linked trace data covers at least one full day of travel.
- The workplace and school location of the person can be determined from the data or are available from other sources. It is possible that workplaces and school locations may be available from data sets such as HAZUS (http://www.fema.gov/hazus).
- Land use data are available for the region being modeled; this experiment used data from the Census Transportation Planning Package (CTPP).

Any data set meeting these limited assumptions should allow for the models in Experiment B to be applied. If further assumptions could be made regarding the input data (for example, "all household members are tracked and can be linked"), then the resulting person models could have more explanatory power; however, this was not done as part of this exercise so as to keep the models as general as possible.

The overall process flow diagram for Experiment B was shown in Figure 3-2. The procedure starts with a data processing step (Stage 0 in Figure 3-2), which is used to derive a series of person-level travel and tour characteristics used as input to the later model stages. The demographic characterization procedure for the GPS traces then proceeds in three stages (Stages 1–3 in Figure 3-2). First, aggregate-level person-type clusters are developed for later use. These clusters serve as the dependent variable for the later stages, so that in the demographic characterization process, the major type of the person is selected first, then more detailed demographics are modeled depending on the type of the person. In the second stage, the travel pattern data and local land use data are used to select one of the major person types. Finally, in Stage 3, all of the input data (i.e., the travel pattern data, land use data, and the selected major person type) are used to determine the various socio-demographic details of interest.

The outputs of Experiment B include open-source computer code, which processes the mode segment data resulting from Experiment A into person-travel records, and a set of model files that can be applied to the processed travel records using various open-source modeling packages, including WEKA and BIOGEME. Finally, a set of findings regarding the use of these procedures, limitations of the procedures, and areas for further development are given.

Data Processing

The primary output of Experiment A (i.e., the mode segments, which aggregate the GPS traces into trip records) is used as the

Table 3-19. Experiment B data processing routine input variables.

Variable	Data Type	Description
HH number	Unique identifier	Combined with person number to form unique ID
Person number	Unique identifier within HH	Combined with HH number, if no household data, set to 1
Activity ID	Unique identifier within person	Activity record number for person
Location type	String	Required location types: “home, work, school, other”
Location ID	Unique identifier	Unique identifier for physical location
Mode	Integer 1–10	Walk, bike, drive, pass, transit, paratransit, taxi, school bus, carpool
Duration	Integer	Trip duration
Activity duration	Integer	Defined as the time spent at trip end

Note: HH = household.

primary input to Experiment B. These data are approximated in the model estimation stage by using the Chicago Household Travel Survey trip record file and stripping out all information not found in the Experiment A results. The main variables in the input trip file are shown in Table 3-19. A sample of individuals who had recorded 2 days of travel where neither travel day was a weekend day was extracted from the full Chicago survey data for use in model estimation. This sample also excluded individuals who recorded no travel since these will not appear in GPS trace data. The data file described in Table 3-19 was then recreated from this sample. The demographic information for the sample was retained for model estimation and validation purposes.

As can be seen in Table 3-19, only eight variables, three of which are identifiers, are used in the development of the person-type and person-attribute models. The five trip record variables are converted using the data processing routine to the set of person-level aggregate travel characteristics in Table 3-20. These routines first look for tour patterns in the trip records, based on the home and work anchor points, and repeated stops at the same destination. Then, the characteristics of the tours are determined, including the number and type of stops, the modes used on tour legs, and the length of time spent in activities and traveling. The full set of derived travel characteristics is shown in Table 3-20.

Finally, these derived travel characteristics are combined with a set of simplified land use variables derived from the CTTP at the census-tract level. These variables describe the built environment and basic land use characteristics such as employment, housing, and population density. The characteristics of the land use variables for the Chicago data set are listed in Table 3-21.

These data sets, taken together, form the basis of the model estimation procedure described in the following section.

Model Estimation

The first task after data processing in Experiment B was the development of the primary demographic clusters in the Stage 1 analysis shown in Figure 3-2. During the completion of the first stage of Experiment B, several subtasks were performed. First, the tour and daily activity pattern variables extracted from the survey data were transformed into a set of primary factors using PCA. This was done to better understand the high levels of interdependence that existed between many of the variables. These factors were then clustered using the *K*-means clustering algorithm. Finally, the major person demographic types were developed using a partial decision tree classification algorithm (PART), with the travel pattern cluster membership distribution serving as the univariate dependent variable.

The classification algorithm has the effect of choosing 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). Using these person clusters helps with the Stage 2 analysis, since it was determined that the person types vary with the travel characteristics. So, when the travel characteristics are used as the independent variables in Stage 2, stronger differences in person clusters (now the dependent variable) are likely to be seen. A description of the most useful demographic clustering model found in the Stage 1 analysis is shown in Table 3-22. These clusters will serve as the primary dependent variables for the Stage 2 analysis.

A variety of models are used to classify individuals in terms of the five primary person-level attributes, which are person/worker type derived from the Stage 1 analysis, educational attainment, gender, possession of 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. A variety of modeling procedures from machine learning, choice modeling, and so forth were evaluated, and the final models are discussed in the following. The final models were selected for performance reasons, but also to give a representation of both joint and non-joint modeling methods, as well as regression-type models versus machine

Table 3-20. Processed person-level travel characteristics.

Variable Name	Description	Avg	Min	Max
total_tours	Number of total tours per day	2.850	1	13
num_subtours	Number of subtours per day	0.017	0	2
work_tours	Number of work tours	0.743	0	4
school_tours	Number of school tours	0.303	0	5
other_tours	Number of other tours	1.804	0	13
avg_stops_per_tour	Average number of stops per tour	2.351	1	12
avg_stops_per_work_tour	Average number of stops per work tour	1.070	0	10
avg_stops_per_school_tour	Average number of stops per school tour	0.308	0	8
avg_stops_per_other_tour	Average number of stops per other tour	1.689	0	12
avg_tour_ttime	Average travel time per tour	57.843	0	841
avg_work_tour_ttime	Average travel time per work tour	32.233	0	1,050
avg_school_tour_ttime	Average travel time per school tour	6.029	0	423
avg_other_tour_ttime	Average travel time per other tour	35.903	0	1,160
at_home_duration	Total time spent at home	1970.328	0	8,571
num_acts_work	Number of work activities	0.775	0	6
total_dur_work	Total duration of all work activities	377.092	0	2,878
avg_dur_work	Average duration of work activities	208.874	0	1,439
num_acts_school	Number of school activities	0.328	0	6
total_dur_school	Total duration of all school activities	123.748	0	2,878
avg_dur_school	Average duration of school activities	65.177	0	1,439
num_acts_pickdrop	Number of pickup/drop-off activities	0.234	0	10
total_dur_pickdrop	Total duration of all pickup/drop-off activities	1.418	0	90
avg_dur_pickdrop	Average duration of pickup/drop-off activities	0.587	0	20
num_acts_other	Number of other activities	3.920	0	32
total_dur_other	Total duration of all other activities	290.168	0	2,878
avg_dur_other	Average duration of other activities	79.994	0	1,439
auto_total	Percentage of tours by auto mode	0.769	0	1
auto_work	Percentage of work tours by auto mode	0.326	0	1
auto_school	Percentage of school tours by auto mode	0.095	0	1

Table 3-21. Census-tract–level information for the Chicago data set from CTPP.

Variable	Description	Avg	Min	Max
Transit use	Percentage of residents using transit	0.122	0.00	0.61
Road density	Miles/sq mile of road	17.248	2.12	43.66
Intersection density	Intersections/sq mile	161.203	5.36	650
Block size	Average block size (intersection density/road density)	0.108	0.05	0.40
Employment density	Jobs/sq mile	4.237	0.01	68.42
Population density	Inhabitants/sq mile	8.561	0.03	92.95
Housing density	Housing units/sq mile	3.823	0.01	78.95

Table 3-22. Optimal person-type clusters.

Class	Description	% of Sample
1	Part-time workers	10.4%
2	Full-time workers	45.0%
3	Retirees	13.0%
4	Young children	6.4%
5	School children	11.8%
6	Others	13.4%

learning approaches. An overview of the modeling components and procedures is shown in Table 3-23. Actual model specifications are included in Appendix E.

The first model estimated and applied when performing the demographic characterization process is the person-type choice model, which combines broad person-type categories (i.e., full-time worker, part-time worker, retiree, children, schoolchildren, and others) with basic educational attainment, including no high school, high school graduate, and college graduate. This is referred to as the Stage 2 model in Figure 3-2. The model is estimated as a nested logit model jointly with the educational attainment to capture the substantial correlation between work/student status and education level. The person-type choice model was generally consistent with expectations. Higher numbers of work activities and longer duration of work activities are associated with being a worker, individuals with even longer durations being more likely to be employed full time. Those with more school activities are more likely to be either children or schoolchildren and are least likely to be retirees. The models show that land use characteristics are also related to work status, to some extent. Workers and children are more likely to live in high-density employment areas with lower housing density, while part-time workers, retirees, and others have more likelihood to live in denser housing areas, probably because these person types are less likely to live in family units. Finally, the inclusive value terms of the joint model demonstrate the strong correlation between educational attainment and work status. The education sub-models similarly conform to expectations. The full results are shown in Appendix E.

The ordinal logit model for age categories excludes the retiree, child, and schoolchild person categories from input since these categories also define age categories (65+, >16, and <16, respectively). Therefore, the model applies to part-time and full-time workers as well as other persons. The ordinal logit model was selected because there is a natural ordering to the age categories, with the various person-travel characteristics shifting the probability of being older (positive coefficients) or younger. For example, someone who lives in a densely populated, high-employment area and has more work, school, and pickup/drop-off activities is likely to be younger. This is intuitive as younger individuals tend to work or be in school. Conversely, those living in high housing density areas and high transit use areas are likely to be older. See Appendix E for a complete specification.

The gender binary logit model was necessary for the gender choice because very little in the daily travel patterns seems to discriminate between males and females, and the decision tree and rule-based classifiers tended to overestimate the presence of females in the sample. Some minor differences exist in that one who has more pickup/drop-off activities, and more numerous—but shorter—work activities is more likely to be female. Meanwhile, males are more likely to have longer work activities, to live in higher population density areas, and to travel further for discretionary tours. However, the differences observed are relatively minor, making this a difficult category to predict.

The possession of a driver's license is another difficult trait to model since the non-possession of a driver's license in adults is a somewhat rare characteristic, especially as children were excluded from the model. Such individuals represented less than 10% of the total sample. The PART decision list, while not providing the overall highest model fit, performs well at identifying the persons with no license, and since these persons are often of interest in travel demand modeling, this

Table 3-23. Model components and procedures.

#	Variable	Values	Model Procedure
1	Person/worker status	Part-time worker, full-time worker, retiree, child, schoolchild, other person type	Nested logit
2	Educational attainment	No high school, high school, college	with above
3	Age	0–16, 16–25, 25–45, 45–65, 65+	Ordered logit
4	Gender	Male, female	Binary logit
5	License	Yes, no	PART decision list
6	HH size	1, 2, 3+	J48 tree
7	Num vehicles	0, 1, 2+	with above
8	Has children	No, yes	with above

Note: HH = household.

criterion was important in the selection of the final model. The decision rules show that the individuals least likely to have driver's licenses are those who make relatively few tours (≤3) and have short durations for their discretionary activities. This is intuitive since those without driver's licenses are more likely to be constrained in their ability to engage in multiple tours and likely have to plan tours to work around mobility constraints. Conversely, individuals with long work durations and, naturally, those observed making trips by auto mode are most likely to have a driver's license. The rules in the decision list are shown in Appendix E.

Finally, a joint household characteristics model was estimated based on the individual tour patterns. The household-type model was estimated jointly for household size, number of vehicles, and presence of children. A joint model was selected here to maintain consistency between models since the selection of household size, for example, implies that certain values of the presence of children and number of vehicles variables are not available. The household-type joint model was estimated using the J48 decision tree to demonstrate joint modeling with a machine learning approach. The full tree is shown in Appendix E.

Model Application

The various models estimated for the person and household characteristics have been applied to both the Chicago data set, which was used as the training data, and the Portland household survey data set. The latter was processed using the data processing procedure and combined with the census-tract land use variables. The fit results for each model are discussed in the following paragraphs.

To evaluate the performance of the person-type model, the prediction matrices for the Chicago (CMAP) training data set and the unused test data set from the Portland household (HH) survey data were used (see Table 3-24). These results are from a probabilistic application of the model, where the person type is assigned to an individual randomly according to the modeled probability distribution. The table first shows the confusion matrix, which contains the counts of correctly and incorrectly classified examples. The second half of the table then shows the percentage correctly predicted for each person-type category, for both the null model and the probabilistic selection. The null model results are obtained by assigning each observation to a category with a probability

Table 3-24. Probabilistic application of the person-type model.

		CMAP – TRAINING RESULTS (67% split)								Prediction Results	
		Simulated Person Type									
		1.part	2.full	3.retiree	4.child	5.student	6.other	Total	%	Null model % correct	Model % correct
Observed Person Type	1	53	155	44	17	19	45	333	10%	1%	16%
	2	152	1,177	87	23	15	73	1,527	45%	20%	77%
	3	43	80	156	39	27	113	458	14%	2%	34%
	4	19	18	37	37	50	55	216	6%	0%	17%
	5	20	14	21	53	205	70	383	11%	1%	53%
	6	48	82	113	47	66	109	465	14%	2%	23%
	Total	335	1,526	458	216	382	465	3,382		27%	51%
	%	10%	45%	14%	6%	11%	14%				

		TEST RESULTS – PORTLAND HH SURVEY (100%)								Prediction Results	
		Simulated Person Type									
		1.part	2.full	3.retiree	4.child	5.student	6.other	Total	%	Null model % correct	Model % correct
Observed Person Type	1	152	464	184	91	83	192	1,166	10%	1%	13%
	2	430	3,267	331	144	72	294	4,538	41%	17%	72%
	3	110	184	413	181	83	312	1,283	12%	1%	32%
	4	45	44	94	100	145	134	562	5%	0%	18%
	5	88	27	50	206	775	268	1,414	13%	2%	55%
	6	214	341	492	285	328	508	2,168	19%	4%	23%
	Total	1,039	4,327	1,564	1,007	1,486	1,708	11,131		25%	47%
	%	9%	39%	14%	9%	14%	15%				

equal to the observed distribution (i.e., if 45% of people are full-time workers, then each observation has a 45% chance of being a full-time worker). The expected percent correct for the null model for each category is then the square of this value. The probabilistic assignment lowers the performance of the model somewhat as compared to a deterministic application of the model (i.e., where the highest probability category is always assigned for each observation) but produces more realistic distributions and provides better fit to infrequent classes. For comparison purposes, the deterministic application results are also shown in Table 3-24 and Table 3-25.

The results show that the person-type choice model, which forms the core of the demographic characterization process, is generalizable at least to the Portland data set, where it performs approximately as well as on the training data set estimation (Chicago model). This is significant since the Portland data were not used in model estimation, and it shows that the model is not likely to have been overfitted and is potentially transferable. In each case, the model can correctly predict between 47% and 51% of person types correctly, which for both applications is substantially higher than the null model expectation of 25% and 27%. One observation from the test results is that the probabilistic application of the model for Portland does not exactly replicate the observed distribution as in the training results for the Chicago model, since the constants in the model are fitted to the Chicago data. However, a calibration process could be undertaken when applying the estimated model to other areas where the alternative specific constants for each person type could be adjusted until a known distribution of the person types is matched. In the case of the Portland model, however, this is not particularly necessary since the distributions for Chicago and Portland are fairly similar.

The deterministic results shown in Table 3-25 are obtained by assigning the category with the highest probability for each example. The deterministic model results show similar characteristics to the probabilistic results, with the training model using the Chicago survey data somewhat outperforming the Portland model, as expected, although the difference in overall percent correctly predicted is not substantial. While the fit results appear better under the deterministic model application, as is often the case with this type of selection process, there is a substantial distortion in the simulated person classification distribution when compared to the observed distribution, which is why the deterministic assignment is not preferred when performing the demographic characterization process.

The educational attainment variable is estimated jointly with the person classification in the nested logit model formulation described previously. Once the results in the person classification process are obtained, the final educational attainment

Table 3-25. Deterministic application of the person-type model.

		CMAP – TRAINING RESULTS (67% split)								Prediction Results	
		Simulated Person Type									
		1.part	2.full	3.retiree	4.child	5.student	6.other	Total	%	Null model % correct	Model % correct
Observed Person Type	1	15	181	93	2	17	25	333	10%	0%	5%
	2	4	1,333	158	4	1	27	1,527	45%	45%	87%
	3		50	401	2	1	4	458	14%	0%	88%
	4	3	6	95	13	59	40	216	6%	0%	6%
	5		9	49	2	300	23	383	11%	0%	78%
	6	2	62	271	10	66	54	465	14%	0%	12%
	Total	24	1,641	1,067	33	444	173	3,382		45%	63%
	%	1%	48%	32%	1%	13%	5%				

		TEST RESULTS – PORTLAND HH SURVEY (100%)								Prediction Results	
		Simulated Person Type									
		1.part	2.full	3.retiree	4.child	5.student	6.other	Total	%	Null model % correct	Model % correct
Observed Person Type	1	24	521	442	14	68	97	1,166	10%	0%	2%
	2	24	3,634	756	24	14	86	4,538	41%	41%	80%
	3	3	152	1,090	6	1	31	1,283	12%	0%	85%
	4	1	28	245	34	177	77	562	5%	0%	6%
	5	2	16	135	25	1,167	69	1,414	13%	0%	83%
	6	17	309	1,250	47	333	212	2,168	19%	0%	10%
	Total	71	4,660	3,918	150	1760	572	11,131		41%	55%
	%	1%	42%	35%	1%	16%	5%				

Table 3-26. Educational attainment model results for training and test data.

CMAP – TRAINING RESULTS (67% split)							Prediction Results	
		Simulated Education Level						
		No High School	High School	College	Total	%	Null model % correct	Model % correct
Observed	No High School	489.2	120.0	150.8	760.0	22%	5%	64%
	High School	120.0	250.9	464.1	835.0	25%	6%	30%
	College	173.1	467.8	1146.1	1,787.0	53%	28%	64%
	Total	782.3	838.7	1,761.0	3,382.0		**39%**	**56%**
	%	23%	25%	52%				

PORTLAND – TEST RESULTS							Prediction Results	
		Simulated Education Level						
		No High School	High School	College	Total	%	Null model % correct	Model % correct
Observed	No High School	1,853	314	357	2,524	23%	5%	73%
	High School	569	817	1,446	2,832	25%	6%	29%
	College	767	1,560	3,448	5,775	52%	27%	60%
	Estimated	3,189	2,691	5,251	11,131		**39%**	**55%**
	%	29%	24%	47%				

model, which is conditional based on the person-type classification, is applied. The fit results are shown in Table 3-26.

The model results for the educational attainment classification show both the Chicago training model and the Portland test model performing very well when compared to the null model results, with a prediction potential of 56% versus 39%. The models both perform very well in identifying individuals without a high school degree and individuals with a college degree, although the model has some trouble identifying individuals with only a high school degree.

Next, the results for the age categorization model are shown in Table 3-27 for the Chicago and Portland data sets. The age categorization model classifies the sample data into five broad age categories, which are children (0–16), young adults (17–25), young middle age (26–45), older middle age (46–65), and seniors (66+), using ordinal logit regression. The results show the model performing reasonably well, with an approximately 50% improvement over the null model. Interestingly, the model performs marginally better for the test application using Portland data, although the differences are slight. The results show that the children and middle-age categories are relatively easy to predict, while the young adult category is difficult, which is expected since it is the most infrequently observed category. Additionally, when the classification for age is incorrect, it is generally only off by one category in either direction, with over 75% for both training and test applications within one level of the observed category.

The results of the gender classification model for Chicago and Portland are shown in Table 3-28. The gender classification model was a simple binary logit model. The results show that the model only slightly outperforms the null model for both cases, demonstrating that gender differences are not strongly reflected in differing travel patterns. The model does a better job of predicting classification as female, possibly reflecting the existence to a certain extent of unique travel pattern identifiers for females.

The last person-level classification model is for the possession of a driver's license, which is modeled using a PART decision rule set. The model is applied to all individuals in the sample over age 16. The results for Chicago and Portland are shown in Table 3-29. The possession of a driver's license is difficult to model since it is an unbalanced distribution, with the vast majority of individuals in the sample possessing a license. The results in the table show the difficulty of predicting which samples do not have licenses based only on observed travel patterns. It is relatively easy to identify which individuals have licenses, especially if the mode classification is included from Experiment A. However, individuals not having a license will appear similar to individuals who have a license and choose to use public transport, individuals who happened to not travel much on the survey day, and so forth. The model, however, does improve on the null model for both training and test data sets, and is highly sensitive to identifying travelers with licenses.

Table 3-27. Ordered logit age category model results for training and test data.

CMAP – TRAINING RESULTS (67% split)									Prediction Results	
		Simulated Age Category							Null model % correct	Model % correct
		0–16	16–25	25–45	45–65	65+	Total	%		
Observed	0–16	314	46	88	93	46	587	17%	3%	53%
	16–25	80	14	39	45	19	197	6%	0%	7%
	25–45	65	48	221	315	136	785	23%	5%	28%
	45–65	68	59	315	519	251	1,212	36%	13%	43%
	65+	28	22	128	266	157	601	18%	3%	26%
	Total	555	189	791	1,238	609	3,382		25%	36%
	%	16%	6%	23%	37%	18%				

TEST RESULTS – PORTLAND HH SURVEY (100%)									Prediction Results	
		Simulated Age Category							Null model % correct	Model % correct
		0–16	16–25	25–45	45–65	65+	Total	%		
Observed	0–16	1,338	155	257	249	118	2,117	19%	4%	63%
	16–25	253	45	128	166	77	669	6%	0%	7%
	25–45	233	152	683	948	407	2,423	22%	5%	28%
	45–65	253	211	1,086	1,732	810	4,092	37%	14%	42%
	65+	118	73	408	795	436	1,830	16%	3%	24%
	Total	2,195	636	2,562	3,890	1,848	11,131		25%	38%
	%	20%	6%	23%	35%	16%				

Table 3-28. Gender classification model results for training and test data.

CMAP – TRAINING RESULTS (67% split)						Prediction Results	
		Simulated Gender				Null model % correct	Model % correct
		Male	Female	Total	%		
Obs.	Male	768	807	1,575	47%	22%	49%
	Female	807	994	1,801	53%	28%	55%
	Total	1,575	1,801	3,376		50%	52%
	%	47%	53%				

PORTLAND – TEST RESULTS						Prediction Results	
		Simulated Gender				Null model % correct	Model % correct
		Male	Female	Total	%		
Obs.	Male	2,294	3,001	5,295	48%	23%	43%
	Female	2,418	3,400	5,818	52%	27%	58%
	Total	4,712	6,401	11,113		50%	51%
	%	42%	58%				

Table 3-29. Possession of driver's license classification model results for training and test data.

CMAP – TRAINING RESULTS (67% split)						Prediction Results	
		Simulated					
		Yes	No	Total	%	Null model % correct	Model % correct
Obs.	Yes	2,497	44	2,541	91%	83%	98%
	No	151	91	242	9%	1%	38%
	Total	2,648	135	2,783		84%	93%
	%	95%	5%				

PORTLAND – TEST RESULTS						Prediction Results	
		Simulated					
		Yes	No	Total	%	Null model % correct	Model % correct
Obs.	Yes	8,066	210	8,276	92%	84%	97%
	No	651	87	738	8%	1%	12%
	Total	8,717	297	9,014		85%	90%
	%	97%	3%				

Finally, the household-type joint model shows a similar pattern in performance as the person-level models, with the model performing nearly as well in Portland as it did in Chicago. The fit results can be seen in Table 3-30. Note here that the full misclassification matrix is not shown for the joint model since it is an 18 × 18 matrix of values and is difficult to interpret. Therefore, only the prediction potential versus null model is shown, which is equivalent to the right-hand side of the tables for the personal characteristics. The overall fit, which represents the number of observations with all three variables predicted correctly, is 45% for Chicago against 40% in Portland, both of which are higher than the null model expectation of 35%. The difference here is less substantial than in the other models, but this is as expected since fitting three categories simultaneously is a difficult task. Individual attribute models perform somewhat better than the null model expectation for Chicago, although there is no difference for the household size and number of vehicles categories when applied to Portland. However, the identification of the presence of children has a substantial increase, showing that the presence of children is related to differences in travel patterns, as expected.

Table 3-30. Training versus test model fit results for joint household-type model.

	Training – Chicago		Test – Portland	
	Model %	Null %	Model %	Null %
% correct overall	45%	35%	40%	34%
% correct hh size	56%	51%	55%	56%
% correct # vehicle	73%	68%	71%	72%
% correct has children	74%	43%	68%	58%

The remaining sub-models for person attributes perform similarly. Fit results for the remaining models are shown in Appendix E, along with the model estimates.

Findings

The results of the Experiment B demographic characterization process appear promising in that the models generally show substantial improvement over null model expectations, appear to be transferable to some degree, and are able to generate consistent person-characteristic estimates. These results are significant in light of the minimal data used as input to the demographic characterization process, meaning that 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.

Several key findings were observed in the performance of the demographic characterization experiment that can help improve the application of such a procedure and can help to guide the data collection process used to gather input traces.

- Multiday data collection is preferable to single-day data collection since it helps to average out intrapersonal day-to-day variation, which can be greater than interpersonal variation. This variability tends to confound the demographic characterization procedure (e.g., if a worker is surveyed on a nonworking day, the person will be very hard to identify as a worker). This finding aligns with previous research (Pas and Sundar 1995).

- It is also important to have reasonable estimates of workplace and school locations, either from access to more detailed location databases or from longer-term observation that can identify recurrent travel patterns.
- If it was possible to ensure that all household members were tracked and linked together, much better 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 early versions of the model effort, which did not conform to the assumptions made in 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. In other words, people travel the way they do because of who they are, but people are generally not who they are because of the way 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., a young child's 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 short-term collection efforts (most part-time workers work full shifts for fewer days).
- Finally, the joint modeling of attributes is difficult but provides benefits over modeling attributes separately, outside of just consistency, which is also important.

References

AirSage Inc. 2011. Origin–Destination Study for the Capital Area Metropolitan Planning Organization. Raleigh, NC: AirSage.

Alvarez-Garcia, J. A., J. A. Ortega, L. Gonzalez-Abril, and F. Velasco. 2010. "Trip Destination Prediction Based on Past GPS Log Using a Hidden Markov Model." *Expert Systems with Applications* 37 (12): 8166–8171.

Anderson, T., V. Abeywardana, J. Wolf, and M. Lee. 2009. *National Travel Survey GPS Feasibility Study.* London: National Centre for Social Research. http://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&ved=0CCIQFjAA&url=http%3A%2F%2Fwebarchive.nationalarchives.gov.uk%2F%2B%2Fhttp%3A%2Fwww.dft.gov.uk%2Fpgr%2Fstatistics%2Fdatatablespublications%2Fpersonal%2Fmethodology%2Fntsreports%2Fn.

Asakura, Y., and E. Hato. 2004. "Tracking Survey for Individual Travel Behaviour Using Mobile Communication Instruments." *Transportation Research Part C* 12 (3–4): 273–291.

Ashbrook, D., and T. Starner. 2003. "Using GPS to Learn Significant Locations and Predict Movement across Multiple Users." *Personal and Ubiquitous Computing* (Springer Science+Business Media) 7 (5): 275–286.

Atlanta Regional Commission. 2012. *ARC Regional On-Board Transit Survey – Final Report.* Atlanta: Atlanta Regional Commission. Accessed August 15, 2012. http://www.atlantaregional.com/File%20Library/Transportation/Travel%20Demand%20Model/tp_arcregionalonboardsurveyfinalreport_063010.pdf.

Auld, J. A., and A. Mohammadian. 2012. "Activity Planning Process in the Agent-based Dynamic Activity Planning and Travel Scheduling (ADAPTS) Model." *Transportation Research Part C* 8 (46): 1386–1403.

Auld, J. A., C. Williams, A. Mohammadian, and P. Nelson. 2009. "An Automated GPS-Based Prompted Recall Survey with Learning Algorithms." *Transportation Letters: The International Journal of Transportation Research* 1 (1): 59–79.

Axhausen, K. W., A. Zimmermann, S. Schönfelder, G. Rindsfüser, and T. Haupt. 2002. "Observing the Rythms of Daily Life: A Six-Week Travel Diary." *Transportation* 29 (2): 95–124.

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.

Bachu, P., T. Dudala, and S. Kothuri. 2001. "Prompted Recall in a GPS Survey: A Proof-of-Concept Study." Presented at the 80th 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.* Lexington, KY: FHWA.

Battelle Memorial Institute. 1999. *Heavy-Duty Truck Activity Data.* FHWA.

Battelle Memorial Institute. 2012. "Minnesota Road Fee Test." Proceedings of the ITS Minnesota 18th Annual Meeting and Information Exchange. ITS Minnesota.

Ben-Akiva, M., and D. Bolduc. 1987. "Approaches to Model Transferability and Updating: The Combined Transfer Estimator." *Transportation Research Record*, No. 1139, Washington, D.C.: TRB, National Research Council: 1–7.

Bernardin, Jr., V. L., J. Avner, J. Short, L. Brown, R. Nunnally, and S. Smith. 2012. "Using Large Sample GPS Data to Develop an Improved Truck Trip Table for the Indiana Statewide Model." *Proceedings of the 4th Annual Conference on Innovations in Travel Modeling.* Tampa: Transportation Research Board.

Bierlaire, M. 2003. "BIOGEME: A Free Package for the Estimation of Discrete Choice Models." *Proceedings of the 3rd Swiss Transportation Research Conference.* Ascona, Switzerland.

Blazquez, C. A., A. Ponce, and P. A. Miranda. 2010. "A Topological Decision-Rule Map - Matching Algorithms in Transportation Applications." *Proceedings of the 16th PANAM.* Lisbon.

Blazquez, C. A., and A. P. Vonderohe. 2005. "Simple Map-Matching Algorithm Applied to Intelligent Winter Maintenance Vehicle Data." *Transportation Research Record: Journal of the Transportation Research Board, No. 1935*, Washington, D.C.: Transportation Research Board of the National Academies, 68–76.

Bohte, W., and K. Maat. 2009. "Deriving and Validating Trip Purpose and Travel Modes for Multi-Day GPS-Based Travel Surveys: A Large-Scale Application in the Netherlands." *Transportation Research Part C* 17 (3): 285–297.

Bradley, M., J. Wolf, and S. Bricka. 2005. "Using GPS Data to Investigate and Adjust for Household Diary Data Non-Response." Presented at the 10th National Transportation Planning Applications Conference. Portland, OR: Transportation Research Board of the National Academies.

Bricka, S. 2008. "Non-Response Challenges in GPS-based Surveys." *Proceedings of the 8th International Conference on Survey Methods in Transport.* Annecy, France: ISCTSC.

Bricka, S., and C. Bhat. 2007. "Comparative Analysis of Global Positioning System-Based and Travel Survey-Based Data." *Transportation Research Record: Journal of the Transportation Research Board, No. 1972*, Washington, D.C.: Transportation Research Board of the National Academies, 9–20.

Bricka, S., and E. Murakami. 2012. "Advances in Travel Survey Technology." *Proceedings of the 13th International Conference on Travel Behaviour Research.* Toronto: ISCTSC.

Bricka, S., J. Zmud, J. Wolf, and J. Freedman. 2009. "Household Travel Surveys with GPS: An Experiment." Presented at the 88th Annual Meeting of the Transportation Research Board. Washington, D.C.: Transportation Research Board of the National Academies.

Broach, J., J. Gliebe, and J. Dill. 2009. "Development of a Multi-class Bicyclist Route Choice Model Using Revealed Preference Data." *Proceedings of the 12th International Conference on Travel Behavior Research.* Jaipur, India: IATBR.

Bullock, D. M., R. J. Haseman, and J. S. Wasson. 2010. "Real Time Measurement of Work Zone Travel Time Delay and Evaluation Metrics Using Bluetooth Probe Tracking." Presented at the 89th Annual Meeting of the Transportation Research Board. Washington, D.C.: Transportation Research Board of the National Academies.

Business Wire. 2012. *Virginia DOT Chooses TomTom Historical Traffic Data for I-95/I-64 Corridor Project.* May 21. Accessed July 04, 2012. http://www.businesswire.com/news/home/20120521005760/en/Virginia-DOT-Chooses-TomTom-Historical-Traffic-Data.

Butler, H., M. Daly, A. Doyle, S. Gillies, T. Schaub, and C. Schmidt. 2008. *The GeoJSON Format Specification.* June 16. Accessed October 25, 2012. http://www.geojson.org/geojson-spec.html.

Byon, Y., B. Abdulhai, and A. Shalaby. 2007. "Impact of Sampling Rate of GPS-enabled Cell Phones on Mode Detection and GIS Map Matching Performance." Presented at the 86th Annual Meeting of the Transportation Research Board. Washington, D.C.: Transportation Research Board of the National Academies.

Cambridge Systematics. 2007. *PSRC 2006 Household Activity Survey Analysis Report: Final Report.* Seattle: Puget Sound Regional Council.

Cambridge Systematics, Inc., 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.* Transportation Research Board of the National Academies, Washington, D.C.

Carpenter, C., M. Fowler, and T. Adler. 2012. "Generating Route Specific Origin–Destination Tables Using Bluetooth Technology." Presented at the 91st Annual Meeting of the Transportation Research Board. Washington, D.C.: Transportation Research Board of the National Academies.

Casas, J., and C. Arce. 1999. "Trip Reporting in Household Travel Diaries: A Comparison to GPS-Collected Data." Presented at the 78th Annual Meeting of the Transportation Research Board. Washington, D.C.: Transportation Research Board of the National Academies.

Center for Urban Transportation Research, University of South Florida. 2012. *TRAC-IT | Location Aware Information Systems Laboratory.* Accessed August 24, 2012. http://www.locationaware.usf.edu/ongoing-research/projects/trac-it/.

Chakirov, A. and A. Erath. 2012. "Activity Identification and Primary Location Modeling Based on Smart Card Payment Data for Public Transport." *Proceedings of the 13th International Conference on Travel Behaviour Research.* Toronto: IATBR.

Chang, F., J. Dean, S. Ghemawat, W. Hsieh, and D. Wallach. 2006. "Bigtable: A Distributed Storage System for Structured Data." *Proceedings of the 7th Symposium on Operating System Design and Implementation.* Seattle: Operating Systems Design and Implementation (OSDI).

Chen, C., G. Hongmian, C. Lawson, and E. Bialostozky. 2010. "Evaluating the feasibility of a passive travel survey collection in a complex urban environment: Lessons learned from the New York City case study." *Transportation Research Part A* 830–840.

Chiao, K., J. Argote, J. Zmud, K. Hilsenbeck, M. Zmud, and J. Wolf. 2011. "Continuous Improvement in Regional Household Travel Surveys: New York Metropolitan Transportation Council Experience." *Transportation Research Record: Journal of the Transportation Research Board, No. 2246*, Washington, D.C.: Transportation Research Board of the National Academies, 74–82.

Chicago Metropolitan Agency for Planning. 2008. *Travel Tracker Survey.* Accessed September 7, 2012. http://www.cmap.illinois.gov/travel-tracker-survey.

Chung, E-H., and A. S. Shalaby. 2005. "Development of a Trip Reconstruction Tool for GPS-Based Personal Travel Surveys." *Journal of Transportation Planning and Technology* 28 (5): 384–401.

Clark, A., and S. T. Doherty. 2010. "A Multi-Instrumented Approach to Observing the Activity Rescheduling Decision Process." *Transportation* 37 (1): 165–181.

Dalumpines, R., and D. M. Scott. 2011. "GIS-Based Map Matching: Development and Demonstration of Postprocessing Map-Matching Algorithm for Transportation Research." Presented at the 90th Annual Meeting of the Transportation Research Board. Washington, D.C.: Transportation Research Board of the National Academies.

de Jong, R., and W. Mensonides. 2003. *Wearable GPS Device as a Data Collection Method for Travel Research.* Institute of Transport Studies, Victoria, Australia: Monash University.

Doherty, S. T., and P. Oh. 2012. "A Multi-sensor Monitoring System of Human Physiology and Daily Activities." *Telemedicine and e-Health* 18 (3): 185–192.

Doherty, S. T., N. Noël, M. Lee-Gosselin, and C. Sirois. 2000. "Moving Beyond Observed Outcomes: Integrating Global Positioning Systems and Interactive Computer-based Travel Behaviour Surveys." *Transportation Research Circular E-C026*, Washington, D.C.: TRB, National Research Council, 449–466.

Elango, Vetri Venthan, and Randall 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.

Fasihozaman, M., T. Rashidi, and A. Mohammadian. 2013. "Investigating the Transferability of Individual Trip Rates: A Decision Tree Approach." Presented at the 92nd Annual Meeting of the Transportation Research Board. Washington, D.C.: Transportation Research Board of the National Academies.

Flamm, M., C. Jemelin, and V. Kaufmann. 2007. "Combining Person-based GPS Tracking and Prompted Recall Interviews for a Comprehensive Investigation of Travel Behaviour Adaptation Processes during Life Course Transitions." *Proceedings of the 11th World Conference on Transport Research.* Berkeley, CA: WCTRS.

Forrest, T. L., and D. F. Pearson. 2005. "A Comparison of Trip Determination Methods in GPS-Enhanced Household Travel Surveys." Presented at the 84th Annual Meeting of the Transportation Research Board. Washington, D.C.: Transportation Research Board of the National Academies.

Foster, D. 2004. *GPX: the GPS Exchange Format.* 8 9. Accessed August 31, 2012. http://www.topografix.com/gpx.asp.

Frignani, M., J. Auld, A. Mohammadian, C Williams, and P. Nelson. 2010. "Urban Travel Route and Activity Choice Survey (ULTRACS): Internet-Based Prompted Recall Activity Travel Survey Using GPS Data." *Transportation Research Record: Journal of the Transportation Research Board, No. 2183*, Washington, D.C.: Transportation Research Board of the National Academies, 19–28.

Furth, P. G., B. Hemily, T. H. J. Muller, and J. G. Strathman. 2006. *TCRP Report 113: Using Archived AVL-APC Data to Improve Transit Performance and Management.* Washington, D.C.: Transportation Research Board of the National Academies.

Giaimo, G., R. Anderson, L. Wargelin, and P. Stopher. 2010. "Will It Work? Pilot Results from the First Large-Scale GSP-Based Household

Travel Survey in the United States." Presented at the 89th Annual Meeting of the Transportation Research Board. Washington, D.C.: Transportation Research Board of the National Academies.

Gonder, J., T. Markel, M. Thornton, and A. Simpson. 2007. "Using Global Positioning System Travel Data to Assess Real-World Energy Use of Plug-In Hybrid Electric Vehicles." *Transportation Research Record: Journal of the Transportation Research Board, No. 2017*, Washington, D.C.: Transportation Research Board of the National Academies, 26–32.

Gonzalez, P. A., J. Weinstein, S. Barbeau, M. Labrador, P. Winters, N. Georggi, and R. Perez. 2008. "Automating Mode Detection Using Neural Networks and Assisted GPS Data Collected Using GPS-Enabled Mobile Phones." *Proceedings of the 15th World Congress on Intelligent Transportation Systems.* New York.

Google. 2012. *KML Reference - Key Hole Markup Language.* April 18. Accessed August 31, 2012. https://developers.google.com/kml/documentation/kmlreference.

Goulias, K. G., C. R. Bhat, and R. M. Pendyala. 2012. "Simulator of Activities, Greenhouse Emissions, Networks, and Travel (SimAGENT) in Southern California." Presented at the 91st Annual Meeting of the Transportation Research Board. Washington, D.C.: Transportation Research Board of the National Academies.

Greenfeld, J. S. 2002. "Matching GPS Observations to Locations on a Digital Map." Presented at the 81st Annual Meeting of the Transportation Research Board. Washington, D.C.: Transportation Research Board of the National Academies.

Griffin, T., and Y. Huang. 2005. "A Decision Tree Classification Model to Automate Trip Purpose." *Proceedings of the 8th International Conference on Computer Applications in Industry and Engineering.* Honolulu: ISCA. 44–49.

Guensler, R., and J. Wolf. 1999. "Development of a Handheld Electronic Travel Diary for Monitoring Individual Tripmaking Behavior." Presented at the 78th Annual Meeting of the Transportation Research Board. Washington, D.C.: Transportation Research Board of the National Academies.

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, 145–151.

Habib, K. and E. Miller. 2008. "Modeling Daily Activity Program Generation Considering Within-Day and Day-to-Day Dynamics in Activity-Travel Behaviour." *Transportation* 35: 467–484.

Hackney, J., F. Marchal, and K. Axhausen. 2005. "Monitoring a Road System's Level of Service: The Canton Zurich Floating Car Study." Presented at the 84th Annual Meeting of the Transportation Research Board. Washington, D.C.: Transportation Research Board of the National Academies.

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. 2009. "The WEKA Data Mining Software: An Update." *SIGKDD Explorations* 11 (1).

Hoh, B., M. Gruteser, H. Xiong, and A. Alrabady. 2006. "Enhancing Security and Privacy in Traffic-Monitoring Systems." *Pervasive Computing* (IEEE) 5 (4): 38–46.

Hood, J., E. Sall, and B. Charlton. 2011. "A GPS-Based Bicycle Route Choice Model for San Francisco, California." *Transportation Letters: The International Journal of Transportation Research* 3: 63-75.

Jan, O., A. J. Horowitz, and Z. R. Peng. 2000. "Using Global Positioning System Data to Understand Variations in Path Choice." *Transportation Research Record: Journal of the Transportation Research Board, No. 1725*, Washington, D.C.: Transportation Research Board of the National Academies, 37–44.

Karim, W. 2004. "The Privacy Implications of Personal Locators: Why You Should Think Twice Before Voluntarily Availing Yourself to GPS Monitoring." *Washington University Journal of Law & Policy* 14: 485–515.

Koh, A., X. Nguyen, and C. Woodard. 2010. *Using Hadoop and Cassandra for Taxi Data Analytics: A Feasibility Study.* Singapore: Singapore Management University. http://ink.library.smu.edu.sg/sis_research_smu/15.

Krygsman, S. C., and P. Schmitz. 2005. "Deriving Transport Data with Cellphones: Methodological Lessons From South Africa." *Proceedings of the 24th Annual Southern African Transport Conference.* Pretoria, South Africa. 696–705.

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

Lawson, C., C. Chen, H. Gong, S. Karthikeyan, and A. Kornhauser. 2008. *GPS Pilot Project, Phase Two: GPS Unit Comparison, Field Tests, and Market Analysis.* New York: New York Metropolitan Transportation Council.

Lee, J., P. Agnello, and J. Chen. 2011. "Origin–Destination Survey Data Collection - A Comparison of Bluetooth vs. Traditional Methods." Presented at the 13th National Transportation Planning Applications Conference. Reno, NV: Transportation Research Board of the National Academies.

Lee, M., A. Fucci, P. Lorenc, and W. Bachman. 2012. "Using GPS Data Collected in Household Travel Surveys to Assess Physical Activity." Presented at the 91st Annual Meeting of the Transportation Research Board. Washington, D.C.: Transportation Research Board of the National Academies.

Lee-Gosselin, M., S. Doherty, and D. Papinski. 2006. "An Internet-based Prompted Recall Diary with Automated GPS Activity-trip Detection: System Design." Presented at the 85th Annual Meeting of the Transportation Research Board. Washington, D.C.: Transportation Research Board of the National Academies.

Li, H., R. Guensler, and J. Ogle. 2005. "An Analysis of Morning Commute Route Choice Patterns Using GPS Based Vehicle Activity Data." Presented at the 84th Annual Meeting of the Transportation Research Board, Washington, D.C.: Transportation Research Board of the National Academies.

Li, Z. J., and A. S. Shalaby. 2008. "Web-based GIS System for Prompted Recall of GPS-assisted Personal Travel Surveys: System Development and Experimental Study." Presented at the 87th Annual Meeting of the Transportation Research Board. Washington, D.C.: Transportation Research Board of the National Academies.

Liu, L., C. Andris, and C. Ratti. 2010. "Uncovering Cabdrivers' Behavior Patterns from Their Digital Traces." *Computers, Environment, and Urban Systems* 34: 541–548.

Long, L., Lin J., and W. Pu. 2009. "Model-Based Synthesis of Household Travel Survey Data in Small and Midsize Metropolitan Areas." *Transportation Research Record: Journal of the Transportation Research Board, No. 2015*, Washington, D.C.: Transportation Research Board of the National Academies, 64–70.

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.

Mahmassani, H. S., and K. C. Sinha. 1981. "Bayesian Updating of Trip Generation Parameters." *Journal of Transportation Engineering* 107 (TE5): 581–589.

Marca, J. E. 2002. The Design and Implementation of an On-Line Travel and Activity Survey. Irvine: Center for Activity Systems Analysis.

Marchal, F., J. Hackney, and K. W. Axhausen. 2005. "Efficient Map Matching of Large Global Positioning System Data Sets: Tests on Speed-monitoring Experiment in Zurich." *Transportation Research Record: Journal of the Transportation Research Board, No. 1935*, Washington, D.C.: Transportation Research Board of the National Academies, 93–100.

McGowen, P., and M. G. McNally. 2006. "Predicting Activity Types from GPS and GIS Data." Presented at the 86th Annual Meeting of the Transportation Research Board. Washington, D.C.: Transportation Research Board of the National Academies.

Mobile MPO. 2011. *Mobile Origin–Destination Study.* Mobile: South Alabama Regional Planning Commission.

Mohammadian, K., K. Kawamura, K. Sturm, and Z. Pourabdollahi. 2013. *GPS Based Pilot Survey of Freight Movements in the Midwest Region.* Madison, WI: National Center for Freight & Infrastructure Research & Education.

Moiseeva, A., J. Jessurun, and H. Timmermans. 2010. "Semi-Automatic Imputation of Activity-Travel Diaries Using GPS Traces, Prompted Recall and Context Sensitive Learning Algorithms." Presented at the 89th Annual Meeting of the Transportation Research Board. Washington, D.C.: Transportation Research Board of the National Academies.

Muckell, J., J. Hwang, V. Patil, C. Lawson, F. Ping, and S. Ravi. 2011. "SQUISH: An Online Approach for GPS Trajectory Compression." *Proceedings of the 2nd International Conference on Computing for Geospatial Research & Applications.* New York: ACM.

Munizaga, M., F. Devillaine, and M. Amaya. 2012. "Analyzing Travel Behaviour of Public Transport Users Observed through Smartcard Data Use." *Proceedings of the 13th International Conference on Travel Behaviour Research.* Toronto, Ontario: IATBR.

Murakami, E., and D. Wagner. 1999. "Can Using Global Positioning System (GPS) Improve Trip Reporting?" *Transportation Research Part C* 7 (2–3): 149–165.

Murakami, E., and S. Bricka. 2012. *Travel Survey Manual Update - Chapter 26.* Accessed August 31, 2012. http://www.travelsurvey-manual.org/Chapter-26-3.html.

Murakami, E., J. Morris, and C. Arce. 2003. "Using Technology to Improve Transport Survey Quality." *Transport Survey Quality and Information.*

Nemala, V. 2009. "Efficient Clustering Techniques for Managing Large Datasets." *UNLV Master's Thesis.* Las Vegas, NV. http://digital scholarship.unlv.edu/thesesdissertations/72/.

Nijland, E., T. Arentze, A. Borgers, and H. Timmermans. 2011. "Modeling Complex Activity-Travel Scheduling Decisions: Procedure for the Simultaneous Estimation of Activity Generation and Duration Functions." *Transport Reviews* 31 (3): 399–418.

NuStats. 2002. 2000-2001 California Statewide Household Travel Survey: Final Report. Sacramento: California Department of Transportation.

NuStats. 2004. *Kansas City Regional Household Travel Survey Final Report.* Kansas City: Mid-America Regional Council.

NuStats. 2010. "Front Range Travel Counts Final Report." Denver, CO.

Oliveira, M., and J. Casas. 2010. "Improving Data Quality, Accuracy, and Response in On-Board Surveys." *Transportation Research Record: Journal of the Transportation Research Board, No. 2183*, Washington, D.C.: Transportation Research Board of the National Academies, 41–48.

Oliveira, M. G. S., P. Vovsha, J. Wolf, Y. Birotker, D. Givon, and J. Paasche. 2011. "GPS-Assisted Prompted Recall Household Travel Survey to Support Development of Advanced Travel Model in Jerusalem, Israel." Presented at the 90th Annual Meeting of the Transportation Research Board. Washington, D.C.: Transportation Research Board of the National Academies.

Oliveira, M., P. J. Troped, J. Wolf, C. E. Mathews, E. K. Cromley, and S. J. Melly. 2006. "Mode and Activity Identification Using GPS and Accelerometer Data." Presented at the 85th Annual Meeting of the Transportation Research Board, Washington, D.C.: Transportation Research Board of the National Academies.

Open Geospatial Consortium. 2008. *OGC KML.* Project OGC 07-147r2, Open Geospatial Consortium.

Papinski, D., D. M. Scott, and S. T. Dougherty. 2008. "Exploring the Route Choice Decision-Making Process: A Comparison of Pre-planned and Observed Routes Obtained Using Person-based GPS." Presented at the 87th Annual Meeting of the Transportation Research Board. Washington, D.C.: Transportation Research Board of the National Academies.

Pas, E. I., and S. Sundar. 1995. "Intra-personal variability in daily urban travel behavior: Some additional evidence." *Transportation* (22): 135–150.

Pendyala, R. M., K. C. Konduri, and Y. C. Chiu. 2012. "An Integrated Land Use-Transport Model System with Dynamic Time-Dependent Activity-Travel Microsimulation." Presented at the 91st Annual Meeting of the Transportation Research Board, Washington, D.C.: Transportation Research Board of the National Academies.

Pew Research Center. 2012. *Digital Differences.* Washington, D.C.: Pew Research Center.

Pierce, B., J. Casas, and G. Giaimo. 2003. "Estimating Trip Rate Under-Reporting: Preliminary Results from the Ohio Travel Survey." *82nd Annual Meeting of the Transportation Research Board.* Washington.

Pierce, B., R. Zimmer, M. Burns, and C. Johnson. 2011. "Thick or Thin, Maximizing Data While Protecting Privacy of Participants: The Minnesota Solution." *Proceedings of the 18th Annual ITS World Congress.* Orlando, FL.

Piscitelli, A. 2008. "A Double Imputation Method for Data Fusion." *Quaderni di Statistica* 10: 35–52.

PTV NuStats. 2011. *Atlanta Regional Commission - Regional Travel Survey - Final Report.* Project Memorandum, Atlanta, GA: Atlanta Regional Commission.

Puget Sound Regional Council. 2008. *Traffic Choices Study - Summary Report.* Seattle, WA: PSRC.

Pyo, J., D. Shin, and T. Sung. 2001. "Development of a Map Matching Method Using the Multiple Hypothesis Technique." *Proceedings of the 2001 IEEE Intelligent Transportation Systems Conference.* Oakland, CA: IEEE. 23–27.

Qiu, Z., and P. Cheng. 2007. "State of the Art and Practice: Cellular Probe Technology Applied in Advanced Traveler Information Systems." Presented at the 86th Annual Meeting of the Transportation Research Board, Washington, D.C.: Transportation Research Board of the National Academies.

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* 10 (3): 103–115.

Quddus, M. A., W. Y. Ochieng, L. Zhao, and R. B. Noland. 2003. "A General Map Matching Algorithm for Transport Telematics Applications." *GPS Solutions* 7 (3): 157–167.

Quiroga, C. 2004. "Traffic Monitoring Using GPS." In *Handbook of Transport Geography and Spatial Systems*, (D. Hensher, K. Button, K. Haynes and P. Stopher, eds.). Emerald Group Publishing Limited.

R Core Team. 2013. "R: A Language and Environment for Statistical Computing." Accessed May 23, 2013. http://www.R-project.org/.

Rashidi, T. H., and A. Mohammadian. 2011. "Household Travel Attributes Transferability Analysis: Application of Hierarchical Rule Based Approach." *Transportation* 38 (4): 697–714.

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* 33 (5): 727–748.

Ratti, C., A. Sevtsuk, S. Huang, and R. Pailer. 2007. "Mobile Landscapes: Graz in Real Time." In *Location Based Services and Telecartography*, Springer, Berlin, Germany, pp. 433–444.

Reuscher, T. R., R. L. Schmoyer, and P. S. Hu. 2002. "Transferability of Nationwide Personal Transportation Survey Data to Regional and Local Scales." *Transportation Research Record: Journal of the Transportation Research Board, No. 1817*, Washington, D.C.: Transportation Research Board of the National Academies, 25–35.

Saporta, G. 2002. "Data fusion and data grafting." *Computational Statistics & Data Analysis* 38: 465–473.

Saporta, G., and V. Co. 1999. "Fusion de fichiers: une nouvelle méthode basée sur l'analyse homogéne." In *Enquêtes et sondages*, (G. Brossier and A.M. Dussaix, eds.), Paris: Dunod, Paris. 81–93.

Schaller, B. 2005. *TCRP Synthesis 63: On-Board and Intercept Transit Survey Techniques.* Washington, D.C.: Transportation Research Board of the National Academies.

Schenk, A., B. Witbrodt, C. Hoarty, R. Carlson, E. Goulding, J. Potter, and S. Bonasera. 2011. "Cellular Telephones Measure Activity and Lifespace in Community-Dwelling Adults: Proof of Principle." *The American Geriatrics Society*, 345–352.

Schönfelder, S., and U. Samaga. 2003. "Where do you want to go today? More observations on daily mobility." *3rd Swiss Transport Research Conference.* Monte Verita/Ascona, Italy.

Schüssler, N., and K. 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 Reseach Board of the National Academies.

Schüssler, N., and K. W. Axhausen. 2009a. *Map Matching of GPS Points on High Resolution Navigation Networks Using Multiple Hypothesis Techniques.* Zurich: IVT-ETH. Accessed October 25, 2012. https://edit.ethz.ch/ivt/vpl/publications/reports/ab568.pdf.

Schüssler, N., and K. W. Axhausen. 2009b. "Processing GPS Raw Data without Additional Information." *Transportation Research Record: Journal of the Transportation Research Board, No. 2105*, Washington, D.C.: Transportation Research Board of the National Academies, 28–36.

Shen, L., and P. Stopher. 2012. "An Improved Process for Trip Purpose Imputation from GPS Travel Data." *Travel Behaviour Research: Current Foundations, Future Prospects.* Toronto: IATBR.

Smith, A. 2012. *Pew Internet & American Life Project - 46% of American Adults Are Smartphone Owners.* Pew Research Center. Accessed August 22, 2012. http://pewinternet.org/~/media//Files/Reports/2012/Smartphone%20ownership%202012.pdf.

Steer Davies Gleave and GeoStats. 2003. *The Use of GPS to Improve Travel Data Study Report.* London: London Department for Transport.

Stopher, P., R. Alsnih, C. Wilmot, C. Stecher, J. Pratt, J. Zmud, W. Mix, et al. 2008. *NCHRP Report 571: Standardized Procedures for Travel Surveys.* Washington, D.C.: Transportation Research Board of the National Academies.

Stopher, P., P. Bullock, and F. Horst. 2002. *Exploring the Use of Passive GPS Devices to Measure Travel.* Sydney: Institute of Transport and Logistics Studies, University of Sydney.

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

Stopher, P., and A. Collins. 2005. "Conducting a GPS Prompted Recall Survey over the Internet." Presented at the 84th Annual Meeting of the Transportation Research Board. Washington, D.C.: Transportation Research Board of the National Academies.

Stopher P., C. Fitzgerald, and T. Biddle. 2006. *Pilot Testing a GPS Panel for Evaluating TravelSmart.* Sydney: Institute of Transport and Logistics Studies, University of Sydney.

Stopher, P., and S. Greaves. 2007. "Household Travel Surveys: Where Are We Going?" *Transportation Research A* 41: 367–381.

Stopher, P., S. Greaves, and P. Bullock. 2003. "Simulating Household Travel Survey Data: Application to Two Urban Areas." Presented at the 82nd Annual Meeting of the Transportation Research Board. Washington, D.C.: Transportation Research Board of the National Academies.

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

Stopher, P., K. Kockelman, S. Greaves, and E. Clifford. 2008. "Reducing Burden and Sample Sizes in Multi-Day Household Travel Surveys." *Transportation Research Record: Journal of the Transportation Research Board, No. 2064*, Washington, D.C.: Transportation Research Board of the National Academies.

Stopher, P., N. Swann, and C. Fitzgerald. 2007. "Using an Odometer and a GPS Panel to Evaluate Travel Behavior Changes." Presented at the 11th National Transportation Planning Applications Conference. Daytona Beach, FL: Transportation Research Board of the National Academies.

Stopher, P., L. Wargelin, J. Minser, K. Tierney, M. Rhindress, and S. O'Connor. 2012. *GPS-Based Household Interview Survey for the Cincinnati, Ohio Region.* Cincinnati, Ohio: Abt SRBI, Incorporated.

Thompson, R. G., and H. Kayak. 2011. "Estimating Personal Physical Activity from Transport." *Proceedings of the 34th Australian Transport Research Forum.* Adelaide, Australia.

Tierney, K., S. Decker, K. Proussaloglou, T. Rossi, E. Ruiter, and N. McGuckin. 1996. *Travel Survey Manual How to do a Survey.* FHWA, Washington, D.C.: U.S. Department of Transportation.

Troped, P., M. Oliveira, C. Matthews, E. Cromley, and B. C. S. Melly. 2008. "Prediction of Activity Mode with Global Positioning System and Accelerometer Data." *Medicine & Science in Sports & Exercise* 40 (5): 972–978.

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, 38–45.

U.S. Bureau of Transportation Statistics. 2005. "BTS Statistical Standards Manual." October. Accessed July 7, 2012. http://www.bts.gov/programs/statistical_policy_and_research/bts_statistical_standards_manual/pdf/entire.pdf.

U.S. Department of Transportation. 2012. *The Transportation Planning Capacity Building Program Focus Areas.* August 22. Accessed August 22, 2012. http://www.planning.dot.gov/documents/primer/intro_primer.asp.

U.S. Office of Management and Budget. 2006. http://www.whitehouse.gov/omb/inforeg_statpolicy. *www.whitehouse.gov/omb.* Accessed July 26, 2012. http://www.whitehouse.gov/sites/default/files/omb/assets/omb/inforeg/statpolicy/standards_stat_surveys.pdf.

Velaga, N. R., M. A. Quddus, and A. Bristow. 2011. "Improving the Performance of a Topological Map-Matching Algorithm through Error Detection and Correction." Presented at the 90th Annual Meeting of the Transportation Research Board. Washington, D.C.: Transportation Research Board of the National Academies.

Vincenty, T. 1975. “Direct and Inverse Solutions of Geodesics on the Ellipsoid with Application of Nested Equations.” *Survey Review.* XXIII (misprinted as XXII) (176): 88–93. http://www.ngs.noaa.gov/PUBS_LIB/inverse.pdf.

Voigt, A. 2011. “Collecting External Data Using Bluetooth Technology.” Presented at the 90th Annual Meeting of the Transportation Research Board. Washington, D.C.: Transportation Research Board of the National Academies.

White, C. E., D. Bernstein, and A. L. Kornhauser. 2000. “Some Map Matching Algorithms for Personal Navigation Assistants.” *Transportation Research Part C* 8: 91–108.

Wilhelm, J., J. Wolf, and M. Oliveira. 2012. “Application of GPS-based Prompted Recall Methods in Two Household Travel Surveys.” Presented at the 91st Annual Meeting of the Transportation Research Board. Washington, D.C.: Transportation Research Board of the National Academies.

Wilmot, C. G. 1995. “Evidence of Transferability of Trip Generation Models.” *Journal of Transportation Engineering* 9: 405–410.

Wolf, J. 2000. “Using GPS Data Loggers to Replace Travel Diaries in the Collection of Travel Data.” Atlanta: Georgia Institute of Technology, School of Civil and Environmental Engineering.

Wolf, J., S. Bricka, T. Ashby, and C. Gorugantua. 2004. “Advances in the Application of GPS to Household Travel Surveys.” Presented at the National Household Travel Survey Conference. Washington, D.C.: Transportation Research Board of the National Academies.

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., R. Guensler, L. Frank, and J. Ogle. 2000. “The Use of Electronic Travel Diaries and Vehicle Instrumentation Packages in the Year 2000 - Atlanta Regional Household Travel Survey: Test Results, Package Configurations, and Deployment Plans.” *Proceedings of the 9th International Association of Travel Behavior Research Conference.* Queensland, Australia.

Wolf, J., and M. Lee. 2009. “Combining Travel Surveys and Physical Activity Studies.” Presented at the 14th National Transportation Planning Applications Conference. Washington, D.C.: Transportation Research Board of the National Academies.

Wolf, J., M. Loechl, M. Thompson, and C. Arce. 2003. “Trip Rate Analysis in GPS-Enhanced Personal Travel Surveys.” In *Transport Survey Quality and Innovation*, (P. Stopher and P. M. Jones, eds.), 483–498. Oxford, UK: Pergamon.

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, 189–198.

Wolf, J., S. Schönfelder, U. Samaga, M. Oliveira, and K. M. Axhausen. 2004. “80 Weeks of GPS Traces: Approaches to Enriching Trip Information.” *Transportation Research Record: Journal of the Transportation Research Board, No. 1870*, Washington, D.C.: Transportation Research Board of the National Academies, 46–54.

Wolf, J., and S. Trost. 2009. “Objective Measurement of Physical Activity and Location Using Accelerometers and Geographic Positioning Systems (GPS).” *Proceedings of the 7th International Conference on Diet and Activity Methods (ICDAM 7).* Washington, D.C.

Zhang, Y., and A. Mohammadian. 2008. “Bayesian Updating of Transferred Household Travel Data.” *Transportation Research Record: Journal of the Transportation Research Board, No. 2049*, Washington, D.C.: Transportation Research Board of the National Academies, 111–118.

Zhao, H. 2000. “Comparison of Two Alternatives for Trip Generation.” Presented at the 79th Annual Meeting of the Transportation Research Board. Washington, D.C.: Transportation Research Board of the National Academies.

Zmud, J., and J. Wolf. 2003. “Identifying the Correlates of Trip Misreporting - Results from the California Statewide Household Travel Survey GPS Study.” *Proceedings of the 10th International Conference on Travel Behaviour Research.* Lucerne, Switzerland.

Abbreviations

ATIS	Advanced traveler information system
CAMPO	Capital Area Metropolitan Planning Agency
CAPI	Computer-assisted personal interview
CASI	Computer-assisted self-interview
CATI	Computer-assisted telephone interview
CMP	Congestion management program
DRCOG	Denver Regional Council of Governments
ETC	Electronic toll collection
GIS	Geographical information system
GPS	Global Positioning System
GSM	Groupe Spécial Mobile
HTS	Household travel survey
ISTDM	Indiana State Transportation Demand Model
LBS	Location-based services
LOS	Level of service
LSS	Location sharing services
MAC	Media access control
MCOG	Mendocino Council of Governments
MPO	Metropolitan planning organization
OD	Origin–destination
PND	Personal navigation device
POI	Point of interest
PR	Prompted recall
QS	Quantified self
RFID	Radio frequency identification
RP	Revealed preference
SP	Stated preference
TAZ	Traffic analysis zone
TDM	Travel demand model
TOS	Terms of service
VMS	Variable message sign

APPENDIX A

Example Data Sets Delivered in Recent Household Travel Surveys

The tables in this appendix list all variables included in the data deliverables of three recently completed household travel surveys: the 2012 California Household Travel Survey, the 2011 (Atlanta) Regional Travel Survey, and the 2012 Northeast Ohio (Cleveland) Regional Travel Survey. The variables are listed based on the four typical tables that make up a household travel survey data set; these are household, person, vehicle, and place/trip tables.

Permission to list these variables was given by each of the sponsoring agencies (the California Department of Transportation, the Atlanta Regional Commission, and the Northeast Ohio Areawide Coordinating Agency).

Household Variables

Description	California Statewide	Atlanta	Cleveland
Sample Number	X	X	X
Recruit Mode	X	X	X
Retrieval Mode	X	X	X
MPO	X	X	X
Sample Type	X	X	X
GPS Sample Type	X	X	
GPS Type	X	X	X
Number of Household Workers	X		X
Number of Household Students	X	X	X
Number of Household Driver License Holders	X	X	X
Number of Household Trips on Travel Day	X	X	X
Flag for Household Who Reported LD Trip	X		
Number of Household Children		X	
Household Life Cycle [COMPUTED]		X	X
Home Travel Analysis Zone		X	
Partial Completed Households. 4+ HHMEM completed survey		X	
Household Vehicles vs. Household Workers			X
Household Vehicles vs. Household Size			X
Household Size vs. Household Workers			X
K thru X2 Student Present in Household			X
Post-Secondary Student Present in Household			X
Diary Collected for Household Non-GPS Members			X
One or more Household members Completed PR survey			X
Incentive Flag	X	X	X
Interview Language	X	X	
Residential County	X	X	
Assigned Travel Date	X	X	X
Assigned Travel Day	X	X	X
Transit Use At Least Once Per Week	X		
New vehicle	X		
Buyer X-8	X		
Residence Type	X	X	
Description of home-OTHER	X	X	
Home Ownership	X	X	
Home ownership-other	X	X	
TENURE	X		
Previous address	X		
Previous suit	X		
Previous address city	X		
Previous address state	X		
Previous address zip code	X		
Number of land line phones	X	X	X
Household income	X	X	X
Household Size	X	X	X
Non-related Household Flag	X	X	
Number of Household Vehicle	X	X	X
Number of Household Bicycle	X		
Home Address	X	X	X
Home suit#	X	X	X
Home city	X	X	X
Home state	X	X	X

(continued on next page)

Description	California Statewide	Atlanta	Cleveland
Home zip	X	X	X
Home x-coordinate	X	X	X
Home y-coordinate	X	X	X
Reason of no possession of a vehicle	X		
Number of operational household vehicle	X	X	
Number of Vehicles with Power Outlet	X	X	
Number of newer vehicles with power outlet	X		
Willingness to Participate in Future Study	X	X	
Hispanic household flag	X		
Household Complete Flag	X		X
Hispanic Origin		X	
Hispanic household flag		X	
Shop & purchase items online via Internet			X
Places Regularly Visit Weekly [Multiple Response]			X
Overnight Guest During Travel Day			X
Number of Guests			X
Total	**50**	**38**	**32**

Person Variables

Description	California Statewide	Atlanta	Cleveland
Sample Number	X	X	X
Person Number	X	X	X
Relationship to Head of House	X	X	X
Gender	X		X
Age	X	X	X
HISPANIC OR LATINO	X		X
ETHNICITY OR RACE	X	X	X
ETHNICITY OR RACE, other	X	X	X
NATIVITY	X		
COUNTRY OF BIRTH	X		
Valid license	X	X	X
Vehicle driven by Respondent	X		
Transit pass	X		
Type of transit pass	X		
Type of transit pass, other	X		
Type of Clipper Card	X		
Type of Compass Card	X		
Type of TAP/EZ Pass Card	X		
Toll Pass	X		
Car Sharing	X		
Employed?	X	X	X
Employment status	X	X	
Employment status	X	X	
Work Location	X	X	X
How many jobs	X	X	X
Primary Work name	X	X	X
Primary Work Address (including suit#)	X	X	X
Primary Work City	X	X	X
Primary Work State	X	X	X
Primary Work Zip	X	X	X
Primary Work Cross StreetX	X	X	X
Primary Work Cross Street2	X	X	X

Description	California Statewide	Atlanta	Cleveland
Primary Work X-coordinates	X	X	X
Primary Work Y-coordinates	X	X	X
Days at Primary work	X	X	X
Work Days	X		
Work Days	X		
Work Days	X		
Work Days	X		
Work Days	X		
Work Days	X		
Hours Worked at Primary Job	X	X	X
Flexible work schedule	X		
Flexible work programs offered	X	X	
Work mode	X	X	X
Industry	X		X
Industry, Other	X	X	X
Occupation	X	X	X
Occupation, Other	X	X	X
Work Location	X		X
Secondary Work name	X		X
Secondary Work Address(including suit#)	X		X
Secondary Work City	X		X
Secondary Work State	X		X
Secondary Work Zip	X		X
Secondary Work Cross StreetX	X		X
Secondary Work Cross Street2	X		X
Days at secondary work	X		
Disability Status	X	X	X
Disability Type	X	X	X
Other, Disability Type	X	X	X
Disabled license plate	X		
Disabled transit registration	X		
Transit trips used in past week	X		
Transit Subsidy	X		
Subsidized amount	X		
fair unit	X		
Other, fair unit	X		
Walk in the last week	X		
Bicycle in the last week	X		
Student	X	X	X
School grade level attends	X	X	X
School grade level attends. Other	X	X	X
Home School	X	X	X
Online School	X	X	X
School name	X	X	X
School address	X	X	X
School City	X	X	X
School State	X	X	X
School Zip	X	X	X
School Cross Streets	X	X	X
School Cross Streets	X	X	X
School X-coordinates	X	X	X
School Y-coordinates	X	X	X
Pre-school location	X		
Pre-school location, other	X		

(continued on next page)

Description	California Statewide	Atlanta	Cleveland
School mode	X	X	X
Level of education completed	X	X	X
Level of education completed, other	X	X	X
ARE YOU INTERVIEWING THIS PERSON?	X	X	X
WHICH PERSON SERVED AS PROXY?	X	X	X
Did [NAME] complete the travel log?	X	X	X
Have diary to refer to	X	X	X
Person Trips	X	X	X
Did you use a toll	X		X
Toll Road Used	X		
Toll Bridge Used	X		
HOV lane used	X		
Why no trips on travel day	X		X
Why no trips on travel day, other	X		X
Person Retrieval Incomplete Flag	X		
Gender		X	
Cell Phone		X	
Volunteer Worker?		X	X
Yes if employed or Volunteer		X	X
Hours Worked at Secondary Job		X	X
Hours Worked at Tertiary Job		X	
Telecommuting Offered at Workplace		X	
Telecommute Hours		X	X
Work Start Time		X	
Work End Time		X	
Work Schedule		X	
Other, Compressed Work Week		X	
Employer		X	
Other, Employer		X	
Employer Provided Parking		X	
Employer Subsidized Parking		X	
Use of Employer Subsidized Parking		X	
Employer Subsidized Transit		X	
Work Travel Analysis Zone		X	
Other, Mode of Transport to Work		X	X
School Travel Analysis Zone		X	
Other, Mode of Transport to School		X	X
Rides Transit		X	
Breeze Card		X	
Value Added to Breeze Card		X	
Other, Value Added to Breeze Card		X	
GRTA Xpress Bus Pass		X	
Type of GRTA Fare		X	
Cobb or Gwinett County Transit Pass		X	
Type of Cobb or Gwinett County Transit Pass		X	
Discounted Fare		X	
Other, Discounted Fare Program		X	
Frequency of Bike to Work		X	X
Reason for No Trips		X	
Other, Reason for No Trips		X	
Person Belongs To Partial Complete		X	
County FIPS		X	
Refused AGE - Age range asked	X	X	X

Description	California Statewide	Atlanta	Cleveland
Over or Under X6	X	X	X
Person-level GPS qualification			X
Type of Pass Owned [Multiple Response]			X
Transit Agency pass is from			X
Type of Transit Pass Owned			X
Pay Full price, Discounted, Student or Employer Subsidy for Transit Pass			X
Amount of Employer Subsidy for Transit Pass			X
Own or Access to a Bike			X
Employment Status			X
Which Days are Worked from Home			X
Location Number for Primary Work			X
Location Number for Secondary Work			X
Primary Work Location Parking			X
Primary Work Location Parking - Other			X
Parking Location Distance from Work			X
How often Walk to Work			X
Location Number for Primary School (K-X2)			X
School Name for Secondary School (>X2th)			X
Location Number for Secondary School			X
How often Bike to School			X
How often Walk to School			X
Total	**104**	**92**	**93**

Vehicle Variables

Description	California Statewide	Atlanta	Cleveland
Sample Number	X	X	X
Vehicle number	X	X	X
Year of vehicle	X	X	X
Vehicle make	X	X	X
Other, Vehicle make	X	X	X
Vehicle model	X	X	X
Series	X		
Other, Series	X		
Body type	X	X	
Other, Body type	X		
Vehicle Transmission	X		
Power train	X		
Power train, other	X		
Cylinders	X		
Cylinders, other	X		
Electrical outlet	X		
Outlet volts	X		
Vehicle Type	X		
Fuel type	X	X	
Other, Fuel type	X	X	
Working power outlet	X	X	
Vehicle acquired	X		
Vehicle ownership	X	X	
Other, Vehicle ownership	X	X	
Vehicle insurance	X		
Vehicle devices	X		

(continued on next page)

Description	California Statewide	Atlanta	Cleveland
Vehicle used on travel day	X	X	
Reason why not	X		
Other, Reason why not	X		
Primary Driver			X
EZPass Tag		X	
Reason Not Used		X	
Total	**29**	**15**	**7**

Location Variables

Description	California Statewide	Atlanta	Cleveland
Sample number	X	X	X
Person number	X	X	X
Place number	X	X	X
Location ID from the Location Table			X
Location Name			X
Place Type	X	X	X
Mode of Trip - Other		X	X
Trip Purpose - Other		X	X
Use a Transfer on Trip			X
Parking Location		X	X
Pay a Toll			X
Amount Paid for Toll			X
When Trip was Scheduled			X
Flexibility of trip timing			X
Trip Speed			X
Total People traveling on trip	X	X	X
Number of household members on trip	X	X	X
Person number on trip	X	X	X
Number of non-household members on trip	X	X	X
Mode of trip	X	X	X
Household vehicle number used on trip	X	X	
Get out of vehicle	X	X	X
Parking location type	X	X	X
Parking location type, other	X	X	X
Parking location address	X		
Time(Mins) walking from Park to destination	X		
Pay to park	X	X	X
Parking amount	X	X	X
Parking unit	X	X	X
How did you pay for parking	X		
How did you pay for parking, other	X		
Parking cost not reimbursed by employer	X		
Transit system	X		
Transit system, other	X		
Transit Route	X	X	
Number of activities	X		
Arrival time - hour	X	X	X
Arrival time - minute	X	X	X
Departure time - hour	X	X	X
Departure time - minute	X	X	X
Travel time to place in minutes.	X		X
Activity duration at place in minutes.	X	X	X

Description	California Statewide	Atlanta	Cleveland
Travel distance (air distance)	X	X	X
Place Name	X	X	
Address	X		
City	X		X
State	X		X
Zip	X		X
X-coordinate	X		X
Y-coordinate	X		X
Primary Trip Purpose		X	X
Secondary Trip Purpose		X	
Other, Trip Purpose		X	
Used HOV Lane		X	
Used TOLL Lane		X	
Transit Service		X	
Other, Transit Service		X	
Transit Fare Type		X	
Transit Fare Cost		X	X
Trip Duration in Minutes		X	
Place Travel Analysis Zone		X	
Trip Number		X	
Number of Person Trips		X	
Transit Access Mode		X	
Transit Access Mode - Other		X	
Transit Egress Mode		X	
Transit Egress Mode - Other		X	
Origin Place Name		X	
Destination Place Name		X	
Origin Travel Analysis Zone		X	
Destination Travel Analysis Zone		X	
Origin Longitude		X	
Origin Latitude		X	
Destination Longitude		X	
Destination Latitude		X	
Departure Time		X	
Destination Arrival Time		X	
Street Number and Street Name of Address			X
Run Street & Cross Street			X
Traffic Analysis Zone			X
Activity number	X		
Household member on this activity?	X		
Activity Number of HH/family Members	X		
Activity Number of other relatives	X		
Activity - Number of people from your Work	X		
Activity - Number of people from your School	X		
Activity - Number of people from same rel/social org.	X		
Activity - Number of Friends	X		
Activity - Number of Other relations	X		
Activity - Purpose	X		
Activity - Purpose, Other	X		
Activity - Start Time	X		
Activity - End Time	X		
Activity - Place number	X		
Total	**53**	**54**	**43**

Long-Distance Travel Variables

Description	California Statewide	Atlanta	Cleveland
Complete Long-distance Log	X	N/A	N/A
HH member who completed LD log	X	N/A	N/A
Last eight weeks	X	N/A	N/A
SAMPLE NUMBER for Long-distance trip	X	N/A	N/A
Long-distance trip number	X	N/A	N/A
Long-distance Date	X	N/A	N/A
Long-distance Origin	X	N/A	N/A
Long-distance Origin Place Name	X	N/A	N/A
Long-distance Origin Address	X	N/A	N/A
Long-distance Origin City	X	N/A	N/A
Long-distance Origin State	X	N/A	N/A
Long-distance Origin Zip code	X	N/A	N/A
Long-distance Origin Country	X	N/A	N/A
Long-distance Origin X-coordinate	X	N/A	N/A
Long-distance Origin Y-coordinate	X	N/A	N/A
Long-distance Destination Place Name	X	N/A	N/A
Long-distance Destination Address	X	N/A	N/A
Long-distance Destination City	X	N/A	N/A
Long-distance Destination State	X	N/A	N/A
Long-distance Destination Zip code	X	N/A	N/A
Long-distance Destination Country	X	N/A	N/A
Long-distance Destination X-coordinate	X	N/A	N/A
Long-distance Destination Y-coordinate	X	N/A	N/A
Long-distance Trip Purpose	X	N/A	N/A
Long-distance Trip Purpose, other	X	N/A	N/A
Long-distance - People on trip	X	N/A	N/A
Long-distance - Household members on trip	X	N/A	N/A
Person who made Long-distance tripX	X	N/A	N/A
Person who made Long-distance trip2	X	N/A	N/A
Person who made Long-distance trip3	X	N/A	N/A
Person who made Long-distance trip4	X	N/A	N/A
Person who made Long-distance trip5	X	N/A	N/A
Person who made Long-distance trip6	X	N/A	N/A
Person who made Long-distance trip7	X	N/A	N/A
Person who made Long-distance trip8	X	N/A	N/A
Long-distance Mode 1	X	N/A	N/A
Long-distance Mode 2	X	N/A	N/A
Long-distance Mode 3	X	N/A	N/A
Long-distance Mode 4	X	N/A	N/A
Latest Long-distance Trip Flag	X	N/A	N/A
Long-distance Time start	X	N/A	N/A
Long-distance Departure Place Name	X	N/A	N/A
Long-distance Departure Mode 1	X	N/A	N/A
Long-distance Departure Mode 2	X	N/A	N/A
Long-distance Departure Mode 3	X	N/A	N/A
Long-distance Departure Mode 4	X	N/A	N/A
Long-distance Arrival Place Name	X	N/A	N/A
Long-distance Arrival Mode 1	X	N/A	N/A
Long-distance Arrival Mode 2	X	N/A	N/A
Long-distance Arrival Mode 3	X	N/A	N/A
Long-distance Arrival Mode 4	X	N/A	N/A
Total	**51**	**N/A**	**N/A**

APPENDIX B

Industry Expert Questionnaires

Travel Survey Practitioners and Travel Survey Researchers

Please answer the following questions within the context of any travel behavior surveys in which you, your agency, or your organization may have been involved (including household travel surveys, transit studies, visitor surveys, and establishment surveys).

1. Please describe the use of technologies in the collection of travel behavior information at your organization or firm. Specifically, please include details of how GPS data are collected, processed, and leveraged. For the purpose of this project, it would be helpful to understand what GPS technologies or methods were used in the past as well as are used in the present.
2. What plans do you or your organization have for near-term or longer-term collection or use of GPS data (or other location-based data) for better understanding travel behavior? Please provide as much detail as possible regarding future technology needs or plans.
3. If you have used GPS devices and/or smart phones with GPS in your travel behavior studies, please discuss how the use of these technologies may have affected recruitment rates, retrieval rates, and sample representativeness.
4. Please explain how you or your firm processes GPS data or other location-based data to answer your research questions or the research questions of your clients.
5. Please describe the GPS-based or other location-based travel behavior details that are included in your data sets or in deliverables to your clients.
6. Please explain how confidentiality is maintained for survey participants who provide GPS data in your studies. Describe any issues or concerns regarding participant privacy that you have encountered in a GPS-specific, cell phone, or other location-based technology study.
7. Overall, how does the coverage and accuracy of GPS data and/or cell phone data compare with previously used methods? (Please elaborate on previous methods as well.) If you are collecting or using mobile phone data, please elaborate on the source of the location information (i.e., from the GPS chip, from cell phone tower locations or handoffs, or from participant self-report) as well as the coverage and accuracy levels obtained.
8. Given your experience to date, what do you consider to be the advantages of using GPS or other location-based technologies for the collection of travel behavior?
9. Given your experience with using GPS or other location-based technologies to measure travel behavior, what do you consider to be the limitations or concerns for these technologies? (Please be specific.)
10. If applicable, please provide cost information that could be used by agencies that are considering adding GPS or other location-based technologies to their travel behavior surveys—either as an addition or instead of traditional methods. Types of costs could include cost per household, per person, or per travel day. Please explain as much as possible.

Transportation Planners and Travel Demand Modelers

1. Please describe the current role of GPS data within your organization or within your daily work responsibilities (for example, is GPS data leveraged to determine transportation network speeds, for congestion management, or for travel behavior analysis); please provide details.
2. Please describe the specific role of GPS data in your model development process (for example, to calculate trip rate correction factors or to evaluate baseline networks).
3. Please describe any secondary applications of GPS data in which you or your organization acquired GPS data for one purpose but then used for another/additional purpose (such as for congestion management planning, bike/pedestrian planning, or transit planning).

4. Please describe any future plans you or your organization/firm have for using GPS data to support travel demand modeling data needs.
5. Please describe others sources of measured origin–destination data that you have used or plan to use to support model development (such as data provided by INRIX, TomTom, NAVTEQ, AirSage, or other similar sources).
6. Have you purchased any advanced travel behavior data (GPS, Bluetooth, cell-phone–based data, or other origin–destination data) in the last 12 months? If so, can you briefly describe what you purchased, the intended purpose, and any challenges you faced with its integration?
7. Please describe your plans for using short-term modeling tools for evaluating transportation improvements (i.e., dynamic transportation models) and identify what you see as the key data needs for short-term forecasts.
8. If you are involved in long-term transportation forecasts, what do you believe to be the key data needs in this modeling area?
9. If you are using or building an activity-based model, what travel behavior data quality issues are most critical from your perspective?
10. What benefits do you believe can be gained by leveraging GPS data to better understand travel behavior?
11. What disadvantages or limitations do you believe exist with using GPS data for modeling travel behavior?
12. How accurate do you believe GPS data should be for modeling travel behavior? What resolution or frequency of GPS locations do you believe is needed?
13. Do you have a plan/idea for using advanced travel behavior data (such as GPS or Bluetooth/sensor data) for some purpose but lack the technical capabilities for its application?
14. Describe your level of knowledge regarding advanced travel behavior data collection and provide the primary sources of your knowledge and training (such as formal classes, conference presentations, research papers, meetings with vendors, or personal exploration). If you need training or reference materials, please explain what training methods or materials you prefer.
15. Is there anything else you would like to say about the use of GPS data for understanding travel behavior?

Traffic Data Providers

1. Describe the primary and secondary markets for your data products (i.e., real-time traffic, transportation planning, etc.)
2. List and describe the different data-generating technologies that are used to build your company's data products (for example, in-dash navigation devices, personal navigation devices, non-GPS cell phones, GPS cell phones, truck GPS/AVL). Is one technology primary; if yes, please identify.
3. If you are using personal mobile devices (such as cell phones or smart phones) as a data source, please describe the penetration rate of this data source.
4. Describe your company's plans for future vehicle or personal technology and data product development.
5. What is your current geographic coverage? Do you have plans to expand? If so, please elaborate on the planned geography as well as implementation timeline.
6. Describe any of your data products' limitations that are relevant to road functional classification or other transportation system characteristics. Are expected error ranges provided by road segment or TMC [Traffic Message Channel] location code?
7. Describe how your products are packaged to specifically serve the needs of transportation planners for travel demand forecast modeling or for congestion management programs.
8. For each of your specific data products (copy and repeat this question and categories as needed), please describe the following:
   - Data product name and description
   - Raw data frequency and accuracy
   - Data cleaning process
   - Level of aggregation or disaggregation
   - Cost structure
9. Describe any potential demographic bias that exists in your data sources.
10. Describe any data usage clauses of agreements that come with new vehicle/device purchases that enable or authorize your firm to use personal mobility data.
11. Describe how data source privacy and location/time of day details are protected.

APPENDIX C

Questionnaire Responses from Traffic Data Providers

AirSage

1. Describe the primary and secondary markets for your data products (i.e., real-time traffic, transportation planning, etc.).

   AirSage's primary market is government MPO's for transportation planning and origin–destination studies. AirSage is also a leader in nationwide traffic flow analysis and an innovative product called FastCache—the next-generation marketing and LBS strategy.

2. List and describe the different data-generating technologies that are used to build your company's data products (for example, in-dash navigation devices, personal navigation devices, non-GPS cell phones, GPS cell phones, truck GPS/AVL). Is one technology primary; if yes, please identify.

   AirSage has the exclusive ability to mine signaling data from any type of mobile device from two major carriers (roughly 70% of the U.S. population), and from those location points we can infer traffic speeds, phone/user origins and destinations, and travel trends.

3. If you are using personal mobile devices (such as cell phones or smart phones) as a data source, please describe the penetration rate of this data source.

   Because we collect data points as a mobile device communicates with a tower, any type of mobile device is seen. We have 100% penetration rate of mobile devices.

4. Describe your company's plans for future vehicle or personal technology and data product development.

   AirSage has spent years getting the technology in place and continues to fine-tune the accuracy of our location data. We built the geolocation systems used by major wireless carriers in North America. We are now commercializing the insights and information that can be realized, all while protecting the privacy rights of mobile consumers. AirSage is the leading mobile Big Data provider and can provide clients with solutions in anytime population studies, origin–destination studies, Fast-Cache (geofencing) reporting and mobility attributes. AirSage has partnered with a firm that specializes in demographic and geographic analysis and through this partnership we will be able to garner much more insight into the users' lifestyle.

5. What is your current geographic coverage? Do you have plans to expand? If so, please elaborate on the planned geography as well as implementation timeline.

   AirSage has a geographic reach of the entire U.S., including Hawaii. Wherever a cell phone with one of our two carriers has service, we see it.

6. Describe any of your data products' limitations that are relevant to road functional classification or other transportation system characteristics. Are expected error ranges provided by road segment or TMC location code?

   For real-time and historical traffic data, AirSage uses TMC codes from FC Class 1-4. Data points are triangulated based on cell signals and assigned a TMC code. Data points are (immediately) tagged as Transient Points (moving) or Activity Points (stationary).

7. Describe how your products are packaged to specifically serve the needs of transportation planners for travel demand forecast modeling or for congestion management programs.

   Projects are defined using our standard PO, with the following options:

   OD Matrix
   A Trip Distribution table shows the number of people that made trips between a specific Zone pairs during the given time period.

Productions
A Productions table shows the number of people that departed from a specific Zone during a specific hour of the day for each hour of each day of the time frame for each of the Zones in the Study Area.

Attractions
An Attractions table shows the number of people that arrived in a specific Zone during a specific hour of the day for each hour of each day of the time frame for each of the Zones in the Study Area.

Traffic Counts
A Traffic Count table shows the number of people that traveled a road segment during the given time period(s) for each of the road segments in the Study Area.

Home-Work Matrix
A Home-Work Matrix shows the number of people who live and work in a specific pair of zones.

Sub-Area Analysis
A Corridor Study shows the number of people who traveled on a specific road segment (corridor) that started and ended their trips in a specific origin, destination, and entered and exited the corridor at specific entry and exit points.

Select Link Analysis
A Select Link Analysis shows the number of people who traveled a specific road segment (link) while making a trip between specific origin and destination pairs.

External Station Analysis
An External Station Analysis shows the number of people that traveled through a given metropolitan area, grouped by given count station pairs.

Trip Chains
Trip Chains show the number of trips within a given geographic area that contain a specific sequence of stops and/or Activity Points for a given time period.

Trip Frequency
A Trip Frequency table shows the number people who took a specific number or number range of trips from a given origin to a given destination over a specified time period.

Trip Duration
A Trip Duration table shows the number of people who traveled between the specific Origin and Destination pairs during the specified time frame and stayed for a certain number of days.

Anytime Population
An Anytime Population table shows the number of people who were present or passed through a given geographic area during a given time period.

Traffic Flow
A Traffic Flow table shows average travel times for specific road segments during a given time period. This data includes, by individual road segment, traffic flow by daypart at 5-minute intervals.

8. For each of your specific data products (copy and repeat this question and categories as needed), please describe the following:

a) Data product name and description
b) Raw data frequency and accuracy
c) Data cleaning process
d) Level of aggregation or disaggregation
e) Cost structure

Origin–Destination
This AirSage suite of analysis and reporting tools lets you analyze billions of location points derived each day from our exclusive network of over 100 million mobile consumer devices. Study traffic zone-to-zone, by trip or travel patterns, even by demographic characteristics. You'll gain powerful insights that fuel the creation of new businesses, additional locations, or the development of a new product or service.

AirSage's suite of Origin–Destination reports is based on billions of location points derived every day from our exclusive access to over 100 million mobile devices. Data is collected in real time and at lat/long accuracy level but aggregated up to 1,000-m grid cells. Smaller grid cells can be utilized.

AirSage's patent-protected Wireless Signal Extraction (WiSE™) technology aggregates and analyzes signaling data from individual handsets throughout a broad cellular network. In essence, each individual handset becomes a mobile location sensor, allowing AirSage to identify how phones move over time. AirSage validates the information and converts it into real-time location data, including traffic speed and time of day. Anywhere our wireless partners provide coverage, AirSage can provide location information in a cost-effective manner, customized for your business.

Patented technology and multiple layers of privacy protection ensure that no proprietary, customer-identifying data is accessed or released from within the carrier's secure environment. We report at the census tract level with at least 10 mobile devices present. All data is balanced with time.

Cost structure varies from project to project and is based on multiple variables such as population, study area, study length, etc.

FastCache

FastCache, as a part of an integrated marketing strategy, can drive greater consumer engagement. FastCache is the quickest, most affordable way to get unlimited location fixes for all your mobile opt-in subscribers any time, all the time. You'll receive continual location updates whenever a device communicates with the network. AirSage FastCache gives you access to multiple device locations with a single request. It provides you with the ability to take location information, store it, analyze it, and act on the trends and patterns you uncover.

Location updates are speedy—with latency of less than a second for single or multiple device locations. Information is time-stamped and provides a "last seen" location for all your opted-in mobile devices.

FastCache works with all mobile devices on any supported network: smartphone or feature phone, with GPS or without. No special software, configuration, or user intervention is required. And, unlike GPS-based location services, FastCache won't drain the batteries on mobile devices. Instant, time-stamped locations wrapped up tight inside a layer of anonymity are sent as often as the client needs them.

Patented technology and multiple layers of privacy protection ensure that no proprietary, customer-identifying data is accessed or released from within the carrier's secure environment.

FastCache is available in an easy-to-understand flat rate. Rates begin at one penny per ping, with a max of $.03/per day and/or $0.45/per month, per subscriber. The volume discount of $0.45 per subscriber/per month allows for unlimited "pinging."

Traffic Flow

AirSage gives you real-time and historical traffic information on over 500,000 miles of highway and arterial roads throughout the U.S. With over 100 million mobile devices continually reporting in, AirSage offers up-to-date information for vehicle movement, traffic speeds, patterns and location.

Clients get traffic data updated every 1–2 minutes, every day, 24 hours a day. AirSage will deliver detailed, minute-by-minute speed and travel time history for the same nationwide coverage. Clients have access to archived traffic information derived from over 100 million mobile devices, 24 × 7. Testers manually collected GPS data points for 3 major cities and compared them to AirSage's results. AirSage scored 91%–93% accuracy on real-time congestion reports.

AirSage's patent-protected Wireless Signal Extraction (WiSE™) technology aggregates, and analyzes signaling data from individual handsets throughout a broad cellular network. In essence, each individual handset becomes a mobile location sensor, allowing AirSage to identify how phones move over time. AirSage validates the information and converts it into real-time location data, including traffic speed and time of day. Anywhere our wireless partners provide coverage, AirSage can provide location information in a cost-effective manner, customized for your business.

Patented technology and multiple layers of privacy protection ensure that no proprietary, customer-identifying data is accessed or released from within the carrier's secure environment. We report at the census tract level with at least 10 mobile devices present. All data is balanced with time.

Cost structure varies from project to project and is based on multiple variables such as population, study area, study length, etc.

9. Describe any potential demographic bias that exists in your data sources.

   According to CTIA US Wireless; "There are now more wireless devices being used in the United States than there are people, and Americans have doubled the amount of Internet data traffic they generate on smart phones, according to the trade group CTIA." Because we can see any and all mobile devices on our two partner networks, there are no demographic biases.

10. Describe any data usage clauses of agreements that come with new vehicle/device purchases that enable or authorize your firm to use personal mobility data.

    All of our data is collected anonymously behind a firewall; therefore it does not affect the user's privacy or mobile device capabilities. Through our FastCache product, a person must consent to opt-in for the phones identity to be revealed.

11. Describe how data source privacy and location/time of day details are protected.

    AirSage spends a great deal of time and effort ensuring the privacy and security of mobile device users. Timestamp information is available and necessary to our studies, but individual location data is always aggregated up to the census block and census tract level. Unless a phone has opted in, we will never follow an individual phone for a study.

    AirSage's privacy module ensures that sensitive personal information is not compromised. It protects consumers from unwanted intrusions, yet enables them to interact with the brands and services they choose.

    Our technology is fully server-based so there's no impact to the carrier, the consumer, or your business—no software to load, no stress on systems, and no impact to cell phone battery life. AirSage is regularly tested by independent security auditors to ensure the data coming in and going out is fully anonymous.

In addition, AirSage is compliant with the Telecommunications Act of 1996; the Wireless Communications and Public Safety Act of 1999; FCC Proposed Rule-making following the CTIA petition to the FCC on Wireless Location Privacy Principles, November 22, 2000; the European Union Location Privacy, Article 9, amended July 12, 2000; and the individual privacy policies of our carrier partners.

INRIX Inc.

1. Describe the primary and secondary markets for your data products (i.e., real-time traffic, transportation planning, etc.).

   INRIX has a full range of services specifically developed for public transportation agencies. Following is an outline of available services, aligned with the three comprehensive strategies typically aimed at fighting congestion.

   Services for Efficient Operations
   - Real-Time and Predictive Traffic Flow
   - Traffic Information Network
   - Regional Incident Integration Platform
   - Dynamic Route Travel Times

   Services for Effective Capacity Planning
   - Historical Traffic Flow Statistics
   - Historical Five-Minute Archives
   - Dynamic Route Travel Time Archive
   - Analytics and Performance Measures

   Services to Optimize Demand
   - White Label Mobile Applications
   - Custom Mobile Application
   - Advanced Routing
   - Consumer Alerting and Planning

2. List and describe the different data-generating technologies that are used to build your company's data products (for example, in-dash navigation devices, personal navigation devices, non-GPS cell phones, GPS cell phones, truck GPS/AVL). Is one technology primary; if yes, please identify.

   INRIX is able to provide extensive traffic information because of the integration of a large variety of sources to calculate the real-time speed on the roadways. INRIX evaluates and uses advanced fusion technologies to intelligently integrate the full range of potential traffic data source types in consideration of creating high-quality traffic services. No single data source type can provide the accuracy, coverage or scalability that is required in the market today.

   GPS-based Probe Vehicles and Devices: While INRIX was not the first company to use GPS vehicle probes for traffic information, INRIX has built the world's largest GPS probe network using real-time data from nearly 100 million probes. These vehicles include cars and commercial vehicles such as taxis, limos, airport shuttles, service delivery vehicles, long-haul trucks, and less than truckload vehicles, plus a rapidly growing number of consumer vehicles. This floating car data is the single biggest input to the INRIX traffic model.

   Mobile Devices: GPS-enabled smartphones are also becoming an important component of the network. As an example, iPhone users in San Francisco report to INRIX via the INRIX Traffic application as well as other GPS-enabled applications that include INRIX traffic information. Mobile consumers using INRIX Traffic apps on the iPhone and Android-based smartphones have been regularly contributing probe data since the app availability in mid-2009.

   Road Sensors: More than 30,000 (and growing) road sensors across the U.S.

   Other Flow Sources: Other traffic flow sources including cellular probe data and toll tags.

3. If you are using personal mobile devices (such as cell phones or smart phones) as a data source, please describe the penetration rate of this data source.

   N/A—This information is proprietary.

4. Describe your company's plans for future vehicle or personal technology and data product development.

   INRIX data services include product offerings developed specifically for the public sector, auto OEMs, consumer mobile applications, Internet and media. Detail of future products is proprietary information.

5. What is your current geographic coverage? Do you have plans to expand? If so, please elaborate on the planned geography as well as implementation timeline.

   From a road coverage standpoint, INRIX provides traffic information on over 850,000 miles across the United States with over 600,000 individual TMC links. INRIX currently provides coverage in the U.S., Canada, and Western Europe. In 2013, the company plans to expand to Brazil, Russia, China, South Africa, and Scandinavia, as well as preparing for operations in Eastern Europe, Central America, Australia, India, and northern Africa.

6. Describe any of your data products' limitations that are relevant to road functional classification or other transportation system characteristics. Are expected error ranges provided by road segment or TMC location code?

   N/A

7. Describe how your products are packaged to specifically serve the needs of transportation planners for travel demand forecast modeling or for congestion management programs.

   INRIX has several products developed specifically for planners and performance managers. Planners and those interested in model calibration are most interested in aggregation of historical data while performance managers may want to see current conditions in addition to the ability to review historical performance. INRIX provides choices in the type of historical aggregation as well as a new suite of analytic tools and visualizations that can display both real-time and historical performance measures.

   INRIX Historical Speed Statistics use comprehensive data from the INRIX Smart Driver Network, including billions of historical data points, to provide true historical average speeds and statistical distribution on individual road segments covering nearly 1 million miles on major freeways, highways, urban and rural arterials and side streets throughout North America. This data is specific to every day of the week, every hour or quarter hour of the day, and is reported at the Traffic Message Channel (TMC) link level or at the smallest road segment for Tele Atlas and NAVTEQ map databases.

   Using INRIX's proprietary technology, the data is analyzed and normalized to account for the impact of major events, seasonal traffic patterns, typical weather conditions and other variables that can impact traffic flow—ensuring the highest degree of accuracy. The data is updated regularly, incorporating both changes to map databases as well as additional historical data from the INRIX Smart Driver Network.

   INRIX statistical parameters provided for each time bin include the average speed, standard deviation, the 5th, 10th, 15th, 20th, 25th, 30th, 40th, 50th, 60th, 70th, 75th, 80th, 85th, 90th, and 95th percentiles of speed values in mph, and the 30, 40, 50, 60, 70, 80, and 90 mph "failure rates," or the percentage of data points that fall below the specific speed threshold for the given segment/time.

   In addition to typical annual files, monthly and quarterly files are also available. Monthly and quarterly files are available starting January, 2009. Future monthly files for 2012 and beyond can be made available by the 15th of the following month.

   INRIX Historical Archive is a running archive of real-time speeds provided by INRIX for all TMCs in service at that time. Data provided on each segment for each time slice includes the Speed, Travel Time, and the Score. The archive data is available in 5 minute increments beginning July 1, 2008, through December 31, 2010. Starting January 1, 2011, the running archive is available in one minute increments.

   INRIX data, both real-time and historic, is now available as a subscription service bundled with state-of-the-practice analytics and visualization tools. INRIX has partnered with the University of Maryland to expand the Vehicle Probe Project Analytics Suite currently in use by the I-95 Corridor Coalition to provide national coverage. The Suite provides a real-time dashboard as well as instant access to historical data and the ability to compute common mobility performance measures on demand along with visualization tools. An overview of the VPP Analytics Suite is available in the form of a short video tutorial at: http://vpp.ritis.org/suite/screencast/.

   The analytics suite includes:
   - A System Dashboard indicating current congestion levels and bottlenecks
   - Raw data query tool for instant access to archived data based on user specified locations and date ranges
   - Historical Analytic Tools to instantly calculate common performance measures for user defined corridors and date ranges, including:
     - Average Speed
     - Travel Time Index
     - Travel Time
     - Buffer Index
     - Buffer Time
     - Planning Time Index
     - Planning Time
   - Visualization Tools for defined performance measures, including: charts, contour plots, and tabular summaries
   - Bottleneck Ranking Tool to identify system bottlenecks for user defined date ranges

8. For each of your specific data products (copy and repeat this question and categories as needed), please describe the following:

   a) Data product name and description
   b) Raw data frequency and accuracy
   c) Data cleaning process
   d) Level of aggregation or disaggregation
   e) Cost structure

Following is an overview of specific INRIX Data Services specifically developed for public-sector transportation agencies. Cost of services costs are based on annual terms but can be customized. Cost structure is proprietary as most agencies contract for specific services through a competitive process.

INRIX Real-Time Flow Data Service

Real-Time Flow is INRIX's full suite of traffic data which is available via an API call as often as once per minute. The data provided via XML includes road segment code; roadway name and cross streets of roadway; time; current speed in mph ("speed"); typical speed in mph ("average"); free flow speed in mph ("reference"); and travel time along segment in minutes ("traveltimeminutes"). Predicted speeds and travel times are also available via an API call. Predicted times include 15 minute, 30 minute, 60 minute, 24 hours, and 48 hours into the future at a minimum.

INRIX Historical Traffic Flow Statistics

INRIX Historical Traffic Flow Statistics use comprehensive data from the INRIX Smart Driver Network, including billions of historical data points, to provide true historical average speeds and statistical distribution on individual road segments covering over nine thousand centerline miles on major freeways, highways, urban and rural arterials and side streets throughout Washington State. This data is specific to every day of the week, every hour or quarter hour of the day, and is reported at the Traffic Message Channel (TMC) link level.

Using INRIX's proprietary technology, the data is analyzed and normalized to account for the impact of major events, seasonal traffic patterns, typical weather conditions and other variables that can impact traffic flow—ensuring the highest degree of accuracy. The data is updated regularly, incorporating both changes to map databases as well as additional historical data from the INRIX Smart Driver Network.

INRIX statistical parameters provided for each time bin include the average speed, standard deviation, the 5th, 10th, 15th, 20th, 25th, 30th, 40th, 50th, 60th, 70th, 75th, 80th, 85th, 90th, and 95th percentiles of speed values in mph, and the 30, 40, 50, 60, 70, 80, and 90 mph "failure rates," or the percentage of data points that fall below the specific speed threshold for the given segment/time.

In addition to typical annual files, monthly and quarterly files are also available. Monthly and quarterly files are available starting January, 2009. Future monthly files for 2012 and beyond can be made available by the 15th of the following month.

INRIX Historical Archive

INRIX Historical Archive is a running archive of Real-Time speeds provided by INRIX for all TMCs in service at that time. Data provided on each segment for each time slice includes the Speed, Travel Time, and the Score. The archive data is available in 5-minute increments beginning July 1, 2008, through December 31, 2010. Starting January 1, 2011, the running archive is available in one minute increments.

INRIX Real-Time Traffic Information Network

INRIX's Real-Time Traffic Information Network provides a visualization of real-time traffic information via a hosted statewide traffic monitoring website of the real-time coverage so agencies can monitor conditions on roadways in real time at Traffic Operations Centers and for other internal agency viewing/use. The monitoring website includes access for 20 simultaneous users (more users are available for additional amount depending on number of users).

INRIX Traffic Tile Overlays

Traffic Tile Overlays allow for a web services API request to INRIX and, in response, agencies receive a semi-transparent, match projection (Mercator) overlay image of INRIX Traffic Flow for display on common mapping platforms (Google, Microsoft, Yahoo!, etc.).

INRIX Traffic Tile Overlays are fully configurable and are available in various tile sizes (256 x 256 standard) with broad browser support (IE, Firefox, WebKit [Safari, Chrome]) and token-based security (access via token with timeout, configurable to a user).

INRIX Dynamic Route Travel Times

INRIX Dynamic Route Travel Times provide the current travel time (based on real-time traffic conditions) along a precisely-specified route between any Origin (starting point) and Destination (ending point) in either direction. Travel times can be queried in real time with archive rights to provide information for both instantaneous and historical analytic purposes This information may be used for: corridor management; planning and modeling purposes; reliability and performance measures; making operational decisions; and disseminating traveler information like travel times on dynamic message signs (DMS).

INRIX Analytical Tools

INRIX data, both real-time and historic, is now available as a subscription service bundled with state-of-the-practice analytics and visualization tools. See full description and figures in response to Question 7.

9. Describe any potential demographic bias that exists in your data sources.

   N/A

10. Describe any data usage clauses of agreements that come with new vehicle/device purchases that enable or authorize your firm to use personal mobility data.

    N/A

11. Describe how data source privacy and location/time of day details are protected.

    INRIX assures complete privacy by requiring all data providers remove all personally identifiable information from all data prior to sending to INRIX.

Nokia/NAVTEQ

1. Describe the primary and secondary markets for your data products (i.e., real-time traffic, transportation planning, etc.)

   Nokia is successfully delivering traffic data to government agencies and private sector companies globally. Nokia's primary market is real-time data for traveler information and secondary markets are traffic operations and transportation planning, including archived data. Nokia North American clients include over 20 State Departments of Transportation, Verizon Wireless, Sirius XM Satellite Radio, Microsoft Bing, Garmin, The Weather Channel, and Comcast. We also provide traffic data services and applications to all of the major United States mobile phone carriers: AT&T, Sprint, T-Mobile, and Verizon.

   Collectively, more than 220 million people in North America are served by Nokia traffic. Nokia Traffic powers:
   - 20 out of the 20 top ranked in-vehicle navigation systems, supporting over 150 vehicle models;
   - 75% of traffic enabled personal navigation devices (PNDs), including 3.1 million PND users with Ad-supported free lifetime traffic; and 3.5 million paying subscribers
   - Major online mapping applications including Bing Maps
   - Wireless providers including Nokia, Verizon's VZ Navigator, RIM/Blackberry, and Gokivo the first traffic enabled Turn-By-Turn app for the iPhone

2. List and describe the different data-generating technologies that are used to build your company's data products (for example, in-dash navigation devices, personal navigation devices, non-GPS cell phones, GPS cell phones, truck GPS/AVL). Is one technology primary; if yes, please identify.

   Nokia's traffic is powered by a variety of originating probe devices including GPS-enabled personal navigation devices, commercial fleets and commercial third-party applications. Many of the clients who use our navigation applications provide data back to Nokia. The mix of data, coupled with fixed point traffic monitoring devices, creates extensive coverage on freeway and arterial roadways using sophisticated algorithms to produce reliable data for Nokia's private and public clients.

3. If you are using personal mobile devices (such as cell phones or smart phones) as a data source, please describe the penetration rate of this data source.

   Nokia receives billions of GPS data points monthly globally from cell phones and other smart devices. The penetration rate varies by geographic area, time of day and day of week.

4. Describe your company's plans for future vehicle or personal technology and data product development.

   Nokia is constantly working to expand and enhance the data, coverage and applications. A few examples of our latest products include Nokia Drive, Nokia Transit, and Mirrorlink. Nokia Drive and Nokia Transit applications allow users to interact more naturally with the world surrounding them and helps travelers determine their best routing options. Mirrorlink enables the phone to interact with in-vehicle devices integrating the smartphone into the dashboard to provide safe and seamless access within the connected vehicle.

5. What is your current geographic coverage? Do you have plans to expand? If so, please elaborate on the planned geography as well as implementation timeline.

   In the United States the Nokia coverage includes all the Interstate highways and limited access roadways, in addition to thousands of miles of arterial roadways in metropolitan areas. We continue to expand the arterial footprint to more rural areas and will soon be expanding to include additional major arterial roadways nationwide.

6. Describe any of your data products' limitations that are relevant to road functional classification or other transportation system characteristics. Are expected error ranges provided by road segment or TMC location code?

   The roadways with higher volume and higher functional class roadways also carry more available probe vehicles. Lower functional classification roadways have many complex movements and require more sophisticated algorithms to match data and determine the speeds associated with a smaller arterial. To provide users with a mechanism to manage data that is distributed in time and space and calculated for each roadway segment, Nokia provides a confidence level in its real-time traffic data feed. The confidence value provides an indication of the quality and density of the data in time and space. There are also specific challenges in very complex urban conditions including tunnels and bridges structures with multiple levels that we are continually working to make as precise as possible.

7. Describe how your products are packaged to specifically serve the needs of transportation planners for travel demand forecast modeling or for congestion management programs.

   Nokia has an application called Traffic Patterns that includes archived data in a variety of aggregations. The most commonly used format is 15-minute average speed and travel time for each TMC at 15-minute increments for each day of the week.

   Nokia currently provides real-time and archived data to Michigan DOT. This data will be integrated into the RITIS software application provided by the University of Maryland for performance measure and other transportation planning.

8. For each of your specific data products (copy and repeat this question and categories as needed), please describe the following:

   a) Data product name and description
   b) Raw data frequency and accuracy
   c) Data cleaning process
   d) Level of aggregation or disaggregation
   e) Cost structure

Real-Time Data Feed

Nokia Traffic Satellite—Nokia is the exclusive provider of real-time traffic services for all satellite radio providers who offer traffic information via satellite radio in North America.

Nokia Traffic RDS—Real-time traffic delivered over FM radio using a radio data system (RDS) sub-carrier channel. RDS is well suited for auto companies and PND manufacturers.

Nokia Traffic ML—Real-time traffic designed for mobile and server-based navigation, as well as mapping applications.

Nokia Traffic Online—Real-time traffic available via consumer traffic websites.

Nokia Traffic Digital—Real-time traffic delivered over digital radio's high bandwidth will mark a major leap forward as additional data services beyond traffic become available.

Real-time data is continuously collected, models are updated every minute and the files containing the updates are provided every two minutes. Accuracy for an individual TMC depends on the number of data sources the age of the data and the variability of traffic conditions.

The data is quality checked prior to data integration, map matched in real time, cleaned for erroneous values, and filtered extensively. The cleaning and matching process is more extensive on arterial roadways due to increased variability of travel patterns requiring more sophisticated map-matching and algorithm processing to support quality results.

The data is collected from consumer and commercial probes, toll tags and sensors. The average speed and travel time is aggregated for each TMC using a sophisticated algorithm.

Nokia's diverse user community of public and private sector clients demands a highly flexible and adaptable licensing program. Several licensing options are available to ensure scalable and effective use of Nokia data. Licensing fees are determined based on the number of users, licensing term (i.e., number of years), geographic extent, and data delivery mechanism (e.g., desktop, web based). Fees are not based on the miles of roads covered in a region.

Archived Data

Nokia maintains several types of historic traffic data that is summarized in the Nokia Traffic Patterns and Traffic Analytics products. Traffic Patterns applies a weighting algorithm to multiple years of traffic data to provide the best possible assignment of the typical speeds experienced on all roads and highways. Traffic Analytics is an annual summary of traffic speeds contain 15-minute intervals by day, month and season. Both models also represent holiday travel conditions. Confidence scores are provided with all data summaries. Nokia also summarizes data directionally at the link level and TMC level for all roads.

Raw GPS data is used to develop the archived products.

During the data cleaning process for the data archive, erroneous values are removed and a raw probe is reformatted to enable map matching. The output of our clean process normalizes the data and ensures uniformity and quality prior to data aggregation. The data cleaning occurs on data obtained from the archive of the real-time traffic services as well as from probe archives received directly from third-party vendors.

The data is collected from individual probe points, then it is aggregated into 15-minute speeds and available at the TMC level and detailed Nokia link level for all mapped roads.

Nokia's diverse user community of public and private sector clients demands a highly flexible and adaptable licensing program. Several licensing options are available to ensure scalable and effective use of Nokia data. Licensing fees are determined based on the number of users, licensing term (i.e., number of years), geographic extent, and data delivery mechanism (e.g., desktop, web based). Fees are not based on the miles of roads covered in a region.

9. Describe any potential demographic bias that exists in your data sources.

   We receive data from a variety of data source types that likely mitigates bias in the data sources. Approximately half of the probe data is consumer, approximately half is commercial. The consumer mix includes mobile phones and mobile phone applications, personal navigation devices and navigation systems.

10. Describe any data usage clauses of agreements that come with new vehicle/device purchases that enable or authorize your firm to use personal mobility data.

    Many of Nokia's clients provide data back to Nokia and therefore provide the data usage clause to the user. Some of Nokia's products simply ask "(Product name) would like to use your current location, OK?"

11. Describe how data source privacy and location/time of day details are protected.

    All data collected has a unique ID and there is no way to track the information back to an individual user. As an example, Nokia uses a completely anonymous id that resets periodically to further ensure anonymity.

TomTom

1. Describe the primary and secondary markets for your data products (i.e., real-time traffic, transportation planning, etc.)

   TomTom markets served include consumer navigation systems, commercial vehicle navigation systems, in-dash solutions for automotive OEMs, traffic information systems (e.g., 511), transportation planning and modeling, and GIS analysis.

   Historic Traffic
   Transportation Planning/Traffic Engineering: Private/Public-Government
   - Network performance monitoring
   - Road network bottleneck reporting/analysis
   - Noise and emission hotspot identification
   - Before and after studies—changes to road infrastructure, construction
   - Performance analysis of intersections
   - Traffic model calibration

   Geo-Marketing
   - Site location
   - Advertising display location
   - Demographic travel patterns

   Logistics
   - Fleet management (See also Navigation below)
   - Supply chain optimization (See also Navigation below)
   - Delivery scheduling (See also Navigation below)

   Navigation (Automotive, Personal Navigation, Internet, Mobile)
   - Estimated Arrival Time calculation based on day of week, time of day
   - Time specific route selection: Routing based on day of week, time of day

   Insurance
   - Risk assessment—accident hotspot and high risk area identification

   Live Traffic
   Transportation Planning/Traffic Engineering: Private/Public-Government
   - Active traffic management
   - Traffic Monitoring—Traffic Control Centers
   - Flow Data—Speed, Travel Time
   - Incident Data—Traffic Jam, accident, closure etc.
   - Traffic website—511 etc.
   - Variable message sign display—travel time, delay time

   Navigation
   - Dynamic live navigation/routing
   - Dynamic pre-trip route planning
   - Dynamic Estimated Time of Arrival calculation

2. List and describe the different data-generating technologies that are used to build your company's data products (for example, in-dash navigation devices, personal navigation devices, non-GPS cell phones, GPS cell phones, truck GPS/AVL). Is one technology primary; if yes, please identify.

   TomTom data sources include connected (GSM enabled) and non-connected after-market GPS devices, in-dash GPS systems, commercial vehicle GPS systems, GPS smart phones, and third-party incident data. After-market GPS devices represent TomTom's primary data source.

3. If you are using personal mobile devices (such as cell phones or smart phones) as a data source, please describe the penetration rate of this data source.

   The specific penetration rate is unknown. There have been in excess of 1 million downloads of the TomTom smartphone GPS navigation application globally—but TomTom only collect traffic traces from devices docked in a car holder as a better indication that they are being used on a vehicle journey as opposed to train/bus etc.

   TomTom also uses cellular probe data as an input source. By looking at the activity of cell phones moving near GSM network antennae, the (anonymous) handset location can be matched to the road network and speed information calculated. Over 80 million GSM probes contribute to the TomTom real-time traffic system in Belgium, France, Germany, Italy, Netherlands, Portugal, Switzerland, and the UK—but always as a supplement to GPS probe data not as a substitute for GPS data.

4. Describe your company's plans for future vehicle or personal technology and data product development.

   TomTom will continue to focus on smartphone applications, in-dash systems, and after-market navigation devices to ensure that the richest set of accurate GPS real-time traces from a broad range of vehicles are available for creating traffic information.

   There will also be a continued focus on traffic content, quality, coverage and geo-expansion to make traffic information available in more markets globally. TomTom will add increased probe quantity both in real time and for the historical traffic database by adding 3rd party partner GPS data and connected device information starting from second half of 2012 to improve the accuracy and confidence in the data. The project to be undertaken which will facilitate addition of the connected, (live,) and 3rd party probes to the historical database will not only increase sample size but also improve the freshness of the data available for analysis. This will take place across all markets where TomTom operates including North America.

   TomTom has technology for multiple platforms. These include after-market consumer navigation devices, smartphone applications, fleet management systems and in-dash solutions. Predictive (based on historical information) and real-time route information is also made available on TomTom's web platform. All our systems are supported by a customer care division, which has won a series of JD Power awards in recent years.

   TomTom's back-office systems include our own data fusion technology and provide updated traffic information every minute. The data fusion uses anonymous crowd-sourced GPS speed measurements from devices and smartphone applications as well as third-party information on road closures and accidents. Industry standards for data transmission are used and the traffic information can be provided to public and private entities using standard protocols.

   Individual privacy is of paramount importance in TomTom's systems. Because a large proportion of real-time information is crowd-sourced, the crowd has to be able to trust that their information will not be misused or shared inappropriately. Protecting privacy goes beyond legal limitations: if the crowd does not perceive us as trustworthy, the crowd will no longer be a source. TomTom has developed methods of safeguarding the privacy of individuals and the information they provide. These continue to evolve as the number of channels providing crowd-sourced information grows.

5. What is your current geographic coverage? Do you have plans to expand? If so, please elaborate on the planned geography as well as implementation timeline.

   TomTom provides historical traffic data for 45 countries, and real-time traffic data in 23 countries. Coverage is expanding to include China and Russia this year and will continue to expand in the future.

6. Describe any of your data products' limitations that are relevant to road functional classification or other transportation system characteristics. Are expected error ranges provided by road segment or TMC location code?

   Limitations: probe-based speed measurements rely on there being vehicles on the road to provide data. Small, local roads with little or no traffic have therefore the lowest coverage in our system, yet we do provide speeds based on multiple years of measured, historical speed information by time of day and day of week. Some of our products include a 'confidence value' for each road section to indicate the number of recent observations taken into account with that updated value in the file. Due to limitations of TMC table coverage, TomTom has implemented new location referencing which can provide traffic information at any location (see OpenLR.org).

7. Describe how your products are packaged to specifically serve the needs of transportation planners for travel demand forecast modeling or for congestion management programs.

   TomTom Traffic Stats
   TomTom lets you access the largest historic traffic database in the world available for governments and enterprises via our web-based Traffic Stats portal. Measure location

accessibility for site selection, identify road network bottlenecks and noise emission hotspots, perform before and after studies relating to infrastructure changes or analyze intersection design and performance. All you need is an Internet connected computer, and you will receive 24 × 7 access to TomTom traffic data.

- Access to TomTom historical traffic products anywhere & anytime.
- Tailor-made reports available within 24 hours.
- Data can be downloaded for use in other applications/traffic modeling tools, etc.

Through the Traffic Stats website three products, Custom Area Analysis, Custom Travel Times and Custom Probe Counts can be accessed. Below please find brief descriptions of these three products, with a more detailed description to follow in response to question 8. NOTE: Custom Probe Counts has yet to be officially released.

Custom Area Analysis
Custom Area Analysis delivers a Shape file (*.shp) with both the road geometry of the segments analyzed and the data base, (*.dbf) with the related statistical data for each road segment. This data can be readily uploaded into standard GIS tools to visualize and manipulate the data. A sample size is provided upon completion of the query and one can reject a report before accepting if the sample size proves inadequate for a given project. This allows for the time parameters to be adjusted so that a greater sample size might be obtained. NOTE: This is also the case for Custom Travel Times which is outlined immediately below.

Custom Travel Times
Results of a Custom Travel Times query are given in three different types of output:

- Excel
- KML
- Charts

NOTE: Charts are only available for viewing within the Traffic Stats web portal. However, graphs can be made using the Excel output.

Custom Probe Counts
The results are given in an industry standard ESRI Shapefile compatible dBASE file and can be downloaded through a simple web interface, once confirmation has been given that the job has been processed.

Speed Profiles
A fourth historical speed product is also available for the purpose of providing real-world data for the calculation of routes and estimated times of arrival in routing/navigation applications. Other potential uses of the data may also be possible though it was created to specifically improve route and ETA calculation. This product links directly to the TomTom MultiNet map database. This product is based on a two year average of speed per time of day, (5-minute time bins,) and is released quarterly.

8. For each of your specific data products (copy and repeat this question and categories as needed), please describe the following:

   a) Data product name and description:

   TomTom Custom Travel Times
   Custom Travel Times is a traffic solution designed to give government agencies and enterprises more insight into traffic flows on a specific roadway, or series of connected roadways, (created using an A to B route calculation.) Custom Travel Times provides highly granular speed and bottleneck information for roads around the world. TomTom's ever-expanding historical traffic database has over 5 trillion data points with over 6 billion new records being added each day—with some roads having more than 20,000 measurements. This makes it possible to obtain actual driven travel times and speeds on any stretch of road over any period of time and time of day. Custom Travel Times covers all roads, from major highways to local and destination roads, throughout Europe and North America. Please find detailed data output information below.

   - Excel
     - Route Name: Customer defined
     - Time collection: Customer defined
     - Length (meters)—Total length of the route.
     - Sample size: Average per segment.
     - Average Travel Time (hh:mm:ss)—Arithmetic average of travel time over the route.
     - Median Travel Time: (hh:mm:ss)
     - Average speed: (kph/mph)—Harmonic average speed.
     - Travel Time ratio: Comparison Sets
     - Percentile travel times: 5%–95%
   - KML
     - Average speed: KPH/MPH
     - Length of segment: Meters
     - Average travel time: Seconds
     - Median travel time: Seconds
     - Standard deviation: Seconds
     - Sample size
     - Travel Time Ratio: Comparison Sets
   - Charts
     - Available online for viewing. Not downloadable.
     - Charts/Graphs can be made from the data contained in the Excel output.

TomTom Custom Area Analysis

Custom Area Analysis is a traffic solution designed to give government agencies and enterprises more insight into traffic flows for complete road networks. Custom Area Analysis provides highly granular speed and bottleneck information for large and small road networks around the world across all road classes. TomTom's ever-expanding historical traffic database has over 5 trillion data points with over 6 billion new records being added each day—with some roads having more than 20,000 measurements. This makes it possible to obtain actual driven travel times and speeds on any stretch of road over any period of time and time of day. Custom Area Analysis covers all roads, from major highways to local and destination roads, throughout Europe and North America. No need to purchase road collection hardware such as cameras, and loop sensors. With Custom Area Analysis it is possible to gather statistics on the road network as a whole, from Interstates through arterials to the local street level.

- Shapefile (*.dbf) output fields
  - ID: Segment ID number
  - Average Travel Time: Harmonic average travel time. kph or mph
  - Median Travel Time: Seconds
  - Average Speed: Arithmetic Average. kph—mph
  - Median Speed: kph or mph
  - Standard Deviation of Speed: kph or mph
  - Sample Size: Per segment
  - Travel Time ratio: Comparison Sets. Seconds
  - Percentile travel times: 5%–95%

Custom Probe Counts

Not all studies of road usage require speed and congestion information. Sometimes it is interesting to understand the relative traffic volumes that use the roads around sites being considered for locating new buildings or shopping areas—or even which locations might attract more viewers for an advertising campaign. Custom Probe Counts can provide data on the number of devices that were recorded on individual MultiNet road segments for any given period back to 2008. Highly granular data from TomTom's ever-expanding historical traffic database has over 5 trillion data points with over 6 billion new records being added each day—with some roads having more than 20,000 measurements. Using this data, analysts can quickly gather an indication of the relative traffic volume on the individual roads surrounding the location of interest.

Speed Profiles

Speed Profiles is different. It is a comprehensive database of actual historic roadway speeds. These are attained by aggregating billions of GPS measurements to offer precise speeds for specific times of day and days of the week. With Speed Profiles routes adapt dynamically to the time of departure and incorporate local knowledge. The optimal route on Monday morning may differ on Sunday afternoon, just as the travel time on a Tuesday in December will be longer than in May. Armed with Speed Profiles, ETAs are highly accurate and travel time is reduced along with stress levels, travel costs and environmental impact. Deployment effort is minimal thanks to a compact data footprint and with wide coverage of highways, urban, rural and secondary roads it delivers a seamless country by country experience.

TomTom Enterprise Traffic

TomTom Enterprise Traffic provides precise locations and delays caused by congestion on the road network, allowing routing programs to provide the fastest route based on actual current travel times. By incorporating TomTom Enterprise Traffic into a navigation solution, drivers can determine the quickest route to their destinations by considering "live" road conditions. The data in each Enterprise Traffic file includes road delays allowing routing programs to evaluate the true travel time to each destination. This product can also be deployed for display purposes such as in traffic control center or even embedded into a website for A to B dynamic routing. The Enterprise Traffic XML feed is made available for consumption via the client pull method every minute in standard formatting and can be implemented using either TMC segments or OpenLR.

TomTom HD Flow

TomTom HD Flow delivers a real-time, detailed view of traffic speeds on the entire road network, designed for easy integration into traffic management systems or calculating current routing travel times. The data output is generated from TomTom's proprietary fusion engine and is refreshed every minute. For each road segment this delivers the road's identification, the total travel time under current and ideal conditions and the average speed and quality for that segment. This enables traffic control centers to determine relative levels of road service over wide areas, and personal navigation device and smartphone manufacturers to benefit from dynamic routing and display. With HD Flow we help you in the placement of real-time traffic signs showing the fastest road choice to a common destination. A quality/confidence value is also provided in the output. The data is made available in XML feed, client pull method, every minute, DATEX2 format, TMC application.

TomTom HD Route Times

TomTom HD Route Times is a turnkey solution providing highly accurate real-time travel and delay times for a specific route either on a temporary basis or for permanent solutions. Key to this data is its flexibility. Travel times and delay statistics are delivered without the need to build infrastructure or install and maintain hardware. Data can be sent continuously to roadside information systems or on a temporary basis to mobile systems used during road works. The data is refreshed every minute allowing traffic control centers to deploy variable message signs (VMS) suggesting alternative routes. Event managers can schedule travel information to be displayed around special events. Even kiosks and corporate offices can take advantage of TomTom HD Route Times' detailed, granular data for specific local route data.

b) Raw data frequency and accuracy:

TomTom GPS data

In raw form, TomTom GPS data has the following characteristics:

- Time stamp: year, month, day, hour, minutes, seconds, hundredths of seconds
- Position stamp: latitude and longitude
- Identification number (randomly generated)

Data from real-time GPS devices are transmitted between every 40 and 60 seconds to a server after which it is processed through the proprietary TomTom fusion engine.

For privacy protection the identification number is changed for every 24-hour period.

Further processing makes it possible to calculate:

- Speed
- Travel time
- Speed and travel time variance
- Acceleration rate
- Deceleration rate
- Trip origin, destination, route choice

To calculate these statistics, the GPS data points must first be 'map-matched'. This is a process developed by TomTom to match each GPS point to a specific road segment as defined in our own map database. The map-matching process also performs filtering of unreliable GPS measurements.

Data presentation:

- Data made available every minute
- XML Feed
- Datex2 format
- Location referenced by TMC code or OpenLR

Real-Time Output Quality:

- TUV Certified
- Regular QKZ tested (Enterprise Traffic)
- HD Flow tested by methodology developed by the Texas Transportation Institute

c) Data cleaning process:

All data inputs are processed through the proprietary TomTom applications to remove possible errors such as map-matching anomalies, outliers etc. Filters are applied to remove possible data from devices used on public transport/pedestrians and also where vehicles are temporarily in gas stations etc.

d) Level of aggregation or disaggregation:

With respect to aggregation at the road element level TomTom offers different levels of aggregation. In Custom Area Analysis and Custom Travel Times road elements can be quite short, (multiple road elements within one city block for example,) to road elements on highways which can be measured in terms of miles. In Live Traffic offerings Traffic Messaging Channel links can be utilized which result in much longer road elements and cover only the TMC network. Real-time data can also be made available in the above mentioned shorter road element lengths via our open-source OpenLR format which allows for the application of traffic information across the road network wherever information is available. With respect to Live Traffic updates are made available via the client pull method every minute.

The historical products Custom Travel Times and Custom Area Analysis data can be sliced by one hour combinable time bins. Future iterations of these two products will offer more flexibility with respect to time bins the details of which are not available at this time. Speed profile data represents two year averaged data which is applied to the TomTom MultiNet map at the road element level.

e) Cost structure

Custom Travel Times: Price per route.
Custom Area Analysis: Price per geographic scope, road class and number of queries.
Custom Probe Counts: Price per mile.
Real-Time Traffic: Price per mile.

9. Describe any potential demographic bias that exists in your data sources.

When TomTom first entered the mass market for navigation we saw a bias toward men between the ages of 25 and 35 who drive often. We now know from our own market research that our customer base is now much more representative of the general population, but we still have some underrepresentation of women and of people older than 65 years of age or short-distance commuters. There may also be some bias against very low household income groups.

10. Describe any data usage clauses of agreements that come with new vehicle/device purchases that enable or authorize your firm to use personal mobility data.

    After-market devices include an opt-in question. In-dash systems vary. Opt-in depends on the actual OEM and the specific agreement.

11. Describe how data source privacy and location/time of day details are protected.

    Individual privacy is of paramount importance in TomTom's systems. Because a large proportion of real-time information is crowd-sourced, the crowd has to be able to trust that their information will not be misused or shared inappropriately. Protecting privacy goes beyond legal limitations: if the crowd does not perceive us as trustworthy, the crowd will no longer be a source. TomTom has developed methods of safeguarding the privacy of individuals and the information they provide. These continue to evolve as the number of channels providing crowd-sourced information grows. TomTom uses various data sources to create its maps and map related products and services, such as real-time and historic traffic information. These data sources include geolocation data obtained from individuals, who need to be able to trust TomTom to use their data in a responsible way that does not violate their privacy. To obtain the highest possible yield, resulting in the highest possible quality, it is paramount for TomTom to foster this trust. As such TomTom is committed to acting above and beyond the OECD privacy principles (notice, purpose, consent, security, disclosure, access and accountability), also laid down in legislation and enforced by independent regulators in the various countries in which TomTom operates.

    More specifically, in all cases where TomTom obtains geolocation data from its customers, this is done based on prior, explicit, informed consent, which can be withdrawn at any time. Effectively this means that users of TomTom products are informed about the use of their geolocation information and **voluntarily opt-in, before** any geolocation information is captured to be sent to TomTom for further use. Users are informed about the data being captured and the fact that it will only be used for map, road, traffic and traffic related purposes, under the moniker "we profile roads, not people." Users also are informed that TomTom will use the data **anonymously**, i.e., elements, such as account names, email addresses or unique serial numbers, potentially allowing (re-)identification of the user are destroyed either immediately or within 20 minutes after their device or car has been shut down. Users also are informed about the fact that TomTom uses their information in an anonymous way to enhance its products and services, which also are made available to business and governments.

    In those cases geolocation information is obtained from third-party sources, TomTom ensures, technically, organizationally and contractually, the data it receives from the third party does not allow TomTom to identify or even single out the individual contributing the data: TomTom obtains anonymous data only and uses it only for the agreed purpose after which the data is destroyed. TomTom applies advanced Privacy Enhancing Technologies and organizational measures to subsequently live up to the agreement with its users and third parties. All geolocation data is protected against unauthorized access (i.e., anyone except TomTom) with strong encryption while stored on end user devices and while in transit. To avoid identification or singling out individuals, TomTom irreversibly destroys any unique identifiers immediately upon reception of the data from its users. It those cases where this is not possible (specifically: generating traffic information), one-way pseudonyms with a short lifetime are used and data is kept in volatile memory to not create potentially recoverable copies of identifiable geolocation data. Pseudonym lifetime is capped at a maximum of 20 minutes after the device contributing the data has been switched off or 24 hours, whichever is shorter. In those cases where TomTom retains a copy of the geolocation data, this always is done without any identifying elements (such as device unique serial numbers) and on a per trip basis only, i.e., without maintaining the relationship between trip originating from the same device.

    Please also refer to www.tomtom.com/yourdata for our publicly available information regarding the way TomTom treats information obtained from its customers.

TrafficCast

1. Describe the primary and secondary markets for your data products (i.e., real-time traffic, transportation planning, etc.)

   Primary markets:

   - Provide real-time traffic data (Dynaflow and incidents) to support online and mobile applications and to TV stations for broadcasting.
   - Provide BlueTOAD services to public agencies for workzone impact study, special events traffic flow

monitoring, OD study, and real-time travel speed and travel time reporting.

Secondary markets:

- Provide real-time and historical speed data to public agencies for transportation planning and modeling.

2. List and describe the different data-generating technologies that are used to build your company's data products (for example, in-dash navigation devices, personal navigation devices, non-GPS cell phones, GPS cell phones, truck GPS/AVL). Is one technology primary; if yes, please identify.

For Dynaflow product, major input data come from two main sources:

- GPS data from fleet and mobile device that TCI purchases and data exchanges from fleet management companies and business partners that provide location-based service through application installed on GPS-enabled mobile phone.
- Data collected by BlueTOAD installed across more than 20 states in the U.S.A. and in other countries such as Canada, Chile, and Brazil. See http://trafficcast.com/products/view/blue-toad/ for the detailed information about BlueTOAD.
- The majority of mobile data used to produce Dynaflow come from GPS data collected through fleets.

For incident data service:

- More than 120 Java programs are developed to retrieve and parse publically available incident data from city, state DOTs and 511 systems.
- Data collected by TCI operators at National Operation Center based in Madison, Wisconsin, and Philadelphia through TrafficCaster, an interactive map and table driven web-based application developed by TrafficCast. Operators watch the traffic camera and listen to police radio scanner to identify and confirm traffic incident during morning and evening rush hours.

3. If you are using personal mobile devices (such as cell phones or smart phones) as a data source, please describe the penetration rate of this data source.

This information is not available to TCI due to data from mobile devices are indirectly obtained through business partners that TrafficCast doesn't have information about the penetration rate.

4. Describe your company's plans for future vehicle or personal technology and data product development.

TCI has the opportunity to take a close look at the mobile data collected by telecommunication carriers and is in the process of investigating its critical attributes required for generating useful traffic information. The preliminary findings indicate the needs of adopting theory developed in the linear dynamic system to make this data useful. The final goal of this exercise is to come out a software product that runs on either device side or server side, which ever mobile data is available, so the traffic information can be produced on the fly to feed mobile application needs.

Another product plan that is directly related to end users and has already been in the testing phase is the traffic broadcaster application that is currently under reviewed by media companies. This application is designed to produce graphical and narrative traffic information for live TV and radio broadcast. The requirement of this initiative also calls for providing mobile traffic application for smartphone so media can use it as name branding and promotion vehicle. In return, the data collected by these applications will be used to enhance TrafficCast's product offering.

As for the BlueTOAD, TCI has developed several related products to meet both real-time traffic management and offline transportation study needs. Please see http://trafficcast.com/products/view/blue-toad/ for the details.

5. What is your current geographic coverage? Do you have plans to expand? If so, please elaborate on the planned geography as well as implementation timeline.

Currently TCI Dynaflow and incident data service coverage for functional class 1 to 4 roads in the U.S. BlueTOAD deployment covers major road and arterial in 20+ states in the U.S. as well as in Vancouver, Calgary, and Kelowna, Canada, Sao Paulo, Brazil, Santiago and Puerto Montt, Chile, and Hong Kong, China.

The planned geographic expansion includes providing top 20 markets of Dynaflow in Canada by 2013 Q1 and continuously expands BlueTOAD deployment to major metropolitan areas in the U.S. and Central and South America in 2013 and 2014.

6. Describe any of your data products' limitations that are relevant to road functional classification or other transportation system characteristics. Are expected error ranges provided by road segment or TMC location code?

Dynaflow: It covers functional class 1 to 4 roads in the U.S. The error rate varies from 10% to 35% depending on the markets and road functional class. Freeway and major arterial in an urban area usually have a higher accuracy than that of minor roads. This is probably due to flow disruption caused by the deployment of unsynchronized traffic control mechanism such as traffic signal or stop sign.

BlueTOAD: It mainly covers arterial roads in urban and suburban areas. The speed accuracy is about 85% to 93% measured by the ground truth done by TCI team, TCI customers, and independent consulting firms.

7. Describe how your products are packaged to specifically serve the needs of transportation planners for travel demand forecast modeling or for congestion management programs.

   Dynaflow-historical is produced and updated once a year based on GPS data to provide historical trend that can be utilized for traffic demand modeling and, service level assessment, and transportation planning. It is packaged in season, day of week, and time of day in 15-minute interval for each TMC for the entire U.S.

   BlueTOAD data has been used to provide real-time travel time delivery through variable message sign, and information to DOT web site. For planning purpose, BlueTOAD is also used by several public agencies to produce route traffic assignment percentile for a small network, which is essential for roadwork impact study and the measurement of effectiveness of detour advisory.

8. For each of your specific data products (copy and repeat this question and categories as needed), please describe the following:

   a) Data product name and description:

   Fused from a variety of sources including historical road speed trends, real-time GPS probe, speed data from public agencies, and anticipated traffic impacts such as incident, construction, weather and upcoming events from both public and private sources, Dynaflow provides accurate historical, real-time and forecast road speeds for the functional class 1 to 4 roads in the United States.

   - Dynaflow-real-time: update every minute, Dynaflow-real-time provides speed either based on NAVTEQ link ID or TMC location code depending on client's needs.
   - Dynaflow-predictive: Although there are seven days of predictive speed stored in the database, at any given time, TrafficCast only allows customer come to retrieve up to 48-hour of prediction. The reason is the quality of weather forecast, the dominant factor for the long-term traffic prediction, degrades drastically beyond 48 hours. To reduce the data size, this data is made available in TMC segment.
   - Dynaflow-historical: TrafficCast packages historical traffic trend data for traffic study or planning purpose in every 15-minute interval for each day in the week and by season. Its basic attributes include season, day of week, time of day, TMC segment ID, and speed. This data is updated annually and being used by some public agencies and navigation device manufactures.
   - Incident: gathered from public agencies and by TCI NOC (National Operation Center in both Philadelphia, PA, and Madison, WI), the content of traffic incident data include accident, planned roadwork, and emergency events such as flooding, hurricane evacuation, etc. This information has been provided to TV stations and major telematics service providers such as Google and OnStar.
   - BlueTOAD data: By detecting Bluetooth signal, BlueTOAD detects the MAC address of the device. With known BlueTOAD location and the time stamp a device is detected, travel time and travel speed between two BlueTOAD units are calculated. BlueTOAD is also used to assess the percentile of traffic distribution for a small network.

   b) Raw data frequency and accuracy:

Dynaflow-real-time and Dynaflow-predictive: The raw data used by Dynaflow-real-time, Dynaflow-predictive come from many sources and each with different update frequency and quality.

- Speed data from DOTs: TrafficCast has access to 45 markets of speed data collected by city or state DOTs with update frequency generally between 3 to 5 minutes. However, the technology used to collect such data imposes a challenge of regular maintenance of those detectors such that its accuracy is certainly questionable. In last few years, TrafficCast conducted two major quality assessments for DOT speed data collected from all available markets by comparing it to GPS probe data. The finding is somewhat discouraging due to many DOT data never change throughout a day probably because of malfunctioning.
- BlueTOAD data: Sold to public agencies for corridor and regional traffic data collection, BlueTOAD data has been integrated into Dynaflow as part of its inputs. It is one unique data source only available to TrafficCast. BlueTOAD scans passing by Bluetooth-enabled devices every 5 seconds and aggregate the data for speed calculation every minute. Its accuracy, which is in the range of 90% to 95%, has been verified by several DOTs and independent consulting firms through ground truth.
- GPS probe data: used to create Dynaflow products, GPS probe is updated in a frequency varies from one data provider to another. Some update every

10 to 30 seconds, and the others update every 1 to 3 minutes. It is difficult to measure raw GPS accuracy, but there are indicators, such as GPS fix time and number of satellites observed, in some of data sources that can tell if the data is reliable so Dynaflow model can decide whether they should be used as input.

- Incident data: incident data from private source usually update within one minute during rush hours. While generally the public sources only report incident at the beginning and the end of an event, except California Highway Patrol. It is expected that the quality of incident data is not consistent across the entire country with some include detailed event description but some only provide a short phrase such as "traffic accident" or "traffic collision" such that it is almost impossible to derive traffic impact from such limited information.

Dynaflow-real-time is updated every minute, Dynaflow-predictive is updated every 15 minutes, and Dynaflow-historical is updated annually.

Accuracy of Dynaflow-historical is difficult to measure, but the accuracy of Dynaflow-real-time and Dynaflow-predictive is between 65% and 90% depending on market and road functional class.

c) Data cleaning process:

- GPS probe data: criteria used to filter GPS probe data include its location (on the roadway or inside a building), quality index (for the data sources possess such information), heading (within 5 degrees compared to road geometry calculated from the digital map), and data latency (need to be within 5 minutes).
- DOT speed data: TrafficCast assigns a quality index 1 to 4, with 4 is the highest quality, to each DOT sensor dynamically based on offline data analysis, real-time speed trend (to make sure data changes with time),and comparison to GPS probe data. Quality index lower than 3 is not used by TrafficCast's traffic models.
- Incident data: The quality of incident data is inconsistent due to they come from different sources. Among three key attributes of an incident—time (when), location (where), and what happens, location is usually the most challenging one to deal with. For the data to be provided to customers and be used by traffic impact model, knowing the exact location of incident data is crucial (so a ramp closure won't be mistakenly treated as a highway closure). This is particularly true for the data used to support mobile navigation. Therefore, tremendous effort has been spent to refine traffic event location by a sophisticated map-matching algorithm. Data lack of sufficient information is not provided to customer or used by traffic impact model.

d) Level of aggregation or disaggregation:

Internally, TrafficCast builds Dynaflow model based on NAVTEQ link identifier. Therefore, one level of aggregation in spatial domain is to aggregate multiple GPS probe points fall into the same link ID within each model iteration cycle (which is usually one minute, the aggregation in temporal domain inside the model) to produce one single speed output for that particular link.

When the speed is produced based on TMC segment to support application needs, another layer of spatial domain is performed by merging multiple link speeds to a single TMC speed. Both Dynaflow-historical and Dynaflow-predictive are based on TMC and use 15-minute as basic time interval.

e) Cost structure:

The cost structure varies from one customer to another and it also depends on market segment such as public agencies, media (TV station), web portal, and mobile location-based services, etc. The table below shows how TrafficCast data products are packaged and sold to different market segments.

Product\Customer	public agency	media	web portal	mobile LBS
Dynaflow-real-time	city or state	city	nation wide	nation wide
Dynaflow-predictive	state	-	nation wide	nation wide
Dynaflow-historical	state	-	nation wide	nation wide
incident	-	city	nation wide	nation wide
BlueTOAD	city, county	-	-	-

- Dynaflow-real-time: For nationwide web portal and mobile LBS, TrafficCast charges client a lump sum annual loyal fee plus a small monthly subscription fee for each end user. If data is not use to support nationwide service, it is mainly charged based on the number of markets and the tier of market (top 20, 20–40, 40 after, etc.) they belong to. For public sector that requires statewide coverage, the price is based on total road mile.
- Dynaflow-historical: It is charged based on number of road mile and data update frequency.
- Dynaflow-predictive: It is charged based on the tier of market (top 20, 20-40, 40 after, etc.) and number of markets. If it is for nationwide traffic service, then

the pricing structure is similar to that of Dynaflow-real-time.

Incident data: It is charged by the tier of the market for media. For web portal and mobile LBS, TrafficCast charges a lump sum annual loyal fee plus a small monthly subscription fee for each end user.

BlueTOAD: it is sold by number of BlueTOAD units. Data is included. Separate charges for solar panel and/or wireless communication fee are applied if power and/or Ethernet connection is not available at the site of installation.

9. Describe any potential demographic bias that exists in your data sources.

GPS probe data: the bias could be (1) it mainly comes from fleet (delivery truck and long-haul truck), (2) depending on data providers, the data could be limited to a certain region. Hence, data from multiple vendors is required to create a solid nationwide flow database, and (3) some data is collected from handheld devices, which is limited to those smart phones with GPS tracking capability enabled.

DOT speed: Generally, it is only available for the urban Interstate freeway system. Arterial may not be covered.

Incident data: Often time data from state DOTs or 511 systems doesn't cover traffic event occurs at local arterial. Therefore, the impact (delay and queue length) on minor roads may not always be available to Dynaflow.

10. Describe any data usage clauses of agreements that come with new vehicle/device purchases that enable or authorize your firm to use personal mobility data.

TrafficCast does not own any product or application to collect mobility data directly from end users.

11. Describe how data source privacy and location/time of day details are protected.

Privacy is a major concern to those GPS probe data providers; therefore, before data made available to TrafficCast, their unique identifiers (either, vehicle ID, device ID or device series number) have been converted to a set of random numbers using hash function that is unknown to TrafficCast.

APPENDIX D

Experiment A Models

Mode Identification MNL Model

Utility Parameters

Name	Value	Std Error	t-test	p-value
asc_auto	7.13	1.80e+308	0	1
asc_bike	-7.65	1.80e+308	0	1
asc_bus	-7.78	1.80e+308	0	1
asc_train	6.44	1.80e+308	0	1
asc_walk	1.86	1.80e+308	0	1
auto_highaccel	16.1	1.18E+03	0.01	0.99
auto_higheraccel	0.785	0.77	1.02	0.31
auto_highspeed	1.44	2.73E+04	0	1
auto_midspeed	-0.997	3.23E+03	0	1
bike_lowspeed	-0.671	2.80E+04	0	1
bike_midaccel	19	1.76E+03	0.01	0.99
bike_midspeed	13.3	2.79E+04	0	1
bus_highaccel	31.1	1.70E+03	0.02	0.99
bus_higheraccel	0.468	0.737	0.64	0.52
bus_highspeed	1.51	2.73E+04	0	1
bus_midspeed	-0.368	3.23E+03	0	1
train_highaccel	-4.46	6.26E+03	0	1
train_higherspeed	3.97	1.06	3.75	0
train_highspeed	20.5	2.79E+04	0	1
walk_lowaccel	0.916	1.3	0.7	0.48
walk_lowspeed	7.47	1.44E+04	0	1

Utility Functions

ID	Name	Specification
3	auto	asc_auto * one + auto_midspeed * midspeed + auto_highspeed * highspeed + auto_highaccel * highaccel + auto_higheraccel * higheraccel
2	bike	asc_bike * one + bike_lowspeed * lowspeed + bike_midspeed * midspeed + bike_midaccel * midaccel
5	bus	asc_bus * one + bus_midspeed * midspeed + bus_highspeed * highspeed + bus_highaccel * highaccel + bus_higheraccel * higheraccel
7	train	asc_train * one + train_highspeed * highspeed + train_higherspeed * higherspeed + train_highaccel * highaccel
1	walk	asc_walk * one + walk_lowspeed * lowspeed + walk_lowaccel * lowaccel

Trip Purpose MNL Model

Utility Coefficients (Excluded Items Equal to Zero)

Beta	Value	Std Error	t-test	p-value
asc_cra_lu_institutional	1.27	0.0605	20.95	0
asc_cra_nearchurch	1.21	0.0837	14.42	0
asc_ctm_iswalkorwheelchair	2.59	0.0618	41.85	0
asc_ctm_transfervariableatleastonenonauto	0.969	0.131	7.38	0
asc_ctm_transfervariablebothnonauto	3.35	0.121	27.74	0
asc_drp_actdurless10min	1.19	0.0473	25.29	0
asc_drp_dropoffvariable	1.39	0.0745	18.69	0
asc_drp_someonedropped	2.56	0.0805	31.77	0
asc_dth_tripdistgreater10mi	0.0837	0.0222	3.77	0
asc_emo_actdurgreater10min	0.55	0.0389	14.12	0
asc_emo_actdurgreater150min	-0.514	0.0775	-6.63	0
asc_emo_actdurless120min	0.413	0.057	7.25	0
asc_emo_adultparty12pmto2pm	0.172	0.0339	5.08	0
asc_emo_adultpartyactdur20to40min	0.238	0.0298	7.98	0
asc_emo_complexsubtour	0.377	0.0777	4.85	0
asc_emo_groupeatoutduration	0.69	0.0389	17.75	0
asc_emo_iswalkorwheelchair	0.375	0.0609	6.16	0
asc_emo_simplesubtour	0.954	0.0609	15.66	0
asc_grc_actdurgreater10min	0.61	0.0289	21.15	0
asc_grc_actdurgreater90min	-0.571	0.0413	-13.82	0
asc_grc_adultparty12pmto2pm	0.113	0.0262	4.32	0
asc_grc_groupgroceryduration	0.179	0.0288	6.21	0
asc_hcr_actdur30to90min	0.105	0.0235	4.48	0
asc_her_actdurgreater45min	-0.321	0.0334	-9.6	0
asc_lu_parks	1.58	0.0756	20.9	0
asc_orc_nearbigbox	0.76	0.0632	12.02	0
asc_orc_nonmand	0.898	0.141	6.36	0
asc_pbs_complexsubtour	0.213	0.0653	3.26	0
asc_pbs_simplesubtour	0.199	0.0627	3.17	0
asc_pkp_actdurless10min	0.955	0.0455	20.98	0

Beta	Value	Std Error	t-test	p-value
asc_pkp_pickupvariable	0.899	0.0818	10.99	0
asc_pkp_someonepicked	3.12	0.0903	34.59	0
asc_rec_grouprecreationduration	0.54	0.0545	9.91	0
asc_sch_schoolbusmode	2.38	0.185	12.86	0
asc_sch_schoollocationmatch	6.99	0.0895	78.12	0
asc_shp_actdurless60min	0.793	0.0329	24.07	0
asc_soc	3.15	0.0676	46.64	0
asc_soc_groupsocialvisitduration	0.554	0.0488	11.34	0
asc_srv_actdurgreater30min	-0.421	0.0368	-11.45	0
asc_wrk_worklocationmatch	6.35	0.0709	89.65	0
asc_wrl_complexsubtour	1.46	0.1	14.57	0
asc_wrl_ftworkeractdurless120min	-0.953	0.0609	-15.64	0
asc_wrl_ptworkeractdurless120min	-0.642	0.105	-6.1	0
child_dis_discretionary_starttime3pmto7pm	1.06	0.0823	12.91	0
child_dis_eating_out_starttime5pmto7pm	0.401	0.0792	5.06	0
child_dis_maintenance_starttime2pmto6pm	0.156	0.0585	2.67	0.01
child_dis_visiting_starttime2pmto7pm	0.478	0.0699	6.84	0
cra_all	1.54	0.0787	19.57	0
ctm_child	2.43	0.0928	26.2	0
ctm_drvchild	1.36	0.163	8.35	0
ctm_ftw	2.57	0.0725	35.4	0
ctm_nonw	1.21	0.0825	14.67	0
ctm_presc	0.825	0.176	4.7	0
ctm_ptw	1.61	0.0947	16.98	0
ctm_retr	0.803	0.112	7.18	0
ctm_ustu	2.39	0.148	16.17	0
dis_discretionary_highincome	0.219	0.0231	9.45	0
dis_discretionary_lu_commercial	0.482	0.0265	18.22	0
dis_discretionary_zerocars	-0.572	0.1	-5.72	0
dis_eating_out_highincome	0.137	0.0231	5.95	0
dis_eating_out_lu_commercial	1.38	0.0455	30.37	0
dis_eating_out_zerocars	-0.377	0.0948	-3.97	0
dis_escorting_female	0.176	0.032	5.49	0
dis_escorting_highschoolenrollment	-0.442	0.114	-3.87	0
dis_escorting_k8enrollment	-0.49	0.126	-3.89	0
dis_escorting_nondrivingchildren	0.549	0.0363	15.14	0
dis_maintenance_female	0.128	0.0186	6.88	0
dis_maintenance_lu_commercial	0.959	0.0301	31.84	0
dis_maintenance_lu_institutional	0.431	0.0266	16.24	0
dis_shopping_female	0.164	0.0197	8.31	0
dis_shopping_lu_commercial	1.49	0.0443	33.56	0
drp_child	1.21	0.175	6.91	0
drp_drvchild	1.08	0.165	6.56	0
drp_ftw	1.7	0.11	15.4	0
drp_nonw	1.24	0.116	10.62	0
drp_presc	1.21	0.135	8.92	0

(continued on next page)

Beta	Value	Std Error	t-test	p-value
drp_ptw	1.39	0.12	11.62	0
drp_retr	1.37	0.117	11.73	0
drp_ustu	1.26	0.167	7.56	0
drvchild_dis_discretionary_starttime3pmto7pm	0.661	0.155	4.27	0
drvchild_dis_eating_out_starttime5pmto7pm	0.24	0.149	1.62	0.11
emo_child	1.65	0.139	11.88	0
emo_drvchild	1.44	0.153	9.4	0
emo_ftw	1.85	0.123	14.95	0
emo_nonw	1.27	0.132	9.57	0
emo_presc	0.91	0.17	5.35	0
emo_ptw	1.55	0.131	11.81	0
emo_retr	1.45	0.131	11.09	0
emo_ustu	1.39	0.17	8.19	0
ftw_dis_discretionary_starttime5pmto7pm	0.812	0.0524	15.48	0
ftw_dis_eating_out_starttime11amto1pm	0.408	0.0473	8.63	0
ftw_dis_eating_out_starttime5pmto7pm	0.532	0.0462	11.49	0
ftw_dis_maintenance_starttime3pmto6pm	0.105	0.034	3.1	0
ftw_dis_shopping_starttime3pmto7pm	0.39	0.0362	10.77	0
ftw_dis_visiting_starttime3pmto8pm	0.487	0.0502	9.69	0
mnt_child	3.62	0.084	43.07	0
mnt_drvchild	3.32	0.0946	35.13	0
mnt_ftw	3.83	0.0705	54.24	0
mnt_nonw	3.34	0.0727	45.99	0
mnt_presc	3.3	0.0934	35.35	0
mnt_ptw	3.57	0.0757	47.19	0
mnt_retr	3.48	0.0752	46.22	0
mnt_ustu	3.37	0.11	30.68	0
nonw_dis_discretionary_startttime8amto11am	0.42	0.0687	6.11	0
nonw_dis_eating_out_starttime11amto1pm	0.452	0.0616	7.34	0
nonw_dis_eating_out_starttime5pmto7pm	0.541	0.0692	7.82	0
nonw_dis_maintenance_startttime8amto2pm	0.412	0.0376	10.97	0
nonw_dis_shopping_startttime9amto3pm	0.205	0.0393	5.21	0
presc_dis_eating_out_starttime11amto1pm	0.758	0.149	5.1	0
presc_dis_eating_out_starttime5pmto7pm	0.958	0.137	6.98	0
presc_dis_maintenance_starttime8amto2pm	0.259	0.0784	3.3	0
ptw_dis_discretionary_starttime5pmto7pm	0.507	0.0879	5.76	0
ptw_dis_eating_out_starttime11amto1pm	0.422	0.0697	6.06	0
ptw_dis_eating_out_starttime5pmto7pm	0.314	0.0765	4.11	0
ptw_dis_shopping_starttime11amto3pm	0.0984	0.0466	2.11	0.03
ptw_dis_visiting_starttime4pmto7pm	0.143	0.0782	1.83	0.07
rec_child	2.76	0.104	26.62	0
rec_drvchild	2.37	0.15	15.77	0
rec_ftw	2.92	0.0822	35.46	0
rec_nonw	2.5	0.0845	29.6	0
rec_presc	2.82	0.0948	29.75	0
rec_ptw	2.67	0.0919	29.03	0
rec_retr	2.51	0.096	26.18	0

Beta	Value	Std Error	t-test	p-value
rec_ustu	2.4	0.159	15.08	0
retr_dis_discretionary_starttime8amto12pm	0.274	0.0858	3.19	0
retr_dis_eating_out_starttime11amto1pm	0.413	0.0654	6.32	0
retr_dis_eating_out_starttime5pmto7pm	0.478	0.0826	5.78	0
retr_dis_maintenance_starttime8amto2pm	0.348	0.0447	7.77	0
retr_dis_shopping_starttime9amto3pm	0.217	0.0475	4.57	0
sch_ftw	-1.16	0.189	-6.14	0
sch_nonw	-1.38	0.188	-7.33	0
sch_ptw	-0.403	0.177	-2.28	0.02
sch_retr	-3.02	0.593	-5.09	0
sch_ustu	0.38	0.287	1.32	0.19
shp_child	1.04	0.19	5.47	0
shp_drvchild	0.74	0.199	3.72	0
shp_ftw	1.12	0.184	6.11	0
shp_nonw	0.885	0.186	4.76	0
shp_presc	0.961	0.189	5.09	0
shp_ptw	0.985	0.187	5.26	0
shp_retr	1	0.187	5.38	0
shp_ustu	0.905	0.204	4.44	0
ustu_dis_eating_out_starttime5pmto7pm	0.244	0.197	1.24	0.21
wrk_ftw	1.56	0.0808	19.36	0
wrk_ptw	1.05	0.104	10.1	0
wrl_ftw	4.02	0.0794	50.65	0
wrl_ptw	3.19	0.108	29.47	0

Nesting Coefficients

Beta	Value	Std Error	t-test	p-value
agg_non_work	1.88	0.0525	35.85	0.00
dis_escorting	1.18	0.0467	25.24	0.00
dis_work	1.00	1.8e+308	0.00	1.00

Utility Equations

ID	Name	Specification
4	change_travel_mode_transfer	asc_ctm_iswalkorwheelchair * iswalkorwheelchair + asc_ctm_transfervariablebothnonauto * transfervariablebothnonauto + asc_ctm_transfervariableatleastonenonauto * transfervariableatleastonenonauto + ctm_ftw * finalftworker + ctm_ptw * finalptworker + ctm_ustu * finalunivstud + ctm_nonw * finalnonworker + ctm_retr * finalretiree + ctm_drvchild * finaldrivingagechild + ctm_child * finalpredriving + ctm_presc * finalpreschool
5	dropped_off_passenger_from_car	dis_escorting_female * female + dis_escorting_nondrivingchildren * nondrivingchildren + dis_escorting_k8enrollment * k8enrollment + dis_escorting_highschoolenrollment * highschoolenrollment + asc_drp_dropoffvariable * dropoffvariable + asc_drp_someonedropped * someonedropped + asc_drp_actdurless10min * actdurless10min + drp_ftw * finalftworker + drp_ptw * finalptworker + drp_ustu * finalunivstud + drp_nonw * finalnonworker + drp_retr * finalretiree + drp_drvchild * finaldrivingagechild + drp_child * finalpredriving + drp_presc * finalpreschool
6	picked_up_passenger_from_car	dis_escorting_female * female + dis_escorting_nondrivingchildren * nondrivingchildren + dis_escorting_k8enrollment * k8enrollment + dis_escorting_highschoolenrollment * highschoolenrollment + asc_pkp_pickupvariable * pickupvariable + asc_pkp_someonepicked * someonepicked + asc_pkp_actdurless10min * actdurless10min + drp_ftw * finalftworker + drp_ptw * finalptworker + drp_ustu * finalunivstud + drp_nonw * finalnonworker + drp_retr * finalretiree + drp_drvchild * finaldrivingagechild + drp_child * finalpredriving + drp_presc * finalpreschool
7	maintenance	ftw_dis_maintenance_starttime3pmto6pm * ftw_starttime3pmto6pm + ptw_dis_maintenance_startttime11amto6pm * ptw_startttime11amto6pm + nonw_dis_maintenance_startttime8amto2pm * nonw_startttime8amto2pm + retr_dis_maintenance_starttime8amto2pm * retr_starttime8amto2pm + drvchild_dis_maintenance_starttime2pmto6pm * drvchild_starttime2pmto6pm + child_dis_maintenance_starttime2pmto6pm * child_starttime2pmto6pm + presc_dis_maintenance_starttime8amto2pm * presc_starttime8amto2pm + dis_maintenance_female * female + dis_maintenance_lu_commercial * lu_commercial + dis_maintenance_lu_institutional * lu_institutional + asc_pbs_actdurgreater120min * actdurgreater120min + asc_pbs_actdurless30min * actdurless30min + asc_pbs_complexsubtour * complexsubtourstop + asc_pbs_simplesubtour * simplesubtourstop + asc_dth_actdurgreater10min * actdurgreater10min + asc_dth_tripdistgreater10mi * tripdistgreater10mi + asc_dth_actdurless30min * actdurless30min + asc_srv_actdurgreater30min * actdurgreater30min + asc_srv_actdurless30min * actdurless30min + asc_her_actdurgreater10min * actdurgreater10min + asc_her_tripdistgreater10mi * tripdistgreater10mi + asc_her_actdurgreater45min * actdurgreater45min + asc_her_actdurless30min * actdurless30min + asc_hcr_actdur30to90min * actdur30to90min + mnt_ftw * finalftworker + mnt_ptw * finalptworker + mnt_ustu * finalunivstud + mnt_nonw * finalnonworker + mnt_retr * finalretiree + mnt_drvchild * finaldrivingagechild + mnt_child * finalpredriving + mnt_presc * finalpreschool
8	work_doing_my_job	asc_wrk_worklocationmatch * worklocationmatch + wrk_ftw * finalftworker + wrk_ptw * finalptworker + wrk_ustu * finalunivstud + wrk_nonw * finalnonworker

ID	Name	Specification
		+ wrk_retr * finalretiree + wrk_drvchild * finaldrivingagechild + wrk_child * finalpredriving + wrk_presc * finalpreschool
11	school	asc_sch_schoolbusmode * schoolbusmode + asc_sch_schoollocationmatch * schoollocationmatch + sch_ftw * finalftworker + sch_ptw * finalptworker + sch_ustu * finalunivstud + sch_nonw * finalnonworker + sch_retr * finalretiree + sch_drvchild * finaldrivingagechild + sch_child * finalpredriving + sch_presc * finalpreschool
13	work_related	asc_wrl_simplesubtour * simplesubtourstop + asc_wrl_complexsubtour * complexsubtourstop + asc_wrl_ftworkeractdurless120min * ftworkeractdurless120min + asc_wrl_ptworkeractdurless120min * ptworkeractdurless120min + asc_wrl_univstudactdurless120min * univstudactdurless120min + wrl_ftw * finalftworker + wrl_ptw * finalptworker + wrl_ustu * finalunivstud + wrl_nonw * finalnonworker + wrl_retr * finalretiree + wrl_drvchild * finaldrivingagechild + wrl_child * finalpredriving + wrl_presc * finalpreschool
15	shopping	ftw_dis_shopping_starttime3pmto7pm * ftw_starttime3pmto7pm + ptw_dis_shopping_starttime11amto3pm * ptw_starttime11amto3pm + nonw_dis_shopping_startttime9amto3pm * nonw_startttime9amto3pm + retr_dis_shopping_starttime9amto3pm * retr_starttime9amto3pm + drvchild_dis_shopping_starttime3pmto5pm * drvchild_starttime3pmto5pm + child_dis_shopping_starttime2pmto7pm * child_starttime2pmto7pm + presc_dis_shopping_starttime9amto11am * presc_starttime9amto11am + dis_shopping_female * female + dis_shopping_lu_commercial * lu_commercial + asc_grc_adultparty12pmto2pm * adultparty12pmto2pm + asc_grc_actdurgreater10min * actdurgreater10min + asc_grc_tripdistgreater10mi * tripdistgreater10mi + asc_grc_actdurgreater90min * actdurgreater90min + asc_grc_groupgroceryduration * groupgroceryduration + asc_orc_nonmand * nonmand + asc_orc_nearbigbox * nearbigbox + asc_shp_actdurless60min * actdurless60min + shp_ftw * finalftworker + shp_ptw * finalptworker + shp_ustu * finalunivstud + shp_nonw * finalnonworker + shp_retr * finalretiree + shp_drvchild * finaldrivingagechild + shp_child * finalpredriving + shp_presc * finalpreschool
21	eat_meal_out_at_restaurant_diner	ftw_dis_eating_out_starttime11amto1pm * ftw_starttime11amto1pm + ftw_dis_eating_out_starttime5pmto7pm * ftw_starttime5pmto7pm + ptw_dis_eating_out_starttime11amto1pm * ptw_starttime11amto1pm + ptw_dis_eating_out_starttime5pmto7pm * ptw_starttime5pmto7pm + ustu_dis_eating_out_starttime11amto1pm * ustu_starttime11amto1pm + ustu_dis_eating_out_starttime5pmto7pm * ustu_starttime5pmto7pm + nonw_dis_eating_out_starttime11amto1pm * nonw_starttime11amto1pm + nonw_dis_eating_out_starttime5pmto7pm * nonw_starttime5pmto7pm + retr_dis_eating_out_starttime11amto1pm * retr_starttime11amto1pm + retr_dis_eating_out_starttime5pmto7pm * retr_starttime5pmto7pm + drvchild_dis_eating_out_starttime5pmto7pm * drvchild_starttime5pmto7pm + child_dis_eating_out_starttime11amto1pm * child_starttime11amto1pm + child_dis_eating_out_starttime5pmto7pm * child_starttime5pmto7pm + presc_dis_eating_out_starttime11amto1pm * presc_starttime11amto1pm + presc_dis_eating_out_starttime5pmto7pm * presc_starttime5pmto7pm + dis_eating_out_zerocars * zerocars + dis_eating_out_highincome * highincome + dis_eating_out_lu_commercial * lu_commercial + asc_emo_iswalkorwheelchair *

(continued on next page)

ID	Name	Specification
		iswalkorwheelchair + asc_emo_adultparty12pmto2pm * adultparty12pmto2pm + asc_emo_adultpartyactdur20to40min * adultpartyactdur20to40min + asc_emo_simplesubtour12pmto2pm * simplesubtour12pmto2pm + asc_emo_complexsubtour12pmto2pm * complexsubtour12pmto2pm + asc_emo_simplesubtour * simplesubtourstop + asc_emo_complexsubtour * complexsubtourstop + asc_emo_actdurgreater10min * actdurgreater10min + asc_emo_tripdistgreater10mi * tripdistgreater10mi + asc_emo_groupeatoutduration * groupeatoutduration + asc_emo_actdurgreater150min * actdurgreater150min + asc_emo_actdurless120min * actdurless120min + emo_ftw * finalftworker + emo_ptw * finalptworker + emo_ustu * finalunivstud + emo_nonw * finalnonworker + emo_retr * finalretiree + emo_drvchild * finaldrivingagechild + emo_child * finalpredriving + emo_presc * finalpreschool
22	civic_or_religious_activities	cra_all * one + cra_notretr * notretiredadult + dis_discretionary_zerocars * zerocars + dis_discretionary_highincome * highincome + dis_discretionary_lu_commercial * lu_commercial + asc_cra_nearchurch * nearchurch + asc_cra_lu_institutional * lu_institutional
23	entertainment	ftw_dis_discretionary_starttime5pmto7pm * ftw_starttime5pmto7pm + ptw_dis_discretionary_starttime5pmto7pm * ptw_starttime5pmto7pm + nonw_dis_discretionary_startttime8amto11am * nonw_startttime8amto11am + retr_dis_discretionary_starttime8amto12pm * retr_starttime8amto12pm + drvchild_dis_discretionary_starttime3pmto7pm * drvchild_starttime3pmto7pm + child_dis_discretionary_starttime3pmto7pm * child_starttime3pmto7pm + dis_discretionary_zerocars * zerocars + dis_discretionary_highincome * highincome + dis_discretionary_lu_commercial * lu_commercial + asc_lu_parks * lu_parks + asc_rec_iswalkorwheelchair * iswalkorwheelchair + asc_rec_grouprecreationduration * grouprecreationduration + rec_ftw * finalftworker + rec_ptw * finalptworker + rec_ustu * finalunivstud + rec_nonw * finalnonworker + rec_retr * finalretiree + rec_drvchild * finaldrivingagechild + rec_child * finalpredriving + rec_presc * finalpreschool
25	social_visit_friends_relatives	asc_soc * one + ftw_dis_visiting_starttime3pmto8pm * ftw_starttime3pmto8pm + ptw_dis_visiting_starttime4pmto7pm * ptw_starttime4pmto7pm + nonw_dis_visiting_starttime10amto1pm * nonw_starttime10amto1pm + retr_dis_visiting_starttime2pmto5pm * retr_starttime2pmto5pm + drvchild_dis_visiting_starttime2pmto7pm * drvchild_starttime2pmto7pm + child_dis_visiting_starttime2pmto7pm * child_starttime2pmto7pm + presc_dis_visiting_starttime2pmto7pm * presc_starttime2pmto7pm + dis_visiting_highincome * highincome + asc_soc_bikemode * bikemode + asc_soc_groupsocialvisitduration * groupsocialvisitduration + soc_ftw * finalftworker + soc_ptw * finalptworker + soc_ustu * finalunivstud + soc_nonw * finalnonworker + soc_retr * finalretiree + soc_drvchild * finaldrivingagechild + soc_child * finalpredriving + soc_presc * finalpreschool

Trip Purpose Decision Tree

```
J48 pruned tree
------------------

schoollocationmatch = 0
|  worklocationmatch = 0
|  |  dropoffvariable = 0
|  |  |  someonepicked = 0
|  |  |  |  nonauto = 0
|  |  |  |  |  someonedropped = 0
|  |  |  |  |  |  actdurgreater150min = 0
|  |  |  |  |  |  |  lu_commercial = 0
|  |  |  |  |  |  |  |  lu_parks = 0
|  |  |  |  |  |  |  |  |  actdurless30min = 0
|  |  |  |  |  |  |  |  |  |  arrhour <= 16
|  |  |  |  |  |  |  |  |  |  |  finalftworker = 0
|  |  |  |  |  |  |  |  |  |  |  |  lu_institutional = 0
|  |  |  |  |  |  |  |  |  |  |  |  |  finalptworker = 0
|  |  |  |  |  |  |  |  |  |  |  |  |  |  actdurgreater30min = 0
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  arrhour <= 12: 7 (34.0/22.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  arrhour > 12: 15 (43.0/27.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  actdurgreater30min = 1
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  k8enrollment = 0
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  finalpreschool = 0
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  starttime11amto3pm = 0
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  starttime4pmto7pm = 0
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  groupeatoutduration = 0
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  tripdistgreater10mi = 0
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  actdurless60min = 0: 23 (149.0/96.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  actdurless60min = 1
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  tripdistance <= 2.736193: 23 (30.0/20.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  tripdistance > 2.736193: 7 (32.0/22.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  tripdistgreater10mi = 1: 7 (56.0/36.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  groupeatoutduration = 1: 7 (26.0/16.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  starttime4pmto7pm = 1: 25 (39.0/25.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  starttime11amto3pm = 1
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  arrhour <= 12
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  age <= 51: 25 (40.0/23.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  age > 51: 21 (89.0/57.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  arrhour > 12
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  adultpartyactdur20to40min = 0
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  hhmem <= 0: 25 (115.0/68.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  hhmem > 0: 7 (60.0/32.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  adultpartyactdur20to40min = 1: 7 (28.0/16.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  finalpreschool = 1: 25 (58.0/30.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  k8enrollment = 1
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  arrhour <= 14: 7 (37.0/21.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  arrhour > 14: 23 (44.0/16.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  finalptworker = 1
|  |  |  |  |  |  |  |  |  |  |  |  |  |  adultparty = 0: 23 (26.0/19.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  adultparty = 1
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  young = 0: 13 (201.0/125.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  young = 1: 25 (51.0/38.0)
|  |  |  |  |  |  |  |  |  |  |  |  lu_institutional = 1
|  |  |  |  |  |  |  |  |  |  |  |  |  groupgroceryduration = 0
|  |  |  |  |  |  |  |  |  |  |  |  |  |  arrhour <= 15: 7 (333.0/149.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  arrhour > 15: 23 (36.0/15.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  groupgroceryduration = 1: 7 (30.0/4.0)
|  |  |  |  |  |  |  |  |  |  |  finalftworker = 1
|  |  |  |  |  |  |  |  |  |  |  |  starttime4pmto7pm = 0
|  |  |  |  |  |  |  |  |  |  |  |  |  mixedparty = 0
|  |  |  |  |  |  |  |  |  |  |  |  |  |  simplesubtour12pmto2pm = 0
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  hhmem <= 0
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  lu_institutional = 0: 13 (463.0/207.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  lu_institutional = 1
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  tripdistgreater10mi = 0
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  female = 0: 7 (52.0/33.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  female = 1: 13 (47.0/20.0)
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  tripdistgreater10mi = 1: 7 (31.0/19.0)
```

(continued on next page)

```
| | | | | | | | | | | | | | | hhmem > 0
| | | | | | | | | | | | | | | | starttime11amto3pm = 0: 23 (34.0/23.0)
| | | | | | | | | | | | | | | | starttime11amto3pm = 1: 21 (26.0/16.0)
| | | | | | | | | | | | | | simplesubtour12pmto2pm = 1: 21 (30.0/16.0)
| | | | | | | | | | | | | mixedparty = 1: 7 (32.0/18.0)
| | | | | | | | | | | | starttime4pmto7pm = 1
| | | | | | | | | | | | | actdurless60min = 0: 23 (26.0/15.0)
| | | | | | | | | | | | | actdurless60min = 1: 7 (28.0/18.0)
| | | | | | | | | | arrhour > 16
| | | | | | | | | | | lu_institutional = 0
| | | | | | | | | | | | groupgroceryduration = 0
| | | | | | | | | | | | | groupeatoutduration = 0
| | | | | | | | | | | | | | hhmem <= 2
| | | | | | | | | | | | | | | k8enrollment = 0
| | | | | | | | | | | | | | | | ptype = 1: 25 (27.0/12.0)
| | | | | | | | | | | | | | | | ptype = 2: 25 (0.0)
| | | | | | | | | | | | | | | | ptype = 3: 25 (0.0)
| | | | | | | | | | | | | | | | ptype = 7
| | | | | | | | | | | | | | | | | adultpartyactdur20to40min = 0
| | | | | | | | | | | | | | | | | | tottr <= 2
| | | | | | | | | | | | | | | | | | | starttime3pmto8pm = 0: 25 (41.0/17.0)
| | | | | | | | | | | | | | | | | | | starttime3pmto8pm = 1
| | | | | | | | | | | | | | | | | | | | starttime2pmto5pm = 0
| | | | | | | | | | | | | | | | | | | | | lowincome = 0
| | | | | | | | | | | | | | | | | | | | | | actdurless60min = 0
| | | | | | | | | | | | | | | | | | | | | | | mixedparty = 0
| | | | | | | | | | | | | | | | | | | | | | | | hhmem <= 0
| | | | | | | | | | | | | | | | | | | | | | | | | adultactdurless100min = 0: 25 (87.0/54.0)
| | | | | | | | | | | | | | | | | | | | | | | | | adultactdurless100min = 1: 23 (81.0/50.0)
| | | | | | | | | | | | | | | | | | | | | | | | hhmem > 0: 22 (49.0/36.0)
| | | | | | | | | | | | | | | | | | | | | | | mixedparty = 1: 23 (35.0/16.0)
| | | | | | | | | | | | | | | | | | | | | | actdurless60min = 1: 25 (26.0/19.0)
| | | | | | | | | | | | | | | | | | | | | lowincome = 1: 25 (42.0/24.0)
| | | | | | | | | | | | | | | | | | | | starttime2pmto5pm = 1: 25 (48.0/32.0)
| | | | | | | | | | | | | | | | | | tottr > 2
| | | | | | | | | | | | | | | | | | | arrhour <= 17: 23 (35.0/15.0)
| | | | | | | | | | | | | | | | | | | arrhour > 17
| | | | | | | | | | | | | | | | | | | | actdur <= 82: 23 (28.0/16.0)
| | | | | | | | | | | | | | | | | | | | actdur > 82: 21 (32.0/19.0)
| | | | | | | | | | | | | | | | | adultpartyactdur20to40min = 1: 15 (56.0/37.0)
| | | | | | | | | | | | | | | k8enrollment = 1: 23 (71.0/26.0)
| | | | | | | | | | | | | | hhmem > 2: 23 (59.0/29.0)
| | | | | | | | | | | | | groupeatoutduration = 1
| | | | | | | | | | | | | | lowincome = 0
| | | | | | | | | | | | | | | nondrivingchildren = 0: 21 (41.0/25.0)
| | | | | | | | | | | | | | | nondrivingchildren = 1: 23 (41.0/21.0)
| | | | | | | | | | | | | | lowincome = 1: 21 (26.0/13.0)
| | | | | | | | | | | | groupgroceryduration = 1
| | | | | | | | | | | | | arrhour <= 18: 15 (39.0/16.0)
| | | | | | | | | | | | | arrhour > 18: 21 (38.0/19.0)
| | | | | | | | | | | lu_institutional = 1
| | | | | | | | | | | | groupeatoutduration = 0
| | | | | | | | | | | | | arrhour <= 17
| | | | | | | | | | | | | | tottr <= 1: 22 (37.0/25.0)
| | | | | | | | | | | | | | tottr > 1: 23 (47.0/20.0)
| | | | | | | | | | | | | arrhour > 17: 22 (235.0/106.0)
| | | | | | | | | | | | groupeatoutduration = 1: 23 (29.0/11.0)
| | | | | | | | | actdurless30min = 1
| | | | | | | | | | arrhour <= 7
| | | | | | | | | | | tottr <= 1
| | | | | | | | | | | | finalftworker = 0: 7 (25.0/18.0)
| | | | | | | | | | | | finalftworker = 1: 4 (104.0/51.0)
| | | | | | | | | | | tottr > 1
| | | | | | | | | | | | highincome = 0: 6 (47.0/26.0)
| | | | | | | | | | | | highincome = 1: 5 (25.0/12.0)
| | | | | | | | | | arrhour > 7
| | | | | | | | | | | subtourdummy = 0
| | | | | | | | | | | | hhmem <= 1
| | | | | | | | | | | | | mixedparty = 0
| | | | | | | | | | | | | | finalftworker = 0
| | | | | | | | | | | | | | | lu_institutional = 0
| | | | | | | | | | | | | | | | actdur <= 6: 7 (240.0/106.0)
| | | | | | | | | | | | | | | | actdur > 6
| | | | | | | | | | | | | | | | | age <= 29: 7 (45.0/33.0)
| | | | | | | | | | | | | | | | | age > 29
| | | | | | | | | | | | | | | | | | finalptworker = 0
| | | | | | | | | | | | | | | | | | | volunactdurless60min = 0
| | | | | | | | | | | | | | | | | | | | age <= 71
| | | | | | | | | | | | | | | | | | | | | lowincome = 0: 15 (154.0/92.0)
| | | | | | | | | | | | | | | | | | | | | lowincome = 1: 7 (70.0/37.0)
| | | | | | | | | | | | | | | | | | | | age > 71: 7 (43.0/25.0)
| | | | | | | | | | | | | | | | | | | volunactdurless60min = 1
| | | | | | | | | | | | | | | | | | | | adultparty12pmto2pm = 0: 25 (30.0/22.0)
| | | | | | | | | | | | | | | | | | | | adultparty12pmto2pm = 1: 15 (27.0/17.0)
| | | | | | | | | | | | | | | | | | finalptworker = 1
| | | | | | | | | | | | | | | | | | | actdur <= 13: 7 (61.0/39.0)
| | | | | | | | | | | | | | | | | | | actdur > 13: 15 (54.0/36.0)
| | | | | | | | | | | | | | | lu_institutional = 1: 7 (166.0/78.0)
| | | | | | | | | | | | | | finalftworker = 1
| | | | | | | | | | | | | | | tottr <= 2
| | | | | | | | | | | | | | | | actdurgreater10min = 0: 7 (319.0/176.0)
| | | | | | | | | | | | | | | | actdurgreater10min = 1
| | | | | | | | | | | | | | | | | starttime1amto6pm = 0: 15 (50.0/30.0)
| | | | | | | | | | | | | | | | | starttime1amto6pm = 1
| | | | | | | | | | | | | | | | | | starttime5pmto7pm = 0
| | | | | | | | | | | | | | | | | | | starttime4pmto7pm = 0: 13 (166.0/90.0)
| | | | | | | | | | | | | | | | | | | starttime4pmto7pm = 1: 7 (37.0/23.0)
| | | | | | | | | | | | | | | | | | starttime5pmto7pm = 1: 15 (43.0/29.0)
| | | | | | | | | | | | | | | tottr > 2: 13 (36.0/8.0)
| | | | | | | | | | | | | mixedparty = 1
| | | | | | | | | | | | | | actdur <= 5
| | | | | | | | | | | | | | | starttime2pmto6pm = 0: 6 (36.0/24.0)
| | | | | | | | | | | | | | | starttime2pmto6pm = 1: 7 (40.0/19.0)
| | | | | | | | | | | | | | actdur > 5
| | | | | | | | | | | | | | | lu_institutional = 0
| | | | | | | | | | | | | | | | groupgroceryduration = 0
| | | | | | | | | | | | | | | | | starttime2pmto6pm = 0: 7 (31.0/13.0)
| | | | | | | | | | | | | | | | | starttime2pmto6pm = 1: 15 (34.0/17.0)
| | | | | | | | | | | | | | | | groupgroceryduration = 1: 15 (45.0/15.0)
| | | | | | | | | | | | | | | lu_institutional = 1: 6 (37.0/24.0)
| | | | | | | | | | | | hhmem > 1
| | | | | | | | | | | | | actdur <= 24
| | | | | | | | | | | | | | tripdistance <= 8.796565
| | | | | | | | | | | | | | | starttime8amto2pm = 0
| | | | | | | | | | | | | | | | starttime3pmto8pm = 0: 5 (33.0/16.0)
| | | | | | | | | | | | | | | | starttime3pmto8pm = 1: 7 (95.0/40.0)
| | | | | | | | | | | | | | | starttime8amto2pm = 1: 7 (33.0/13.0)
| | | | | | | | | | | | | | tripdistance > 8.796565: 25 (34.0/21.0)
| | | | | | | | | | | | | actdur > 24: 7 (25.0/16.0)
| | | | | | | | | | | subtourdummy = 1
| | | | | | | | | | | | starttime8amto11am = 0
| | | | | | | | | | | | | actdurgreater10min = 0: 7 (52.0/32.0)
| | | | | | | | | | | | | actdurgreater10min = 1: 13 (49.0/25.0)
| | | | | | | | | | | | starttime8amto11am = 1: 13 (28.0/13.0)
| | | | | | | | lu_parks = 1
| | | | | | | | | actdurless30min = 0: 23 (245.0/26.0)
| | | | | | | | | actdurless30min = 1: 7 (32.0/12.0)
| | | | | | | lu_commercial = 1
| | | | | | | | actdurgreater10min = 0
| | | | | | | | | simplesubtour = 0
| | | | | | | | | | actdur <= 5: 7 (2318.0/786.0)
| | | | | | | | | | actdur > 5
| | | | | | | | | | | mixedparty = 0
| | | | | | | | | | | | starttime1amto6pm = 0
| | | | | | | | | | | | | tottr <= 1
| | | | | | | | | | | | | | female = 0: 15 (74.0/33.0)
| | | | | | | | | | | | | | female = 1
| | | | | | | | | | | | | | | actdurless10min = 0: 15 (38.0/21.0)
| | | | | | | | | | | | | | | actdurless10min = 1: 7 (57.0/21.0)
| | | | | | | | | | | | | tottr > 1: 15 (87.0/39.0)
```

(continued on next page)

```
| | | | | | | | | | | | starttime1amto6pm = 1
| | | | | | | | | | | | | finalretiree = 0
| | | | | | | | | | | | | | starttime4pmto7pm = 0: 7 (1047.0/486.0)
| | | | | | | | | | | | | | starttime4pmto7pm = 1
| | | | | | | | | | | | | | | actdurless10min = 0: 15 (120.0/52.0)
| | | | | | | | | | | | | | | actdurless10min = 1
| | | | | | | | | | | | | | | | highincome = 0: 7 (78.0/33.0)
| | | | | | | | | | | | | | | | highincome = 1: 15 (39.0/21.0)
| | | | | | | | | | | | | finalretiree = 1
| | | | | | | | | | | | | | actdur <= 7: 7 (71.0/20.0)
| | | | | | | | | | | | | | actdur > 7
| | | | | | | | | | | | | | | highincome = 0
| | | | | | | | | | | | | | | | lowincome = 0
| | | | | | | | | | | | | | | | | starttime3pmto5pm = 0: 7 (90.0/33.0)
| | | | | | | | | | | | | | | | | starttime3pmto5pm = 1: 15 (31.0/15.0)
| | | | | | | | | | | | | | | | lowincome = 1: 15 (59.0/31.0)
| | | | | | | | | | | | | | | highincome = 1: 15 (34.0/10.0)
| | | | | | | | | | | mixedparty = 1
| | | | | | | | | | | | arrhour <= 12: 7 (112.0/38.0)
| | | | | | | | | | | | arrhour > 12
| | | | | | | | | | | | | tripdistance <= 1.317402
| | | | | | | | | | | | | | starttime11amto3pm = 0: 7 (82.0/40.0)
| | | | | | | | | | | | | | starttime11amto3pm = 1: 15 (25.0/8.0)
| | | | | | | | | | | | | tripdistance > 1.317402: 7 (179.0/79.0)
| | | | | | | | | simplesubtour = 1
| | | | | | | | | | female = 0
| | | | | | | | | | | actdur <= 5: 7 (29.0/15.0)
| | | | | | | | | | | actdur > 5: 21 (32.0/17.0)
| | | | | | | | | | female = 1: 7 (56.0/18.0)
| | | | | | | | actdurgreater10min = 1
| | | | | | | | | simplesubtour = 0
| | | | | | | | | | actdurless60min = 0
| | | | | | | | | | | arrhour <= 16
| | | | | | | | | | | | starttime7amto12am = 0: 23 (102.0/40.0)
| | | | | | | | | | | | starttime7amto12am = 1
| | | | | | | | | | | | | finalftworker = 0
| | | | | | | | | | | | | | starttime4pmto7pm = 0
| | | | | | | | | | | | | | | nearbigbox = 0
| | | | | | | | | | | | | | | | starttime11amto1pm = 0
| | | | | | | | | | | | | | | | | finalptworker = 0
| | | | | | | | | | | | | | | | | | childparty = 0
| | | | | | | | | | | | | | | | | | | highschoolenrollment = 0
| | | | | | | | | | | | | | | | | | | | actdur <= 115
| | | | | | | | | | | | | | | | | | | | | lu_institutional = 0
| | | | | | | | | | | | | | | | | | | | | | finalpredriving = 0
| | | | | | | | | | | | | | | | | | | | | | | tripdistgreater10mi = 0
| | | | | | | | | | | | | | | | | | | | | | | | arrhour <= 8
| | | | | | | | | | | | | | | | | | | | | | | | | tripdistance <= 3.341837
| | | | | | | | | | | | | | | | | | | | | | | | | | tripdistance <= 1.760107: 15 (25.0/17.0)
| | | | | | | | | | | | | | | | | | | | | | | | | | tripdistance > 1.760107: 23 (27.0/15.0)
| | | | | | | | | | | | | | | | | | | | | | | | | tripdistance > 3.341837: 7 (26.0/10.0)
| | | | | | | | | | | | | | | | | | | | | | | | arrhour > 8
| | | | | | | | | | | | | | | | | | | | | | | | | mixedparty = 0
| | | | | | | | | | | | | | | | | | | | | | | | | | tottr <= 1
| | | | | | | | | | | | | | | | | | | | | | | | | | | highincome = 0: 15 (185.0/92.0)
| | | | | | | | | | | | | | | | | | | | | | | | | | | highincome = 1: 7 (70.0/40.0)
| | | | | | | | | | | | | | | | | | | | | | | | | | tottr > 1
| | | | | | | | | | | | | | | | | | | | | | | | | | | arrhour <= 10: 7 (65.0/37.0)
| | | | | | | | | | | | | | | | | | | | | | | | | | | arrhour > 10: 15 (86.0/45.0)
| | | | | | | | | | | | | | | | | | | | | | | | | mixedparty = 1
| | | | | | | | | | | | | | | | | | | | | | | | | | tripdistance <= 1.752865: 15 (45.0/16.0)
| | | | | | | | | | | | | | | | | | | | | | | | | | tripdistance > 1.752865
| | | | | | | | | | | | | | | | | | | | | | | | | | | actdur <= 69: 15 (30.0/12.0)
| | | | | | | | | | | | | | | | | | | | | | | | | | | actdur > 69
| | | | | | | | | | | | | | | | | | | | | | | | | | | | hhmem <= 1: 7 (25.0/15.0)
| | | | | | | | | | | | | | | | | | | | | | | | | | | | hhmem > 1: 23 (25.0/12.0)
| | | | | | | | | | | | | | | | | | | | | | | tripdistgreater10mi = 1: 7 (100.0/48.0)
| | | | | | | | | | | | | | | | | | | | | | finalpredriving = 1: 7 (47.0/23.0)
| | | | | | | | | | | | | | | | | | | | | lu_institutional = 1: 7 (40.0/15.0)
| | | | | | | | | | | | | | | | | | | | actdur > 115
| | | | | | | | | | | | | | | | | | | | | starttime11amto3pm = 0: 7 (117.0/64.0)
| | | | | | | | | | | | | | | | | | | | | starttime11amto3pm = 1: 15 (26.0/10.0)
| | | | | | | | | | | | | | | | | | | highschoolenrollment = 1: 7 (33.0/11.0)
| | | | | | | | | | | | | | | | | | childparty = 1: 23 (32.0/18.0)
| | | | | | | | | | | | | | | | | finalptworker = 1
| | | | | | | | | | | | | | | | | | tottr <= 1
| | | | | | | | | | | | | | | | | | | adultactdurless100min = 0: 23 (42.0/30.0)
| | | | | | | | | | | | | | | | | | | adultactdurless100min = 1
| | | | | | | | | | | | | | | | | | | | highincome = 0
| | | | | | | | | | | | | | | | | | | | | starttime9amto11am = 0: 7 (52.0/32.0)
| | | | | | | | | | | | | | | | | | | | | starttime9amto11am = 1: 23 (26.0/15.0)
| | | | | | | | | | | | | | | | | | | | highincome = 1: 15 (31.0/19.0)
| | | | | | | | | | | | | | | | | | tottr > 1
| | | | | | | | | | | | | | | | | | | arrhour <= 12: 7 (25.0/12.0)
| | | | | | | | | | | | | | | | | | | arrhour > 12: 15 (34.0/19.0)
| | | | | | | | | | | | | | | | starttime11amto1pm = 1
| | | | | | | | | | | | | | | | | grouprecreationduration = 0
| | | | | | | | | | | | | | | | | | hhmem <= 1
| | | | | | | | | | | | | | | | | | | actdurless120min = 0: 7 (47.0/32.0)
| | | | | | | | | | | | | | | | | | | actdurless120min = 1
| | | | | | | | | | | | | | | | | | | | starttime8amto11am = 0
| | | | | | | | | | | | | | | | | | | | | arrhour <= 12
| | | | | | | | | | | | | | | | | | | | | | finalptworker = 0
| | | | | | | | | | | | | | | | | | | | | | | hhmem <= 0: 15 (164.0/100.0)
| | | | | | | | | | | | | | | | | | | | | | | hhmem > 0
| | | | | | | | | | | | | | | | | | | | | | | | adultparty12pmto2pm = 0: 21 (47.0/18.0)
| | | | | | | | | | | | | | | | | | | | | | | | adultparty12pmto2pm = 1: 15 (34.0/17.0)
| | | | | | | | | | | | | | | | | | | | | | finalptworker = 1: 21 (58.0/34.0)
| | | | | | | | | | | | | | | | | | | | | arrhour > 12
| | | | | | | | | | | | | | | | | | | | | | tottr <= 1: 15 (43.0/20.0)
| | | | | | | | | | | | | | | | | | | | | | tottr > 1: 7 (29.0/18.0)
| | | | | | | | | | | | | | | | | | | | starttime8amto11am = 1: 15 (32.0/18.0)
| | | | | | | | | | | | | | | | | | hhmem > 1: 15 (28.0/7.0)
| | | | | | | | | | | | | | | | | grouprecreationduration = 1: 7 (42.0/23.0)
| | | | | | | | | | | | | | | nearbigbox = 1: 15 (55.0/15.0)
| | | | | | | | | | | | | | starttime4pmto7pm = 1
| | | | | | | | | | | | | | | age  <= 60: 23 (71.0/44.0)
| | | | | | | | | | | | | | | age > 60: 15 (26.0/7.0)
| | | | | | | | | | | | | finalftworker = 1
| | | | | | | | | | | | | | hhmem <= 0
| | | | | | | | | | | | | | | tripdistgreater10mi = 0
| | | | | | | | | | | | | | | | tottr <= 1
| | | | | | | | | | | | | | | | | starttime4pmto7pm = 0
| | | | | | | | | | | | | | | | | | female = 0
| | | | | | | | | | | | | | | | | | | nondrivingchildren = 0
| | | | | | | | | | | | | | | | | | | | starttime11amto1pm = 0: 7 (79.0/55.0)
| | | | | | | | | | | | | | | | | | | | starttime11amto1pm = 1: 21 (25.0/18.0)
| | | | | | | | | | | | | | | | | | | nondrivingchildren = 1: 13 (64.0/37.0)
| | | | | | | | | | | | | | | | | | female = 1
| | | | | | | | | | | | | | | | | | | arrhour <= 10: 7 (44.0/30.0)
| | | | | | | | | | | | | | | | | | | arrhour > 10: 15 (82.0/53.0)
| | | | | | | | | | | | | | | | | starttime4pmto7pm = 1: 7 (33.0/23.0)
| | | | | | | | | | | | | | | | tottr > 1: 21 (34.0/19.0)
| | | | | | | | | | | | | | | tripdistgreater10mi = 1: 13 (113.0/54.0)
| | | | | | | | | | | | | | hhmem > 0: 7 (92.0/48.0)
| | | | | | | | | | | arrhour > 16
| | | | | | | | | | | | nearchurch = 0
| | | | | | | | | | | | | lu_institutional = 0
| | | | | | | | | | | | | | starttime3pmto5pm = 0
| | | | | | | | | | | | | | | lowincome = 0
| | | | | | | | | | | | | | | | tottr <= 1
| | | | | | | | | | | | | | | | | actdur <= 68: 15 (48.0/29.0)
| | | | | | | | | | | | | | | | | actdur > 68
| | | | | | | | | | | | | | | | | | young = 0
| | | | | | | | | | | | | | | | | | | arrhour <= 17: 23 (49.0/28.0)
| | | | | | | | | | | | | | | | | | | arrhour > 17: 21 (148.0/100.0)
| | | | | | | | | | | | | | | | | | young = 1: 23 (40.0/23.0)
| | | | | | | | | | | | | | | | tottr > 1
| | | | | | | | | | | | | | | | | starttime3pmto8pm = 0: 21 (54.0/24.0)
| | | | | | | | | | | | | | | | | starttime3pmto8pm = 1
```

(continued on next page)

```
| | | | | | | | | | | | | | | | | | adultparty = 0
| | | | | | | | | | | | | | | | | | | tottr <= 2: 23 (118.0/59.0)
| | | | | | | | | | | | | | | | | | | tottr > 2
| | | | | | | | | | | | | | | | | | | | actdur <= 66
| | | | | | | | | | | | | | | | | | | | | tripdistance <= 3.75533: 21 (32.0/8.0)
| | | | | | | | | | | | | | | | | | | | | tripdistance > 3.75533: 15 (27.0/10.0)
| | | | | | | | | | | | | | | | | | | | actdur > 66
| | | | | | | | | | | | | | | | | | | | | tripdistance <= 3.969448
| | | | | | | | | | | | | | | | | | | | | | tripdistance <= 2.141056: 21 (73.0/22.0)
| | | | | | | | | | | | | | | | | | | | | | tripdistance > 2.141056: 23 (49.0/27.0)
| | | | | | | | | | | | | | | | | | | | | tripdistance > 3.969448
| | | | | | | | | | | | | | | | | | | | | | highincome = 0: 21 (40.0/11.0)
| | | | | | | | | | | | | | | | | | | | | | highincome = 1: 22 (50.0/27.0)
| | | | | | | | | | | | | | | | | | adultparty = 1: 21 (266.0/91.0)
| | | | | | | | | | | | | | | lowincome = 1
| | | | | | | | | | | | | | | | starttime2pmto7pm = 0: 15 (35.0/20.0)
| | | | | | | | | | | | | | | | starttime2pmto7pm = 1
| | | | | | | | | | | | | | | | | actdur <= 84: 15 (47.0/16.0)
| | | | | | | | | | | | | | | | | actdur > 84
| | | | | | | | | | | | | | | | | | actdur <= 99: 23 (25.0/15.0)
| | | | | | | | | | | | | | | | | | actdur > 99: 22 (45.0/32.0)
| | | | | | | | | | | | | | starttime3pmto5pm = 1
| | | | | | | | | | | | | | | tottr <= 1
| | | | | | | | | | | | | | | | highincome = 0: 7 (34.0/21.0)
| | | | | | | | | | | | | | | | highincome = 1: 15 (30.0/20.0)
| | | | | | | | | | | | | | | tottr > 1: 21 (56.0/34.0)
| | | | | | | | | | | | | lu_institutional = 1
| | | | | | | | | | | | | | actdurgreater90min = 0: 23 (29.0/14.0)
| | | | | | | | | | | | | | actdurgreater90min = 1: 22 (25.0/16.0)
| | | | | | | | | | | | nearchurch = 1: 22 (47.0/21.0)
| | | | | | | | | | actdurless60min = 1
| | | | | | | | | | | groupeatoutduration = 0
| | | | | | | | | | | | starttime7amto12am = 0
| | | | | | | | | | | | | actdurgreater30min = 0: 7 (63.0/43.0)
| | | | | | | | | | | | | actdurgreater30min = 1: 23 (26.0/17.0)
| | | | | | | | | | | | starttime7amto12am = 1
| | | | | | | | | | | | | subtourdummy = 0
| | | | | | | | | | | | | | starttime1amto6pm = 0
| | | | | | | | | | | | | | | actdurless30min = 0
| | | | | | | | | | | | | | | | tottr <= 1: 15 (212.0/81.0)
| | | | | | | | | | | | | | | | tottr > 1
| | | | | | | | | | | | | | | | | tripdistance <= 4.168607
| | | | | | | | | | | | | | | | | | highincome = 0
| | | | | | | | | | | | | | | | | | | mixedparty = 0: 21 (44.0/21.0)
| | | | | | | | | | | | | | | | | | | mixedparty = 1: 15 (46.0/6.0)
| | | | | | | | | | | | | | | | | | highincome = 1
| | | | | | | | | | | | | | | | | | | mixedparty = 0: 15 (40.0/16.0)
| | | | | | | | | | | | | | | | | | | mixedparty = 1: 21 (27.0/12.0)
| | | | | | | | | | | | | | | | | tripdistance > 4.168607: 21 (53.0/22.0)
| | | | | | | | | | | | | | | actdurless30min = 1: 15 (614.0/195.0)
| | | | | | | | | | | | | | starttime1amto6pm = 1
| | | | | | | | | | | | | | | arrhour <= 8
| | | | | | | | | | | | | | | | actdurgreater30min  = 0
| | | | | | | | | | | | | | | | | highincome = 0
| | | | | | | | | | | | | | | | | | tripdistance <= 1.202759: 15 (44.0/15.0)
| | | | | | | | | | | | | | | | | | tripdistance > 1.202759: 7 (119.0/68.0)
| | | | | | | | | | | | | | | | | highincome = 1: 7 (69.0/40.0)
| | | | | | | | | | | | | | | | actdurgreater30min  = 1
| | | | | | | | | | | | | | | | | tripdistance <= 4.482676
| | | | | | | | | | | | | | | | | | finalftworker = 0: 15 (50.0/24.0)
| | | | | | | | | | | | | | | | | | finalftworker = 1: 7 (31.0/20.0)
| | | | | | | | | | | | | | | | | tripdistance > 4.482676: 7 (38.0/20.0)
| | | | | | | | | | | | | | | arrhour > 8
| | | | | | | | | | | | | | | | finalftworker = 0
| | | | | | | | | | | | | | | | | nearbigbox = 0
| | | | | | | | | | | | | | | | | | actdur <= 20
| | | | | | | | | | | | | | | | | | | hhmem <= 2
| | | | | | | | | | | | | | | | | | | | highschoolenrollment = 0
| | | | | | | | | | | | | | | | | | | | | finalptworker = 0
| | | | | | | | | | | | | | | | | | | | | | tottr <= 2
| | | | | | | | | | | | | | | | | | | | | | | actdur <= 15
| | | | | | | | | | | | | | | | | | | | | | | | starttim e3pmto5pm = 0
| | | | | | | | | | | | | | | | | | | | | | | | | starttime8amto2pm = 0
| | | | | | | | | | | | | | | | | | | | | | | | | | tottr <= 1: 15 (53.0/22.0)
| | | | | | | | | | | | | | | | | | | | | | | | | | tottr > 1: 7 (43.0/20.0)
| | | | | | | | | | | | | | | | | | | | | | | | | starttime8amto2pm = 1
| | | | | | | | | | | | | | | | | | | | | | | | | | tripdistance <= 0.294596: 15 (42.0/11.0)
| | | | | | | | | | | | | | | | | | | | | | | | | | tripdistance > 0.294596
| | | | | | | | | | | | | | | | | | | | | | | | | | | highincome = 0
| | | | | | | | | | | | | | | | | | | | | | | | | | | | age <= 65
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | starttime10amto1pm = 0: 15 (56.0/25.0)
| | | | | | | | | | | | | | | | | | | | | | | | | | | | | starttime10amto1pm = 1: 7 (96.0/45.0)
| | | | | | | | | | | | | | | | | | | | | | | | | | | | age > 65: 15 (85.0/34.0)
| | | | | | | | | | | | | | | | | | | | | | | | | | | hi ghincome = 1: 15 (57.0/21.0)
| | | | | | | | | | | | | | | | | | | | | | | | starttime3pmto5pm = 1: 15 (87.0/44.0)
| | | | | | | | | | | | | | | | | | | | | | | actdur > 15: 15 (395.0/140.0)
| | | | | | | | | | | | | | | | | | | | | | tottr > 2
| | | | | | | | | | | | | | | | | | | | | | | lowincome = 0
| | | | | | | | | | | | | | | | | | | | | | | | starttime3pmto5pm = 0: 7 (60.0/30.0)
| | | | | | | | | | | | | | | | | | | | | | | | starttime3pmto5pm = 1: 15 (32.0/6.0)
| | | | | | | | | | | | | | | | | | | | | | | lowincome = 1: 7 (42.0/12.0)
| | | | | | | | | | | | | | | | | | | | | finalptworker = 1
| | | | | | | | | | | | | | | | | | | | | | starttime8amto11am = 0: 15 (252.0/109.0)
| | | | | | | | | | | | | | | | | | | | | | starttime8amto11am = 1: 7 (48.0/23.0)
| | | | | | | | | | | | | | | | | | | | highschoolenrollment = 1: 15 (35.0/14.0)
| | | | | | | | | | | | | | | | | | | hhmem > 2: 15 (57.0/19.0)
| | | | | | | | | | | | | | | | | | actdur > 20
| | | | | | | | | | | | | | | | | | | tripdistgreater10mi = 0: 15 (2128.0/797.0)
| | | | | | | | | | | | | | | | | | | tripdistgreater10mi = 1
| | | | | | | | | | | | | | | | | | | | young = 0
| | | | | | | | | | | | | | | | | | | | | tottr <= 1
| | | | | | | | | | | | | | | | | | | | | | fem ale = 0: 7 (36.0/21.0)
| | | | | | | | | | | | | | | | | | | | | | fem ale = 1: 15 (52.0/21.0)
| | | | | | | | | | | | | | | | | | | | | tottr > 1
| | | | | | | | | | | | | | | | | | | | | | tripdistance <= 13.666906: 21 (27.0/12.0)
| | | | | | | | | | | | | | | | | | | | | | tripdistance > 13.666906: 7 (34.0/20.0)
| | | | | | | | | | | | | | | | | | | | young = 1: 15 (31.0/15.0)
| | | | | | | | | | | | | | | | | nearbigbox = 1: 15 (127.0/15.0)
| | | | | | | | | | | | | | | | finalftworker = 1
| | | | | | | | | | | | | | | | | starttime3pmto6pm = 0
| | | | | | | | | | | | | | | | | | groupgroceryduration = 0
| | | | | | | | | | | | | | | | | | | lu_institutional = 0
| | | | | | | | | | | | | | | | | | | | tripdistgreater10mi = 0
| | | | | | | | | | | | | | | | | | | | | starttime11amto1pm = 0
| | | | | | | | | | | | | | | | | | | | | | lowincome = 0
| | | | | | | | | | | | | | | | | | | | | | | highincome = 0
| | | | | | | | | | | | | | | | | | | | | | | | actdur <= 24
| | | | | | | | | | | | | | | | | | | | | | | | | starttime9amto11am = 0: 15 (50.0/28.0)
| | | | | | | | | | | | | | | | | | | | | | | | | starttime9amto11am = 1: 7 (33.0/13.0)
| | | | | | | | | | | | | | | | | | | | | | | | actdur > 24: 15 (83.0/44.0)
| | | | | | | | | | | | | | | | | | | | | | | highincome = 1
| | | | | | | | | | | | | | | | | | | | | | | | tripdistance <= 1.209368: 15 (35.0/16.0)
| | | | | | | | | | | | | | | | | | | | | | | | tripdistance > 1.209368: 7 (79.0/46.0)
| | | | | | | | | | | | | | | | | | | | | | lowincome = 1: 15 (45.0/21.0)
| | | | | | | | | | | | | | | | | | | | | starttime11amto1pm = 1
| | | | | | | | | | | | | | | | | | | | | | actdurgreater30min = 0
| | | | | | | | | | | | | | | | | | | | | | | adultparty12pmto2pm = 0
| | | | | | | | | | | | | | | | | | | | | | | | tripdistance <= 1.884516: 15 (34.0/17.0)
| | | | | | | | | | | | | | | | | | | | | | | | tripdistance > 1.884516: 13 (35.0/22.0)
| | | | | | | | | | | | | | | | | | | | | | | adultparty12pmto2pm = 1
| | | | | | | | | | | | | | | | | | | | | | | | actdur <= 23
| | | | | | | | | | | | | | | | | | | | | | | | | female = 0: 15 (32.0/15.0)
| | | | | | | | | | | | | | | | | | | | | | | | | female = 1: 7 (34.0/17.0)
| | | | | | | | | | | | | | | | | | | | | | | | actdur > 23: 21 (27.0/18.0)
| | | | | | | | | | | | | | | | | | | | | | actdurgreater30min = 1
| | | | | | | | | | | | | | | | | | | | | | | highincome = 0
| | | | | | | | | | | | | | | | | | | | | | | | arrhour <= 11: 7 (27.0/17.0)
| | | | | | | | | | | | | | | | | | | | | | | | arrhour > 11: 21 (28.0/15.0)
| | | | | | | | | | | | | | | | | | | | | | | highincome = 1: 15 (37.0/22.0)
| | | | | | | | | | | | | | | | | | | | tripdistgreater10mi = 1
```

(continued on next page)

```
| | | | | | | | | | | | | | | | | | | | |  age <= 47: 15 (36.0/24.0)
| | | | | | | | | | | | | | | | | | | | |  age > 47: 13 (56.0/32.0)
| | | | | | | | | | | | | | | | | | |  lu_institutional = 1: 13 (32.0/21.0)
| | | | | | | | | | | | | | | | | |  groupgroceryduration = 1: 15 (60.0/26.0)
| | | | | | | | | | | | | | | | |  starttime3pmto6pm = 1: 15 (714.0/271.0)
| | | | | | | | | | | | |  subtourdummy = 1
| | | | | | | | | | | | | |  starttime11amto3pm = 0: 13 (28.0/12.0)
| | | | | | | | | | | | | |  starttime11amto3pm = 1
| | | | | | | | | | | | | | |  actdur <= 23: 15 (55.0/31.0)
| | | | | | | | | | | | | | |  actdur > 23: 21 (90.0/46.0)
| | | | | | | | | | |  groupeatoutduration = 1
| | | | | | | | | | | |  starttime1amto6pm = 0: 21 (253.0/104.0)
| | | | | | | | | | | |  starttime1amto6pm = 1
| | | | | | | | | | | | |  starttime5pmto7pm = 0
| | | | | | | | | | | | | |  adultparty12pmto2pm = 0
| | | | | | | | | | | | | | |  highincome = 0: 15 (419.0/177.0)
| | | | | | | | | | | | | | |  highincome = 1
| | | | | | | | | | | | | | | |  tripdistance <= 2.108016
| | | | | | | | | | | | | | | | |  actdurgreater45min = 0: 21 (36.0/18.0)
| | | | | | | | | | | | | | | | |  actdurgreater45min = 1: 15 (35.0/14.0)
| | | | | | | | | | | | | | | |  tripdistance > 2.108016
| | | | | | | | | | | | | | | | |  tripdistance <= 4.656515: 7 (47.0/26.0)
| | | | | | | | | | | | | | | | |  tripdistance > 4.656515: 15 (38.0/21.0)
| | | | | | | | | | | | | |  adultparty12pmto2pm = 1
| | | | | | | | | | | | | | |  arrhour <= 12: 21 (37.0/13.0)
| | | | | | | | | | | | | | |  arrhour > 12
| | | | | | | | | | | | | | | |  tripdistance <= 4.433623: 15 (33.0/13.0)
| | | | | | | | | | | | | | | |  tripdistance > 4.433623: 21 (26.0/9.0)
| | | | | | | | | | | | |  starttime5pmto7pm = 1: 21 (135.0/61.0)
| | | | | | | | |  simplesubtour = 1
| | | | | | | | | |  arrhour <= 10: 13 (36.0/20.0)
| | | | | | | | | |  arrhour > 10: 21 (431.0/150.0)
| | | | | |  actdurgreater150min = 1
| | | | | | |  finalpreschool = 0
| | | | | | | |  finalpredriving = 0
| | | | | | | | |  arrhour <= 14
| | | | | | | | | |  finalretiree = 0
| | | | | | | | | | |  finalnonworker = 0
| | | | | | | | | | | |  groupsocialvisitduration = 0
| | | | | | | | | | | | |  arrhour <= 3
| | | | | | | | | | | | | |  ptype = 1: 13 (139.0/70.0)
| | | | | | | | | | | | | |  ptype = 2: 13 (0.0)
| | | | | | | | | | | | | |  ptype = 3: 13 (0.0)
| | | | | | | | | | | | | |  ptype = 7
| | | | | | | | | | | | | | |  finalftworker = 0: 25 (48.0/23.0)
| | | | | | | | | | | | | | |  finalftworker = 1
| | | | | | | | | | | | | | | |  female = 0: 13 (100.0/33.0)
| | | | | | | | | | | | | | | |  female = 1: 25 (40.0/23.0)
| | | | | | | | | | | | |  arrhour > 3
| | | | | | | | | | | | | |  ptype = 1: 8 (34.0/16.0)
| | | | | | | | | | | | | |  ptype = 2: 13 (0.0)
| | | | | | | | | | | | | |  ptype = 3: 13 (0.0)
| | | | | | | | | | | | | |  ptype = 7: 13 (657.0/309.0)
| | | | | | | | | | | |  groupsocialvisitduration = 1: 7 (85.0/52.0)
| | | | | | | | | | |  finalnonworker = 1
| | | | | | | | | | | |  ptype = 1
| | | | | | | | | | | | |  age <= 49: 5 (33.0/24.0)
| | | | | | | | | | | | |  age > 49: 7 (27.0/19.0)
| | | | | | | | | | | |  ptype = 2: 7 (0.0)
| | | | | | | | | | | |  ptype = 3: 7 (0.0)
| | | | | | | | | | | |  ptype = 7
| | | | | | | | | | | | |  lu_institutional = 0
| | | | | | | | | | | | | |  tottr <= 0: 25 (41.0/18.0)
| | | | | | | | | | | | | |  tottr > 0
| | | | | | | | | | | | | | |  actdur <= 412
| | | | | | | | | | | | | | | |  age <= 60
| | | | | | | | | | | | | | | | |  lu_commercial = 0: 25 (63.0/35.0)
| | | | | | | | | | | | | | | | |  lu_commercial = 1
| | | | | | | | | | | | | | | | | |  age <= 42: 15 (26.0/17.0)
| | | | | | | | | | | | | | | | | |  age > 42: 7 (41.0/25.0)
| | | | | | | | | | | | | | | | age > 60: 25 (38.0/28.0)
| | | | | | | | | | | | | | | actdur > 412
| | | | | | | | | | | | | | | | arrhour <= 8: 13 (27.0/12.0)
| | | | | | | | | | | | | | | | arrhour > 8: 25 (27.0/18.0)
| | | | | | | | | | | | | lu_institutional = 1: 7 (76.0/34.0)
| | | | | | | | | | finalretiree = 1
| | | | | | | | | | | age <= 54: 25 (26.0/13.0)
| | | | | | | | | | | age > 54
| | | | | | | | | | | | lowincome = 0
| | | | | | | | | | | | | ptype = 1: 23 (29.0/19.0)
| | | | | | | | | | | | | ptype = 2: 7 (0.0)
| | | | | | | | | | | | | ptype = 3: 7 (0.0)
| | | | | | | | | | | | | ptype = 7
| | | | | | | | | | | | | | starttime9amto3pm = 0
| | | | | | | | | | | | | | | arrhour <= 6: 7 (27.0/16.0)
| | | | | | | | | | | | | | | arrhour > 6: 23 (42.0/23.0)
| | | | | | | | | | | | | | starttime9amto3pm = 1
| | | | | | | | | | | | | | | tottr  <= 1: 25 (44.0/29.0)
| | | | | | | | | | | | | | | tottr > 1: 7 (35.0/18.0)
| | | | | | | | | | | | lowincome = 1
| | | | | | | | | | | | | tripdistance <= 2.371755: 25 (32.0/21.0)
| | | | | | | | | | | | | tripdistance > 2.371755: 7 (53.0/32.0)
| | | | | | | | | arrhour > 14
| | | | | | | | | | lu_institutional = 0
| | | | | | | | | | | lu_commercial = 0: 25 (482.0/222.0)
| | | | | | | | | | | lu_commercial = 1
| | | | | | | | | | | | actdur <= 324
| | | | | | | | | | | | | starttime2pmto7pm = 0
| | | | | | | | | | | | | | finalftworker = 0: 21 (28.0/13.0)
| | | | | | | | | | | | | | finalftworker = 1: 25 (30.0/18.0)
| | | | | | | | | | | | | starttime2pmto7pm = 1
| | | | | | | | | | | | | | finalptworker = 0
| | | | | | | | | | | | | | | finalnonworker = 0
| | | | | | | | | | | | | | | | starttime5pmto7pm = 0: 23 (56.0/44.0)
| | | | | | | | | | | | | | | | starttime5pmto7pm = 1
| | | | | | | | | | | | | | | | | tripdistgreater10mi = 0
| | | | | | | | | | | | | | | | | | age <= 47: 21 (32.0/22.0)
| | | | | | | | | | | | | | | | | | age > 47: 23 (29.0/15.0)
| | | | | | | | | | | | | | | | | tripdistgreater10mi = 1: 21 (25.0/15.0)
| | | | | | | | | | | | | | | finalnonworker = 1: 15 (29.0/22.0)
| | | | | | | | | | | | | | finalptworker = 1: 13 (44.0/31.0)
| | | | | | | | | | | | actdur > 324: 25 (38.0/27.0)
| | | | | | | | | | lu_institutional = 1
| | | | | | | | | | | adultparty = 0: 23 (25.0/8.0)
| | | | | | | | | | | adultparty = 1
| | | | | | | | | | | | actdur <= 261: 22 (64.0/41.0)
| | | | | | | | | | | | actdur > 261: 25 (25.0/17.0)
| | | | | | | | finalpredriving = 1
| | | | | | | | | ptype = 1: 23 (25.0/18.0)
| | | | | | | | | ptype = 2: 25 (0.0)
| | | | | | | | | ptype = 3: 25 (0.0)
| | | | | | | | | ptype = 7
| | | | | | | | | | lu_commercial = 0
| | | | | | | | | | | lu_institutional = 0: 25 (140.0/57.0)
| | | | | | | | | | | lu_institutional = 1: 23 (29.0/18.0)
| | | | | | | | | | lu_commercial = 1: 23 (56.0/32.0)
| | | | | | | finalpreschool = 1
| | | | | | | | lu_commercial = 0: 25 (102.0/38.0)
| | | | | | | | lu_commercial = 1: 11 (41.0/27.0)
| | | | | someonedropped = 1
| | | | | | actdurgreater45min = 0
| | | | | | | actdurgreater30min = 0
| | | | | | | | adultpartyactdur20to40min = 0: 5 (959.0/200.0)
| | | | | | | | adultpartyactdur20to40min = 1
| | | | | | | | | lu_commercial = 0: 5 (36.0/18.0)
| | | | | | | | | lu_commercial = 1: 7 (36.0/23.0)
| | | | | | | actdurgreater30min = 1
| | | | | | | | lu_commercial = 0: 5 (29.0/15.0)
| | | | | | | | lu_commercial = 1: 15 (27.0/16.0)
| | | | | | actdurgreater45min = 1
```

(continued on next page)

```
| | | | | | | arrhour <= 14
| | | | | | | | lu_commercial = 0
| | | | | | | | | finalftworker = 0
| | | | | | | | | | starttime11amto3pm = 0: 5 (49.0/34.0)
| | | | | | | | | | starttime11amto3pm = 1: 25 (47.0/25.0)
| | | | | | | | | finalftworker = 1: 13 (44.0/30.0)
| | | | | | | | lu_commercial = 1
| | | | | | | | | starttime11amto1pm = 0
| | | | | | | | | | actdurgreater120min = 0: 7 (26.0/15.0)
| | | | | | | | | | actdurgreater120min = 1: 25 (32.0/22.0)
| | | | | | | | | starttime11amto1pm = 1: 21 (25.0/9.0)
| | | | | | | arrhour > 14
| | | | | | | | lu_institutional = 0
| | | | | | | | | lu_commercial = 0
| | | | | | | | | | mixedparty = 0: 25 (65.0/29.0)
| | | | | | | | | | mixedparty = 1: 23 (50.0/23.0)
| | | | | | | | | lu_commercial = 1
| | | | | | | | | | arrhour <= 17: 23 (47.0/29.0)
| | | | | | | | | | arrhour > 17: 21 (47.0/18.0)
| | | | | | | | lu_institutional = 1: 23 (41.0/13.0)
| | | | nonauto = 1
| | | | | actdurless30min = 0
| | | | | | simplesubtour = 0
| | | | | | | k8enrollment = 0
| | | | | | | | tripdistance <= 51.838412
| | | | | | | | | schoolbusmode = 0
| | | | | | | | | | finalftworker = 0
| | | | | | | | | | | lu_commercial = 0
| | | | | | | | | | | | lu_institutional = 0
| | | | | | | | | | | | | tripdistance <= 1.716877
| | | | | | | | | | | | | | actdurless60min = 0: 25 (79.0/36.0)
| | | | | | | | | | | | | | actdurless60min = 1: 23 (42.0/26.0)
| | | | | | | | | | | | | tripdistance > 1.716877: 7 (28.0/18.0)
| | | | | | | | | | | | lu_institutional = 1: 7 (61.0/22.0)
| | | | | | | | | | | lu_commercial = 1
| | | | | | | | | | | | actdur <= 237
| | | | | | | | | | | | | lowincome = 0
| | | | | | | | | | | | | | tottr <= 1
| | | | | | | | | | | | | | | actdur <= 64: 15 (28.0/16.0)
| | | | | | | | | | | | | | | actdur > 64: 7 (25.0/14.0)
| | | | | | | | | | | | | | tottr > 1: 21 (35.0/23.0)
| | | | | | | | | | | | | lowincome = 1: 15 (154.0/79.0)
| | | | | | | | | | | | actdur > 237: 7 (38.0/28.0)
| | | | | | | | | | finalftworker = 1
| | | | | | | | | | | actdur <= 160
| | | | | | | | | | | | arrhour <= 16
| | | | | | | | | | | | | lu_commercial = 0: 4 (42.0/28.0)
| | | | | | | | | | | | | lu_commercial = 1
| | | | | | | | | | | | | | tripdistance <= 0.297274: 21 (29.0/18.0)
| | | | | | | | | | | | | | tripdistance > 0.297274: 23 (26.0/19.0)
| | | | | | | | | | | | arrhour > 16: 21 (58.0/34.0)
| | | | | | | | | | | actdur > 160: 13 (42.0/25.0)
| | | | | | | | | schoolbusmode = 1: 23 (37.0/24.0)
| | | | | | | | tripdistance > 51.838412
| | | | | | | | | adultactdurless100min = 0: 13 (55.0/26.0)
| | | | | | | | | adultactdurless100min = 1: 4 (26.0/10.0)
| | | | | | | k8enrollment = 1
| | | | | | | | starttime2pmto6pm = 0: 11 (33.0/20.0)
| | | | | | | | starttime2pmto6pm = 1: 25 (61.0/29.0)
| | | | | | simplesubtour = 1: 21 (47.0/18.0)
| | | | | actdurless30min = 1
| | | | | | simplesubtour = 0
| | | | | | | groupgroceryduration = 0
| | | | | | | | hhmem <= 1: 4 (2878.0/464.0)
| | | | | | | | hhmem > 1
| | | | | | | | | mixedparty = 0: 4 (25.0/8.0)
| | | | | | | | | mixedparty = 1: 15 (49.0/25.0)
| | | | | | | groupgroceryduration = 1: 15 (32.0/21.0)
| | | | | | simplesubtour = 1: 21 (42.0/25.0)
| | | someonepicked = 1
| | | | actdurless60min = 0
| | | | | arrhour <= 9
| | | | | | hhmem <= 0: 13 (69.0/45.0)
| | | | | | hhmem > 0: 5 (36.0/27.0)
| | | | | arrhour > 9
| | | | | | lu_commercial = 0
| | | | | | | starttime5pmto7pm = 0
| | | | | | | | actdurgreater120min = 0
| | | | | | | | | young = 0: 6 (60.0/33.0)
| | | | | | | | | young = 1: 25 (46.0/28.0)
| | | | | | | | actdurgreater120min = 1: 25 (107.0/55.0)
| | | | | | | starttime5pmto7pm = 1: 23 (77.0/44.0)
| | | | | | lu_commercial = 1
| | | | | | | lowincome = 0
| | | | | | | | adultparty = 0: 23 (37.0/22.0)
| | | | | | | | adultparty = 1
| | | | | | | | | arrhour <= 17: 7 (38.0/27.0)
| | | | | | | | | arrhour > 17: 21 (31.0/19.0)
| | | | | | | lowincome = 1: 15 (33.0/21.0)
| | | | actdurless60min = 1
| | | | | nonauto = 0
| | | | | | starttime7amto12am = 0
| | | | | | | actdurless10min = 0: 4 (30.0/16.0)
| | | | | | | actdurless10min = 1: 6 (73.0/29.0)
| | | | | | starttime7amto12am = 1: 6 (2932.0/365.0)
| | | | | nonauto = 1
| | | | | | nondrivingchildren = 0: 4 (130.0/43.0)
| | | | | | nondrivingchildren = 1
| | | | | | | schoolageactdurless150min = 0
| | | | | | | | starttime2pmto5pm = 0: 4 (44.0/18.0)
| | | | | | | | starttime2pmto5pm = 1: 6 (91.0/20.0)
| | | | | | | schoolageactdurless150min = 1: 4 (146.0/54.0)
| | dropoffvariable = 1
| | | zerocars = 0
| | | | finalpredriving = 0: 5 (2082.0/93.0)
| | | | finalpredriving = 1
| | | | | iswalkorwheelchair = 0
| | | | | | tottr <= 2: 4 (28.0/9.0)
| | | | | | tottr > 2: 5 (286.0/39.0)
| | | | | iswalkorwheelchair = 1: 4 (30.0/7.0)
| | | zerocars = 1: 4 (26.0/6.0)
| worklocationmatch = 1
| | actdurless10min = 0
| | | volunactdurless60min = 0
| | | | actdurless30min = 0: 8 (8722.0/296.0)
| | | | actdurless30min = 1
| | | | | iswalkorwheelchair = 0: 8 (143.0/48.0)
| | | | | iswalkorwheelchair = 1: 4 (28.0/14.0)
| | | volunactdurless60min = 1
| | | | partysizechange = 0: 8 (25.0/8.0)
| | | | partysizechange = 1: 6 (35.0/15.0)
| | actdurless10min = 1
| | | nonauto = 0
| | | | someonedropped = 0
| | | | | someonepicked = 0: 8 (59.0/32.0)
| | | | | someonepicked = 1: 6 (29.0/6.0)
| | | | someonedropped = 1: 5 (57.0/8.0)
| | | nonauto = 1: 4 (166.0/15.0)
schoollocationmatch = 1
| actdurless10min = 0
| | actdurless30min = 0: 11 (4509.0/173.0)
| | actdurless30min = 1
| | | hhmem <= 0: 11 (28.0/15.0)
| | | hhmem > 0: 6 (25.0/13.0)
| actdurless10min = 1
| | someonedropped = 0
| | | age <= 18: 6 (33.0/18.0)
| | | age > 18: 4 (35.0/3.0)
| | someonedropped = 1: 5 (30.0/13.0)
```

(continued on next page)

```
Number of Leaves:  389

Size of the tree:       765


Time taken to build model: 18.81 seconds
Time taken to test model on training data: 1.26 seconds

=== Error on training data ===

Correctly Classified Instances    36860        70.5387 %
Incorrectly Classified Instances   15395        29.4613 %
Kappa statistic            0.6681
Mean absolute error          0.0689
Root mean squared error         0.1856
Relative absolute error         46.3931 %
Root relative squared error       68.1127 %
Total Number of Instances      52255
```

APPENDIX E

Experiment B Models

Person-Type and Education Model

Upper-level Person-type Model

Parameter	Coeff.	t-stat
ASC_part	-0.848	-3.6
ASC_full	-1.280	-4.9
ASC_retired	1.310	6.5
ASC_child	0.920	4.1
ASC_student	0.000	–
ASC_other	0.600	2.8
B_PART_num_acts_other	0.087	4.6
B_PART_num_acts_work	3.020	16.0
B_FULL_Employment_Density	0.176	4.2
B_FULL_Housing_Density	-0.199	-4.7
B_FULL_avg_stops_per_tour	0.161	3.0
B_FULL_avg_tour_ttime	0.004	4.3
B_FULL_num_acts_school	-1.620	-8.3
B_FULL_num_acts_work	2.760	10.5
B_FULL_total_dur_other	0.001	3.8
B_FULL_total_dur_work	0.002	5.5
B_RETIRE_num_acts_other	-0.051	-2.6
B_RETIRE_num_acts_school	-3.150	-5.8
B_RETIRE_total_dur_pickdrop	-0.075	-3.8
B_CHILD_Employment_Density	0.238	3.6
B_CHILD_Housing_Density	-0.217	-3.1
B_CHILD_avg_tour_ttime	-0.013	-4.7
B_CHILD_num_acts_other	-0.149	-5.0
B_CHILD_total_dur_pickdrop	0.017	2.1
B_STUDE_Employment_Density	0.158	2.2
B_STUDE_Housing_Density	-0.172	-2.3
B_STUDE_num_acts_other	-0.152	-4.7
B_STUDE_total_dur_other	0.000	2.3
B_STUDE_total_dur_pickdrop	-0.096	-2.9
B_STUDE_total_dur_school	0.003	11.5
IV_PART_EDUC	0.116	2.2
IV_FULL_EDUC	0.217	5.6
IV_RET_EDUC	0.052	2.2
IV_OTHER_EDUC	0.125	2.8

Education – full-time worker

Parameter	Coeff.	t-stat
ASC_NOHS	0.000	–
ASC_HS	2.090	3.2
ASC_COLL	2.500	4.1
B_HS_Housing_Density	0.696	2.6
B_HS_Pop_Density	0.251	-3.0
B_HS_avg_tour_ttime	0.002	1.7
B_HS_intersection_density	0.003	1.5
B_HS_num_acts_other	0.234	2.2
B_HS_road_density	0.066	-3.2
B_HS_total_dur_work	0.001	1.6
B_COLL_employment_density	0.252	4.1
B_COLL_Housing_Density	0.713	2.7
B_COLL_Pop_Density	0.399	-4.7
B_COLL_avg_work_tour_ttime	0.004	3.0
B_COLL_num_acts_other	0.287	2.8
B_COLL_total_dur_work	0.001	1.2

Education – part-time worker

Parameter	Coeff.	t-stat
ASC_NOHS	0.000	–
ASC_HS	1.670	6.6
ASC_COLL	2.410	5.4
B_HS_total_dur_school	0.004	-4.7
B_COLL_Employment_Density	0.434	3.6
B_COLL_Pop_Denssity	0.123	-2.7
B_COLL_avg_other_tour_ttime	0.005	1.4
B_COLL_avg_school_tour_ttime	0.012	1.3
B_COLL_avg_tour_ttime	-0.009	-2.0
B_COLL_avg_work_tour_ttime	0.004	1.1
B_COLL_num_acts_other	0.074	2.2
B_COLL_road_density	0.050	-2.0
B_COLL_total_dur_school	0.006	-4.4

Education - retiree

Parameter	Coeff.	t-stat
ASC_NOHS	0.000	–
ASC_HS	2.250	3.4
ASC_COLL	1.920	3.0
B_HS_Employment_Density	0.752	2.8
B_HS_Pop_Density	-0.158	-2.1
B_HS_num_acts_other	0.235	2.2
B_HS_road_density	-0.102	-3.4
B_HS_total_dur_pickdrop	-0.097	-1.9
B_COLL_Employment_Density	0.973	3.4
B_COLL_Pop_Density	-0.257	-3.3
B_COLL_intersection_density	-0.011	-3.6
B_COLL_num_acts_other	0.317	3.0
B_COLL_total_dur_pickdrop	-0.069	-1.8

Education - other

Parameter	Coeff.	t-stat
ASC_NOHS	0.000	–
ASC_HS	0.864	-1.1
ASC_COLL	0.314	0.5
B_HS_Housing_Density	0.134	1.9
B_HS_Pop_Density	0.048	-1.3
B_HS_at_home_duration	0.001	3.3
B_HS_avg_other_tour_ttime	0.005	1.9
B_HS_avg_school_tour_ttime	0.011	1.4
B_HS_avg_stops_per_other_tour	0.387	-2.0
B_HS_avg_stops_per_tour	0.273	1.4
B_HS_num_acts_school	1.320	-4.8
B_COLL_Employment_Density	0.346	4.0
B_COLL_Pop_Density	0.163	-3.9
B_COLL_at_home_duration	0.001	2.7
B_COLL_num_acts_school	1.770	-7.4

Ordinal Logit Model for Age Categories

Variable	Coefficient	t-stat
Constant	3.208	25.2
avg_school_tour_ttime	0.005	2.0
num_acts_work	-0.355	-9.1
num_acts_school	-1.073	-9.1
avg_dur_school	-0.007	-10.9
num_acts_pickdrop	-0.188	-2.7
total_dur_pickdrop	-0.023	-2.5
total_dur_other	0.000	-3.8
auto_tour_percent	0.338	3.4
transituse	1.090	2.6
Employment_Density	-0.091	-3.6
Pop_Density	-0.024	-2.3
Housing_Density	0.125	5.2
Threshold		
Mu(1)	0.703	17.9
Mu(2)	2.365	57.5
Mu(3)	4.351	84.8

Gender Binary Logit Model

Parameter	Coefficient	t-stat
ASCMALE	-0.026	0.2
B_MALE_Employment_Density	0.053	1.9
B_MALE_Housing_Density	-0.068	-2.4
B_MALE_Pop_Denssity	0.021	1.7
B_MALE_avg_other_tour_ttime	0.003	4.0
B_MALE_avg_school_tour_ttime	0.003	1.5
B_MALE_avg_stops_per_tour	-0.102	-2.5
B_MALE_avg_stops_per_work_tour	-0.070	-1.5
B_MALE_avg_work_tour_ttime	0.003	3.1
B_MALE_num_acts_pickdrop	-0.152	-3.2
B_MALE_num_acts_school	0.144	2.4
B_MALE_num_acts_work	-0.218	-1.8
B_MALE_road_density	-0.021	-3.3
B_MALE_total_dur_work	0.001	4.5

Has License PART Decision List Model

auto_total > 0 avg_dur_school <= 261.667 total_tours > 1 auto_total > 0.75: YES (1707.0/38.0)

auto_work <= 0.5 avg_dur_school <= 380 avg_stops_per_work_tour > 2.5: YES (141.0/9.0)

auto_work <= 0.5 auto_total > 0 avg_dur_work <= 373 other_tours > 0 num_acts_other <= 6 at_home_duration <= 2803 total_tours <= 3 block_size <= 0.14 block_size <= 0.12 intersection_density <= 233.33: YES (167.0/35.0)

auto_work > 0.5: YES (127.0/1.0)

auto_total > 0 avg_dur_school <= 386.333 avg_dur_work <= 373 Pop_Denssity <= 5.64: YES (115.0/4.0)

avg_dur_work > 373 auto_total > 0: YES (87.0)

num_acts_other > 6 total_tours <= 5 intersection_density > 232.14: YES (21.0)

avg_dur_work > 467.5: YES (94.0/19.0)

num_acts_other <= 9 num_acts_pickdrop > 0: YES (22.0/5.0)

num_acts_other <= 9 other_tours > 3 auto_total > 0: YES (21.0/2.0)

num_acts_other <= 9 num_primary_tours > 2 Housing_Density <= 11.27: NO (39.0/8.0)

other_tours <= 2 total_dur_work <= 675 avg_dur_other <= 669 total_dur_work <= 262 total_tours <= 1 transitu <= 0.3: YES (54.0/19.0)

num_primary_tours <= 4 avg_dur_other <= 669 total_tours > 2: YES (32.0/8.0)

total_tours <= 3 avg_dur_other <= 669 total_tours > 1 work_tours > 0: YES (24.0/9.0)

total_tours <= 3 avg_dur_other <= 600 total_tours <= 1: NO (21.0/6.0)

total_tours <= 3 avg_dur_other <= 444.667 block_size <= 0.09 avg_other_tour_ttime <= 40: YES (13.0/2.0)

total_tours <= 3 avg_dur_other <= 444.667: NO (75.0/30.0)

Joint Household-Type Decision Tree Model

```
total_dur_school <= 190
  num_acts_pickdrop <= 0
       auto_total <= 0.25
         Pop_Denssity <= 13.17
           at_home_duration <= 2588
             avg_stops_per_tour <= 2.8: H_220 (64.0/50.0)
             avg_stops_per_tour > 2.8: H_321 (30.0/17.0)
           at_home_duration > 2588: H_100 (21.0/14.0)
         Pop_Denssity > 13.17
           Employment_Density <= 5.77: H_100 (25.0/18.0)
           Employment_Density > 5.77
             block_size <= 0.11
               intersection_density <= 230.76: H_100 (32.0/14.0)
               intersection_density > 230.76: H_210 (47.0/29.0)
             block_size > 0.11: H_100 (25.0/21.0)
       auto_total > 0.25
         Housing_Density <= 6.31
           transitu <= 0.18
             auto_work <= 0
               avg_dur_other <= 290
                 Employment_Density <= 5.84
                   work_tours <= 0
                     auto_total <= 0.7: H_220 (48.0/33.0)
                     auto_total > 0.7
                       Pop_Denssity <= 7.16
                         num_acts_other <= 2
                           total_dur_other <= 89
                             avg_tour_ttime <= 29.5: H_321 (24.0/16.0)
                             avg_tour_ttime > 29.5: H_220 (20.0/11.0)
                           total_dur_other > 89
                             at_home_duration <= 2698: H_321 (33.0/23.0)
                             at_home_duration > 2698: H_210 (20.0/10.0)
                         num_acts_other > 2: H_220 (350.0/212.0)
                       Pop_Denssity > 7.16
                         Housing_Density <= 3.54: H_210 (24.0/16.0)
                         Housing_Density > 3.54: H_220 (20.0/10.0)
                   work_tours > 0
                     avg_tour_ttime <= 69: H_321 (21.0/13.0)
                     avg_tour_ttime > 69: H_220 (20.0/14.0)
                 Employment_Density > 5.84: H_321 (28.0/9.0)
               avg_dur_other > 290: H_321 (32.0/18.0)
             auto_work > 0
               num_acts_work <= 2: H_220 (506.0/304.0)
               num_acts_work > 2: H_321 (22.0/10.0)
           transitu > 0.18
             auto_total <= 0.8: H_311 (39.0/29.0)
             auto_total > 0.8: H_220 (112.0/74.0)
         Housing_Density > 6.31: H_110 (160.0/119.0)
     num_acts_pickdrop > 0: H_321 (188.0/52.0)
   total_dur_school > 190: H_321 (344.0/96.0)
```

APPENDIX F

Using the Bundled Scripts and Code

Introduction

The scripts compiled as part of this appendix implement basic aspects of the GPS processing methods tested as part of Experiments A and B. The main goal of these implementations was to test the feasibility of various methods and to assess how well they worked. As such, these implementations do not necessarily produce readily usable results, although they certainly can be modified and extended to meet these types of applications. It should also be noted that the code and procedures provided with this report were developed as part of this project, but NCHRP makes no warranty that the code and procedures will continue to work as written given that the software tools they depend upon are periodically updated.

Prerequisites for Running the Bundled Code

The methods implemented as part of Experiment A were mostly developed using R version 3.0.1, the latest version of R for various platforms can be downloaded from www.r-project.org. The methods implemented in R are the most straightforward to use given their procedural nature (i.e., to get results simply call the appropriate function with the correct data parameters). The fact that R supports multiple computing platforms also makes it easy to use these methods in Microsoft Windows, Mac OS X, and various Linux distributions. To facilitate the use of R, end users are encouraged to download and install RStudio, which provides a nice integrated user environment for running R scripts, editing code, browsing data, and viewing graphical outputs. RStudio install packages for most popular computing platforms can be downloaded from www.rstudio.com.

The WEKA toolkit was also used to implement some of the procedures described; for procedures implemented using WEKA, a simple list of steps to follow is given that should allow practitioners to reproduce the results. Whenever applicable, model configuration files are provided that can be used as a starting point by practitioners. The latest version of WEKA can be obtained from www.cs.waikato.ac.nz/ml/weka. WEKA requires Java Runtime, which can be installed in Microsoft Windows, Mac OS X, and various Linux Distros.

The discrete choice modeling package BIOGEME was used to estimate models for travel mode and trip purpose identification. The original Bison BIOGEME model specification files are provided along with instructions on how to use them with BIOGEME. Source code and pre-compiled binaries of Bison BIOGEME can downloaded from biogeme.epfl.ch. The website makes available pre-compiled Microsoft Windows binaries, but BIOGEME can also be built from source on Mac OS X and most Linux Distros.

Finally, code in Java, SQL, and C++ is also referenced as part of some of the implemented methods. The Java code is invoked directly from R using the rJava package; pre-compiled .jar files are provided so only the Java Run-Time Engine (JRE) is needed. The latest version of the JRE can be downloaded from https://java.com/en/download/index.jsp.

The SQL scripts provided were used with an instance of PostgreSql version 9.1, which can be downloaded from www.postgresql.org. The C++ code implements the tool named NCHRP_GPS_Data_Reduction, which is used to prepare the input data for Experiment B; it can be compiled in the Microsoft Windows platform using free versions of Microsoft Visual Studio, which can be downloaded from http://www.microsoft.com/visualstudio/eng/downloads#d-2013-express. Using the provided C++ source code in other platforms is possible, but may require modifications as well as the creation of make files, which are not included in this package.

Experiment A Instructions

This section covers the loading and use of the procedures implemented as part of five Experiment A methods tested, which include:

1. GPS point noise filtering
2. Trip end identification
3. Mode transition identification

4. Travel mode identification
5. Trip purpose identification

Information on how to use the Experiment A method implementations is organized based on the software tools used. The majority of the methods in Experiment A were implemented using R, with a smaller set done using WEKA and BIOGEME.

Methods Implemented Using R

Before the routines can be loaded, it is necessary to configure the R environment by ensuring that the following packages are installed: geosphere, rJava, ggplot2, and ggmap. This action can be done by issuing the following command:

```
> install.packages("geosphere,rJava,
ggplot2,ggmap");
```

Once these packages are loaded, the methods routines can be loaded into R for use. The simplest way to do this is to load the RStudio project file NchrpScripts.Rproj; this action will set up R's home to the home folder of the script files. Once the project file is open, the following command can be issued to initialize the R environment and preload the implemented routines:

```
> source('Initialize.r')
```

Loading GPS Point Data

The package includes the function loadData() for loading GPS point data in GeoLogger format. The GeoLogger format is a comma-separated values text file that contains the following data fields:

Field	Description

Field 1	A = valid data, GPS ok
	D = valid data, DGPS ok
	V = first valid point after loss of signal or power

Field 2	Latitude (dd.ddddd)

Field 3	N = North of the Equator
	S = South of the Equator

Field 4	Longitude (ddd.dddd)

Field 5	E = East of Greenwich
	W = West of Greenwich

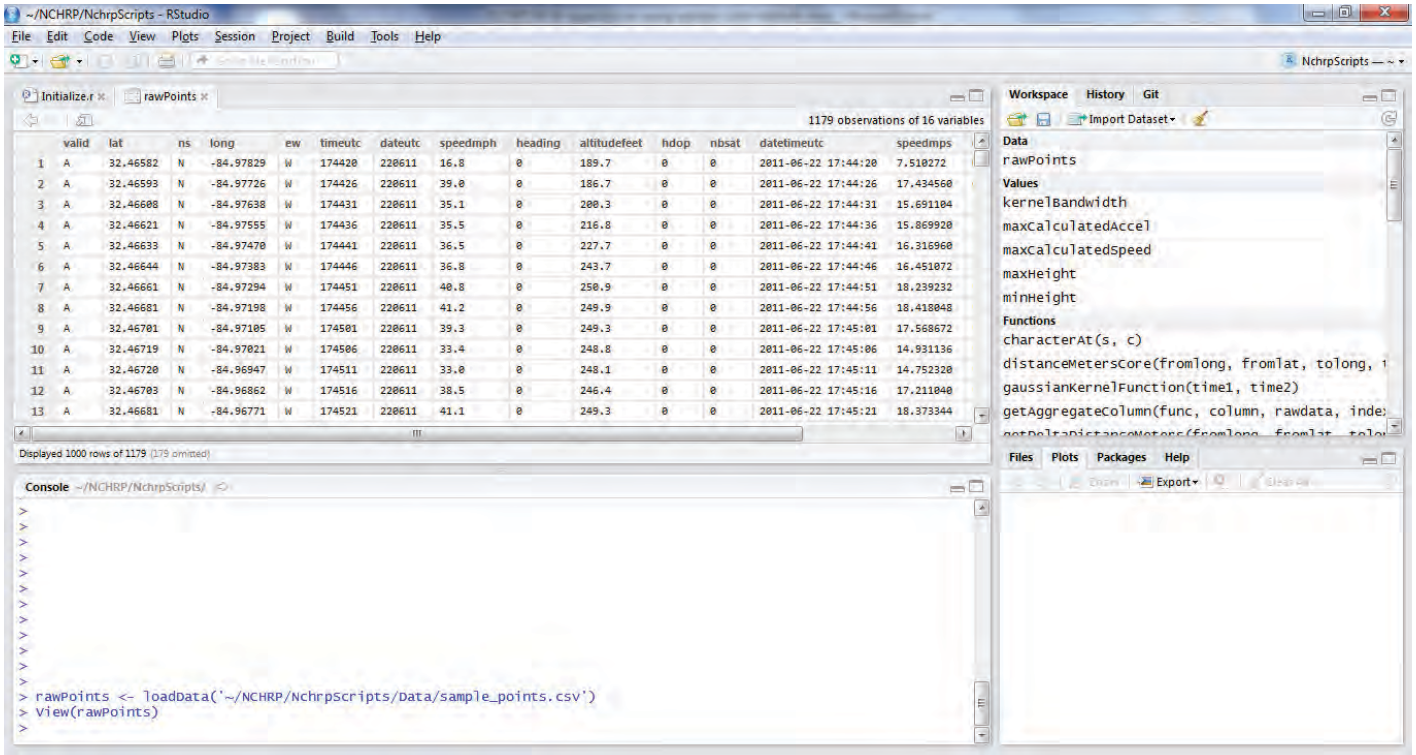
Field 6	Speed in mph (s.s)
Field 7	Time UTC (hhmmss)
Field 8	Date UTC (ddmmyy)
Field 9	Heading – clockwise degrees from north (000 - 259)
Field 10	Altitude in feet (a.a)
Field 11	HDOP (00.5 - 99.9)
Field 12	Number of satellites (00 - 12)

The loadData function returns a data set that can be used as inputs to point-based methods. To invoke the provided function and have the loaded GPS data assigned to a variable, the following command can be issued:

```
> rawPoints <– loadData('~/NchrpScripts/
Data/sample_points.csv')
```

The loaded points can then be viewed using RStudio's built-in data grid browser by issuing the following command:

```
View(rawPoints)
```


~/NCHRP/NchrpScripts - RStudio
File Edit Code View Plots Session Project Build Tools Help
NchrpScripts — ~
Initialize.r
rawPoints
Workspace History Git
Import Dataset
Files Plots Packages Help
Export

1179 observations of 16 variables

	valid	lat	ns	long	ew	timeutc	dateutc	speedmph	heading	altitudefeet	hdop	nbsat	datetimeutc	speedmps
1	A	32.46582	N	-84.97829	W	174420	220611	16.8	0	189.7	0	0	2011-06-22 17:44:20	7.510272
2	A	32.46593	N	-84.97726	W	174426	220611	39.0	0	186.7	0	0	2011-06-22 17:44:26	17.434560
3	A	32.46608	N	-84.97638	W	174431	220611	35.1	0	200.3	0	0	2011-06-22 17:44:31	15.691104
4	A	32.46621	N	-84.97555	W	174436	220611	35.5	0	216.8	0	0	2011-06-22 17:44:36	15.869920
5	A	32.46633	N	-84.97470	W	174441	220611	36.5	0	227.7	0	0	2011-06-22 17:44:41	16.316960
6	A	32.46644	N	-84.97383	W	174446	220611	36.8	0	243.7	0	0	2011-06-22 17:44:46	16.451072
7	A	32.46661	N	-84.97294	W	174451	220611	40.8	0	250.9	0	0	2011-06-22 17:44:51	18.239232
8	A	32.46681	N	-84.97198	W	174456	220611	41.2	0	249.9	0	0	2011-06-22 17:44:56	18.418048
9	A	32.46701	N	-84.97105	W	174501	220611	39.3	0	249.3	0	0	2011-06-22 17:45:01	17.568672
10	A	32.46719	N	-84.97021	W	174506	220611	33.4	0	248.8	0	0	2011-06-22 17:45:06	14.931136
11	A	32.46720	N	-84.96947	W	174511	220611	33.0	0	248.1	0	0	2011-06-22 17:45:11	14.752320
12	A	32.46703	N	-84.96862	W	174516	220611	38.5	0	246.4	0	0	2011-06-22 17:45:16	17.211040
13	A	32.46681	N	-84.96771	W	174521	220611	41.1	0	249.3	0	0	2011-06-22 17:45:21	18.373344

Displayed 1000 rows of 1179 (179 omitted)

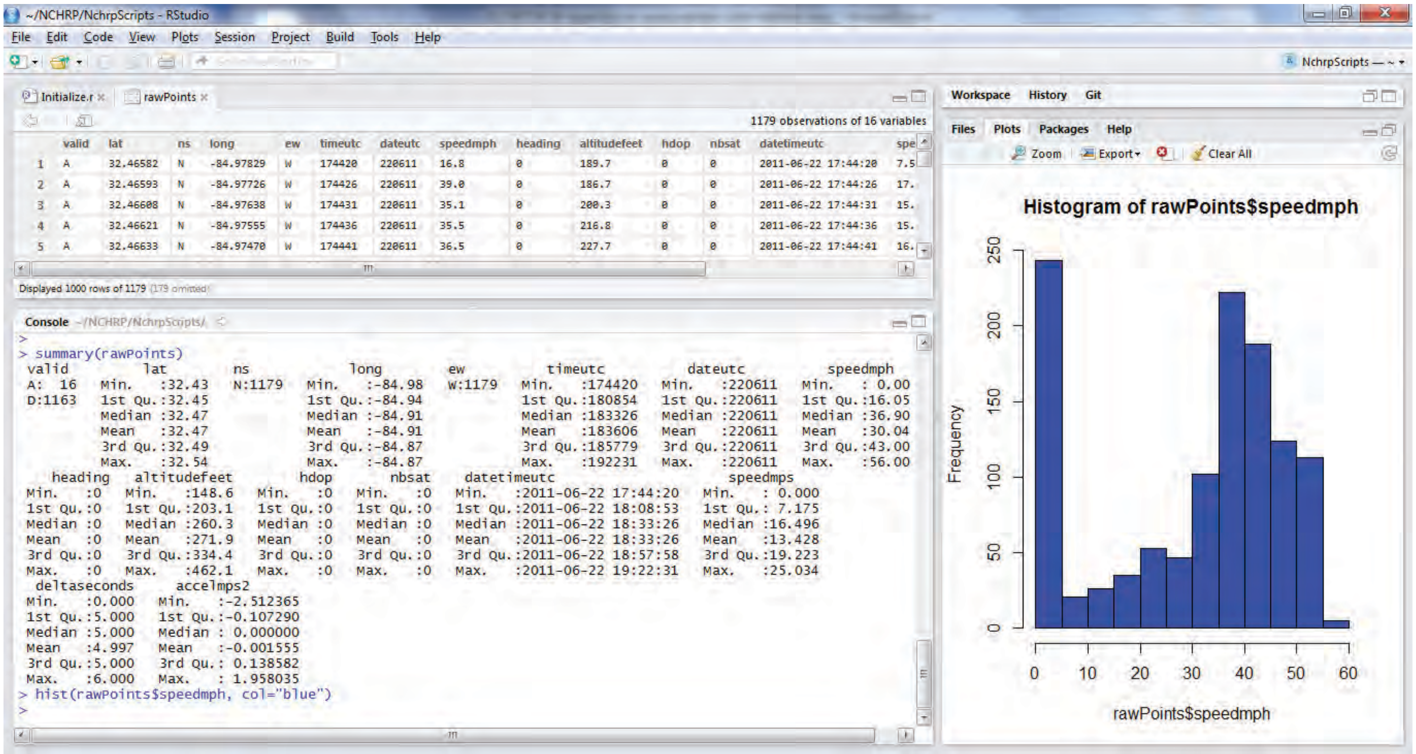
Data
rawPoints

Values
kernelBandwidth
maxCalculatedAccel
maxCalculatedSpeed
maxHeight
minHeight

Functions
characterAt(s, c)
distanceMetersCore(fromlong, fromlat, tolong, t
gaussianKernelFunction(time1, time2)
getAggregateColumn(func, column, rawdata, inde
getDeltaDistanceMeters(fromlong, fromlat, tolo

Console ~/NCHRP/NchrpScripts/

```
>
>
>
>
>
>
>
>
>
>
>
>
> rawPoints <- loadData('~/NCHRP/NchrpScripts/Data/sample_points.csv')
> View(rawPoints)
>
```


~/NCHRP/NchrpScripts - RStudio
File Edit Code View Plots Session Project Build Tools Help
NchrpScripts
Initialize.r
rawPoints
Workspace History Git
Files Plots Packages Help
Zoom Export Clear All
Histogram of rawPoints$speedmph
Frequency
0
50
100
150
200
250
0 10 20 30 40 50 60
rawPoints$speedmph

1179 observations of 16 variables

	valid	lat	ns	long	ew	timeutc	dateutc	speedmph	heading	altitudefeet	hdop	nbsat	datetimeutc	spe
1	A	32.46582	N	-84.97829	W	174420	220611	16.8	0	189.7	0	0	2011-06-22 17:44:20	7.5
2	A	32.46593	N	-84.97726	W	174426	220611	39.0	0	186.7	0	0	2011-06-22 17:44:26	17.
3	A	32.46608	N	-84.97638	W	174431	220611	35.1	0	200.3	0	0	2011-06-22 17:44:31	15.
4	A	32.46621	N	-84.97555	W	174436	220611	35.5	0	216.8	0	0	2011-06-22 17:44:36	15.
5	A	32.46633	N	-84.97470	W	174441	220611	36.5	0	227.7	0	0	2011-06-22 17:44:41	16.

Displayed 1000 rows of 1179 (179 omitted)

Console ~/NCHRP/NchrpScripts/

```
>
> summary(rawPoints)
 valid        lat            ns           long            ew            timeutc           dateutc          speedmph
 A:  16   Min.   :32.43   N:1179   Min.   :-84.98   W:1179   Min.   :174420   Min.   :220611   Min.   : 0.00
 D:1163   1st Qu.:32.45            1st Qu.:-84.94            1st Qu.:180854   1st Qu.:220611   1st Qu.:16.05
          Median :32.47            Median :-84.91            Median :183326   Median :220611   Median :36.90
          Mean   :32.47            Mean   :-84.91            Mean   :183606   Mean   :220611   Mean   :30.04
          3rd Qu.:32.49            3rd Qu.:-84.87            3rd Qu.:185779   3rd Qu.:220611   3rd Qu.:43.00
          Max.   :32.54            Max.   :-84.87            Max.   :192231   Max.   :220611   Max.   :56.00
    heading    altitudefeet         hdop         nbsat       datetimeutc                       speedmps
 Min.   :0   Min.   :148.6   Min.   :0   Min.   :0   Min.   :2011-06-22 17:44:20   Min.   : 0.000
 1st Qu.:0   1st Qu.:203.1   1st Qu.:0   1st Qu.:0   1st Qu.:2011-06-22 18:08:53   1st Qu.: 7.175
 Median :0   Median :260.3   Median :0   Median :0   Median :2011-06-22 18:33:26   Median :16.496
 Mean   :0   Mean   :271.9   Mean   :0   Mean   :0   Mean   :2011-06-22 18:33:26   Mean   :13.428
 3rd Qu.:0   3rd Qu.:334.4   3rd Qu.:0   3rd Qu.:0   3rd Qu.:2011-06-22 18:57:58   3rd Qu.:19.223
 Max.   :0   Max.   :462.1   Max.   :0   Max.   :0   Max.   :2011-06-22 19:22:31   Max.   :25.034
  deltaseconds      accelmps2
 Min.   :0.000   Min.   :-2.512365
 1st Qu.:5.000   1st Qu.:-0.107290
 Median :5.000   Median : 0.000000
 Mean   :4.997   Mean   :-0.001555
 3rd Qu.:5.000   3rd Qu.: 0.138582
 Max.   :6.000   Max.   : 1.958035
> hist(rawPoints$speedmph, col="blue")
>
```

The built-in statistical functions of R can also be used to summarize and graph the data:

```
> summary(rawPoints)
> hist(rawPoints$speedmph)
```

GPS Point Noise Filtering

To run the noise filtering methods, pass in the raw GPS points loaded using the loadData() function. The methods add a Boolean variable (noise) to the passed-in data set that is set to TRUE if the point is considered to be noise. Three noise filtering method implementations are contained in the bundled source code. The sample code below shows how to run them and do a quick summarization of their results:

```
> lwPoints <- noiseFiltering_Lawson
(rawPoints)
> summary(lwPoints$noise)
  Mode FALSE TRUE NA’s
logical  4337  663  0
>
> safPoints <- noiseFiltering_Schuessler_
Axhausen(rawPoints)
> summary(safPoints$noise)
  Mode FALSE TRUE NA’s
logical  4744  256  0
>
> stfPoints <- noiseFiltering_Stopher
(rawPoints)
> summary(stfPoints$noise)
  Mode FALSE TRUE NA’s
logical  4996  4  0
```

The filtered out points can also be visualized using the ggmap library. For example, the following commands will create a map using a bounding box computed based on the points’ coordinates and will apply different colors (or gray levels) based on their speeds:

```
> fgps <- subset(lwPoints, noise)
> bbox <- c(min(fgps $long), min(fgps $lat),
  max(fgps $long), max(fgps $lat))
> map <- qmap(bbox, zoom = 15)
> map + labs(x = “Longitude”, y=”Latitude”)
> map + geom_point(aes(long, lat, colour=
speedmph, size=2, alpha=0.25), data=
fgps) + scale_colour_gradient(low=”red”,
high=”green”)
```

Trip Identification

The trip identification methods implemented take as input a data set of GPS points and return a list of GPS trips, with some basic attributes added. To derive trips using the two implemented methods and the GPS points filtered by the Lawson method as input, the following commands can be used:

```
saTrips <- tripIdentification_Schuessler_
Axhausen(subset(lwPoints, !noise))
```

0.25
● 0.25
2
● 2
speedmph
60
40
20
0
DOWNTOWN
ATLANTA
LUCKIE-MARIETTA DISTRICT
CASTLEBERRY HILL
Ivan Allen Jr Blvd
Ralph McGill Blvd NE
Ellis St NE
Courtland St NE
Downtown Connector
Mitchell St SW
Decatur St SE
Piedmont Ave SE
Memorial Dr SE
Fulton County Government Center
Georgia Aquarium
The World of Coca-Cola
Philips Arena
State Bar of Georgia
Pemberton Place
Mayor's Park
St Luke's Episcopal Church
Map data ©2013 Google, Sanborn

```
wlTrips <- tripIdentification_Wolf(subset
(lwPoints, !noise))
```

The generated trips include the following basic attributes:

- startindex & endindex: point indexes into the passed-in point data for each trip
- starttime & endtime: UTC date and time stamps for the start and end of each trip
- startlat & startlong: latitude and longitude coordinates for the trip start
- endtlat & endlong: latitude and longitude coordinates for the trip end
- distancemeters: total distance accumulated over the trip's points, calculated using the great-circle distance formula and returned in meters
- travtimeminutes: endtime – starttime in fractional minutes
- avgspeedkph: trip's average speed in km/h

Travel Mode Transition Identification

These methods can break a sequence of GPS points into mode segments and are only applicable to GPS data that have been collected using on-person data loggers. Mode segments consist of individual legs in a multimodal trip. The output of these methods is similar to that of the trip identification methods, as input that take in a data set of GPS points returned by the loadData() function. The first implemented mode transition identification methods can be invoked using the following commands:

```
segments <- modeTransitionIdentification_
Oliveira (rawPoints)
```

The second method (modeTransitionIdentification_TsuiShalaby_SchuesslerAxhausen) employed a Fuzzy Logic Engine written in Java by Edward Sazonov that can be found at http://people.clarkson.edu/~esazonov/FuzzyEngine.htm. The authors modified the engine by adding a method to return several variables, one for each mode of interest. (A fuzzy engine's normal operation is to return a decimal value between 0 and 1; the modification allowed an array of values to be returned: 0 to 1 for each mode, such that the sum of all is 1.) Because the Experiment A reference data used a different set of travel modes, further modifications to the engine to generalize the mode handling were made. The modified code and compiled objects are included in the distributed package. The package rJava was used to invoke the Java .jar code directly from R, so the method can be invoked directly using the following command:

```
modeSegments <- travelModeIdentification_
TsuiShalaby_SchuesslerAxhausen(segments,
filteredPoints)
```

Travel Mode Identification

The implementation of the Stopher travel mode identification method was very dependent on the available spatial data regarding the location of roadways and railroads. To implement this method within R, it was necessary to conduct extensive preprocessing of the data using a GIS; the results of this preprocessing were then saved as text files, which are directly referenced by the R code. This makes the implementation unsuitable for uses outside of this project, and, because of this, this method is not covered in the bundled package.

Methods Implemented Using WEKA

The WEKA machine learning tools were used to implement machine learning-based methods, namely neural networks for travel mode identification and decision trees for trip purpose. The packaged model files (.model) can be opened with WEKA to run on other data sources. Note that text files produced by WEKA use UNIX line endings, which Windows Notepad will not display correctly. Use WordPad or a more advanced text editor to view them.

WEKA makes a distinction between numeric values and nominal values. A numeric variable can vary over the whole real line, while a nominal value is one of a set. We use nominal values where there is no "closeness" relationship, and the concept of "for cases with less than a value" does not make sense. For example, the number of household members on a trip (hhmem) is a numeric value because there might be interesting relationships between having one household member and having more than one household member on the trip. On the other hand, the place type (ptype) is nominal, because there is no obvious relationship that home and work share that school does not.

Travel Mode Identification Using Neural Networks

The saved file ForWekaTrain180.model contains the final trained neural net, which can identify between the travel modes walk, bike, car, bus, and train. This file can be opened in WEKA and then applied to input data files containing independent variables that the net can use to estimate travel mode.

Input files for WEKA use the Attribute-Relation File Format (ARFF). These are CSV text files that include metadata that covers such items as all the possible values that nominal variables can take. For more information on ARFF files see http://www.cs.waikato.ac.nz/~ml/weka/arff.html.

You can use WEKA's explorer to prepare input ARFF files from CSVs. In this case the following transformations were done:

1. Remote attributes 2 and 3 (startindex and endindex)
2. Convert attribute 6 (travmode) from numeric to nominal
3. Convert attribute 1 (caseid) from numeric to nominal, and then from nominal to string.

The ARFF file records this as:

```
@relation ForWekaTrain180-weka.filters.
unsupervised.attribute.Remove-R2-3-weka.
filters.unsupervised.attribute.Numeric
ToNominal-R6-weka.filters.unsupervised.
attribute.NumericToNominal-R1-weka.
filters.unsupervised.attribute.Nominal
ToString-C1
```

As part of Experiment A, a random sample of 180 trips was chosen to train the network, and then a separate random sample of 90 trips was used to test the network. The neural net uses the following fields (travmode is the result):

- Average speed in mph
- Max speed in mph
- Standard deviation of the distance between locations in feet
- Dwell time in seconds
- Travel mode [1 = walk, 2 = bike, 3 = car, 5 = bus, 7 = train]

Once an ARFF file is ready, the neural network model can be built by training it on that ARFF. This example script uses a learning rate of 0.1 and runs for 300 epochs:

```
>java -cp weka.jar weka.classifiers.meta.
FilteredClassifier -d ForWekaTrain180.
model -t ForWekaTrain180.csv.arff -F
weka.filters.unsupervised.attribute.
RemoveType -W weka.classifiers.functions.
MultilayerPerceptron — -L 0.1 -N 300 >
ForWekaTrain180.out.txt
```

To validate the model, run the following script on the test ARFF:

```
>java -cp weka.jar weka.classifiers.meta.
FilteredClassifier -l ForWekaTrain180.
model -T ForWekaTest90.csv.arff
> ForWekaTest90.out.txt
```

The FilteredClassifier/RemoveType is necessary to have the input ARFF file have a record identifier, but not allow the neural net to make inferences based on that record identifier. The record identifier allows the mapping of the results back to the original data, but since the neural net's training set and test set must have the same schema, the removal of the record identifier must be done within the WEKA command. It is RemoveType because the record identifier is a string, and it is the only string in the data.

Trip Purpose Identification Using Decision Trees

To apply the trip purpose decision tree, it is first necessary to prepare an input file containing all of the variables referenced in the model, as specified in Appendix D. This process was conducted using a PostgreSql database. A template that can be used to recreate this database's structure is included in the bundled packager. To use it, first install PostgreSql and start the service. Once you connect to the server, create a new database and load the PostGIS extension on it. More information on how to configure and use PostgreSql and PostGIS can be found at http://www.postgresql.org/docs/manuals/ and http://postgis.net/documentation, respectively.

Once the database is created, open a query window to it using a client like pgAdmin (http://www.pgadmin.org/) and then load the purpose_template.sql file and execute it. This action will create a blank database structure as well as a series of functions that can be called to prepare the data.

The database structure uses places to store trip information. Each place record has a reference to a location record, where the actual destination addresses and coordinates are stored. Similarly, household, person, and vehicle information should be entered into the supporting tables. Once the database is populated, the function prepare_data() can be invoked with the following command:

```
SELECT prepare_data();
```

This command will populate the table placestripimputevariablesints, which can be exported to CSVs and used as input to the trip purpose identification models.

The saved decision tree for estimating trip purpose is included in the file arc_agg_sample.model. This file can be opened in WEKA and used to classify trip purpose files. Use WEKA's Explorer to prepare an ARFF file with input data and then follow these steps to make the data usable by WEKA:

1. Remove columns ID, finalpersoncategory, tpurp, stpurp, distancetohome, airportdestnotflying, orig_taz, dest_taz, airport, outoftown, airportpurpose, longitude, latitude
2. Move apurp to the last column (WEKA prefers that training validation/output columns be last)
3. When preparing the validation file, remove nearschool, lu_name
4. Turn all columns into nominal values except arrhour, age, actdur, tripdistance, tottr, hhmem
5. When preparing the validation file, remove home purposes with
   a. RemoveWithValues index 99 (apurp)
   b. Nominal indexes 1,2,3,16,17 modifyheader (16 and 17 are codes 96 and 97)
6. Rename apurp to tpurp by editing the ARFF file in a text editor

This is recorded in the resulting ARFF as:

```
@relation 'arc_agg_sample-weka.filters.
unsupervised.attribute.Remove-R1,5,15-16,
21,33,49-51,53-54,111-112-weka.filters.
unsupervised.attribute.Reorder-Rfirst-
12,14-last,13-weka.filters.unsupervised.
attribute.NumericToNominal-R2,3,12,15,
16,36-V'
```

And for the validate file:

```
@relation 'arc_agg_validate-weka.
filters.unsupervised.attribute.
Remove-R1,5,15-16,21,33,49-51,
53-54,113-114-weka.filters.unsupervised.
attribute.Reorder-Rfirst-12,14-last,
13-weka.filters.unsupervised.attribute.
Remove-R96,100-weka.filters.unsupervised.
attribute.NumericToNominal-R2,3,12,15,
16,36-V-weka.filters.unsupervised.
instance.RemoveWithValues-S0.0-Clast-
L1,2,3,16,17-H'
```

The steps described here can also be done in the graphical user interface (GUI) tool but it is simpler and faster to explain them as command-line interface (CLI) statements. First make a .model file, which contains all the information about the tree. This can be done by issuing the following command:

```
>java -cp weka.jar weka.classifiers.
trees.J48 -t arc_agg_sample.csv.arff
-M 25 -x 10 -d arc_agg_sample.model
-i > arc_agg_sample.output.txt
```

This command gives WEKA an ARFF file and produces a J48 decision tree with a minimum leaf of 25 instances and 10 folds of cross-validation. The output file gives a textual description of the decision tree, its success rate on the training sample, and the confusion matrix on the training sample (which is a grid comparing model predictions with actual values). The tree can be visualized by creating a DOT file, which can be done using the following command:

```
>java -cp weka.jar weka.classifiers.
trees.J48 -t arc_agg_sample.csv.arff
-M 25 -x 10 -g > arc_agg_sample.dotty
```

The open-source program "dot" can then be used to create a flowchart using the .dotty file. The easiest way to get this program is to install "GraphViz" from www.graphviz.org. The output from the dot program will be scalable vector graphics (SVG) flowcharts, which can be opened in most modern browsers. Alternatively, WEKA's built-in visualization of trees can be used, but it does not produce as clean of an image. To obtain a SVG flowchart, issue the following command:

```
>dot.exe -Tsvg -o -Kdot arc_agg_sample.
dotty > arc_agg_sample.svg
```

Finally, the created tree can be applied to an existing data set (arc_agg_validate.csv.arff) to generate aggregate results using the following command:

```
>java -cp weka.jar weka.classifiers.
trees.J48 -T arc_agg_validate.csv.arff
-l arc_agg_sample.model > arc_agg_
validate.output.txt
```

The following command can be used to obtain a file with the predicted values for each input record:

```
>java -cp weka.jar weka.classifiers.
trees.J48 -T arc_agg_validate.csv.arff
-l arc_agg_sample.model -p 0 > arc_agg_
validate.predictions.txt
```

Methods Implemented Using Bison BIOGEME

BIOGEME is very strict about its input variables; so, to save time and to avoid major headaches, ensure that the input data contain only numbers (except for the header row) and that no empty values are present in the file. Bison BIOGEME uses a CLI; to obtain a CLI console in Windows, click on the start menu, type in cmd.exe, and hit enter. Once the console window is open, navigate to the path where the BIOGEME model file is using the console's cd command. Finally, to run Bison BIOGEME with the provided model specification file (assumes extension .mod) and, passing in the input file name (tab-delimited text file), issue the following console command:

```
> biogeme mymodel sample.dat
```

Once a satisfactory model estimation is obtained, the BIOSIM tool can be used to simulate choices using the Monte Carlo method. Several output files will be created by the program as it runs. The two most important files are an HTML document with a detailed report on the model estimation results and a .res text file, which follows the same format as the input model file, with the final estimated model.

The final model .res file can be used as input along with a tab-delimited text file containing the independent variables of the model. To do this, create a copy of the file and change its extension to .mod, then open the file and update the [SampleEnum] value to the desired number of simulated outcomes.

A GUI is also available. More information on using the GUI and BIOGEME can be found on the biogeme.epfl.ch website; a helpful tutorial is also available at http://biogeme.epfl.ch/v18/tutorialv18.pdf.

Probabilistic Travel Mode Identification

To use the included discrete choice travel mode identification model, prepare an input tab-delimited text file with the appropriate speed attributes, as described in Appendix D. The file header should contain the following columns:

```
Caseid startindex endindex minspeedmph
maxspeedmph avgspeedmph sdspeedmph
minaccelmps2 maxaccelmps2 avgaccelmps2
sdaccelmps2 distancemiles travmode
```

It is important to have all records populated with a valid value for travmode as shown in Appendix D; otherwise the record will be ignored by BIOGEME. Once the input file is assembled, an enumeration file can be generated by issuing the following command:

```
>biosim final test.dat
```

This command will produce an enumeration file (.enu) with utilities for all the choices, probabilities, and simulated outcomes.

Trip Purpose Identification

To apply the trip purpose discrete choice model, it is first necessary to prepare an input file containing all of the variables referenced in the model, as specified in Appendix D. The database procedure described previously in the section named "Trip Purpose Identification Using Decision Trees" can be used for this purpose. Once the data are prepared, BIOSIM can be used to simulate choices based on an input file. To simulate choices using the aggregate purpose model, the following command can be used:

```
> biosim agg_purpose input.dat
```

The resulting enumeration results can then be related back to the input data for analysis.

Experiment B Instructions

This section provides instructions for applying the modeling process for identifying demographic characteristics of GPS sample data described in Experiment B. The process here assumes that the sample models estimated during the development of the modeling process are to be applied to sample data that have been generated through the trip imputation process.

Prerequisites

Obtain the 'NCHRP_GPS_Data_Reduction.exe' executable from the bundled files. Alternatively the project can be compiled from source code. This can be done using free software such as Microsoft Visual C++ express, or free open-source software such as Eclipse.

Steps

1. Generate the Trip Input File from the results of the GPS trace analysis in the format shown in the table below. There should be one record for each travel episode for each person. If no unique linkage between persons in the sample is known (i.e., no household connections), set the PERNO variable = 1, and the SAMPN variable as the unique identifier. Ideally, multiple days of input data will be used here, but a minimum of 1 day is required. Save the file as a tab-delimited text file.
2. Run the NCHRP_GPS_Data_Reduction tool and enter the inputs requested. These include the filepath to the trip input file described above, and the length of data collection for the trips in the filepath. For a one day survey enter "1."
3. After pressing Enter, the program will then run for a short time and create three output files, trip_info.xls, tour_info.xls, and person_info.xls. The tour_info and trip_info files show the trips/tours identified by the algorithm, while the person_info.xls file contains the estimated travel/tour characteristics for the individuals in the sample. This file forms the basis for further analysis.
4. From the original input file resulting from the Trip Imputation process, identify the Home Location for each individual in the sample, which will include the coordinates of the home location. Using suitable GIS software, create a Shapefile of the sample home locations and perform an overlay analysis with Census Tract Shapefiles from TIGER Line or other sources to identify the home Census Tract. This can be approximated without GIS by calculating straightline distances from the home location coordinates to Census Tract centroids and assigning the sampled individual to the nearest tract. Add the home census tract for each individual to the person_info.xls file.
5. For each census tract in the study area, create the following variables using the Census Transportation Planning Package data:

Variable	Description
transituse	% of residents in tract using transit
road_density	length of roads in CT / area (miles/sq mile)
intersection_density	intersections / area (#/sq mile)
block_size	avg block size (road density/intersection density)
employment_density	employees per sq mile
pop_density	population per sq mile

6. Join the land use variables to the person_info.xls file using the Census Tract ID.
7. First, apply the four education models conditional on the work status to the sample (full-time worker, part-time worker, retiree, and other – students and children are excluded). This gives the conditional log-sum values to be used in the upper-level work status model. To calculate the logsums for each work-status category, follow this procedure:
   a. Rename person_info.xls to person_info.dat.
   b. Select this file as the input file in the biogeme GUI.
   c. Select the "educ_<workstatus>_sim.mod" file for the workstatus for which the IV is being estimated, as the "model" input in BIOGEME GUI. Mod files are included in the Report appendix and on the GitHub repository.
   d. Alternatively, steps b and c can be input at once using BIOGEME command line.
   e. Press "Simulate" in BIOGEME. This will result in the creation of a *.enu file, which contains the utility estimates for each education alternative.
   f. Calculate the inclusive value (IV) to use in work status model from the utility estimates: IV_<workstatus> = $\ln \sum_i e^{V_i}$, where utility V_i is given in the *.enu file for each alternative.
   g. Append the education IV to the person_info.dat file for each sample.
   h. Repeat for each work status.

Data Requirements for Tour Identification

Variable	Data Type	Description
SAMPN	Integer	Unique Identifier of Person (or Household if HH level analysis)
PERNO	Integer	Identifier of Person in Household (if HH level analysis), otherwise set to 1
PLANO	Integer	Identifier of Activity, unique within SAMPN-PERNO combination
LOCATION_TYPE	String	Required Location types: 'Home, Work, School, Other'
LOCATION_ID	Integer	Location identifier - unique within SAMPN-PERNO combination
MODE	Integer 1-10	Walk=1, Bike, Drive, Pass, Transit, Paratransit, Taxi, School bus, Carpool, Other
TRPDUR	Integer	Trip duration in minutes
ACTDUR	Integer	Activity at trip end duration in minutes

8. Next, apply the upper-level work status model using BIOGEME in the same manner as described in 7 a–e, using “wkstat_simulation_FINAL.mod” to estimate the work status utilities.
9. Calculate work status probabilities for each sample using the estimated utilities in the *.enu file and the MNL formula: $P_i = e^{V_i} / \sum_j e^{V_j}$ and choose a realization of the work status using simulation. Append the work status to the person_info.dat file as a new column.
10. Estimate the education status conditional on work status as follows:
    a. Split the person_info.dat file into four temporary files based on work status (ignore child and student work status in this step.
    b. Repeat steps 7 b–e to estimate the utilities for each education alternative.
    c. Follow step 11 using the *.enu file in the previous step to estimate probabilities for each educational status and choose a realization. Append to the temporary data file.
    d. Education status values have been selected for each work-status data file; append the results back to the person_info.dat main input file. At this point the work status and education status conditional on the work status have been fully specified for each individual in the sample.
11. Estimate the gender of the sample using the process described in 7 a–e using the “gender_ALL_simulation.mod” BIOGEME model file. Generate the gender for the sample using the process described in 11 with the resulting *.enu file.
12. The person age model is an ordered logit model developed using the NLOGIT software. The probabilities for age categories can be estimated in Excel as follows:
    a. Add a column to the “person_info.dat” file called utility.
    b. Calculate the utility by multiplying all of the estimated coefficients shown in the final report by the appropriate columns and summing the results.
    c. Add five columns called cumulative probability to calculate the probability of each category.
    d. Calculate the cumulative probability of each category (0–4) using the formula $P_i = e^{\mu_i - V} / (e^{\mu_i - V} + 1)$, where $\mu_0 = 0$ and $\mu_4 = \infty$, the remaining μ-values are as shown as shown in “Ordinal Logit Model for Age Categories” table in Appendix E.
    e. Finally, calculate the probabilities for each category by subtracting the cumulative probability of the previous category (except for category 0, where the cumulative probability equals the category probability).
13. Estimate the possession of a driver’s license for the individual using the WEKA software.
    a. Download “license_model.model” from the code repository.
    b. Save a copy of the “person_info.dat” file as a csv file (using Excel or other file converter).
    c. Open WEKA (if the “Weka GUI Chooser” window appears, select “Explorer”).
    d. Under the “Preprocess” tab click “Open file . . . ” and select the csv file.
    e. In the lower left corner labeled “Attributes,” select all variables not in the list in the table below, then click the “Remove” button. It is critical that the remaining attributes shown in the window match EXACTLY the variables shown in the table, including variable names, otherwise the simulation will fail. If no “LIC” (license) variable exists, create the column and set all values equal to “NO” in the .DAT file and repeat steps a–c.
    f. Click the “Save . . . “ button at the top of the form and save as an ARFF file.
    g. Navigate to the “Classify” tab, and click the “Set..” button next to “Supplied test set” and choose the file saved in the previous step.
    h. Right click in the “Result list” area and select the “license_model.model” file previously downloaded.
    i. Right click on the model that was loaded in the “Result list” area and select “Re-evaluate model on current test set.”
    j. Right click again on the model and select “Visualize Classifier Errors”; when the window appears, click “Save.” The saved file will then contain all of the attributes as well as the column “predictedLIC,” which contains the license prediction. This column can then be joined back to the original “person_info.dat” file.
14. The process for running the Household-type joint model is nearly identical to the process for the license model. Differences in the steps are listed based on the corresponding letter.
    a. Download “household_type_j48.model” from the code repository.
    e. Create the “HHTYPE” column if necessary and default values to “H_100.”
    h. Select “household_type_j48.model.”
    j. Save the file that will contain the “predictedHHTYPE” column. Create new columns in the original “person_info.dat” file for the household size, number of vehicles, and presence of children, then populate the columns using the first, second, and third digit after the “H_” in the “predictedHHTYPE” column.

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