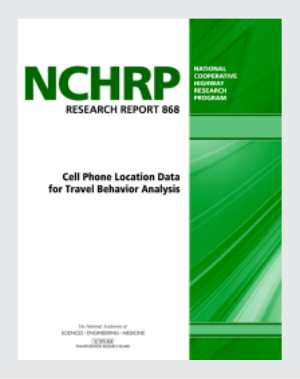


Cell Phone Location Data for Travel Behavior Analysis

DETAILS

152 pages | 8.5 x 11 | PAPERBACK
ISBN 978-0-309-39035-4 | DOI 10.17226/25189

CONTRIBUTORS

Cambridge Systematics and Massachusetts Institute of Technology; National Cooperative Highway Research Program; Transportation Research Board; National Academies of Sciences, Engineering, and Medicine

NATIONAL COOPERATIVE HIGHWAY RESEARCH PROGRAM

NCHRP RESEARCH REPORT 868

Cell Phone Location Data for Travel Behavior Analysis

Cambridge Systematics, Inc.
Chicago, IL

WITH

Massachusetts Institute of Technology
Cambridge, MA

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

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

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

TRB
TRANSPORTATION RESEARCH BOARD

2018

NATIONAL COOPERATIVE HIGHWAY RESEARCH PROGRAM

Systematic, well-designed research is the most effective way to solve 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 results in increasingly complex problems of wide interest to highway authorities. These problems are best studied through a coordinated program of cooperative research.

Recognizing this need, the leadership of the American Association of State Highway and Transportation Officials (AASHTO) in 1962 initiated an objective national highway research program using modern scientific techniques—the National Cooperative Highway Research Program (NCHRP). NCHRP is supported on a continuing basis by funds from participating member states of AASHTO and receives the full cooperation and support of the Federal Highway Administration, United States Department of Transportation.

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

The program is developed on the basis of research needs identified by chief administrators and other staff of the highway and transportation departments, by committees of AASHTO, and by the Federal Highway Administration. Topics of the highest merit are selected by the AASHTO Special Committee on Research and Innovation (R&I), and each year R&I's recommendations are proposed to the AASHTO Board of Directors and the National Academies. Research projects to address these topics are defined by NCHRP, and qualified research agencies are selected from submitted proposals. Administration and surveillance of research contracts are the responsibilities of the National Academies and TRB.

The needs for highway research are many, and NCHRP can make significant contributions to solving 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 RESEARCH REPORT 868

Project 08-95
ISSN 2572-3766 (Print)
ISSN 2572-3774 (Online)
ISBN 978-0-309-39035-4
Library of Congress Control Number 2018906086

COPYRIGHT INFORMATION

NOTICE

The research 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 National Academies of Sciences, Engineering, and Medicine.

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 Academies of Sciences, Engineering, and Medicine; or the program sponsors.

The Transportation Research Board; the National Academies of Sciences, Engineering, and Medicine; 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 research 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 by going to

http://www.national-academies.org

and then searching for TRB

Printed in the United States of America

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

The **National Academy of Sciences** was established in 1863 by an Act of Congress, signed by President Lincoln, as a private, non-governmental institution to advise the nation on issues related to science and technology. Members are elected by their peers for outstanding contributions to research. Dr. Marcia McNutt is president.

The **National Academy of Engineering** was established in 1964 under the charter of the National Academy of Sciences to bring the practices of engineering to advising the nation. Members are elected by their peers for extraordinary contributions to engineering. Dr. C. D. Mote, Jr., is president.

The **National Academy of Medicine** (formerly the Institute of Medicine) was established in 1970 under the charter of the National Academy of Sciences to advise the nation on medical and health issues. Members are elected by their peers for distinguished contributions to medicine and health. Dr. Victor J. Dzau is president.

The three Academies work together as the **National Academies of Sciences, Engineering, and Medicine** to provide independent, objective analysis and advice to the nation and conduct other activities to solve complex problems and inform public policy decisions. The National Academies also encourage education and research, recognize outstanding contributions to knowledge, and increase public understanding in matters of science, engineering, and medicine.

Learn more about the National Academies of Sciences, Engineering, and Medicine at **www.national-academies.org**.

The **Transportation Research Board** is one of seven major programs of the National Academies of Sciences, Engineering, and Medicine. The mission of the Transportation Research Board is to increase the benefits that transportation contributes to society by providing leadership in transportation innovation and progress through research and information exchange, conducted within a setting that is objective, interdisciplinary, and multimodal. The Board's varied committees, task forces, and panels annually engage about 7,000 engineers, scientists, and other transportation researchers and practitioners from the public and private sectors and academia, all of whom contribute their expertise in the public interest. The program is supported by state transportation departments, federal agencies including the component administrations of the U.S. Department of Transportation, and other organizations and individuals interested in the development of transportation.

Learn more about the Transportation Research Board at **www.TRB.org**.

COOPERATIVE RESEARCH PROGRAMS

CRP STAFF FOR NCHRP RESEARCH REPORT 868

Christopher J. Hedges, *Director, Cooperative Research Programs*
Lori L. Sundstrom, *Deputy Director, Cooperative Research Programs*
Lawrence D. Goldstein, *Senior Program Officer*
Anthony P. Avery, *Senior Program Assistant*
Eileen P. Delaney, *Director of Publications*
Natalie Barnes, *Associate Director of Publications*
Janet M. McNaughton, *Senior Editor*

NCHRP PROJECT 08-95 PANEL
Field of Transportation Planning—Area of Forecasting

Kermit W. Wies, *Northwestern University Transportation Center, Evanston, IL* (Chair)
Rebekah S. Anderson, *Ohio DOT, Columbus*
Tae-Gyu Kim, *North Carolina DOT, Raleigh*
Guy Rousseau, *Atlanta Regional Commission, GA*
Erik E. Sabina, *Colorado DOT, Denver*
Reginald R. Souleyrette, *University of Kentucky, Lexington*
Fang Yuan, *Delaware Valley Regional Planning Commission, Philadelphia, PA*
Sarah Sun, *FHWA Liaison*
Michael L. Cohen, *NAS Committee on National Statistics (CNSTAT) Liaison*
Jennifer L. Weeks, *TRB Liaison*

AUTHOR ACKNOWLEDGMENTS

In NCHRP Project 08-95, the research team evaluated the use of cell phone location data for travel behavior analysis and developed guidelines for practitioners. The team, led by Cambridge Systematics, Inc. staff Kimon Proussaloglou, Daniel Beagan, and Anurag Komanduri, analyzed travel data derived from call detail records (CDRs) and contrasted them with estimates derived from survey data and models to provide practical guidance on the value and uses of CDR data.

We want to acknowledge the key research contributions of our colleagues at the Massachusetts Institute of Technology (MIT). The case study analysis and results presented in Chapters 4 through 8 and parts of Chapter 2 were contributed by Professor Marta González and Dr. Shan Jiang and based on their original research and data analysis. The case study also reflects work undertaken in prior studies by Professor González and her research group at MIT.

Cambridge Systematics developed the background approach and set the stage for the analysis and the guidelines by adopting a practitioner's perspective in the summary and Chapters 1 through 3. We contributed to the literature review and added our insights to the case study discussed in Chapters 4 through 8 through the lens of a practitioner's perspective focusing on the inference of stay activity locations and the practical comparisons presented. Finally, we developed Chapter 9 to distill the findings and develop practical guidance for transportation practitioners. This concluding chapter discusses administrative, data-related, and modeling considerations for using cell phone data and presents recent research aimed at improving the industry's best practices.

FOREWORD

By Lawrence D. Goldstein
Staff Officer
Transportation Research Board

NCHRP Research Report 868: Cell Phone Location Data for Travel Behavior Analysis presents guidelines for transportation planners and travel modelers on how to (1) evaluate the extent to which cell phone location data and associated products accurately depict travel; (2) identify whether and how these extensive data resources can be used to improve understanding of travel characteristics and the ability to model travel patterns and behavior more effectively; and (3) support practitioners' evaluation of the strengths and weaknesses of anonymized call detail record locations from cell phone data.

The report includes guidelines for transportation practitioners and agency staff with a vested interest in developing and applying new methods of capturing travel data from cell phones to enhance travel models. This is an emerging field of interest subject to complexities linked to acquiring data and applying these data while maintaining privacy in a complex legal and practical framework. The emergence of these data constitutes a significant opportunity for change in the travel modeling community, with access to detail and volume not previously available. A better understanding of the strengths and weaknesses of these data is an important step in this direction.

Information on billions of locations is generated every day from mobile devices. Over the past decade, cell phone location data have become commercially available for transportation planning purposes. Mobile signaling can provide a detailed picture showing how people move throughout the day. Cell phone location data used and analyzed in this study correspond to "call detail records," which include location information every time a call is made or answered, a text message is sent or received, or the Internet is accessed.

Call detail records from cell phone location data offer the potential to provide information about activity location, frequency of repeated travel, travel outside of a study area, and origin–destination data for travel to special events. This information can be used to model, evaluate, and analyze the flow patterns of both residents and visitors in a given study area. With the emergence of large amounts of data, research is needed to explore and evaluate methods used for processing cell phone location data to generate travel behavior information and provide guidelines for the use of the information by transportation planning practitioners.

In tackling this problem, the research team led by Cambridge Systematics, Inc., with support from the Massachusetts Institute of Technology, had to address a broad range of significant questions to evaluate and test opportunities for use of cell phone data. The study approach used travel in the Boston, Massachusetts, region as a case study to compare and contrast traditional travel survey data and regional models with Census data and with

cell phone-derived data describing regional travel characteristics. The questions that were addressed included the following:

- What are the best options for using cell phone data for travel estimates in support of modeling techniques?
- What are the strengths and weaknesses of cell phone data that can support travel behavior analysis and policy decision making?
- How can cell phone data be used to enhance access to information on travel behavior characteristics necessary for effective model applications?
- What tools and techniques are available for collecting and analyzing cell phone data?
- How can travel modelers overcome practical and legal problems associated with data acquisition, and how can this acquisition process respond to privacy requirements?

These questions represent only a portion of the detailed considerations required to evaluate the potential application of cell phone data in improving travel behavior modeling as input to overall transportation planning efforts.

The first three chapters of *NCHRP Research Report 868* set the stage for the analysis and guidelines for practitioners by adopting a practitioner's perspective and discussing how the strengths and weaknesses of cell phone data are likely to influence planning for transportation projects. Chapters 4 through 6 provide an in-depth discussion of the types of data available and procedures that can be used to apply these data to primary issues affecting travel modeling. The report describes data available; how call detail record data are analyzed to extract daily trajectories; steps involved in identifying activity types encompassing home, work, and "other"; how to derive trip purpose by time of day; and, finally, methods used to develop trip tables using cell phone data. Chapters 7 and 8 present a case study approach to compare the inferred trip tables extracted from the cell detail record data with trip tables from Boston, household travel surveys, and the Boston regional travel demand model. In Chapter 9, the report concludes with guidelines for practitioners, summarizing key administrative, data-related, and modeling considerations about the potential uses of cell phone data and applications by planning and modeling practitioners.

CONTENTS

SUMMARY

Cell Phone Location Data for Travel Behavior Analysis

The objectives of this research were (*a*) to evaluate the extent to which cell phone data accurately reflect daily travel and (*b*) to develop guidelines on how to best use these data to understand and model travel behavior. Given the interest in new sources of locational data, it is critical to evaluate the strengths and weaknesses of this new stream of data and its applicability to the development and application of travel demand models.

This report is for transportation practitioners and agency staff interested in the value of and potential applications for cell phone–derived travel data. Other interested parties may include transportation modelers and planners and their peers in metropolitan planning organizations (MPOs), state departments of transportation (DOTs), and federal agencies. The research provides insights on the strengths and weaknesses of cell phone–derived travel data and provides advice on how best to harvest its value.

The Behavioral Paradigm and Disruptive Technologies

New and innovative approaches to data and analytical methods may be indicative of disruptive technologies that provide opportunities and challenges to an industry. The emergence of locational data represents such a disruptive technology. The availability of new and detailed sources of vast amounts of travel data may signal a shift in how transportation planning and travel demand modeling will be conducted in the future. Although these data sources will affect how well society understands the present and how it should plan for the future, there are competing hypotheses about how these data will affect thinking and practices.

One hypothesis is that these data streams are so powerful and detailed that they can replace the traditional approach to modeling, either in its entirety or for key components of the model system. Another hypothesis is that these new data streams will be harnessed to offer valuable inputs to understanding today's observed travel patterns, both for traditional and for new forms of transportation.

The research team approached these new streams of data as unique opportunities to get better snapshots of travel patterns and to enhance its understanding of the travel behavior of today's drivers. The team believes that these new forms of data will help the evolution of models within a travel behavior framework to help create methodological advancements that will more accurately reflect today's travel patterns and will allow practitioners to evaluate what-if scenarios for more nuanced policy decision-making needs today and in the future.

The question is how these new data sources will become part of planning and modeling practice over the next 10 years. Locational data in an aggregate form offer opportunities to support validation and allow for easier and more frequent model updates. Locational data

in a disaggregate form hold the promise of a better understanding of the drivers of demand for both traditional modes and the emerging sharing economy modes.

These new locational data sources are expected to disrupt the way practitioners think about travel, and these sources can augment traditional data collection methods as long as sampling issues and biases are understood and addressed. The research team believes in the concept of creative collaboration, in which new locational data sources can augment and enhance thoughtful, behaviorally based approaches to planning and modeling.

The research team believes that the value of new data sources will be enhanced if insights from locational data are incorporated into transportation practitioners' understanding of the factors driving travel behavior. Approaches that integrate these new data streams into the behavioral paradigm for travel can fill gaps in existing models and allow for policy-sensitive and sophisticated approaches to answer today's nuanced policy questions.

The individual traveler and the array of travel choices that this individual makes, coupled with the choices made by other members of the individual's household, remain at the center of the behavioral paradigm describing daily travel. The key question is whether and how a better understanding of travel patterns for traditional and new modes can be gained from new sources of locational data. The current behavioral paradigm and modeling framework can benefit from new data to (*a*) become more policy-sensitive and (*b*) forecast traveler choices and travel patterns with greater confidence.

The Types of Questions to Ask

To meet the challenge of interpreting and applying these data, the research team used a case study approach that relied on data and models from Boston. The research team compared and contrasted traditional survey data and models with U.S. Census data and with cell phone–derived estimates for regional travel. The following questions were explored:

- How can cell phone data best be used to derive travel estimates and support modeling analyses? Is a behavioral approach to modeling still important and relevant, or can cell phone estimates of travel replace the traditional approach?
- Which aspects of the behavioral approach to travel demand modeling can be supported by cell phone data? Which elements of model estimation and model validation can benefit the most from these data?
- Is the existing paradigm of travel behavior analysis applicable to this new stream of data? Are new methods required to harvest the value of cell phone data? Are current tools adequate to analyze the new stream of data?
- Do cell phone data affect how survey data collection is approached? Does this new technology offer an alternative standalone option for data collection, or can it augment existing methods of data collection?
- What are the expansion methods used to arrive at a representative sample? How do these methods compare with the industry's state of the practice, and what information is needed to assess the representativeness of the sample?
- Are the assumptions needed to analyze this new stream of data understood? What biases need to be considered, and how should they be addressed? What aspects of the "black box" need to be better understood?

As traditional data and models are contrasted with travel estimates from cell phone data, this report will address the trade-offs between sample sizes, the inferences and assumptions that need to be made, and the richness of information about travel and socioeconomics that characterizes each method.

Concept of the Black Box

The complexity of travel demand models can lead the public and decision makers to think of models as a black box whose internal structure and underlying assumptions are not known or clearly understood. The modeling community has made efforts over the years to better document, explain, and communicate data, methodologies, and assumptions to decision makers and to the public. A similar level of transparency is critical when new methodologies and assumptions are used to analyze cell phone data to support planning decisions and build travel demand models.

The Research Approach and the Comparisons

Transportation agencies have either purchased or evaluated the purchase of products on the basis of anonymized and aggregate cell phone data in order to supplement, enhance, or replace traditional data. Practitioners need to be clear about the terminology used to describe and analyze cell phone call detail records (CDRs) and to understand the methods used to translate locational data into estimates of travel flows. Key to this discussion is a better understanding of how cell phone CDR data and other forms of big data compare with traditional forms of data and models that are commonly available to transportation planners.

The policy needs that planning agencies face will determine how cell phone data may be used to support these analyses. Interviews of agency staff are summarized to document and understand these policy needs. A checklist has been developed for planners to use when thinking about the uses of CDR data, data strengths and weaknesses, and planners' efforts to understand products that may appear as black boxes. The potential utility of cell phone data focuses on the data as a source of travel information, as an input for estimating travel demand matrices, and as a substitute for different modeling components.

The research approach used to shed light into the black box was based on a case study from Boston that allowed the research team to make multiway comparisons by using similar measures of travel from CDR data, traditional household surveys, and regional model results. These transparent multiway comparisons were used to highlight the strengths and weaknesses of each source of data. The difficulty of determining, with certainty, what constitutes ground truth estimates is also discussed.

The research discussed in the case study is unique, in that it used raw CDR data to develop estimates of activities, their location, and the time of day when they happened. These estimates are rolled up to the regional level to produce metrics readily comparable to traditional estimates of travel developed from household surveys and used in regional models. The transparency of the research assumptions, the open discussion of the strengths and weaknesses of the underlying data and methods, and the step-by-step methodological insights are useful in understanding the properties of similar products in today's marketplace.

The results are presented in detail and followed by a brief technical summary of key points in each chapter. The analysis of CDR data suggests their potential for (*a*) supporting frequent and targeted data collection of travel patterns for external, long-distance, and special events and visitor travel; (*b*) identifying seasonal and year-to-year variations in travel; and (*c*) providing a means of assessing trends in regional travel. Although CDR data cannot be used in traditional travel demand models, these data can be used as an additional source

for model validation, to help drive model updates for intermediate years, and to support analyses for model components (e.g., long-distance travel, special generators, visitor travel, special events, and, potentially, corridor studies).

Guidelines

This report concludes with a chapter on the issues practitioners typically consider about data and models in their daily work. The research team reframes these considerations for CDR locational data to help agency staff assess the potential value of these data. The practitioner guidelines are grouped into three categories: administrative considerations, data considerations that staff face, and modeling-related issues that practitioners need to address as they assess the potential value of CDR data.

In an epilogue, the research team recommends that practitioners apply the following principles as they evaluate their policy needs, data options, and modeling tools:

- Be aware of the assumptions made in processing cell phone data to determine locations and infer activities and purposes.
- Recognize that results from traditional surveys and models are built on different sets of assumptions and that ground truth is difficult to establish.
- Expect that increases in the quantity of CDR data, improvements in signal and CDR data quality, and the use of machine-learning algorithms are likely to improve methods for analyzing locational data and inferring travel patterns.
- Appreciate the uncertainty underlying CDR estimates and traditional data and measures of travel patterns.
- Use the conceptual framework based on the behavioral paradigm that examines individuals' travel behavior as a guide.

The research indicates that, together, collective industry experience, academic research over the years, and a collaborative approach linking research to practice have helped refine the data design process, spawn new and more sophisticated analytical methods, and increase understanding of travel behavior and its drivers. As new data and methods are introduced, the interpretation of their value and uses through a behavioral framework lens will help improve the state of the art in travel demand forecasting. Major opportunities will likely emerge to harness new ideas, data, and methods to shape innovative practices that have long-term potential to benefit the transportation research community.

CHAPTER 1

Roadmap to the Report

This report presents cell phone locational data and its use in understanding travel behavior, evaluates the extent to which cell phone data can be used to accurately reflect daily travel, and offers guidelines for planners using these data to understand and model travel behavior. To provide the proper context, the research team compared and contrasted traditional survey data and models with U.S. Census data on regional travel and with cell phone–derived estimates for regional travel in Boston, Massachusetts.

The report can be read in accordance with the background and interests of the reader. The overall value of the data and potential uses are presented at the beginning and the end of the report. Each chapter begins with a roadmap and ends with conclusions highlighting the key points.

Chapters 2 and 3 present a **policy discussion** and the view of cell phone data from a planner's perspective. This discussion provides background on how available cell phone data and the ongoing research are likely to influence planning for transportation projects in the near future:

- Chapter 2 sets the stage for the research and analyses presented, the nature and potential value of cell phone data, and the recommendations that are part of the practitioner guidelines. The research team discusses the need to evaluate the strengths and weaknesses of traditional data and models before focusing on this new stream of data and its applicability to the development and application of a travel demand model. The research team poses questions of interest, discusses current practices regarding survey data and traditional or advanced models, and then introduces cell phone data as a disruptive technology. The chapter concludes with a discussion of the challenge of using "big data" for different aspects of the traditional modeling stream.
- Chapter 3 provides an in-depth look at cell phone data and compares it with the needs that transportation planners have to address. The chapter starts with a definition of call detail records (CDRs) that are at the core of most available industry products, discusses how traces from cell phone data are analyzed to provide locations and to infer activities, and focuses on the assumptions made to draw inferences. How the dynamic nature of cell phone data as technology continues to evolve is also noted.

  The chapter then focuses on the need to measure travel today and to forecast travel in the future with a discussion of how cell phone data can augment or validate existing models and tools. The research team comments on the need for a behavioral mechanism to organize these data. Then the views of transportation planners on cell phone data and the potential value and uses of these data are summarized. The current marketplace for cell phone data products, the planners' need to understand the underlying assumptions, and the types of questions practitioners are likely to ask when using these products are outlined. The chapter concludes with a presentation of how travel measures produced by traditional models and survey data will be compared with measures of travel that were developed on the basis of the analysis of cell phone data.

Concept of the Black Box

The complexity of travel demand models can lead the public and decision makers to think of models as a black box whose internal structure and underlying assumptions are not known or clearly understood. The modeling community has made efforts over the years to better document, explain, and communicate data, methodologies, and assumptions to decision makers and to the public. A similar level of transparency is critical when new methodologies and assumptions are used to analyze cell phone data to support planning decisions and build travel demand models.

Chapters 4 through 8 provide an in-depth **technical discussion** of cell phone data, inference of locations and activity types, development of origin–destination (O-D) matrices, and comparisons with survey data to evaluate the robustness of the methods and results. These chapters provide a unique glimpse into the "black box" where cell phone traces are translated into travel patterns.

The report also presents a case study from the Boston area that provides a unique perspective on how cell phone data are analyzed to develop outputs consistent with traditional model results and presents the strengths and weaknesses of such data. The reader interested in an in-depth literature review, the evolution of research in this area, and practical multiway comparisons of cell phone–derived measures of travel with traditional model output will find Chapters 4 through 8 of particular interest.

- Chapter 4 provides an overview of the raw data and describes the range of spatial and temporal resolutions of cell phone data. It demonstrates the massive and passive nature of raw cell phone data and explains in detail their spatial and temporal characteristics. The analysis relies on CDR data from 2 million cell phones collected over 2 months in the Boston region.
- Chapter 5 presents methods to extract stay locations where individuals conduct daily activities as anchor points. The spatial and temporal patterns of extracted stay locations for the Boston region are shown.
- Chapter 6 focuses on methods, results, and validation to label activity types for "home," "work," and "other" stay locations. These activities provide the cornerstone for the estimation of O-D trip matrices. Also discussed are the factors developed to expand the data from cell phone users to the metropolitan population. The expanded home and work activity estimates are compared with journey-to-work travel data.
- Chapter 7 presents methods for estimating O-D matrices by purpose and by time of day on the basis of identified activity types and expansion factors from cell phone data and Census population data. The results discussed are analogous to outputs of travel demand models and include estimates of trip generation and trip distribution.
- Chapter 8 compares the O-D flows estimated from raw cell phone data for 2 months in 2010 with other sources of travel estimates. These sources include the 2009 National Household Travel Survey; the 2011 Massachusetts Travel Survey; the 1991 Boston Household Travel Survey; the 2007 and 2010 versions of the regional Boston model maintained by the Central Transportation Planning Staff; and a 2015 third-party proprietary data set purchased from a CDR data vendor.

Chapter 9 provides guidelines for planners that summarize key issues that analysts consider and the potential uses of cell phone data. The discussion combines the policy context with the technical analyses to provide guidance to transportation planners and modelers who evaluate different sources of data and models. The chapter also discusses recent trends in cell phone research, including data collection technologies that combine traditional diary-based methods with the power of locational data reflecting cell phone use. Chapter 9 should be of interest to a broad spectrum of transportation practitioners.

The report includes a glossary of terms for this new subject area and references current at the time of writing (late 2016).

CHAPTER 2

Travel Behavior from Cell Phone Data

2.1 Research Objectives

The objectives of the research were (*a*) to evaluate the extent to which cell phone data accurately reflect daily travel and (*b*) to develop guidelines on how to best use such data to understand and model travel behavior. Achieving these objectives depended on evaluating the strengths and weaknesses of this new stream of data and its applicability to the development and application of travel demand models.

This report provides support to transportation practitioners, agency staff, and researchers interested in cell phone–derived travel data. Other interested parties may include transportation modelers and planners and their peers in metropolitan planning organizations (MPOs), state departments of transportation (DOTs), and federal agencies. The research provides insights on strengths and weaknesses of cell phone data and provides advice on how best to harvest their value.

To meet the challenge of collecting, interpreting, and applying these data, the research team used a case study approach that relied on data from Boston, Massachusetts. Traditional survey data and models were compared and contrasted with Census data and with cell phone–derived estimates for regional travel. Specifically, the following questions were explored:

- How can cell phone data best be used to derive travel estimates and to support modeling analyses? Is a behavioral approach to modeling still important and relevant, or can cell phone estimates of travel replace the traditional approach?
- Which aspects of the behavioral-based approach to travel demand modeling can be supported by cell phone data? Which elements of model estimation and model validation can benefit the most from these data?
- Is the existing model of travel behavior analysis applicable to this new stream of data? Are new methods required to harvest the value of cell phone data? Are current tools adequate for analyzing the new stream of data?
- Do cell phone data affect the way the collection of survey data is approached? Does this new technology offer an alternative stand-alone data collection option, or can it augment existing methods of data collection?
- What additional information is needed to ensure that travel estimates are representative of the population as a whole?

In summary, as the research team compared traditional data and models with travel estimates from cell phone data, it considered trade-offs between sample sizes, inferences needed for the analysis required, and richness of information about travel and socioeconomic characteristics of each method.

2.2 Current Practice: Data

The current practice of regional model development and implementation has its origins in the early 1960s with major modeling efforts such as those in Chicago, Illinois, and Detroit, Michigan. The overall approach is rooted in the collection of survey data and their analysis with various statistical modeling techniques. Over time, emphasis on a higher degree of disaggregation has resulted in regional models of greater detail, increased sophistication, and enhanced policy sensitivity.

Household travel surveys have traditionally been used to collect travel behavior data from a sample of a region's households and have been combined with regionwide estimates of population and employment to quantify travel demand. Such surveys are often supplemented by onboard, special-generator, and workplace surveys that focus on different modes, subpopulations of users, and geographies of interest. Other unobtrusive methods of data collection (such as traffic counts and transit ridership estimates) provide a snapshot of travel demand by mode. Measures of transportation supply and indices of system performance are reflected in the highway and transit networks and measures of travel times.

Household travel surveys collect information on the attributes of household members and characteristics of their daily travel and help determine the explanatory variables used in four-step models or activity-based models. Typical information collected includes the times that trips are made, the activities connected by those trips, the number of trips by trip purpose, the mode(s) used, travel distance and time, the number of people traveling together, and travel costs. Such surveys also include socioeconomic information about the household and individual members. Typical data included household income, household size, household life cycle, automobile availability, number of workers, home and work locations, and individual socioeconomic characteristics.

Surveys typically ask for one designated travel day, although some recent efforts have recognized variability within a week and expanded the number of days for which a diary needs to be completed. Households are recruited to complete diaries of the activities and travel on a given travel day by each member of the household.

Typical sampling rates are about 1% and represent a sample of the universe of all trips within the region for which a travel demand model is estimated. Households are weighted so that their travel is statistically expanded to represent the universe of total trips in a region.

Although every effort is made to minimize biases inherent in the sample selection, survey design, and data collection process, certain types of trips may not be fully captured by the survey. Types of trips that may be underreported include short-distance travel, trips by nonmotorized modes, and travel by younger or lower-income respondents who may be clustered in specific areas of a region.

2.3 Survey Data: Strengths and Weaknesses

The increasing sophistication of travel demand models requires detailed, high-quality input data for model development, calibration, and validation. Data requirements are not limited to detailed travel behavior from household surveys but also include data on transportation networks, highway capacities, levels of highway and transit service, and detailed land use, population, and employment data.

Survey diaries are used to record the trips linking activities to locations at specific times of day. Survey data from a regional sample are expanded and analyzed to develop travel models that represent the universe of trips in a region. These models combine methods of statistical sampling in local (Daganzo 1980, Smith 1979) and national household travel surveys (Stopher

and Greaves 2007, Richardson et al. 1995) to process and infer travel at different levels of detail, including cities, regions, interregional corridors, states, and nations.

The cost of household surveys varies but generally ranges between $150 and $300 per completed household, depending on the mode of administration and the technology used. In part because of the survey costs, data collection is often limited to 1% of regional households. The number of completed surveys determined to be needed for the effort is typically the lowest number that can support statistically significant behavioral choice models and produce an adequate sample size for distinct market segments in the region (often at an aggregate level). Low response, nonresponse, and differential response rates are frequent considerations in survey data collection.

Most regions also conduct travel behavior surveys infrequently. It is generally considered good practice to conduct a survey roughly every 10 years to (*a*) capture changes in socioeconomics, development, and travel patterns and (*b*) improve and update the underlying travel demand model. For example, the last two regional household travel surveys in Boston were conducted about 20 years apart—in 1991 and then next in 2010 and 2011 (Massachusetts DOT 2012).

The strengths of the traditional survey approach are the detailed representation of trips and activities at the household and individual traveler levels and the direct tie to the socioeconomic characteristics of trip makers. Trip ends, activities, purposes, modes used, time of day of travel, and socioeconomic characteristics provide unique depth in reflecting a typical day's travel patterns.

The sample drawn is designed to reach a representative cross section of the population, and special care is taken to obtain responses from hard-to-reach segments. Finally, the sample is weighted and expanded by market segment and geography to control for nonresponse patterns and thus arrive at a representative population for the region.

2.4 Current Practice: Models

Travel demand models have played an essential role in managing existing transportation systems and in planning for future development (Manheim 1979, Ben-Akiva and Lerman 1985, Ortúzar and Willumsen 2011). Widely applied models in the transportation planning domain fall into two main categories: traditional four-step models (McNally 2008), and newer activity-based models (Bowman and Ben-Akiva 2001, Castiglione et al. 2015).

Traditional models include the sequential steps of trip generation, trip distribution, mode choice, and trip assignment. These models rely on household travel surveys to generate total travel and travel patterns by purpose. The generated trips are distributed to different destinations, creating origin–destination (O-D) matrices allocated to competing modes and assigned to highway and transit networks. Variants of the traditional modeling approach use different levels of disaggregation in estimation and application and may address differences in travel behavior by time of day and by market segment. Other variants include models that use feedback loops to reflect how changes in travel times influence regional travel and decisions on route and mode choice.

Activity-based models explicitly consider travel as a derived demand in pursuit of activities. These models adopt a more disaggregate framework that incorporates interaction between activities and travel and recognizes interactions between household members. These models also rely on household travel surveys and more detailed time-of-day travel data to construct an entire sequence of activities during a typical day. Surveys are analyzed to model activity episode generation and scheduling processes (Bhat and Koppelman 1999).

Demand estimation outputs from both traditional and activity-based models are crucial for understanding the use of transportation infrastructure and planning for its future. They are used

to develop transportation plans, conduct environmental impact studies, and support infrastructure investment and prioritization decisions (Beimborn and Kennedy 1996, Van Zuylen and Willumsen 1980, Spiess 1987, Maher 1983, Lo et al. 1996, Hazelton 2003, Lu et al. 2013, Cascetta 1984, Bell 1991).

This report discusses the strengths and weaknesses of cell phone data and how such data can best be incorporated into regional models and analysis. The underlying behavior paradigm framework that has guided research and applications in the transportation field is used to evaluate the cell phone data. A systems approach to analyzing observed transportation flows is essential to understanding the links between the underlying need to participate in activities located elsewhere and the travel flows observed.

To a large extent, this behavioral underpinning is guiding the research and analysis of cell phone data. Researchers seek to better understand and quantify how the observed movements of cell phone devices can provide insights about the location of cell phone users' home, work, and "other" activities and how their patterns of cell phone use can be translated into travel flows.

2.5 Cell Phones: Sensors for Data Collection

Sources of urban sensing data and the high penetration of telecommunications in modern societies have transformed cities into repositories for exabytes of digital traces of human activities with fine-grained spatial and temporal information. The pervasive use of cell phones has generated a wealth of data that can be analyzed to reveal travel patterns and flows. Such data present new possibilities for urban transportation planners to examine social–technological ecology in cities (Jiang 2015).

The ubiquity of mobile devices (including cell phones and tablets), accompanied by rapidly advancing mobile computing technology, has made mobile devices increasingly effective sensors of individuals' daily whereabouts (Lane et al. 2010). The 6 billion cell phones in use almost triples the number of Internet users. High penetration rates of cell phones are routine in the developed world and sometimes exceed one cell phone per person (e.g., 104% in the United States and 128% in Europe), while penetration rates of more than 85% are observed in developing countries (GSMA 2011, International Telecommunication Union 2014).

Mobile devices and apps that run on them passively record users' social and mobility behaviors with high spatial and temporal resolution (Toole 2015). With the increasing use of cell phones, each individual generates tens to hundreds of traces daily, and this number is only likely to increase. Through specific agreements or through open-data challenges (de Montjoye et al. 2014), location data on millions of users have been made available to researchers and used to augment traditional travel surveys.

These data sets offer digital footprints at a scale and resolution that cannot be captured by typical travel behavior surveys, which record a few travel days for a sample of households in a metropolitan area. Call detail records (CDRs) are automatically collected by cell phone service providers for billing purposes and contain time-stamped coordinates of anonymized customers every time the customer uses his or her phone in a cellular network. The location and time data provide rich spatial and temporal information about human mobility patterns. These data can be gathered more often and at a much larger scale than traditional travel survey data.

The volume of CDR cell phone data is massive from cross-sectional and longitudinal perspectives. As a result, such data can provide wider geographic coverage and a longer time horizon.

With various degrees of data privacy protection in place, human activities can be observed over longer periods and on a large scale.

Recent work has also found that individuals are generally predictable, unique, and slow to explore new places (González et al. 2008, Brockmann et al. 2006, Song et al. 2010a, Candia et al. 2008, Calabrese et al. 2013, Jiang et al. 2013, Jiang 2015). The availability of similar data nearly anywhere in the world has facilitated comparative studies that show that many of these properties hold across the globe, despite differences in culture, socioeconomic variables, and geography.

2.6 Cell Data: Strengths and Weaknesses

Traditional survey-based methods of collecting data on traveler behavior are becoming costlier, and surveys in a region are often collected 10 years or more apart. Although detailed data on daily travel are collected, the sample of the population is relatively small and the travel information can become dated in a rapidly changing world. On the other hand, cell phones that travelers use daily passively collect a wealth of locational and time-of-day information that can be translated to travel data.

Cell phone–derived data can provide some of the typical outputs of a regional travel demand model. The interpretation of this new stream of data will require development of new analysis tools and different ways to infer total travel, travel by purpose, destinations visited, and modes used, as well as heuristics to translate raw cellular location data into travel volumes. How data and modeling are approached in assessing travel demand will change radically.

Locational data are collected from individuals passively as incidental outputs from daily use of phones for calls, text messages, and data use. Because network operators are prevented by privacy considerations from providing identifying customer information, several challenges are created from the point of view of traditional analysis:

- It is difficult to identify how many trips are unreported because the owner may not have carried the cell phone during some or all of his or her travel; the cell phone may not have been heavily used for calls, texts, or Internet data access; or the cell phone device did not pick up a signal.
- Given that the observation unit is a device, the analyst cannot distinguish between a traveler with multiple devices, a single device used by multiple travelers, or multiple travelers with multiple devices traveling together.
- Traditional market segmentation is not feasible without socioeconomic data.
- The purpose of each trip needs to be inferred, and it is difficult to determine the exact origin or destination land use, especially in a mixed land use scenario.
- The locational observations do not provide information about the mode in which the cell phone user was traveling or the size of the traveling party.
- Heuristic algorithms specific to every region need to be developed to identify true activity stops.

However, if these observations can be obtained for a long period, information about how the same device (traveler) makes multiple trips may be used to construct less invasive travel "diaries" over an extended period. Users who travel regularly to unique work, medical, or shopping locations can be observed over time and provide good travel data.

As compared with diary data collected for a small sample during a limited period, a longer-term observational approach may yield a better travel and activity data set for at least some aspects of daily travel.

2.7 Inferring Trip Ends and Activities

The traditional approach to travel demand modeling relies on developing analytical procedures and making inferences that allow the use of a small survey sample of daily travel to represent the daily travel in a metropolitan region. Given that cell phone data do not record trips, new analytical procedures are needed to make inferences about activities and travel. CDR data can be used to infer trip ends on the basis of the locational and temporal pattern of a sequence of CDRs.

The benefits of cell phone data have been realized in various contexts, such as the spread of disease (Belik et al. 2011, Wesolowski et al. 2012) and population movement (Lu et al. 2012). Methods of analyzing cell phone data for travel and activity behaviors need to be evaluated, documented, and shared with practitioners before such methods can be widely adopted in transportation planning.

The following assumptions need to be made to develop a sequence of trips and activities:

- A rule for inferring a trip end needs to be developed. Such a rule may postulate a trip end if the device did not move, for example, more than 5 meters in 5 minutes.
- The activity that occurs at a trip end can also be inferred. A rule may link the number of times that the same location was observed for a given device over a long period and over repeated observations. The rule may infer that
  - The trip end observed most often during evenings is home,
  - The trip end observed most often during weekday daytime is work, and
  - All other trip ends serve "other" activities.

These simple inferences may produce problems under different scenarios:

- A device that makes or receives no transmission before, during, or after traveling does not generate data. Given that there is no trace of the individual traveler, no information about such trips and activities can be inferred.
- Incorrect inferences will be made if the cell phone is used by someone who (*a*) works a graveyard shift; (*b*) is retired, unemployed, or a student; or (*c*) works from home or telecommutes most of the time.
- Incorrect inferences may also be made if the device that traces the data is used by different people on different days.

However, these differences may not be significant if, at an aggregate level, the expanded results are comparable to expanded trips from household surveys or from regional models.

Finally, the use of CDR data to infer trip ends and activities requires that a sufficient number of CDRs (in the form of calls, texts, or Internet data access) have actually been recorded for a device over a period.

2.8 Inferring Travel Flows

The inference of trip ends and activities is the first step in developing a trip table at a city or regional level. However, location inferences are approximate and the CDR data do not include information on purpose, mode, and socioeconomics. The following data challenges and differences from traditional approaches need to be understood before CDR data are used for detailed analyses:

1. Cell phone data lack the individual and household **socioeconomic characteristics, purposes, modes,** and **travel costs** available from travel surveys.
2. Despite the advantage in data size and lower cost, CDR data contain passive traces of a user at **approximated locations** when a phone connects with cellular networks that provide an inexact, incomplete picture of daily travel.

3. **Mode inference** is difficult. Unlike GPS data, CDR data are sparse in space and time. Therefore, it is not feasible to identify travel mode by relying only on CDR data.

New and innovative methods for extracting meaningful spatial and temporal information from the massive but noisy raw data must be developed before CDR data are used to model travel demand.

Pioneering research has used cell phone data to capture distinct trip-making patterns pertinent for transportation planning applications:

- At the regional level, daily trip chains, trajectories, and activity patterns constructed from cell phone data were found to be consistent with household surveys (Schneider et al. 2013, Jiang et al. 2013, Widhalm et al. 2015).
- Road use patterns inferred from CDR data have been validated by comparison with GPS speed data and road assignment results from travel demand models (Wang et al. 2012, Huntsinger and Donnelly 2014).
- CDR data have been used to infer realistic, cost-effective O-D matrices (Alexander et al. 2015, Colak et al. 2015, Iqbal et al. 2014, Toole et al. 2015) as compared with conventional approaches that rely on travel surveys or traffic counts (Spiess 1987, Cascetta 1984, Bell 1991, Yang et al. 1992).

Chapters 3 through 7 take the preceding considerations into account in explaining how state-of-the-practice research on cell phone data can be used to estimate and validate travel demand models similar to those used by the modeling community. The research team reviews the recent literature that describes how massive, passive, and noisy raw cell phone data can be parsed, filtered, synthesized, and analyzed to extract O-D matrices.

The research team also uses the Boston region as a case study to compare travel demand estimation results based on cell phone data with results from traditional survey data and transportation models. The research presented demonstrates how the movements of cell phone devices can provide insights about the location of cell phone users' home, work, and "other" activities and how patterns of cell phone use can be translated into travel flows.

CHAPTER 3

A Planner's View of Cell Phone Data

Transportation agencies across the country have either purchased or are evaluating the purchase of anonymized and aggregate cell phone data to supplement, enhance, or even replace traditional data sources in support of planning and modeling projects. Agency staff and practitioners need to be clear about the terminology used to describe cell phone data derived from call detail records (CDRs). It is even more important for them to understand the methods used to translate locational data into estimates of travel flows and generate the final data sets used to support transportation planning. This chapter has four major sections:

- Section 3.1 focuses on big data and cell phone CDR data, especially as they pertain to data commonly available to transportation planners.
- Section 3.2 focuses on the policy needs of planning agencies and on how cell phone data may be used to support the analyses required by these policy needs. Interviews of agency staff are summarized so readers can gain from their perspectives. This section includes a checklist for planners to use when thinking about the uses of CDR data, the strengths and weaknesses of these data, and efforts to open up vendor products that may appear to be a black box (see Box 1-1).
- Section 3.3 focuses on the utility of cell phone data as (*a*) a source of travel information, (*b*) an input in estimating travel demand matrices, and (*c*) a substitute for different modeling components. The chapter discusses how traditional and CDR sources of data can address questions related to sampling and expansion, lists how key elements of daily travel are recorded under each option, and presents how aggregate analyses and model components can be built with each data source.
- Section 3.4 describes the research approach used to shed light into the black box and to make multiway comparisons with similar measures of travel from traditional surveys and model results to highlight the strengths and weaknesses of each source of data.

3.1 Cell Phone Data in Transportation Planning

This section focuses on key aspects of cell phone data commercially available to transportation planning agencies and presents commonly used terms, standardized location analytics procedures, and information on sample sizes.

3.1.1 What Is Big Data?

Commercial cell phone data sold to transportation agencies are categorized as big data. Although this term conveys that these data sets are large, other elements are relevant within

16 Cell Phone Location Data for Travel Behavior Analysis

this description. Data streams that are termed "big data" meet one or more of the following "V" criteria:[1]

- **Volume**, the size of the amount of data. In the case of cell phone data, information is obtained from millions of devices nationally, making these databases massive in size.
- **Velocity**, the speed at which data are generated and processed. Most cell phone data vendors obtain live streams of data and handle, process, and store them in near real time.
- **Variety** in the type of transmitted data, which makes processing challenging. In the case of cell phone data used in location analytics, the most relevant information includes geographic elements, alphanumeric device IDs, time stamps, and dates. These may sometimes be combined with additional sources of information (e.g., land use or topography) to support waypoint analyses.[2]
- **Valence**, the measure of the degree of connectedness of the data. This attribute is relevant in cell phone data analytics where device locations are imputed using transportation and land use data and where heuristics are used to understand the difference between traffic stops and actual trips ends. In addition, device IDs are monitored over a longer period and machine learning algorithms are used to impute home and primary daytime locations.
- **Veracity**, a measure of the accuracy and usefulness of the data. Veracity is an important consideration when data elements collected passively are being evaluated. Algorithms described in Chapters 4 through 8 use principles grounded in transportation planning to improve the veracity of passive cell phone data.
- **Value**, that which is gained from processing the data. Value is the center of the new wave of collecting, analyzing, and interpreting data for transportation planning and modeling purposes.

3.1.2 What Are CDRs?

A CDR is a record produced by telecommunications equipment that documents the details of an incoming or outgoing call, an incoming or outgoing text message, or a connection to an app, web browser, or e-mail database. The communication passes through the device, which creates a record of the transactions made but not the content of the communication.

CDR records contain only metadata, not the actual data packets being transferred. This property makes CDRs ideal for use in anonymized data analytics. Key dimensions of financial and operational CDR data are as follows (see also Figure 3-1):

- CDR data capture all metadata that trigger a contact with the telecommunications towers, including calls, text messages, and Internet data access.
- Both active and passive signals trigger the capture of CDR data from a single cell phone:
  - Active signals include cases when users make calls, send text messages, or use the Internet actively.
  - Passive signals are recorded when users receive calls or text messages.
  - Passive signals are also received from apps accessing the phone either continuously or periodically (e.g., e-mail, music, news, social media, and sports and fitness apps).

[1]The earliest reference to "three Vs" was made by Doug Laney (2001), who discussed three dimensions of increasing volume, velocity, and variety in an unpublished research note. Examples of references in the literature include a discussion of four dimensions (IBM 2014) and a discussion of five dimensions (Scott 2015).

[2]Waypoints are intermediate observations between the inferred trip ends while the cell phone device is moving. Trip ends can provide information about the locations of activities supported by travel. Waypoints can provide information about the paths used while the device is traveling between those inferred trip ends.

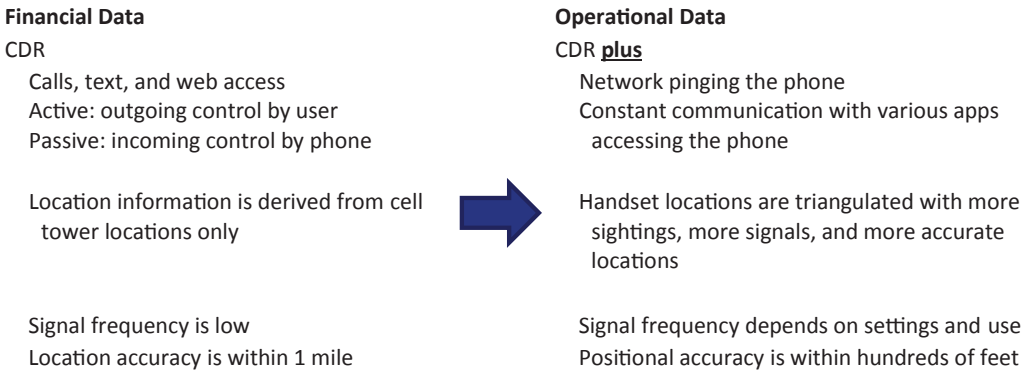


Figure 3-1. Types of call detail record data.

- The frequency and amount of CDR data generation depend on user usage patterns. Frequent users, through repeated use of their devices, provide a lot of information about their daily travel as compared with occasional or infrequent users.
- The amount of CDR data also depends on the number of apps installed by users on their cell phones and the frequency with which each installed app interacts with telecommunications towers.
- Most CDR data (e.g., start and end time of transactions and device ID) are accurate, but the location of a device is less accurate, given that device locations are estimated with methods, such as triangulation, that often produce geographic errors. Vendors have indicated that errors in CDR data range from tens to hundreds of meters and depend on geography and the density of cell towers.
- The CDR data available from cell phone service providers include financial information maintained by the cell phone service providers to support billing to their subscribers. Financial records are the data most likely to be maintained and available for additional processing.
- Providers also collect operational data processed to deliver services efficiently to subscribers.

3.1.3 How Are Locations Determined?

Using trace data from CDRs to identify the locations of devices is a critical step in supporting transportation planning and model development, application, and validation. Several analytical procedures have been implemented to convert trace data into locational and stop data:

- Devices may be traced through analysis of CDR data or use of GPS tracking on devices.
- Cell phone traces may be obtained by evaluating CDR transactions that include information about device location and the start and end times of transactions. Depending on the technology being used by the device and the number of transactions with telecommunications towers, locational traces of a device may either be few and far between or may be fast and furious. The frequency with which these traces are collected does not affect the accuracy of the mapping. However, the accuracy of CDR trace data ranges from tens to hundreds of meters.
- Cell phone traces may also be obtained by tracking GPS outputs from cell phones. These are obtained when GPS options on mobile phones are turned on. The GPS traces may be obtained (*a*) through apps when their location settings are turned on and (*b*) through apps that track location continuously in the background or by using a native GPS system. Although GPS traces are more accurate than CDR data, not all devices have GPS traces turned on all the time. The locational information from GPS-enabled cell phones is much more accurate than locations inferred from triangulated CDR observations.

- Cell phone traces obtained through CDR or GPS data are synthesized through sophisticated algorithms to provide estimates of locations, activities, and travel patterns. To do this, the analyst needs to address the following issues:
  - Qualifying criteria are needed to determine which devices have sufficient trace information to support further algorithmic investigation. These qualifying criteria are defined both in terms of the number of days with trace activity and in terms of the amount of activity itself.
  - A minimum threshold must be defined so as to recognize a true stop as opposed to a traffic stop. With regard to the nature of the stop, an algorithm must be developed to identify the location of the stop by exploring several nearby activity stops and "averaging" them to arrive at a location.
  - Regular nighttime locations (corresponding to home) and regular daytime locations (reflecting work or a school or university) need to be synthesized by using multiday data and a machine learning algorithm.
  - Locations that are not regularly visited and that do not qualify as home or work locations are much more difficult to identify. Several geolocation algorithms can be applied to identify these types of activities and their locations.
  - Algorithms are used to analyze consecutive stops and assign travel between them. The time of travel is determined through the time stamps of waypoint locations or, if there are no waypoints, by using heuristic algorithms based on other observed data.
  - In cases where sufficient waypoints exist, the analyst may be able to use map-matching algorithms to map the route taken by the device. In other cases, standard assignment algorithms can be run to evaluate network loading.[3]

3.1.4 What Are the Challenges of Using CDR Data?

As with every analytical method where inferences about travel need to be drawn on the basis of a data sample, measuring travel and activity patterns by relying on cell phone location data presents challenges. Given that location data are collected on the basis of users' active and passive use of cell phones, end users of these data sources need to consider the following:

- Not all movements may be recorded. Phones may be turned off; devices may not interact with telecommunications systems; individuals may not travel with their cell phones; or multiple users may share the same phone.
- The same movements may be recorded more than once, as individuals may carry more than one cell phone at any given time.
- Devices may travel through urban canyons in dense urban areas with multiple cell towers or through areas with poor cell tower coverage. In such cases, location analytics become less reliable.
- In cases of neighboring cell towers, cell phone signals may jump from one tower to another, so that it may appear that the device traveled between two different points, resulting in a spurious trip. Similarly, short "pseudotrips" may be inferred from changes in triangulation patterns that result in slightly distinct location inferences.
- Movements within a large office campus or hospital need to be identified, given that they do not affect transportation system performance.
- Although anchor locations such as home and work can be accurately located by repeated observations over multiple days and machine learning algorithms, the algorithms are less successful in identifying locations that reflect nonregular activity points outside home and work.
- Current research suggests that cell phone location data cannot readily measure key modeling metrics such as
  - Mode of travel,
  - Party size,

[3]Although this report does not focus on the waypoints between inferred trip ends, such information could be valuable.

  - Activity stops at which the activity duration is shorter than the threshold used in the algorithm, and
  - Activity purposes for locations other than home and work.
- Cell phone carrier restrictions aimed at preserving intellectual property and user privacy do not allow the end user to know how exactly the data are collected.

3.2 Transportation Planner Needs

In the context of population growth in cities and the restructuring of urban economies and societies, the fundamental task of transportation planners, modelers, and engineers—to effectively move people and goods—has become increasingly challenging. Meanwhile, transportation services and infrastructure greatly affect economic growth and quality of life.

Today's planning agencies focus on various issues when evaluating policy options: mobility, access to jobs, safety, environmental justice, land use and zoning, congestion, air quality and transportation emissions, transit utilization and fare policies, bike–pedestrian improvements, roadway construction, asset management, tolling and managed lanes, and socioeconomic zonal data. These issues need to be studied not only under current-day conditions, but also under future scenarios. Agencies use various data to support such analyses, including Census planning tools, custom survey data, and sensor data (from tolling, transit, and roadway agencies).

Regional planning relies on behavioral statistical models that link these data sources and provide a framework that explains today's travel behavior (Ben-Akiva and Lerman, 1985). Regional models provide snapshots of system performance by mode and time of day for different facilities across a region and are validated with observed measures of flows and levels of service.

Regional models are also applied under future estimates of population and employment using different transportation supply and policy scenarios to estimate the effect on travel flows and system performance. The underlying principles of the behavioral approach to modeling are as follows:

- The observed behavior of a traveler as a rational decision maker is key to understanding his or her observed choices in a spatial and temporal context.
- The analyst observes travel decisions and makes inferences about the aspects of traveler behavior that are not observed.
- Individual travelers are rational decision makers who use an "expected utility" approach to make trade-offs and evaluate and choose among alternatives.
- Individuals are assumed to have perfect information about the alternatives available to them and the attributes of each alternative.
- The individual behavioral approach is broad and applies to total trip making, the destinations and stops of individual trips, the timing of trips within a given day, and the mode choice decisions that individuals make.

Within this framework, planning agencies are now tasked with assessing how and where advanced sensor-driven analytics tools (such as trip tables based on cell phone data) fit in.

3.2.1 Perspectives of Transportation Planners

The objective of the interviews with agency staff was to discuss the cell phone data available, how agencies intended to use that data, and how modeling practitioners actually use that data. The research team interviewed modeling practitioners about how cell phone data could be used to develop and enhance transportation models. The interviewees—regional

planning staff, academics, and modeling practitioners—understand the data needed for traditional transportation models and the more extensive data requirements for detailed activity-based models.

Most of the participants have conducted or used traditional household, onboard, and intercept surveys and have considered using nontraditional sources of data to develop, augment, or validate regional and corridor-level models. Participants reflected a mix of backgrounds in Federal research, department of transportation (DOT) and metropolitan planning organization (MPO) experience, and academic research. The questions listed in Table 3-1 were used as a rough interview guide to frame the discussion. This section summarizes the key findings from the discussions under each group of questions.

Table 3-1. Interview outline with CDR data users.

Key Question	Attributes of CDR Data
What are the greatest agency needs that cell phone data can address?	• Passenger versus freight flows • Regional versus corridor-level flows • Insights into internal versus external flows • Representation of all regional travel versus focus on specific individual markets • Understanding of "underreported travel" (e.g., short and nonmotorized trips)
How do agencies perceive the potential value and role of cell phone data?	• Replacement versus augmentation of traditional data • Cell phone data as a source to use in model estimation versus validation • Option to update existing data periodically without a 10-year regional survey • Targeted use of cell data for geographic markets or traveler segments
What was the agency's view of the cell data it acquired or examined?	• Were the cell phone data usable and useful? • What types of cell phone data did the agency use or consider? • What was the greatest value of these data? • What were the weaknesses of the data that might require revisiting?
What was the agency's experience working with these data?	• What was the format of the cell data, and what issues were encountered regarding the ease of use, processing, and storage of the data? • What were the strengths of the cell data? The critical weaknesses? The ways to overcome the weaknesses? • What assumptions was the agency willing to accept when using these data? • What additional analyses did the agency conduct to "validate" these data?
Does the agency believe that the modeling paradigm will change?	• Should cell phone data be fit in current model formulations? • Should model formulations change in response to these data? How?
For which purposes does the agency intend to use these data going forward?	• Willingness to invest in these data? • Interest in additional testing before making a commitment to use these data? • Wait for the products to mature? • Plans for using cell data as part of a new data collection cycle?

Source: Cambridge Systematics, Inc.

3.2.1.1 Agency Needs That Can Be Addressed by Cell Phone Data

Agency staff recognized the need for increasingly complex travel demand models to address today's more nuanced and policy-sensitive questions. Issues included congestion by time of day, the mix of passenger and freight flows, introduction of technology and new modes, optimization of existing facilities, and experimentation with new tools (e.g., congestion pricing and technology) to better manage travel demand.

Participants also recognized the increasing costs of traditional data collection and the time and resources required to collect, analyze, and interpret such data within an advanced modeling framework. They were intrigued by the potential uses of new sources of data that can be harvested faster, more often, and at a lower cost than traditional data sources.

In addition, participants realized that cell phone–derived data cannot address all of an agency's planning needs, especially the nuanced and policy-sensitive questions that activity-based models address. Participants mentioned concerns about sampling bias and their desire to model both passenger and freight flows in their regions.

Also, participants recognized the value of new sources of data as reflected in the data's scale and cost. The staying power of such data will be greater if the data can provide answers regarding individual model components or markets currently not addressed by regional models (e.g., external models, long-distance travel, corridor studies, and travel to special events).

3.2.1.2 Potential Value of Cell Phone Data

Participants assessed the potential value of cell phone data as compared with the pros and cons of traditional data and methods and were skeptical about some of the cell phone–based data products available in the marketplace, given that some of these products represent a black box to agency staff as users of the data.

The paradigm that has served the transportation community well over the years has evolved along with increasing transparency in the methods used by academia and the industry. The state of the art has progressed with healthy debates about the value of different data sources, analysis methods, validation approaches, and forecasting mechanisms. Transportation practitioners are eager to dive into the details of new data to better recognize and appreciate their value. Transportation practitioners are also accustomed to specifying their data needs and have a high level of control that is less amenable to ready-to-use data products and analysis solutions.

In this light, agency staff, practitioners, and academics alike were intrigued by the promise of new data sources. They recognized that improvements in data and in the methods used to analyze them could improve the products available in the marketplace. They also were interested in gaining a better understanding of the underlying assumptions, so as to better evaluate for themselves the strengths and weaknesses of new data product options, and were interested in helping shape the design of these data products to better reflect their own priorities.

3.2.1.3 Staying Power of the Behavioral Modeling Paradigm

There was general consensus about the value of the behavioral paradigm in guiding the development of traditional four-step models and more sophisticated activity-based travel demand models. Industry and academia believe that the study of individuals' travel decisions and their activities are key to the analysis, understanding, and forecasting of travel patterns. The skepticism sometimes expressed about prepackaged data products reflects, at least in part, analysts' preference for specifying their own detailed data requirements and developing customized travel demand models to fit their region's modeling needs for a range of analytical purposes. Although the discussion did not focus on model evolution, the behavioral paradigm itself has changed over time from crude aggregate models to individual-level disaggregate estimation to sophisticated

activity-based models. The behavioral paradigm may evolve again in response to emerging policy and analysis needs and may benefit from current ongoing research on locational data in transportation and related fields.

The participants acknowledged that it is likely that research in refining inferences based on locational data and machine learning methods may spur development of new analytical methods. In this case, it can be expected that the behavioral principles and methods of travel behavior analysis will evolve to best take advantage of cell phone data sources, accompanying land use data, and new flexible forms of personal surveys.

3.2.1.4 Experience with and Future Uses of Cell Phone Data

The participants' experiences with data products derived from cell phone data varied. Most of the uses focused on passenger travel at a regional or corridor level, with one application of cell phone data for travel at a national level. In addition to conducting the interviews, the research team reviewed an in-depth presentation by Ron Milam at the Washington Council of Governments.

Participants agreed that cell phone–derived data cannot address all of an agency's planning needs, especially with regard to nuanced and policy-sensitive questions. Concern was also expressed about the gap between the resolution provided by cell phone data and that provided by today's sophisticated activity-based models with regard to travel for different purposes, at different times of day, and by different members of a household. The following topics were raised:

- The use of cell phone data for estimation of an activity model was mentioned as problematic, given the aggregate nature of the cell phone–derived data.
- Concern about sample bias reflected in differences in ownership and use of cell phones across different market segments was discussed.
- The power of cell phone data in providing origin–destination (O-D) flows was recognized but needs to be coupled with checks against existing data.
- The accuracy of O-D data at different levels of geographic detail is a major consideration in the practitioner community.
- The inability to differentiate between passenger and freight traffic is another limitation present in current cell phone data products.

Participants recognized the value of quick access to travel data outside the long cycle of household surveys. They also indicated that cell phone–derived travel data could provide input to individual model components, as follows:

- External models, a promising area in which cell phone data could replace current methods or allow for more frequent model updates;
- Long-distance travel and corridor studies, which require a lower level of geographic detail, could benefit from cell phone O-D flows; and
- Travel to special events and traffic to special generators were mentioned as examples of promising applications of cell phone–derived travel data.

3.2.2 A Planner's List of Do's and Don'ts for Cell Data

Findings from practitioner interviews can be synthesized into three key areas that agency staff must evaluate before making the decision to purchase and use cell phone data for planning and modeling purposes. Positioning cell phone data within an agency's overall data and modeling program, recognizing the strengths and weaknesses of this data stream, and identifying the questions to ask the data vendor are critical to extracting the most value from these data.

3.2.2.1 Data Collection Program

First and foremost, planners should position cell phone data requests within the bigger picture of the agency's data collection program that responds to the different analysis requests and modeling requirements. In particular, questions that agency staff and practitioners need to ask themselves and discuss in detail with data vendors include the following:

- Where do cell phone data fit within the bigger picture of the agency's overall planning functions? Can cell phone data as packaged by the vendors be used to
  - Support existing behavioral analysis tools and models?
  - Provide a snapshot of travel today and a forecast of future travel?
- What are the agency's data collection and analysis strategies? How can the agency benefit from integrating cell phone data into its data collection program? Can cell phone data be used to
  - Augment or replace the collection of regional household surveys?
  - Replace special purpose surveys at airports and other special generators?
  - Support visitor surveys or external travel data collection efforts?
- More critically, agencies must evaluate whether and how cell phone data may support or augment travel demand models. Specifically, can cell phone data be used to
  - Estimate regional travel demand model systems?
  - Develop models at a level of resolution similar to that of today's tools?
  - Replace individual model components instead of an entire model system?
  - Support freestanding analyses such as time-of-day travel?
  - Validate existing travel behavior models on a regular basis?
  - Refresh model inputs more regularly than every 10 years, which is itself an ambitious standard not met by most planning agencies?

3.2.2.2 Evaluation of CDR Data Products

Second, planning agency staff need to recognize the strengths and the limitations of commercial cell phone data across three dimensions—the data themselves, the procedures vendors use, and the applicability of the data for different planning and modeling uses.

- Data content, technology, and the structure of cell phone–derived data sources are different from traditional survey data sources.
- Vendors use their own proprietary analytical procedures to convert cell phone data into usable data products.
  - Agencies that regularly collect survey data to develop and update sophisticated models to their own exact specifications need to understand the vendors' underlying methods.
  - Smaller agencies that collect data less often and maintain simpler travel demand models may recognize the benefit of these forms of data and also need to appreciate the underlying methods.
  - In both cases, agency staff may be willing to accept that they will rely more on a black box approach to gain the benefit of frequent data and model updates that use cell phone data.
- The utility and value of cell phone data vary according to the intended purposes of different agencies and departments within an agency.
  - Cell phone–derived data products are not likely to be appropriate for model estimation, given that less detail is available regarding travel purposes, sample weighting and expansion, and sensitivity by market segment.
  - Cell phone–derived data may be appropriate to enhance analyses of travel by time of day, special events, and external travel.
  - Cell phone data can also be used to help validate individual model components or to refresh selected model components.

24 Cell Phone Location Data for Travel Behavior Analysis

3.2.2.3 Questions to Ask CDR Data Vendors

As a third and final step, agency staff evaluating cell phone data should ask vendors a range of specific questions and hold discussions with their colleagues to help determine the value of cell phone data packages.

- What is the technology supported in the data product (e.g., 4G versus 3G)? How does the technology affect the frequency of signals and their location accuracy?
- What is the geographic coverage of devices in the study region? What is the density of service or cell towers? What effect does the density have on the spatial accuracy of the data?
- Who are the providers of cell phone data in the region and what is their subscriber base and market share? Are the CDR data from a representative sample of the region's cell phone users that reflects the regional population?
  - What is the profile of devices and users in the marketplace? This issue is especially important in cases where the subscribers' socioeconomic and usage profiles differ across vendors.
  - What are the types of analysis that would be most appropriate given the local sample of cell phone data collected? Vendors can be asked to provide details on why they believe cell phone data are appropriate for different analyses.

3.2.2.4 Understanding the Attributes of the Black Box

- In summary, a discussion that clarifies aspects of a final product that is now a black box will help agency staff and practitioners better understand the underlying assumptions for the product. Such a dialogue will help users of the data evaluate key aspects of the data set and will provide input to influence the design of future products. Agency staff can be part of a two-way educational effort that may inform software vendors in developing and refining future cell phone data products in response to the planning and modeling needs of transportation agencies.
- Increased collaboration between industry, planning agencies, and academia can accomplish this task even if business realities prevent the sharing of proprietary methods by vendors. At a minimum, increased transparency of the black box through input provided by users of the data will increase the usefulness and value of the tool to planning agencies. Recognition of the strengths and weaknesses of CDR data and the underlying analysis assumptions will increase confidence in cell phone data products.
- Agency staff need to formulate their own understanding of the utility of CDR data from a practitioner's perspective. In particular, it is important to identify the agency objectives that can be supported by CDR data and those for which other data sources are more appropriate. Thinking about how CDR data will be implemented and analyzed within an agency's data collection and planning support cycle will help crystallize the agency's approach to purchasing and utilizing these data.

3.3 Utility of Cell Phone Data

Cell phone data can provide metrics with which to understand snapshots of today's travel demand and can provide data that can be analyzed in a modeling framework to forecast future travel. This chapter discusses the utility of cell phone data in providing a bird's-eye view of travel demand patterns. Cell phone data are compared with traditional travel surveys in Section 3.3.1. Typical model outputs from each data source are compared in Section 3.3.2. Finally, the ways in which different model elements can be captured by traditional surveys and cell phone data are addressed in Section 3.3.3.

3.3.1 CDR Data Utility as a Source of Travel Data

Because of the large sample size of cell phone data matrices and their lower unit costs as compared with traditional household travel surveys, there are often discussions about replacing household travel surveys with cell phone data. Table 3-2 contributes to this discussion by showing the strengths and weaknesses of cell phone data as compared with survey data:

- The lower unit cost and lower total cost of cell phone data allow the collection of a much larger data set that often spans multiple days. which offers a key advantage over household surveys

Table 3-2. Sampling and expansion by data source.

	Basis of Travel Data	
Variable of Interest	**Traditional Survey**	**Cell Phone Use**
Sample Size		
Sampling rate (%)	0.5–2	15–35
Sample size for a region with 1,000,000 households 3,000,000 population	 5,000–20,000 households 15,000–60,000 individuals	 150,000–350,000 households 450,000–1,050,000 individuals
Sampling Strata		
Unit of analysis	Individual and household	Cell phone
Sampling unit	Household	Cell phone
Sampling by geography	As fine-grained as block group level	Feasible at aggregate level
Sampling by market segment	Yes	N/A
Socioeconomic Information		
Respondent attributes	Age Gender Worker status Student status Occupation Work hours Ability to telecommute Ethnicity	Nighttime location of cell phone Daytime location of cell phone
Household attributes	Size Number of vehicles Number of workers Income Life cycle Residential location Number of children	N/A
Survey Expansion		
Sampling rate and size	Small sample size requires careful sampling and expansion	Large data set—relatively robust sample sizes for expansion
Household attributes	Used in all household travel survey expansion	No data available for detailed weighting on basis of personal or household attributes
Respondent attributes	Increasingly used in weighting of surveys for activity-based models	
Geography Attributes	Often carried out at county level; smaller geographic levels possible, depending on sample sizes	Some geography-based expansion feasible, but not at individual or household levels

Source: Cambridge Systematics, Inc.

Note: N/A = not available.

26 Cell Phone Location Data for Travel Behavior Analysis

that use a smaller sample focus mostly on travel during a single day, and are administered every 10 years or so.

- The units of the analysis are different. Surveys focus on households and individual travelers within a household, while cell phone data rely on devices. Data products built on cell phone data are weaker for the following reasons:
  - Individuals may own multiple devices, members of a household may share devices, and individuals may carry different devices at different times of day. These patterns are not accounted for and may affect data quality.
  - The use of the device as the sampling unit does not allow for differential sampling by market segment or by detailed geography and prevents the analyst from focusing on specific segments of greater interest to a region.
- Cell phone data do not include any household or individual socioeconomic information linked to each device because of privacy considerations. The absence of this information limits the resolution of cell phone data for model estimation.
  - In traditional surveys, detailed household and individual data are collected and offer critical input to the model estimation.
  - Traveler and household socioeconomics allow the segmentation of the market and the evaluation of travel responses to different pricing and level of service scenarios.
  - Recent research has focused on inferring socioeconomic characteristics at the home end and linking cell phone records to those characteristics.
- Survey expansion methods also differ significantly and provide an advantage to traditional survey methods.
  - Cell phone data most likely need to be expanded by using the device owner's residential address, which is in turn inferred by the nighttime location of a device. There are also potential biases in ownership that are not possible to identify and correct during sample expansion.
  - In contrast, household survey expansion and weighting are much more detailed and take into account socioeconomic and geographic criteria that provide a richer and more balanced sample across different criteria.

3.3.2 CDR Data Utility as a Source of Travel Demand Metrics

The lack of socioeconomic data in trip tables and travel estimates derived from cell phone data limits the use of CDR data as a full-fledged replacement for travel models. Today's activity models provide detailed estimates about the type of traveler who uses different modes and facilities of the transportation system to reach various destinations at different times of day. Table 3-3 lists the elements that traditional models provide and discusses how traditional and cell phone data sources can address each model component.

- **Metrics.** Cell phone data can provide metrics of aggregate measures of residential travel, visitor travel, and travel at external stations. However, they cannot support fine-grained analyses by purpose and market segment.
- **Total travel.** Daily travel is underreported in surveys because respondents may not report short trips that are considered less important (Bricka and Bhat 2006, NuStats 2002, and Zmud and Wolf 2003). Monitoring of cell phone traces over multiple days offers an objective way to observe travel. The drawbacks of CDR data include
  - The device as the unit of the analysis instead of the individual;
  - The need to infer stops, activities, and purposes;
  - Potential gaps in cell phone signals in lower-density areas; and
  - The possibility that travelers do not carry their cell phone at all times.
- **Home, work, and nonwork activities.** Repeated observations of a device during night hours and during a typical workday provide a robust definition of home-based work and school

Table 3-3. Recording of travel elements.

Variable of Interest	Travel Data from Traditional Surveys	Travel Data Based on Cell Phone Use
Total daily travel	Self-reported in survey diaries. Travel may be underreported. Prompted recall offers an improvement.	Passive cell signals over days may offer more robust metrics than surveys. Unit is device-trips rather than person-trips. Quality depends on CDR data density.
Time of travel	Self-reported in survey diaries. Times may be inaccurate and incomplete.	Accurate time stamps. Need to infer activity and link it to the time stamp versus en route travel.
Stops versus activities	Self-reported in survey diaries. Detailed log of stops and activities. Good detail on all travel purposes.	Need to infer stops, activities, segments. Nonwork purposes are difficult to infer.
Location of activities	Self-reported in survey diaries. Smart geocoding needed to match. Prompted recall offers an improvement.	Difficult to infer the location of activities. A challenge in mixed land use areas.
Travel purpose	Self-reported in survey diaries. Prompted recall offers an improvement.	Home and work locations are inferred. Poor inference on nonhome and nonwork.
Joint travel	Self-reported in survey diaries. Risk of underreporting. Prompted recall offers an improvement.	Not feasible to record or capture.
Mode of travel	Self-reported in survey diaries. Good detail by tour and segment. Walk and bike trips may be underreported.	Not readily inferred.
Route assignment	Not usually captured in surveys.	Depends on trace data and algorithm.
Tour generation	Self-reported in detail in a survey. Analysis by using heuristics and rules.	Data products do not include chains. Only aggregate trips are sold.

Source: Cambridge Systematics, Inc.

or university activities. However, cell phone–derived data do not provide detail for activities other than home and work.

- **Spatial resolution.** Locations other than home and work are also subject to reporting errors by respondents in traditional surveys. In the case of cell phone data sources, locations are inferred by analyzing cell phone traces and using assumptions about an activity introducing the potential for spatial error.
- **Purposes and joint travel.** Surveys have a clear advantage, given that they provide detailed information about all types of purposes and joint travel with other members of the household.
- **Temporal resolution.** Cell phone data record traces with accuracy as long as they are linked to a call, message, or Internet data access. Under these conditions, they may be preferable to surveys where recording of time is approximate and less detailed.

- **Travel modes.** Mode choice can be imputed in cell phone data tables by using a combination of travel speeds and perhaps some transit routing information. However, this method is not yet reliable, especially in large urban areas that experience congestion and where the speed difference across modes may be smaller.
- **Tour metrics.** Commercially available CDR data provide aggregate trips between zones and do not include travel at the tour level in contrast to the high level of individual travel detail offered by activity-based models.
- **Traffic assignment.** Traditional surveys do not capture detailed route information, whereas cell phone–derived data do. Cell phone–derived data are advantageous in this regard, as long as the transactions made produce enough signals to reflect the entire route.

3.3.3 CDR Data Utility for Model Components

The labor-intensive efforts and considerable data resources allocated to the development of a regional model do not support frequent updates or reestimation of regional models. In most cases, model updates are limited to reflecting the availability of new sources of control data such as the Census Transportation Planning Products, journey-to-work data, American Community Survey (ACS) data, and up-to-date traffic or ridership counts.

Models are also updated to account for the effect of major transportation investments such as new highways, the introduction of tolls, or the introduction of new transit services. It is also possible that major changes in residential or commercial land use, such as urban revitalization, development of a new major employment cluster, or the introduction of new sports or conference facilities, may motivate the update of the model to better quantify their likely effects on travel.

It is often necessary to develop freestanding modules to examine special event travel; changes in travel that originates outside the region or traverses the region; effect of new facilities on visitor travel; or needs of visitors who travel within a region.

Table 3-4 provides a list of model-related tasks where CDR data can provide input instead of, or in addition to, traditional survey methods. For each module, the key attributes of each data source are outlined to highlight the corresponding strengths and weaknesses of traditional survey data and CDR data.

3.3.3.1 Seasonality of Travel

The granularity of policy questions changes over time, and it may be desirable to address questions related to the seasonality of travel, especially for regions with major differences in travel patterns by season.

Rolling samples of surveys similar in concept to the ACS or multiple snapshots of travel by season based on CDR data can achieve this objective and should be considered. The calculation of the absolute or percentage change from season to season using seasonal data provides an additional off-model estimate of relative change over a typical day model approach.

3.3.3.2 Visitor Travel Patterns

This element may be related to the seasonality of regional travel but it may also affect metropolitan regions where tourism is a key component and driver of the local economy. Visitor and establishment surveys can be supplemented through the use of CDR data.

3.3.3.3 Special Generators

Travel patterns to and from nodes of intense recurring or nonrecurring activity, such as airports, shopping malls, and sports or cultural events, can have a major effect on travel infrastructure

Table 3-4. Travel elements for aggregate analysis and metrics.

Variable of Interest	Travel Data from Traditional Surveys	Travel Data Based on Cell Phone Use
Seasonal variation	A well-thought-out sampling plan. Continuous data collection by season.	CDR data by month of the year. Differences in seasonal travel.
Visitor travel patterns	Targeted detailed visitor surveys. Airports, train stations, highway rest areas, hotels, and popular visitor sites.	Home as nighttime device location. Differentiate visitor from residential devices.
Special generators	Specialized surveys at airports, malls, or special event sites, Supplement to regional surveys. Data on mode, time of day, origin of trip, and socioeconomic detail.	CDR data for "event days". Capture of time of day and trip origin. Mode inference is weak. Socioeconomic data not available.
Year-to-year variation	Longitudinal/panel or rolling sample data. Measurement of change over time.	CDR data sets from different years. Measurement of change in patterns.
External travel	License plate capture at cordon line. Follow-up survey of auto owners. Bluetooth data as an option.	Definition of external cordon line. Number of devices crossing cordon. Home origin to measure external travel.

Source: Cambridge Systematics, Inc.

and display significant peaking. A better understanding of these events can benefit from off-model components that can be readily updated by using periodic surveys or snapshots that use CDR data. Processed CDR data may be used to assess site-specific event studies such as airports, concerts, or sporting events. Limitations in inferring, aggregating, and expanding CDR data and the difficulties in providing path traces make these locations unsuitable for evacuation, emergency response, or other route-based studies.

3.3.3.4 Year-to-Year Variation

A weakness of existing regional models is that the time between model updates may easily exceed 10 years and may therefore miss subtle or more important trends in population and employment growth or stagnation. As a result, planners may under- or overestimate corresponding increases or slowdowns in travel.

Surveys or CDR data that are collected periodically can provide more frequent updates of the factors that affect travel. The identification of trends in data and model results can be valuable tools to account for growth in intermediate years until a new regional survey and regional model are completed.

As with seasonality effects, an analyst can focus on calculating absolute or percentage changes from year to year to provide an estimate of upward or downward trends in travel compared with the base-year model. These estimates can also prove valuable in updating forecast year estimates in cases where observed trends greatly exceed or significantly lag projected patterns.

3.3.3.5 External Travel

Traditionally, external stations or zones are used to supplement a regional model, especially in metropolitan areas that interact heavily with cities, counties, and states outside the model

area boundaries. Traditional license plate number recording and follow-up survey methods are labor intensive and can be supplemented or replaced by CDR data to provide snapshots of total external–internal or through travel. Key assumptions related to the home location of the device need to be accepted as part of this method.

3.4 Research Framework

Raw cell phone data, which include an identifier of the cell device, are exceedingly hard to obtain because of the confidential information that could be inferred from those CDR data. Although this project used raw CDR data for the Boston, Massachusetts, area that had been obtained for research purposes, the research team recognized that it is not likely that such a set of raw CDR data will be made available again.

At the present time, only aggregated results of processed CDR data can be obtained commercially. The methods used by vendors to process this data are proprietary and are not disclosed. From a practitioner's standpoint, it is a challenge to ascertain the quality of the end product without an understanding of the procedures that drive the end product. To bridge this intellectual gap, the research team used the following three-step process:

1. The team analyzed the raw CDR data and described the processing methods used to infer trip ends and activities from the raw data. The open and transparent procedures allow practitioners to get a bird's-eye view of the techniques used in the field of cell phone data processing.
2. The team compared the results of its processing of the raw CDR data with commercially processed CDR data for the same geography. In cases of similar results, it was concluded that the methods used to process the raw CDR data discussed in this report were broadly comparable to those used in preparing the commercial data. In cases of differences, the team documented the differences and discussed how they might be reconciled.
3. The team compared results generated from the commercially available CDR products and the custom analysis of the raw CDR data with two independent transportation sources: regional household travel surveys and the regional travel demand model in Boston. These comparisons were necessary to identify the strengths and weaknesses of the CDR data and to develop a roadmap to enable practitioners to use CDR data effectively in the development of transport modeling and analysis.

This analysis will allow practitioners to assess the value of this new CDR data stream as compared with that of traditional surveys and models. On the one hand, CDR data offer a much larger volume of data on travel observed over a long period. However, despite their sample size and the advantage of repeated observations, CDR data are less detailed and require inferences to be made regarding locations, activities, purpose of travel, and the time of day of travel. CDR data also do not provide information on users' socioeconomic characteristics, which are a key part of traditional and activity-based models.

On the other hand, household surveys are well tested, their strengths and weaknesses are well understood, and they are currently evolving through the use of technology. Surveys are generally more expensive, have a much smaller sample size, and are collected infrequently as compared with CDR data. However, they offer the great advantage of providing household and person-level travel data, including accurate information on activities and purpose. They also provide detailed socioeconomic characteristics for each respondent, which allows for the development of nuanced models of daily travel at a disaggregate level.

If the outcomes of the two data sets are similar and the processes used to infer trip ends and activities are understood and considered acceptable, processed CDR data may offer a suitable and acceptable supplement to household surveys. Under such a scenario, CDR data can be used

for a range of purposes, from estimating travel demand models or model components to providing selected model outputs for estimation or validation to serving as interim data sets between consecutive travel behavior survey efforts.

3.4.1 Research Method

The overarching goal of this research is to present a method that extracts activity locations (**stay points**), labels **activity types** ("home," "work," and "other"), estimates **O-D trip matrices**, and assigns traffic in the road network by analyzing raw cell phone data (Jiang et al. 2013, Alexander et al. 2015). The next sections present a flexible, modular, and computationally efficient software platform built to implement these methods, which are analogous to procedures of traditional travel demand models.

This system enables researchers to import raw cell phone data to produce trip matrices and road usage patterns in any city (Toole et al. 2015). It also visualizes these outputs to communicate mobility patterns effectively to planners, stakeholders, and decision makers. The platform is an alternative to proprietary transportation software packages and has been built specifically to handle massive mobile phone data sets and additional open-source data.

The research team used the Boston metropolitan area as a case study for analyzing cell phone records. The gamut of travel demand estimation using big data is presented through a discussion of methods, validation, implementation, and applications. Given the scope of this analysis, cases from other continents, such as Latin America and Europe, are not included, although these were also tested (Toole et al. 2015) to confirm the flexibility and applicability of the modeling framework.

3.4.2 Multiway Comparisons: A Case Study

To mirror the thinking and approaches used by transportation modelers and planners, the case study compared and contrasted travel demand estimation results from CDR data, traditional survey data, Census data, and the Boston regional model. These comparisons allowed the research team to get an understanding of the strengths and weaknesses of cell phone data and how these data could be incorporated into different aspects of travel demand modeling. The sources compared included the following:

- **Travel purposes**, including home-based work trips (HBW), home-based other trips (HBO), and non-home-based trips (NHB);
- **Time-of-day patterns**, including a.m. peak (6 to 9 a.m.), midday (9 a.m. to 3 p.m.), p.m. peak (3 to 7 p.m.), and early evening/night (7 p.m. to 6 a.m.); and
- **Geographic aggregation**, including Census tracts and towns.

Model comparisons relied on the following data sets and models:

- Travel demand estimates based on **raw CDR data.** The data set included 2 million cell phone subscribers in the Boston metropolitan area for 2 months in 2010. Two different methods were used to extract travel patterns, which were compared with the other data sources for validation and evaluation purposes.
- O-D matrices by a **CDR data provider.** These proprietary results by a third party use 3 months of 2015 CDR data that are adjusted by using the 2010 Census for the Boston region. The estimation methods and procedures are proprietary and not known to the project team.
- **2010 Boston Travel Demand Model.** The Central Transportation Planning Staff model results were used as the baseline for comparisons with the demand estimates from the raw CDR data and the commercially available third-party CDR data.

- **Census Transportation Planning Products.** These data were used to obtain journey-to-work travel flows for 2010 (Federal Highway Administration 2013) and to validate home and work inferences and the commuting flows estimated with the raw CDR data.
- **2009 National Household Travel Survey** (Federal Highway Administration 2009). This survey provides information on the departure time distribution used in analyzing raw cell phone data. The survey data were also compared with the raw CDR data estimation results in terms of trip purpose distribution.
- **2011 Massachusetts Travel Survey** (Massachusetts Department of Transportation 2012). This survey, completed in 2011, was used as another independent source for evaluating the travel demand results obtained with the CDR data. The survey data were compared with the raw CDR estimates on trip purpose and travel departure time.

3.5 Summary

This chapter takes the perspective of transportation agency staff and practitioners who are evaluating the purchase of aggregate cell phone data to supplement, enhance, or complement traditional data sources to support planning and modeling projects.

The properties of big data and how CDR data fit in this picture are discussed in Section 3.1, which outlines the way CDR data are used to determine locations and the challenges of relying on the CDR data commonly available to transportation planners. In Section 3.2, the chapter shifts gears to consider the perspective of transportation planners and academics and discusses interviews with practitioners working in MPOs, DOTs, and federal agencies. This section includes a checklist that planners can use when thinking about the uses of CDR data, the strengths and weaknesses of these data, and efforts to open up the black box.

Section 3.3 describes the utility of CDR data as a source that may replace or augment traditional surveys. The option of using CDR data to develop travel demand modeling metrics is discussed, along with the value of CDR data in developing individual model components. Summary tables are used to compare traditional surveys and CDR data with regard to sampling and expansion; how key elements of daily travel are recorded in surveys and CDR data; and how travel elements for aggregate analysis and model components can be captured by each data source.

Section 3.4 concludes the discussion by presenting the research framework and case study approach used in Chapters 4 through 8. These sections describe how data and models from the Boston region were analyzed to compare and contrast measures derived from traditional sources to evaluate the strengths and weaknesses of CDR data.

CHAPTER 4

Description of Raw Data

4.1 Roadmap to the Chapter

This is the first chapter of a highly technical discussion about cell phone data, inference of locations and activity types, development of origin–destination (O-D) matrices, and comparisons with survey data to evaluate the robustness of the methods and results. The case study used regional survey data from Boston, Massachusetts; regional model outputs; and cell phone data to provide a unique glimpse into the black box where cell phone traces are translated into travel patterns.

This chapter provides an overview of the raw cell phone data used in the case study and describes the range of spatial and temporal resolutions of cell phone data. The massive and passive nature of raw cell phone data is demonstrated and the spatial and temporal characteristics of these data are explained in detail. The analysis relies on call detail record (CDR) data from 2 million cell phones collected over 2 months in the Boston region.

4.2 Context: Rapid Urbanization

Cities are growing at an unprecedented rate in human history. Today, more than 54% of the world's population (3.9 billion) resides in cities, and every one in eight of the world's urban dwellers lives in 28 megacities of more than 10 million inhabitants (United Nations 2014). On the one hand, the density of cities has brought economic productivity and provided cultural amenities and diversity; on the other hand, it is also the root of problems related to congestion, environmental degradation, climate change, decrease in quality of life, and unsustainable development (Dimitriou and Gakenheimer 2011).

Rapid urban growth places enormous strain on already burdened transportation infrastructure, which is critical to providing residents with access to places, people, and goods. Delays and poor levels of service resulting from congestion waste time and money and exacerbate harmful vehicle emissions.

In 2011, energy use in the transport sector alone reached 103 quadrillion British thermal units globally, accounting for 20% of total global energy consumption (U.S. Energy Information Administration 2015). Over the past 15 years, owing to the rapid growth of private vehicle ownership and freight traffic, carbon dioxide emissions from the transportation sector doubled in countries that are not members of the Organisation for Economic Co-operation and Development (International Energy Agency 2015).

In the United States, total vehicle miles traveled increased from 1.79 billion miles in 1986 to a high of 3.17 billion miles in 2016, reflecting an increase of 77% over the past 30 years (Bureau of Transportation Statistics 2017). The total fuel wasted as a result of congestion increased by more

than 400%, from 0.6 billion gallons in 1984 to 3.1 billion gallons in 2014 (Schrank et al. 2015). The total carbon dioxide produced as a result of congestion increased by 460%, from 10 billion pounds in 1982 to 56 billion pounds in 2011 (Schrank et al. 2012).

4.3 General Description of Data

Private technology companies, smart device apps, and telecommunications network providers collect and store enormous quantities of data on users of their products and services. A lot of information needs to be processed to maximize the value of these data. Billions of cell phone transactions must be processed; data from open and crowdsourced repositories must be parsed; and results must be made more accessible to the individuals who generated those data (Toole et al. 2015). Meanwhile, it is critical that measurements from these new data sources be statistically representative and corrected for biases inherent in them. This process requires integration of new pervasive data with traditional data sources.

This report describes the raw input data employed to estimate travel demand from cell phone data. In particular, it focuses on CDR data, including cellular tower–based and triangulated data. Owing to the variation in mobile positioning technologies, the spatial resolution of these technologies differs and has different effects on travel demand estimates. The report also examines the cell phone traces recorded by an individual student's smartphone app.

Advantages of cell phone data are discussed here in terms of their massive size and wide coverage in both space and time. Also explained are the disadvantages of cell phone data, which lack information on individual users' socioeconomic characteristics and details about their daily travel patterns as compared with traditional surveys. The combination of traditional and new data sources illustrates the system architecture of deriving estimates of travel demand with cell phone data.

4.3.1 Traditional Data Sources

Before cell phone data are discussed in detail, the traditional data sources for travel demand models are summarized briefly. These data range from Census data on population and daily commutes to travel diaries filled out by individuals in a household. Traditional travel surveys are typically administered by state or regional planning organizations. During sampling, weighting, model validation, and model application, survey data are integrated with public data such as Census demographics and journey-to-work patterns at different levels of geographic detail.

Household surveys are generally expensive to conduct, as they cover interviews of tens of thousands of residents in each metropolitan area and require intensive manual data encoding. To extract high-resolution data, individuals are asked to recall and report when, where, and how they traveled on a recent day, which makes them prone to recall errors and reporting biases. These challenges make it hard for surveys to cover more than a day or two at a time. Cost considerations limit the sample size to a small portion of the population—usually less than 1% of the households in the region. Typically, household surveys are conducted infrequently, with 10-year survey cycles reflecting industry best practices. For purposes of model estimation, validation, and comparison with cell phone CDR data, the following traditional data sets were included in this research:

- **Census data.** Census data are the only traditional data source necessary to estimate measures of travel demand patterns with cell phone data. The research team obtained population and vehicle usage rates of residents from the 2010 American Community Survey at the Census tract

level or for traffic analysis zones, which contain on average about 5,000 people. Population at the Census tract level was used to develop expansion factors to translate cell phone–derived estimates of travel to person-trips. Given that it is difficult to infer travel mode from CDR data, vehicle usage rates reported in the Census were used to estimate vehicle trips from phone data.

- **Survey data and model comparisons.** The team obtained data sets from different travel surveys to compare and validate methods of estimating travel demand from cell phone data. In particular, the following sources were used:
  - The 2009 National Household Travel Survey to model the travel departure time of cell phone users,
  - The 2010 Census Transportation Planning Products to validate activity inferences of home and work from cell phone data,
  - The 2011 Massachusetts Travel Survey to compare the total number of trips by purpose and by time of day, and
  - The 2010 Central Transportation Planning Staff travel demand model for Boston to compare and evaluate the travel demand estimated from the CDR cell phone data.

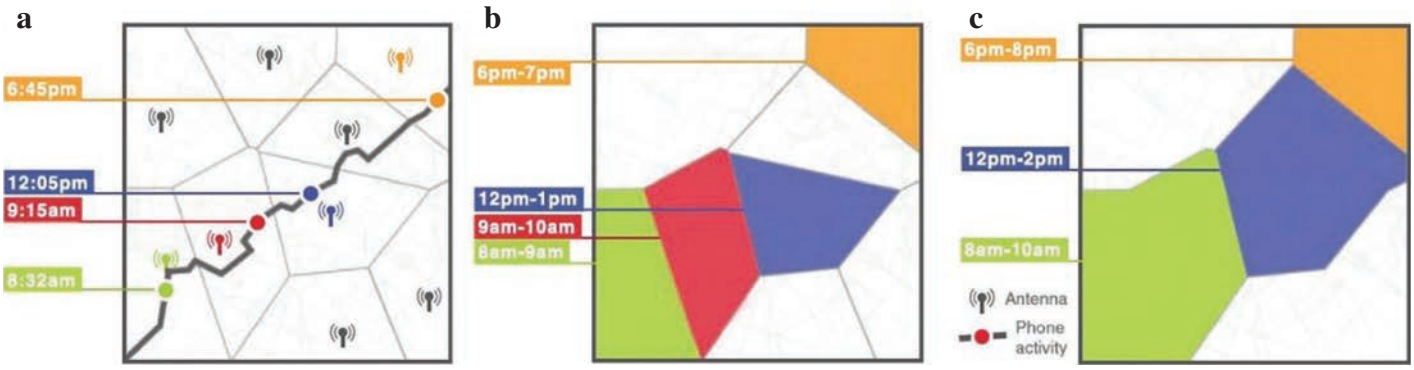
4.3.2 New Sources of Big Data

Cell phones, with their high penetration rates, are extremely useful sensors for human mobility. A large fraction of cell phone data are currently in the form of CDRs collected by carriers when users perform actions on their devices that makes use of telecommunications networks.

The location of each device is recorded when a call, message, or data request is registered by carriers for billing, network performance, and legal purposes. This type of data now forms the core of numerous human mobility studies in the context of U.S. metropolitan areas. However, the methods have been tested and applied successfully to cities in other countries as well (Colak et al. 2015, Toole et al. 2015).

Cell phones have been increasingly used to collect human mobility data. Figure 4-1*a* from de Montjoye et al. (2013) depicts a sequence of phone usage events made by a user at different time stamps and locations. These events are localized to the area served by the cellular tower to the user (Figure 4-1*b*). These events can be aggregated into individual-specific zones where a user is likely to be found at different times of the day or week (Figure 4-1*c*).

Another type of cell phone data, although generally less common than CDR data, is from apps running on smartphones. For example, the Future Mobility Survey app passively maintains



Source: De Montjoye et al. 2013.

Figure 4-1. Cell phone data to measure human mobility.

36 Cell Phone Location Data for Travel Behavior Analysis

activity diaries of users while requiring limited human inputs for validation purposes (Cottrill et al. 2013).

Some smartphone apps may provide even more precise estimates of users' positions than CDRs. Various sensors, from GPS to Wi-Fi, can locate a mobile device with accuracy down to a few meters and can record data every few minutes (Aharony et al. 2011).

- Protocols such as Bluetooth and near field communication allow devices to discover and connect to each other within a radius of a few meters, creating ad hoc sensors and social proximity networks (Eagle and Pentland 2009).
- Some of these apps explicitly add social networks to mobility data. For example, Foursquare invites users to "check in" at specific places and establishments. Twitter automatically geotags tweets with precise coordinates from where they were sent.

While these new sources of big data come with their own privacy challenges (Kosta et al. 2014), they offer planners, engineers, and policy makers great potential in better understanding, managing, and planning urban infrastructure systems efficiently for the public good. With strict privacy protection rules in place, this report presents examples of two types of anonymized cell phone data and open-sourced road network data, as follows:

- **CDR data.** Cell phone locations recorded in CDR data are inferred either by observing the cellular tower through which the phone is connected or by triangulation with nearby towers. In this research, 2-month triangulated CDR data from 2010 for the Boston Metropolitan Area were employed. The data were obtained from a technology company through a research nondisclosure agreement. The data provider provides location services to telecommunications service carriers in the region. Personal information in the data was anonymized by the data provider with ciphered identification strings. The researchers further anonymized the ciphered identification strings through the use of hashed IDs. This data set contains around 1 billion phone usage events made by 1.6 million unique mobile devices (hereafter referred to interchangeably as "cell phone users"), consuming roughly 70 gigabytes of disk space in its raw format. In cities with longer observation periods, data size can quickly become a performance issue.
- **Cell phone data via smartphone apps.** The researchers also included a self-recorded data set donated by a student volunteer, who turned his cell phone into a tracker of his everyday whereabouts for 2 academic years (18 months in 2013 and 2014). This was made possible by a smartphone app. Unlike CDR data, cell phone data recorded via smartphone apps vary depending on their specifications. In this case, the app only recorded a user's triangulated locations when a change in the device's location was detected—it smartly recorded the device user's travel while the app was turned on. Given that the student volunteer also provided ground truth information about his phone traces, this set of data was used as a controlled experiment and an illustrative example to explain key algorithms and methods discussed in this report.
- **Road network data.** For many cities in the United States, detailed road networks are made available by local or state transportation authorities. These geographic information system shapefiles generally contain road characteristics such as speed limits, road capacities, number of lanes, and classifications. In cases in which these properties are incomplete or missing, it is useful to turn to OpenStreetMap (OSM), an open-source community dedicated to mapping the world through community contributions (https://www.openstreetmap.org). For cities where a detailed road network cannot be obtained, it is possible to parse OSM files and infer required road characteristics to build realistic and routable networks (Colak et al. 2015, Toole et al. 2015). At this time, the entirety of the OSM database contains roughly 4 terabytes of geographic features related to roads, buildings, and points of interest, among other features.

4.4 A Closer Look at Cell Phone Data

4.4.1 Typical Data Set Layout

Each time a phone is positioned, it generates a single record in a mobile phone data set, which is the equivalent of a row in the data set. Each record contains at least three basic pieces of information: an ID number, a unique number associated with the device generating the record; a location that indicates the device's location when this record is generated; and a time stamp that indicates when the record is generated (Table 4-1). For privacy purposes, the real ID of a device is always encrypted by network operators. The format of the location information varies, depending on the technique network operators use to perform positioning. The implications of these different technologies on data quality are discussed in the next section. Although the format of the time data can vary, UNIX time is frequently used in mobile phone data sets.

4.4.2 Spatial Resolution

It is common for network operators to record the location of mobile phones in terms of the cell tower to which they are currently connected and other towers that could process transmission between the tower and that cell phone. Yet in some cases, only the ID of the connected tower is provided because of privacy issues.

Mobile users' traces are, therefore, represented by time-ordered sequences of cell tower IDs, which can be used to infer the topology of cell towers (Bayir et al. 2010). However, given that the geographical locations of the towers remain unknown, the spatial resolution cannot be determined in this case. In other cases, the geographical locations of cell towers are known and can be presented either as the coordinates of the tower or the geographical area in which the tower is located. Most of the time, the latitude and longitude of the towers is used (Song et al. 2010b).

The spatial resolution of these data sets is determined by the density of cell towers, which varies from as little as a few hundred meters in metropolitan areas to a few kilometers in rural regions. In other words, an uncertainty level of a few kilometers is possible if the location of users in rural area is considered. In cases in which a geographical area is used, the study area is first divided into smaller zones, each of which is served by one or more cell towers. Any phone activity routed through a tower within a zone will result in a record with the location represented by the location of this zone (e.g., the centroid of the zone). Therefore, the spatial resolution of

Table 4-1. A hypothetical sample mobile phone data set.

ID	Time[a]	Location[b] (longitude | latitude)
3X35E90	1319242582	34.044162 | –112.454400
3X35E90	1319242583	34.044059 | –112.455550
3X35E90	1319301785	34.044392 | –112.453519
3X35E90	1319339560	34.040538 | –112.453760
5YU86I0	1315093092	33.948195 | –112.170318
5YU86I0	1315093145	33.961547 | –112.165304
5YU86I0	1315093169	33.977657 | –112.175295
5YU86I0	1315093992	34.057944 | –112.178316

[a]Time is the UNIX time stamp. The UNIX time stamp (or Epoch time) is the number of seconds that have elapsed since January 1, 1970, 00:00 UTC.
[b]Location is defined by the longitude and latitude coordinates of mobile phones.

location records greatly depends on the size of these zones. Knowing the connected cell tower is important to the network operator for assigning costs and revenue.

It is also possible for network operators to determine the location of a mobile phone by triangulation, transmission delay from multiple base stations, or other more advanced positioning techniques. These techniques can identify the location of phones anywhere in a cell and usually result in a finer spatial resolution than cell-tower-based methods, though the accuracy of their positioning also varies (Rose 2006).

Knowing the exact location of a device and the towers to which it could be connected is only necessary for operational considerations. Once a transmission is actually made through one cell tower, information needed for billing is retained, while operational records may be discarded.

Other approaches for locating mobile phones may require additional infrastructure to be installed or normal mobile devices to be modified. For instance, in the system for traffic information and positioning project (Ygnace et al. 2001), location estimates of mobile phones were obtained by installing monitoring devices along freeway segments to monitor signaling messages exchanged between mobile phones and the cellular network. In other examples, accurate locations of phones were acquired through built-in GPS receivers in the phones (Reddy et al. 2010). Yet, such infrastructure and technologies were developed for specific studies and are not always available.

4.4.3 Temporal Resolution

Temporal resolution of the cell phone data sets also varies substantially, depending on the specific mobile phone data set. A general categorization of these data sets is based on the mechanism that triggers what is recorded.

One type of data set is based on CDR, in which each record corresponds to a call activity of a cell phone user. Studies employing this type of data set identify a burst pattern of time intervals between consecutive records/calls. Although most calls are placed soon after a previous call, it is also possible to identify long periods of time without any call activity. González et al. (2008) identified an average inter-event time as 8.2 hours for 100,000 individuals over a course of 6 months.

A second type of data set can be viewed as a superset of the first type. A record is generated each time an activity is performed on the cell phone, including calling, texting, and Internet browsing. This type of data set has a finer temporal resolution than the first type, which is based only on call activity. Calabrese et al. (2011a) identified an average inter-event time of 260 minutes, which was much lower than the 8.2-hour inter-event time reported by González et al. (2008); they further characterized the time interval between consecutive phone activities by its first, second, and third quartiles. The authors reported the arithmetic average of the medians as 84 minutes and found that the temporal resolution of their data was fine enough to detect changes of location where the user stops for as little as 1.5 hours.

These two types of data are automatically and passively generated for cellular network operators' own purposes, including collection of billing information and network management. Cellular network operators do not maintain positions of users at all times to improve network performance, save bandwidth, and protect users' privacy. Positioning is only considered necessary when a user communicates with the network. When a user initiates a network connection event (e.g., a voice call), the cellular network operator needs to know the user's location to determine the cell tower used to channel the event. Therefore, the positioning data only describe the user's location in space when an event occurs.

4.4.4 Uncertainty in Location Estimates

Advanced positioning techniques, such as triangulation, are capable of estimating the location of a mobile phone within a cell and produce data sets with a finer spatial resolution than the cell-tower-based positioning method. Calabrese et al. (2013) used mobile phone traces to study individual mobility patterns from urban sensing data and reported an uncertainty range with an average of 320 meters and a median of 220 meters. More-sophisticated approaches can further reduce localization errors. Zang et al. (2010) proposed a technique based on Bayesian inference to locate cell phones in cellular networks. They were able to improve the accuracy of localization by 20% as compared with a baseline approach with a randomly selected location.[1] Despite these attempts, uncertainty of location estimation remains. Owing to the uncertainty in location estimation, distinct estimates of multiple neighboring locations can occur, although a device actually remains at the same location. Thus, these location records need to be aggregated.

There are generally two classes of approaches to aggregate spatial points. One is to impose a grid over the space and aggregate points within each grid cell. In a study to infer destinations from partial trajectories, Krumm and Horvitz (2006) divided the Seattle area into cells of 1,681 square kilometers and converted sequences of GPS points to sequences of cells by replacing the coordinates of a point by the index of the cell containing the point. This method depends on the layout of the grid, including the size and shape of the grid cell. Ye et al. (2009) described another problem of this grid-based technique: grid boundaries could be problematic when points corresponding to the same place fall in different grids.

The other class of approaches to aggregating spatial points is through clustering. Clustering-based approaches allow points to be aggregated with arbitrary shape and often require a distance threshold as an input. Ye et al. (2009) aggregated a sequence of points into one location if

- The temporal difference between the first point and the last point was more than 30 minutes and
- All the points were within a range of 200 meters.

Similarly, in a series of studies with cell phone data from the Boston area, Calabrese et al. (2010, 2011a) fused sequences of points into one location if the distance between any two points was less than 1 kilometer.

The general procedure of clustering-based approaches is summarized as follows:

1. The series of location records for an individual is ordered by time stamps, denoted as $\{l_{t_1}, \ldots, l_{t_n}\}$.
2. The first location record (l_{t_1}) is chosen to be the center of the first cluster, and the distance between the second location record (l_{t_2}) and l_{t_1} is calculated.
   - If the distance is less than a threshold k, then l_{t_2} is fused with this cluster and the cluster center is updated as the geometric center of l_{t_1} and l_{t_2}.
   - If the distance is greater than k, then l_{t_2} becomes the center of a new cluster.
3. The second step is repeated for all the remaining location records $\{l_{t_3}, \ldots, l_{t_n}\}$ until all the points are assigned to a cluster. All the points within a cluster are then analyzed as a virtual location for subsequent analysis. This procedure is graphically illustrated in Figure 4-2.

The distance threshold in these studies was determined, to a large extent, heuristically. It is generally recommended that, if clustering-based approaches are to be adopted, sensitivity analysis needs to be performed to fully evaluate the implications of different distance thresholds on location detection.

[1]For a full review of positioning techniques in cellular networks, interested readers are referred to Mao et al. (2007) and to Zhao (2000).

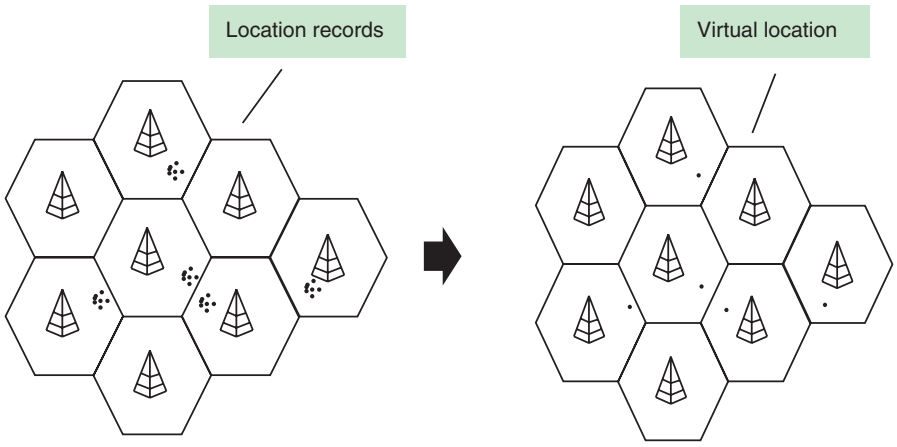


Figure 4-2. Clustering location records.

4.4.5 Device Oscillation

At any given location in a cellular network, there may be several cell towers whose radio signals reach a device. If these multiple cell towers have similar signal strengths, the connection of a device may hop between multiple towers even when the device is stationary. In such a case, it may appear that the user travels for several kilometers in just a few seconds. This phenomenon is known as oscillation in a cellular network. The potential effects of the oscillation phenomenon on the detection of a device's location are illustrated in Figure 4-3.

A device is on the boundary of Cell A and Cell B, and the signal strengths received by this device from Tower A and Tower B are equal. This device can be registered to either Tower A or Tower B, depending on the real-time traffic through these two towers. When it is registered to Tower A, its location may be recorded as Location A. Similarly, its location may be recorded as Location B when it is handed over to Tower B. Distinct location records—Location A and Location B—resulting from oscillation need to be consolidated. A few methods have been proposed to address this oscillation problem.

Iovan et al. (2013) proposed a speed-based method. Oscillation is detected if Location B is recorded in the middle of two records with Location A and if the switch speed from Location A

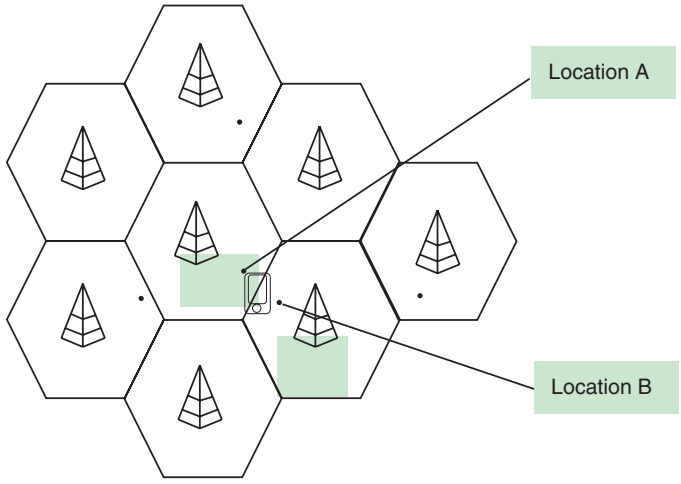


Figure 4-3. Oscillation in a cellular network.

to Location B is larger than a predetermined threshold. This method is based on the observation that oscillation results in a location change characterized by an abnormally high speed. Yet, a critical question in this method is the choice of a speed threshold that distinguishes normal from abnormal speed.

Other studies have applied a pattern-based method. This method recognizes the unique pattern in location updates associated with oscillation—frequent switches between pairs of locations. Lee and Hou (2006) identified the occurrence of oscillation as each time three consecutive mutual switches between a pair of locations is observed. Once oscillation is identified, all the locations involved in these switches are replaced with that location in the pair with which the user has been associated most of the time. A similar method was adopted by Bayir et al. (2010), who discuss a framework for discovering mobility profiles of cell phone users.

The procedure used to perform Lee and Hou's pattern-based method can be described as follows. A sequential scan starts from the beginning of the location records of a cell phone user ordered by time stamps. Oscillation is considered to be present in cases in which a subsequence of location records contains mutual switches between two Locations, A and B, for at least three times, such as

$$\{X_{t_1}, A_{t_2}, B_{t_3}, A_{t_4}, B_{t_5}, Y_{t_6}\}(t_1 < t_2 < \ldots < t_6).$$

This subsequence is then updated so that all location records indicate just one location—the one which the user has been associated with most of the time. In the same subsequence example, if the user is found to be associated with Tower A for a longer time than Tower B, then Location B is replaced by Location A, which results in an updated subsequence:

$$\{X_{t_1}, A_{t_2}, A_{t_3}, A_{t_4}, A_{t_5}, Y_{t_6}\}$$

The pattern-based method has the risk of mistaking the actual movements of a user who travels frequently between two locations for oscillation. The research team believes that a combination of the speed-based and pattern-based approaches may render more reliable results. First, subsequences that seemingly result from oscillation are detected on the basis of the pattern-based approach. Then, switching speeds between pairs of locations are determined for each subsequence. Finally, subsequences are only updated if the switching speed is beyond a speed threshold as determined in the speed-based approach.

4.4.6 Potential Issues with Cellular Data

Several issues apply to the use of positioning data from cell phone activity to study travel behavior:

- **Penetration rate.** Mobile phone data sets can suffer from being unrepresentative, depending on the mobile phone penetration rate in the study population. Though this may not seem to be a problem in developed countries, mobile phones are far from ubiquitous in many developing countries. Individuals who do not own mobile phones are precluded from studies. It is expected, though, that this issue will be resolved as the penetration rate keeps rising throughout the world.
- **Network operator**. Depending on the cellular network operator(s) who provides the positioning data, nonsubscribers are precluded and, thus, underrepresented. The biggest cellular network operator in the United States, Verizon, holds a market share of only 32% (Experian Simmons 2011); there are dozens of other operators in the United States. Little is known about whether there are any systematic differences in the travel behavior of subscribers with different cellular network operators.

42 Cell Phone Location Data for Travel Behavior Analysis

- **Sample selection.** It is common for researchers to select a study sample from all the subscribers included in a raw mobile phone data set provided by the network operator. When this selection is nonrandom, it may render the final sample unrepresentative. Song et al. (2010b) discussed the limits of predictability in human mobility in a study of a sample of mobile phone users who made at least one call every 2 hours.

  Recent studies showed that user mobility had a strong correlation with phone usage, with more-active users being more mobile (Iovan et al. 2013, Ranjan et al. 2012, Couronné et al. 2011). Therefore, sample selection based on phone usage would potentially result in an overestimation of mobility levels. However, a study by Iovan et al. (2013) also suggested that some mobility measures seem to be immune to this sampling bias.

  In summary, past research suggests that caution should be exercised when mobility information derived from cell phone data is generalized to the general population.
- **Socioeconomic information.** Cell phone data do not contain the user's socioeconomic information. If the research objective is to explain mobility measures derived from mobile phone data with socioeconomic variables, mobility measures can be aggregated to a geographic level where the distribution of demographic variables is publicly available.

  Calabrese et al. (2013) derived individuals' daily trip lengths from mobile phone data, aggregated the data to the block group level, and associated socioeconomic information from U.S. Census. More studies are needed to check the validity of such procedures and comparability across regions. Although such procedures can be validated, individual socioeconomic data are not available as required in conventional disaggregate travel modeling.
- **Privacy.** Privacy protection is usually achieved by researchers receiving an anonymous data set from cellular network operators. Also, research results are expected to be published at an aggregated level (Caceres et al. 2008). Researchers also have the choice to adopt an opt-in policy so that individuals' permission is guaranteed before their data are used for research purposes (Rose 2006).

  An opt-in policy could potentially reduce sample size and create questions of sample representativeness. Ahas et al. (2010) asked 576 individuals for their agreement to monitor their phones for research purposes and 231 of them agreed. The main reason for refusal was not related to privacy but rather to the lack of a contract with a specific cellular network operator. Only 10 respondents reported a serious concern about surveillance.

4.5 Evaluation of CDR Data for This Research

4.5.1 Spatial Resolution

The CDR data used in this research include time stamp and location for every use of a phone in the telecommunications service network. This includes information about location every time the device is used to make a phone call, send a text message, or access data on the Internet. The spatial granularity of data varies from cellular towers to triangulated geographical coordinate pairs in which each call has a unique pair of coordinates with an estimated accuracy within a few hundred meters. This information also varies according to the carrier that provides the data.

To demonstrate the spatial resolution of cell phone data with different technologies, the research team used data from San Francisco, California, in addition to Boston. For Boston, triangulated CDR data with a spatial accuracy within 200 to 300 meters are presented. As mentioned above, these CDR data were obtained through a nondisclosure agreement for research purposes from a technology company that provides location services to telecommunications service carriers. For the San Francisco Bay area, where CDR data were not available, only the distribution of cellular towers is shown. Table 4-2 shows the descriptive statistics for the data sources of these two study areas.

Table 4-2. Attributes of the two study areas.

Statistic	Boston	San Francisco Bay Area
Number of tracts	975	1,199
2010 population (millions)	4.46	5.40
Area (thousands of square kilometers)	7.32	8.73
Number of cell phone users (millions)	1.65	0.43
Number of cell phone events (millions)	905	429
Number of cell towers	na	849

Note: na = not applicable.

4.5.1.1 Tower-Based CDR Data

The San Francisco Bay area is used as an example to demonstrate the spatial resolution and coverage of tower-based CDR data. In such CDR data, a cellular tower ID is often recorded with a time stamp when a cell phone connects to a cellular network for a call, message, or data transmission.

Table 4-3 provides an example of tower-based CDR data for a fictitious user. The tower ID identifies the cellular tower to which the cell phone connected when its user made a phone call, sent a text message, or accessed data in the network. Epoch time is a time stamp identifying when such a cell phone usage event occurred. The time stamp is in the UNIX time format, which presents time in seconds that have elapsed since 00:00:00 Coordinated Universal Time, Thursday, January 1, 1970.

The San Francisco Bay area has more than 800 cellular towers. Figure 4-4 shows the Bay Area Census tracts and their boundaries, the population density at the Census tract level, and the location of the cellular towers in this region. The service area of a cellular tower can be represented as a Voronoi polygon whose interior consists of all points in the plane that are closer to a particular lattice point (e.g., cellular tower) than to any other tower.

Figure 4-5 shows the frequency distribution of the Census tract area (orange) and the tower-based service area (blue), with a bin size of 1 square kilometer. The first three quartiles of the tower-based service area are 1.9, 3.0, and 6.2 square kilometers, respectively, while those of the Census tracts are 0.8, 1.3, and 2.5 square kilometers, respectively. This comparison shows that, in general, cellular towers cover areas that are larger than Census tracts.

Figure 4-6 shows the frequency distribution of population density at the Census tract level for the San Francisco Bay area, with a bin size of 500 people per square kilometer of land area. For this region, the first three quartiles of population density at the Census tract level are 1,785, 3,272, and 5,579 people per square kilometer, respectively, with an average population density of 4,607 people per square kilometer.

Table 4-3. Example of tower-based CDR data for a fictitious user.

Tower ID	Epoch Time
2023	1266513700
2050	1266513800
1221	1266513900

Note: Only the first three data points are shown.

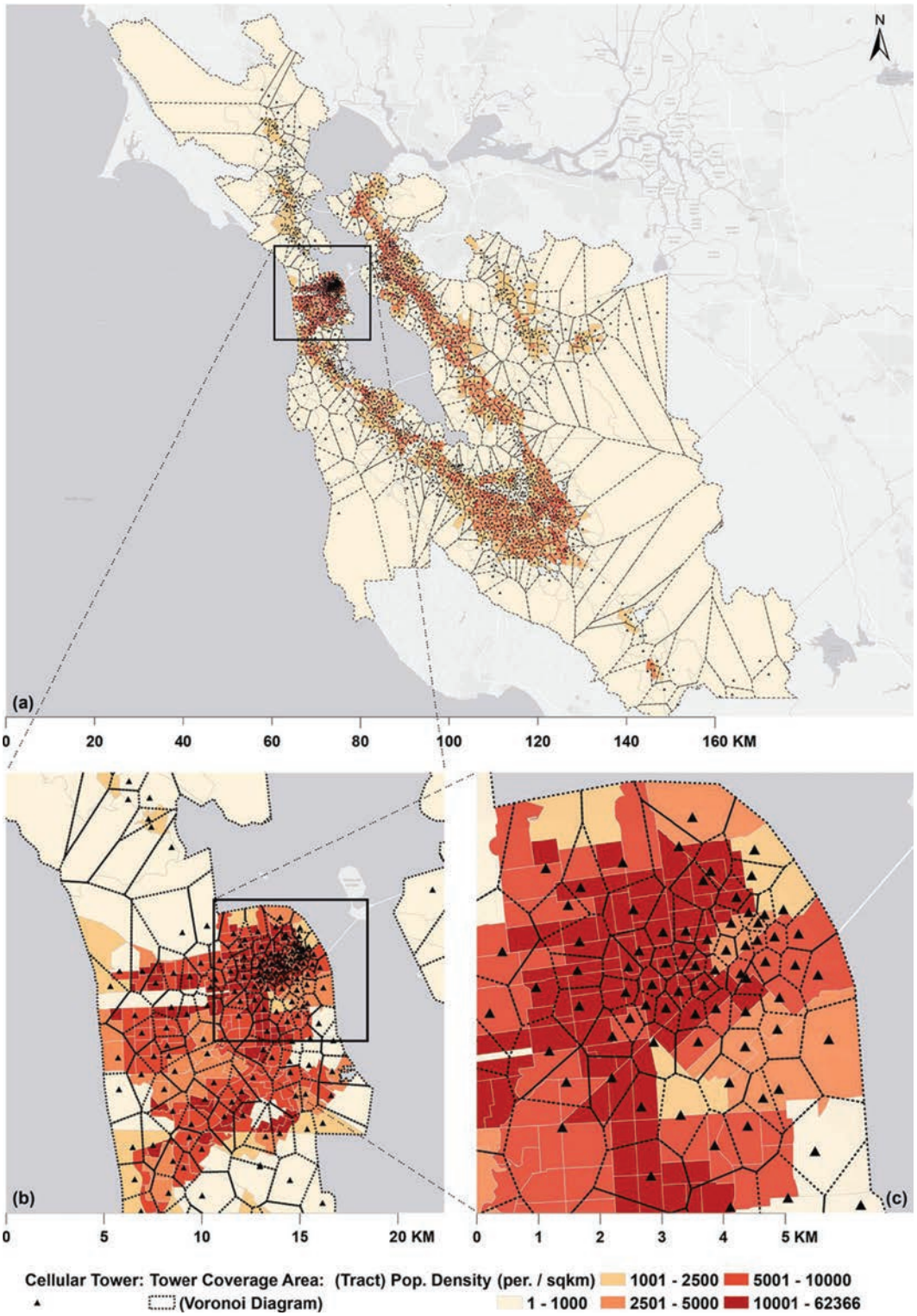


Figure 4-4. Cellular towers and census tracts in the Bay Area: (a) entire Bay area, (b) downtown area and (c) detail of downtown area.

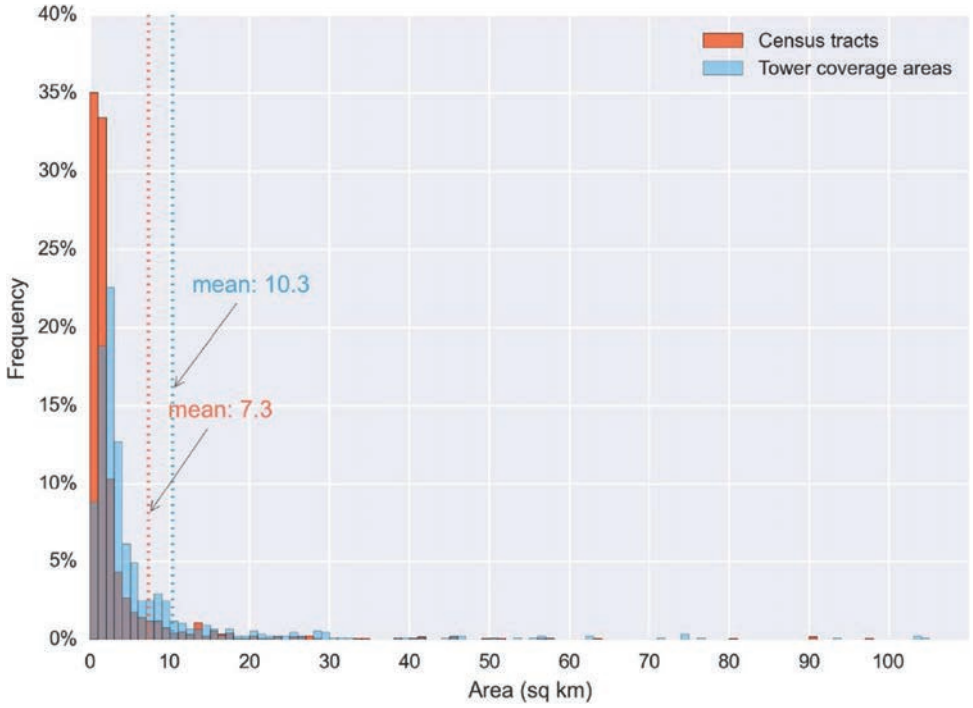


Figure 4-5. Census tract size and cell tower coverage in the San Francisco Bay Area.

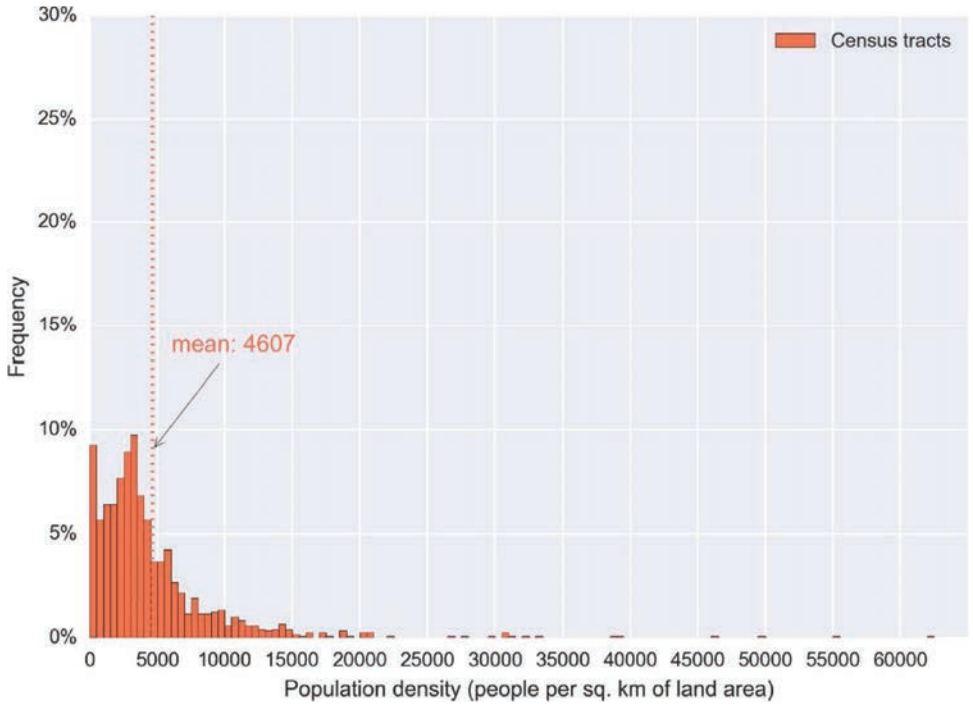


Figure 4-6. Population density at the Census tract level in the San Francisco Bay Area.

Table 4-4. Example of triangulated CDR data for fictitious user.

Longitude	Latitude	Epoch Time
–71.092110	42.359820	1266513700
–71.083856	42.361974	1266513800
–71.094821	42.359168	1266513900

Note: Only the first three data points are shown.

4.5.1.2 Triangulated CDR Data

With more advanced technology, a cell phone's locations can be pinpointed more accurately while it connects to an operator's service network. The triangulated CDR data in Table 4-4 provide an example using the Boston region. Table 4-4 provides an example of triangulated CDR data for a fictitious user. The Epoch time is the time stamp when a cell phone is connected to the network. The longitude and latitude are the pinpointed coordinate pairs of the device, estimated by the technology company with a reported accuracy of 200 to 300 meters.

Figure 4-7 shows the spatial distribution of triangulated cell phone data on a sample day in 2010 for the Boston metropolitan area. Every point in the figure is a triangulated location of a cell phone when it was connected to a cellular network. In the background of the figure, population density is shown at the Census tract level.

Figure 4-8 summarizes Census tract size and population density. The first three quartiles of the Census tract area frequency distribution for the Boston region are 0.8, 2.6, and 9.8 square kilometers, respectively, with an average of 7.5 square kilometers. The first three quartiles of the population density at the Census tract level are 482, 1,624, and 4,971 people per square kilometers respectively, with an average of 3,521 people per square kilometers.

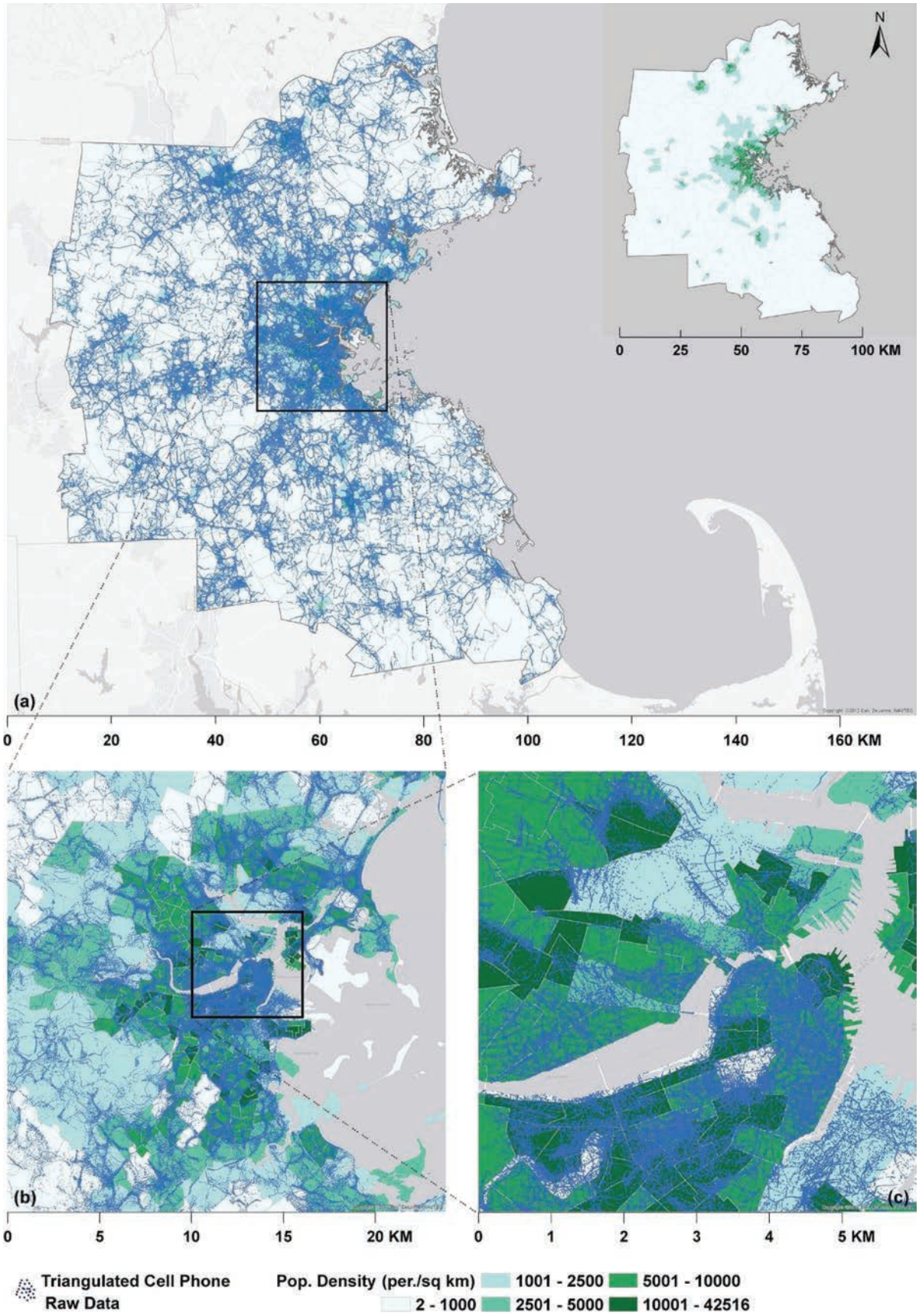
On the basis of the 200- to 300-meter cell phone location accuracy reported by the data provider, the study area was divided into grid cells of 300 by 300 meters. Figure 4-9 shows the event density of cell phone usage (i.e., event count per hour per square kilometer) in an average hour by time of day for the same sample day in 2010. The event density patterns are shown for early morning (midnight to 6 a.m.), morning peak hours (6–9 a.m.), midday (9 a.m. to 3 p.m.), afternoon peak hours (3–6 p.m.), evening hours (6 p.m. to 12 a.m.), and the day as a whole.

Figure 4-9 illustrates changes in the spatial distribution of phone usage at different times of day across the metropolitan area. During the early morning hours, phone usage in the suburban areas was limited, whereas the City of Boston contained spots with the highest density of phone usage (Figure 4-9*a*). In contrast, during midday and the evening peak hours (Figure 4-9, *c* and *d*, respectively), phone usage density in the suburban areas of the region was higher than in the early morning or morning peak hours (Figure 4-9, *a* and *b*, respectively). Finally, it should be noted that the spatial distribution of an average hour in the day, displayed in Figure 4-9*f*, shows a pattern of population density distribution that, in general, is similar to the distribution exhibited in Figure 4-7*a*.

Figure 4-10 shows a side-by-side comparison of population density and the density of cell phone use during an average time of day. This comparison highlights the similarity in the patterns of cell phone use and the distribution of population in a region.

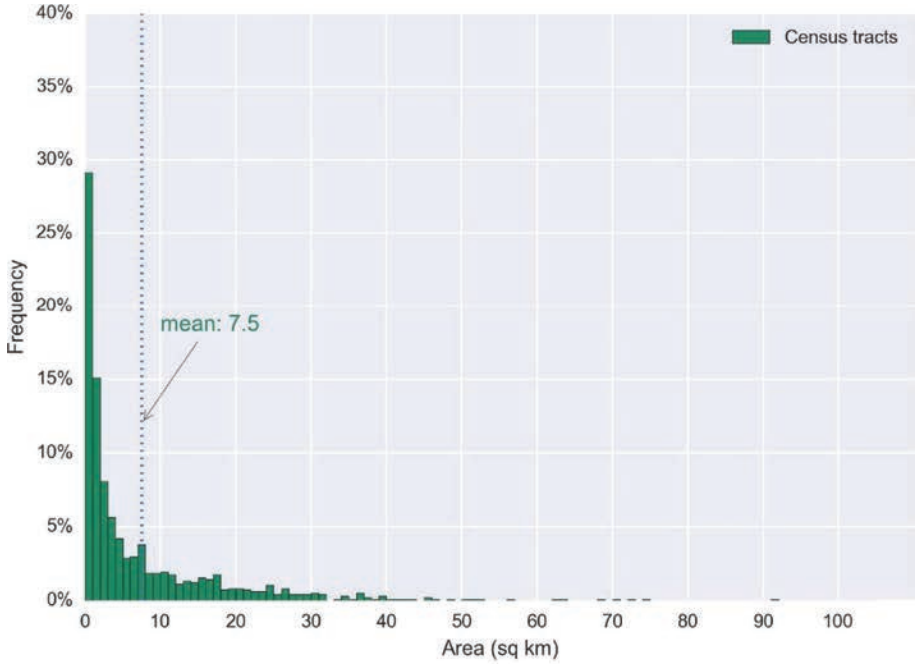
4.5.1.3 Triangulated Cell Phone Data via Smartphone Apps

Figure 4-11 shows the spatial distribution of the 18-month cell phone traces self-recorded via a smartphone app by the student volunteer. This data set is the record of the student's daily traces for 2 academic years in 2013 and 2014. The location accuracy of this data set and its format are similar to those of the triangulated CDR data.

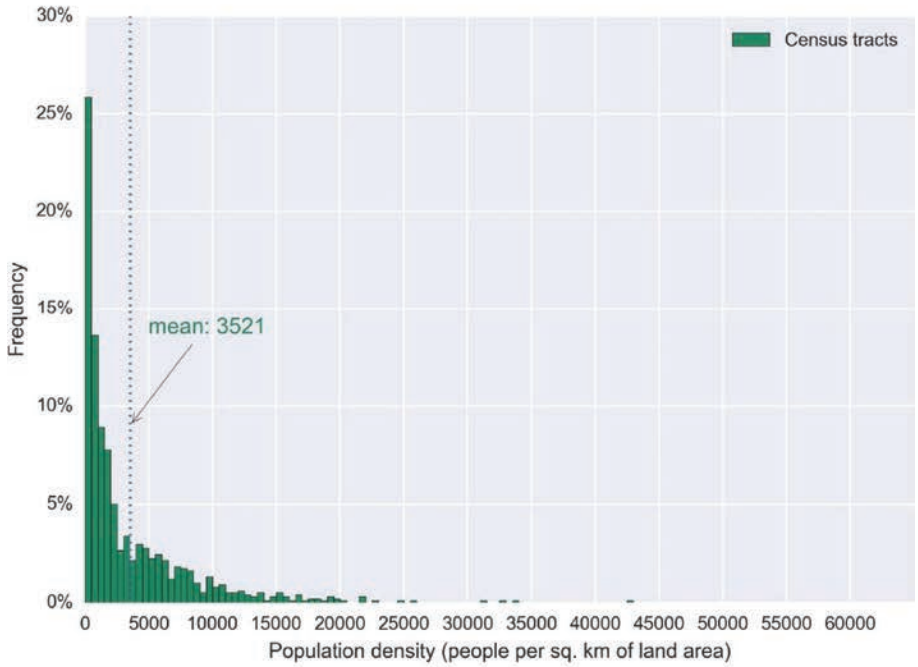


Source: Jiang et al. 2013.

Figure 4-7. Triangulated cell phone data and population density: (a) dots represent triangulated cell phone data at a regional level; (b) zoomed-in comparison between cell phone data and population density; and (c) patterns of cell phone use and population density in downtown Boston and Cambridge, Massachusetts.



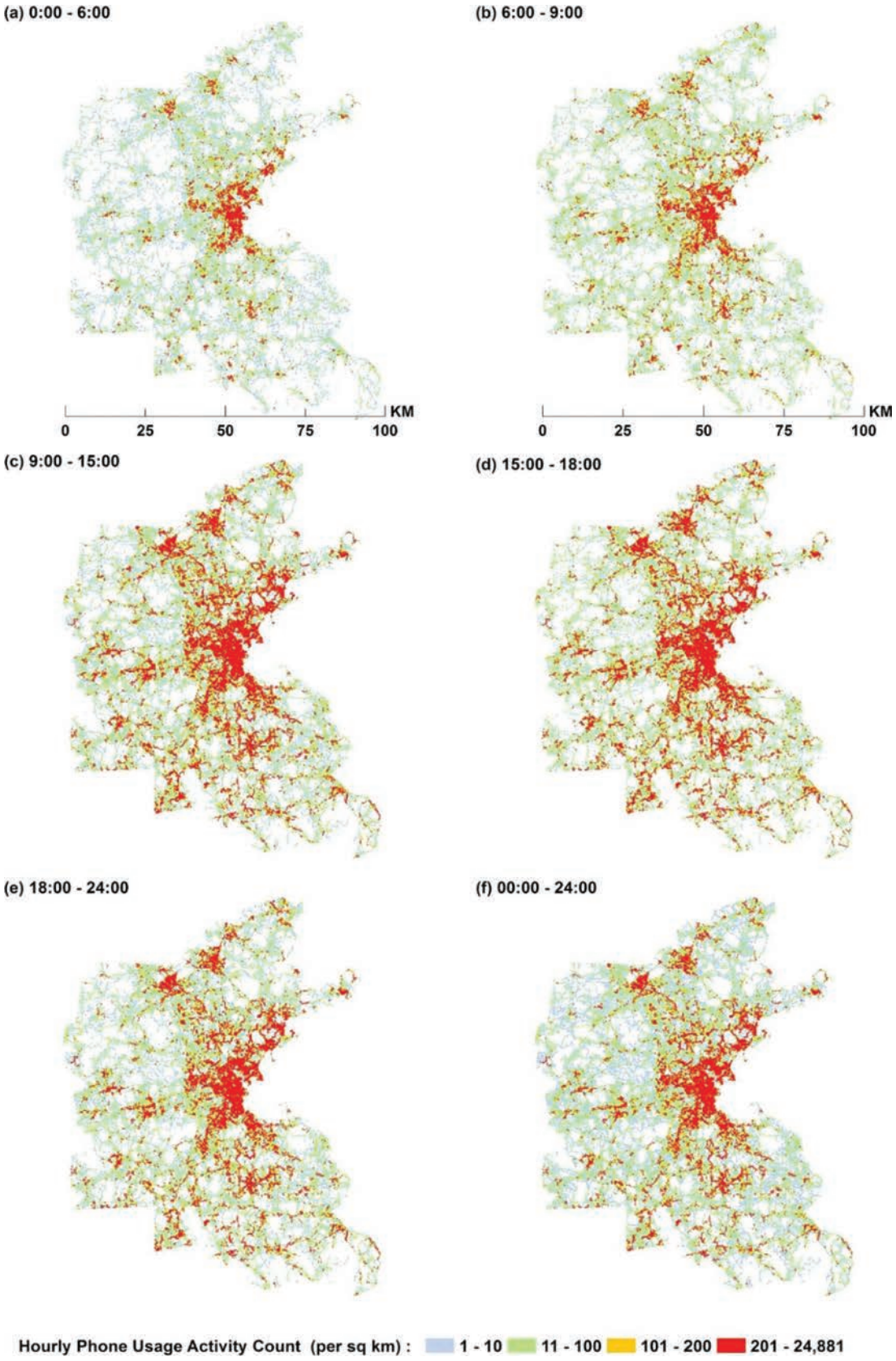
(a) Size distribution of census tracts



(b) Population density distribution (at the census tract level)

Source: Jiang et al. 2013.

Figure 4-8. Size of Census tracts and population density summary.



Source: Jiang et al. 2013.

Figure 4-9. Density of cell phone use by time of day.

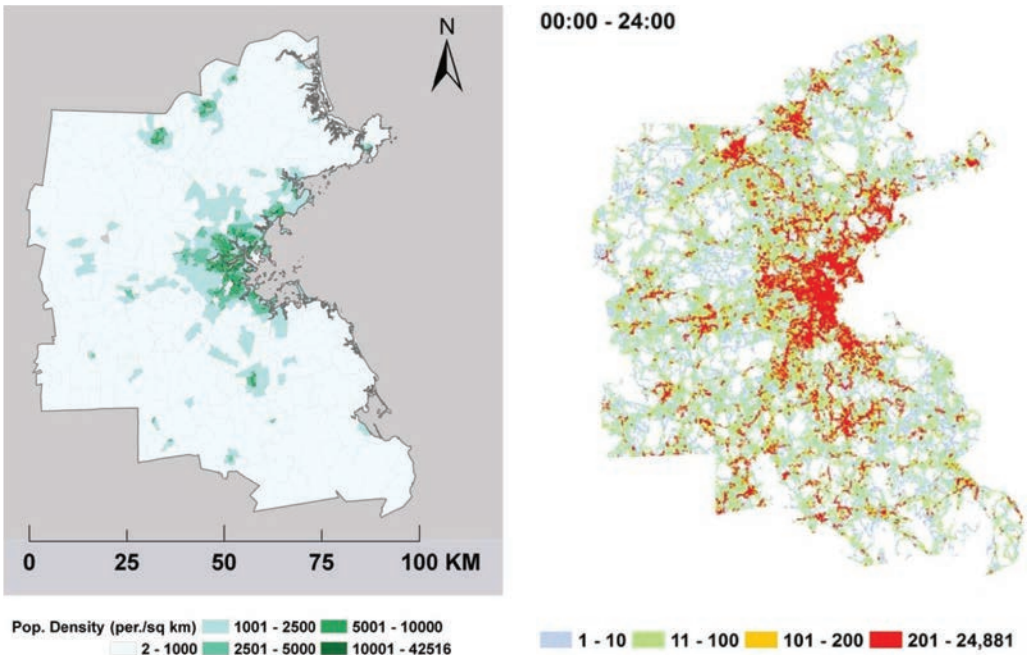


Figure 4-10. Population density and cell phone use patterns.

Although this set of cell phone data does not record calls, messages, or data events as the CDR data do, it captures every movement of the device while the app is on. Therefore, it catalogs the student user's movements in a smart way. This set of cell phone data is used in Chapters 5 and 6 to demonstrate some of the core algorithms used to extract meaningful stay locations of various types of activities, such as "home," "work," and "other."

4.5.2 Temporal Resolution

Equally important as the accuracy of the geographic location data is the amount and quality of temporal information included in cell phone data. The frequency of cell phone use for events such as calls, text messages, and Internet data access; the daily patterns of cell phone use; and the distribution of cell phone use over a typical day are important elements of temporal resolution discussed in this section.

4.5.2.1 Inter-event Time

A key variable that affects the trips that are inferred with cell phone data is the inter-event time distribution of the underlying cell phone data. In essence, more frequent use of the cell phone device for calls, text messages, and Internet data access reduces the inter-event time. Frequent use of the cell phone provides more location data points and therefore allows for a richer data set on travel patterns. Infrequent use of the cell phone provides a more limited set of travel information with fewer location data available.

Figure 4-12*a* shows the frequency distribution of the inter-event time between successive uses of the cell phone for the triangulated CDR data across all users for a 2-month period in the Boston region. Figure 4-12*b* shows the frequency distribution of inter-event time for the self-recorded cell phone data (via the smartphone app) of the student user over 18 months.

For both parts of Figure 4-12, the bin size of the histogram is 1 minute. The distribution has a long tail in the *x*-axis and is shown with a cut-off at 120 minutes. These patterns suggest that

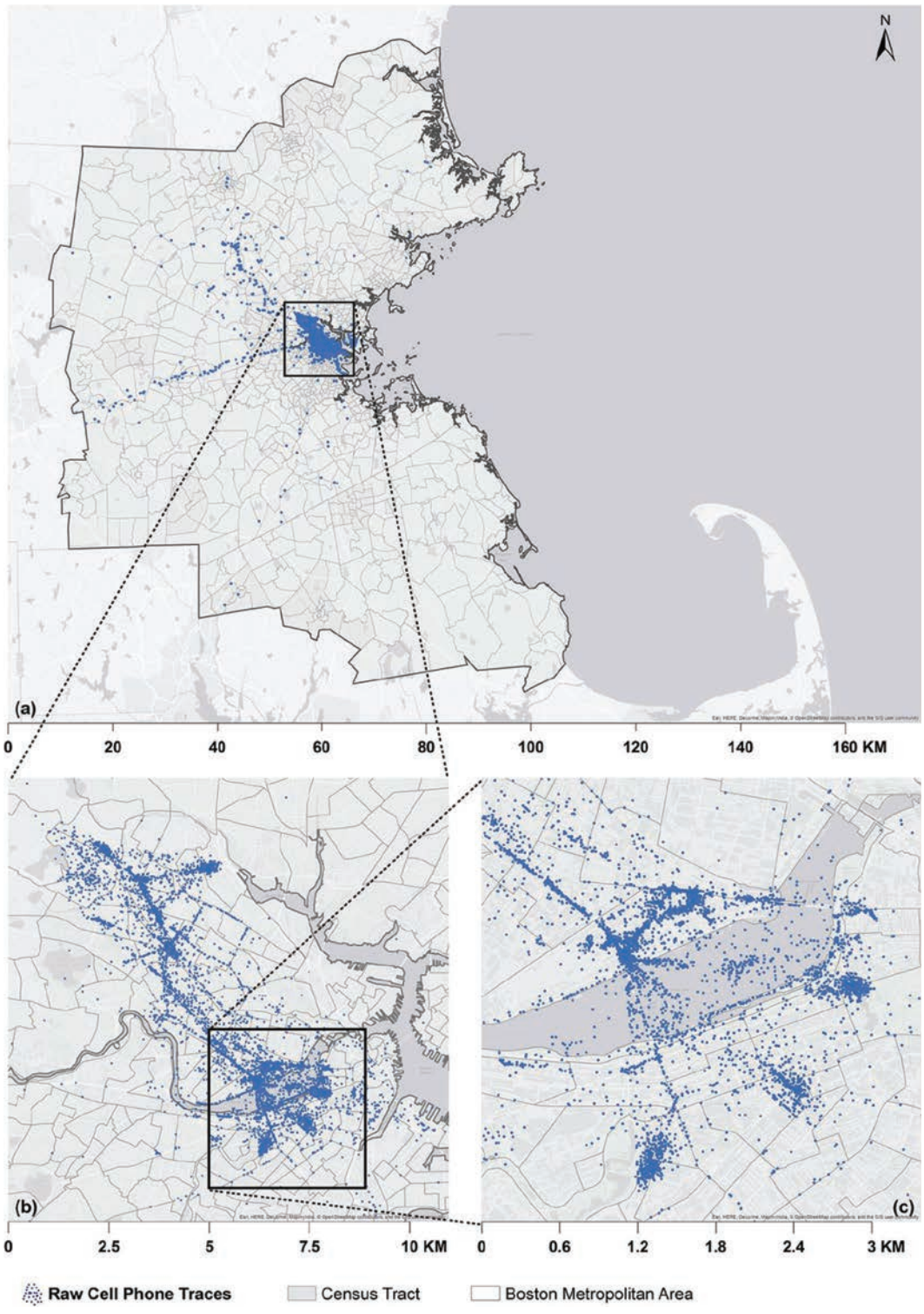


Figure 4-11. Triangulated cell phone traces of a volunteer individual.

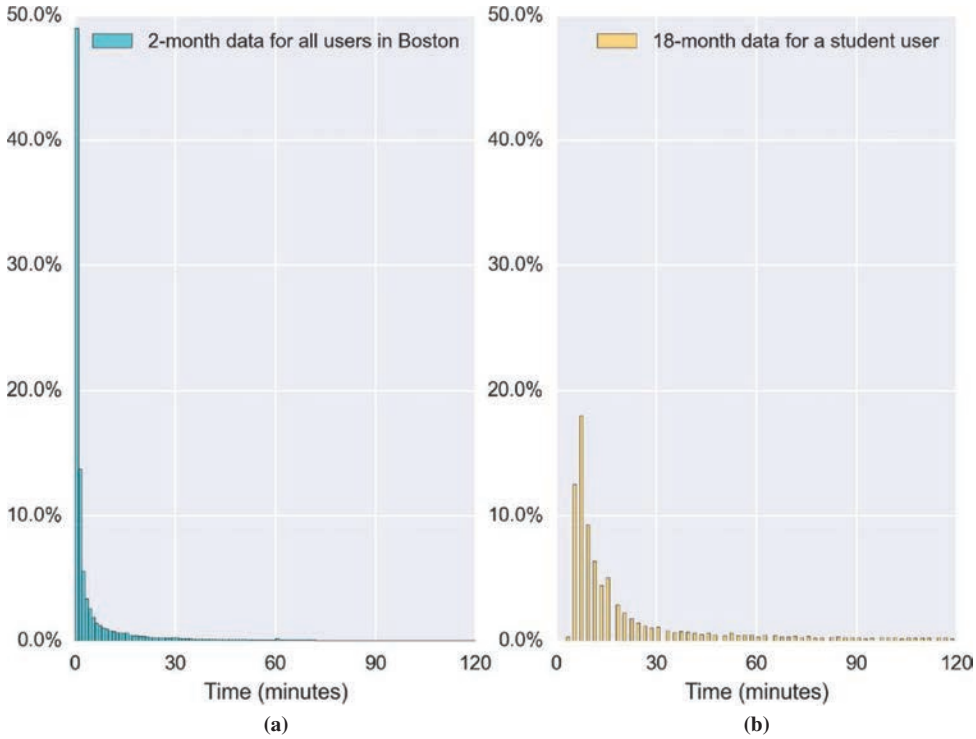


Figure 4-12. Frequency distribution of inter-event time in cell phone use.

just less than half of the cell phone records in the CDR data occurred consecutively within a 1-minute interval, 75% within 6 minutes, and 90% within 30 minutes, which indicates bursts of short events in cell phone usage patterns. This phenomenon has also been observed and discussed in other studies of telecommunication behavior and human dynamics (Vázquez et al. 2006, Barabási 2005, Hidalgo 2006, Malmgren et al. 2008, Karsai et al. 2012).

Given that the student employed a smartphone app that catalogs the device's movements instead of every use of the phone in a cellular network, the average inter-event time of the student user's cell phone data tends to be longer than that of an average person shown in the CDR data. Around half of the student's smartphone app records have inter-event times within 13 minutes, 75% within 56 minutes, and 90% within 233 minutes.

4.5.2.2 Daily Event Distribution

The daily event distribution of cell phone CDR data provides a picture of the incidence of location data throughout a typical day.

The inferred CDR locations were analyzed to identify travel patterns within a region. Figure 4-13*a* presents the frequency distribution of daily cell phone events for the CDR data of all sampled users for 2 months in the Boston region; Figure 4-13*b* presents the self-recorded smartphone records of the student user for 18 months. These records are not directly comparable, given that the self-recorded smartphone records represent movements by the student and are expected to be greater than the inferred trips from the CDR data.

Both parts of Figure 4-13 have a bin size of one event count. The distribution for the regional CDR data has a long tail, shown here with a cut-off at 150 events in the *x*-axis. The 10th, 25th, 50th, 75th, and 90th percentiles of the daily events for the CDR data of all sampled users in Boston are 3, 8, 24, 61, and 129 events, respectively. The median estimate is 24 cell phone events during a typical day, which is likely to yield good travel information for a typical day.

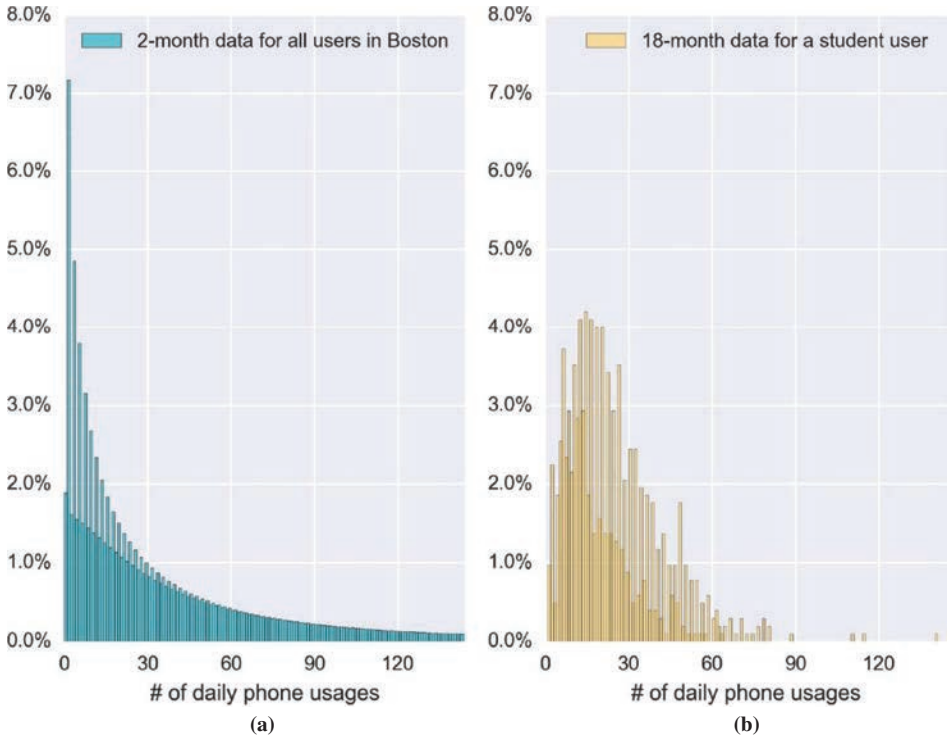


Figure 4-13. Frequency distribution of daily cell phone use patterns.

On the low end, three events and eight events correspond to infrequent cell phone use and are likely to yield either no travel information at all or a low level of inferred travel. On the high end, 129 events represents a high rate of cell phone use during a typical day by users who frequently use their cell phones to talk, text, or access the web.

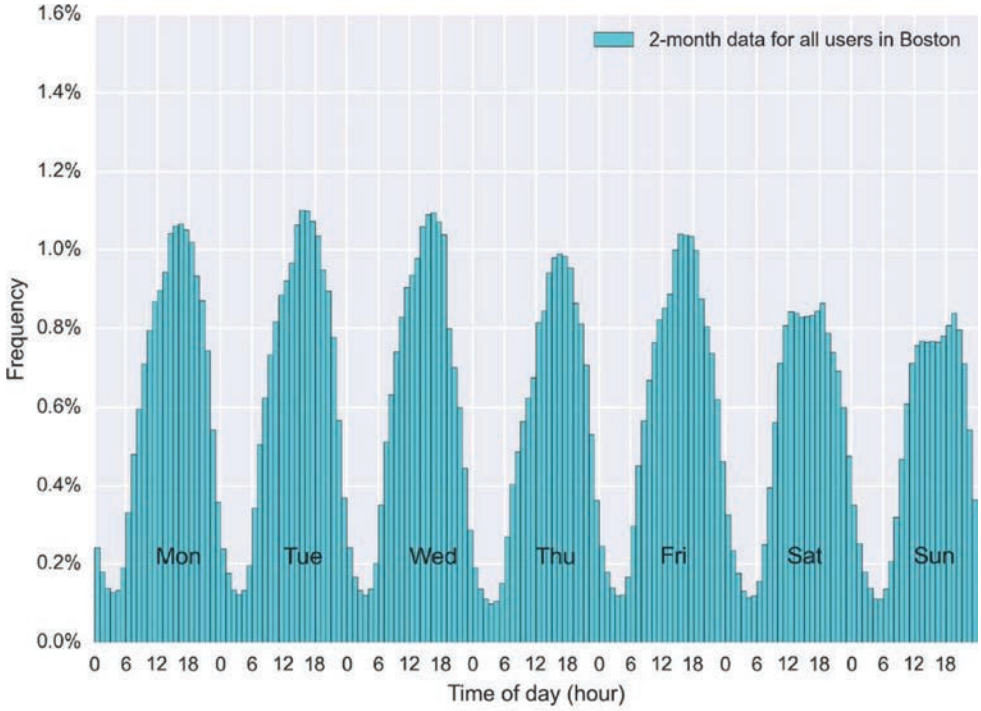
The distribution of the smartphone data of the student user are 6, 11, 19, 31, and 46 events, respectively, for the same percentiles. These are estimates of movements that are not directly comparable to the cell phone events or to estimates of daily travel. The median value of 19 total daily movements includes true trips to activities as well as much shorter movements that do not qualify as travel. On the lower end of the spectrum, six movements may correspond to a low level of travel, while the upper end estimate of 46 movements is almost certainly heavily influenced by outlier observations.

In the CDR data, the proportion of daily phone usage events with even counts is higher than those with odd counts, revealing the symmetric nature of data recording. Presumably, when a service carrier records a user's cell phone usage for billing purposes, it needs to record the start and end time of each usage event for phone calls and for data transmission, thus generating two records for each usage. On the other hand, it may only record one instance for events such as sending or receiving text messages.

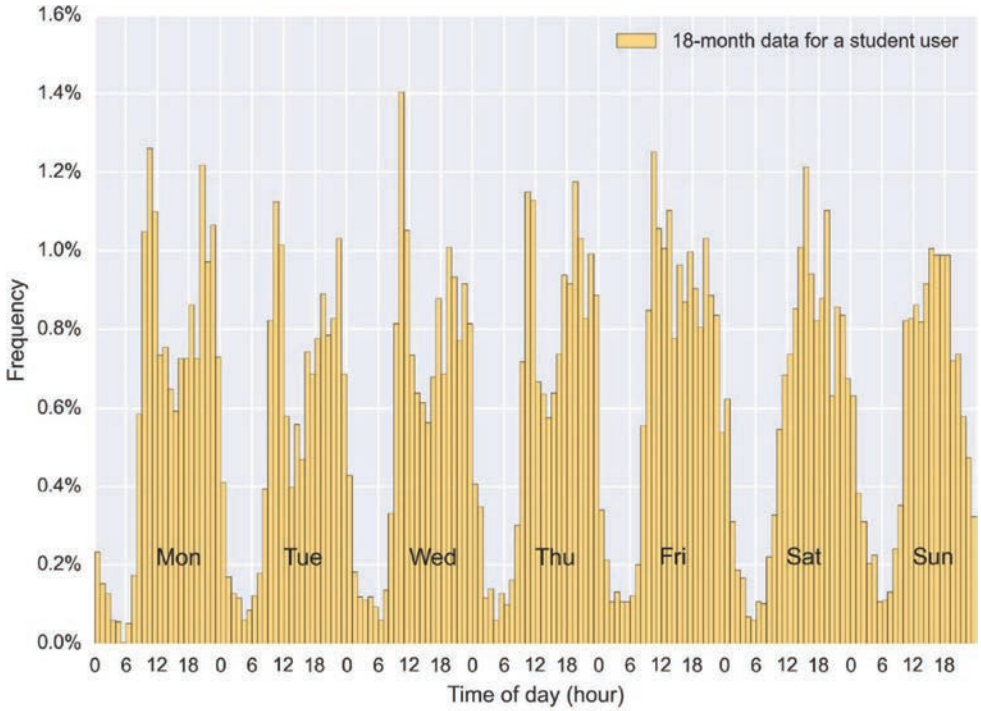
4.5.2.3 Temporal Rhythms of Cell Phone Data

Figure 4-14*a* shows the hourly trend of cell phone usage patterns in the population during a typical week on the basis of the 2 months of CDR records of all sampled users in the Boston region in 2010. The figure shows the temporal pattern of phone usage in the cellular networks for the population in the metropolitan area.

These patterns suggest a single peak in the late afternoon around 4 to 5 p.m. during weekdays for all days of the week. During the weekend, the distribution is flatter, with two peaks on



(a) CDR data of all sampled users in the Boston Metropolitan Area



(b) Self-recorded smartphone data of a student user

Figure 4-14. Time-of-day cell use: sample and individual user patterns.

Saturday (one around noon and one around 6 p.m.) and one peak on Sunday around 6 to 7 p.m. in the early evening.

Figure 4-14*b* shows the hourly trend of the smartphone app records for the student user in a typical week, according to his self-collected data from 2013 and 2014. The figure shows, more or less, the mobility patterns of the student user, given that the smartphone app records movements of the device.

In this instance, the graph displays the temporal rhythm of the student's movements. There are two distinct daily peaks on weekdays: one in the morning and one in the evening, with fewer movements during the midday. During the weekends the pattern changes, with one daily peak in the early afternoon on Saturday and one daily peak in the early evening on Sunday.

4.6 Summary

This chapter discusses key technical concepts related to cell phone data. A comparison of traditional data and the new massive big data sources provides the background for identifying the strengths and weaknesses of using cell phone CDR data in transportation planning and modeling.

A closer look at the CDR data includes the typical layout of the data set and an in-depth discussion of potential issues with cellular data, including the penetration rate of the cell phone market, the network operators in a region, sample selection, privacy considerations, and the lack of socioeconomic information in CDR data.

A concept of key importance to the research is the temporal resolution of cell phone data. The richness and value of CDR data depends directly on the frequency of use of each cell phone device for calls, texts, and Internet data access. The concepts of inter-event time, the daily distribution of events, and the temporal rhythms of cell phone data are discussed to provide a good understanding of the temporal nature of CDR data.

A second key concept is the spatial resolution of cell phone CDR data. Given the reliance on signals sent and received during a typical day, the analyst needs to analyze CDR records from an individual device across multiple days to infer locations and activity types in order to develop O-D matrices. The uncertainty of location estimates, the phenomenon of device oscillation, and the triangulation methods discussed provide the background on spatial resolution.

The analysis presented in Chapters 5 through 8 relies on CDR data collected over 2 months from 2 million cell phones. The research team used Boston as a regional case study to compare CDR data with summaries from traditional surveys and results from the regional model.

CHAPTER 5

Extraction of Daily Trajectories

5.1 Roadmap to the Chapter

This technical chapter discusses how the noisy and massive cell phone call detail record (CDR) data were analyzed to extract daily trajectories. The discussion starts with the motivation for the approach and then focuses on how the noise from CDR data needs to be removed to arrive at meaningful data for assessing travel patterns.

To measure travel in space and through time, some location and activity inferences were made. This chapter describes the need to drop spurious observations in order to parse trajectories and identify stay locations. The analysis steps needed to identify activity types at detected stay locations where individuals carrying cell phones conduct their activities are reviewed. Temporal activity patterns are examined by analyzing start times and activity duration patterns.

The analytical components of this chapter describe how grid-based and point-based algorithms analyze traces, locations, and the time spent at each location to identify and extract stay points and stay regions. Data from the student user are used to present examples of stay extraction results.

The density of stay locations and the duration of activities for different arrival times are presented at the regional level. The results were aggregated to the zone level to facilitate origin–destination (O-D) comparisons of travel patterns.

5.2 Motivation and Purpose

A critical task in urban and transportation planning is the examination and understanding of what people are doing in space and time (Chapin 1974; Lynch 1976; Hägerstrand 1989; Ahmed and Miller 2007; Janelle 2012; Jiang et al. 2012a, 2012b). To answer this question by using billions of cell phone traces in the CDR data, it is crucial to infer the spatial and temporal activities that people engage in and position their travel as reflecting the need to pursue activities that are located elsewhere (Manheim 1979, Pinjari and Bhat 2011).

As discussed already, location data gathered from cell phones, while massive, are often noisy across spatial and temporal dimensions. They can also be biased because of differences in usage of the technology and in penetration rates among different segments of the population. Moreover, CDR data provide an irregular sampling frequency, as they only include information when a cell phone is connected to cellular networks for a call, a text message, or data usage. CDR data also provide no insights into the mode of transportation used.

However, cell phone data provide an opportunity to measure travel more directly because of the sheer volume of data available to analyze. Such data sets are only growing larger and richer each year as the economy and individuals make increasing use of mobile-based systems.

Passively collected cell phone data provide an unparalleled scale of observation. New methods of estimating travel demand need to balance trade-offs between small, but complete, data for a short period as compared with large, but incomplete, data over a longer period (Toole et al. 2015). In both cases, noise and biases must be carefully dealt with to produce valid measurements.

To this end, the research team addressed several challenges presented by the cell phone data. Key insights from past research that tackled these problems to extract meaningful locations from massive and passive cell phone data for estimates of travel demand have been integrated into this body of work.

5.2.1 Motivation

CDR data include spatiotemporal information on people's movements relative to cell towers or triangulated locations, depending on the positioning technology used by the mobile service carrier. An initial study by Wang et al. (2012) used CDR data to estimate travel demand. The study generated transient O-D matrices for different time periods by simply counting as a trip a pair of consecutive calls made within the same hour from two different towers. By assigning the converted intersection-to-intersection transient O-D matrices to the road network and using a bipartite network framework, Wang et al. (2012) presented a method for analyzing road usage patterns and pinpointing areas as driver sources contributing to major traffic congestion in Boston, Massachusetts, and the San Francisco Bay area in California.

Following a similar approach, Iqbal et al. (2014) used CDR data collected in Dhaka, Bangladesh, over 1 month and combined them with traffic count data to estimate intersection-to-intersection transient O-D matrices. By using an optimization-based approach, they generated expansion factors for node-to-node transient O-D flows and compared the results with the limited traffic count data.

While Wang et al. (2012) presented groundbreaking work using CDR data to infer road usage patterns, the method of using transient O-D flows for travel demand estimation could lead to a biased view of movements. Rather than modeling travel flows between activity destinations, transient O-D data capture segments of travel on the basis of the appearance of people in space and time as presented in raw cell phone data.

Transient O-D data are particularly problematic for raw data that contain noise caused by triangulation of mobile positions such as the oscillation problem described in Chapter 4 and cases in which CDR data have low spatial resolution. For example, if the distance between cell towers is more than a few kilometers, transient O-D matrices can introduce biases, even if the road networks within each tower coverage area are dense. Traffic may be detoured to local roads, although the assigned travel path is not necessarily a direct route from the true origin to the destination.

5.2.2 Purpose

To address this issue, the noise from cell phone trajectories must be removed by identifying and dropping spurious points or calls made in the middle of routes rather than at an origin or a destination. It is important to parse the trajectories observed in cell phone data into meaningful locations, termed "stays." To produce meaningful estimates of travel demand, the goal, in general, is to find suitable algorithms with which to extract meaningful stay locations from noisy cell phone data for further analysis. By applying algorithms to parse passive cell phone trajectories into stay locations, researchers can estimate O-D trip tables for an average day or by time period from tower-based or finer-grained triangulated cell phone traces (Alexander 2015, Colak et al. 2015, Toole et al. 2015).

This chapter presents two sets of algorithms that are tailored to address the distinct characteristics of cell phone data triangulated at the 200- to 300-meter accuracy level. These algorithms

58 Cell Phone Location Data for Travel Behavior Analysis

are designed to filter the cell phone data to infer human activities and travel in space and time. Parsing passive cell phone data to extract "stay locations" identifies activity anchor points in an individual's daily travels. This approach allows differentiation of destinations from pass-by points and helps identify an individual's mobility pattern. Similar research is also key in analyzing GPS data from commercial vehicles.

The wide adoption of smartphones and location-based mobile apps has led to a vast body of computer science literature on the topic of trajectory mining. Jiang et al. (2013) provide a review of these techniques. A detailed review on trajectory mining methods can also be found in Zheng's (2015) research.

5.3 Stay Extraction Algorithms

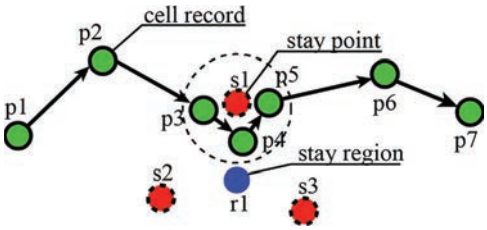
5.3.1 Grid-Based Algorithm

In the preprocessing of the CDR data, the first step is to identify stays, which represent phone records that are registered when users engage in activities. These stays are distinct from pass-by points, which represent records made while traveling along each user's trajectory.

As illustrated in Figure 5-1, a stay point is identified by a sequence of consecutive cell phone records bounded by both temporal and spatial constraints. The spatial constraint is defined by the roaming distance when a user is staying at a location, which should be related to the accuracy of the technology collecting location data. In this study, the spatial accuracy for triangulated CDR data, also known as the roaming distance, was set as 300 meters. This distance was established to approximate the area that might likely be traversed on foot as part of an urban activity.

The temporal constraint is the minimum duration spent at a location, which is measured as the time difference between the first and the last record in a stay location. In this study, records satisfying the spatial constraint (300 meters) and temporal constraint (duration of 10 minutes or more) were counted as stays. Once a stay point was identified, the geographic location was set as the centroid of all records belonging to that stay. In Figure 5-1, Point *s*1 is the centroid of Record Locations *p*3, *p*4, and *p*5. Both constraints may be adjusted on the basis of the data availability and the researchers' understanding of the quality of the data.

The second step is to distinguish stay regions from stay points. Different stay points identified in a user's several different trajectories may refer to a same location, but the coordinates of these stay points are unlikely to be exactly the same. A grid-based clustering method was used to cluster stay points and get stay regions.



Source: Jiang et al. 2013.

Figure 5-1. Illustration of the stay extraction process [green dots = raw triangulated CDR data points (p); red dots = stay points (s); blue dot = grid-based stay region (r) from the cluster of stay points].

As shown by Zheng et al. (2010), the advantage of the grid-based clustering method over the *k*-means algorithm and the density-based ordering points to identify the clustering structure (OPTICS) algorithm is that it can constrain the output cluster sizes. This property is desirable when each location should have a bounded size and the accuracy of the records is within a certain range. The procedure for performing grid-based clustering follows these steps:

- The entire region is divided into rectangular cells of about 100 meters (one-third of the roaming distance of 300 meters).
- All the stay points are mapped to the appropriate cell.
- The unlabeled cell is iteratively merged with the maximum stay points and its unlabeled neighbors to create a new stay region.
- Once a cell is assigned to a stay region, it is marked as labeled.[1]

In Figure 5-1, the three stay points are clustered to one stay region ($r1$).

5.3.2 Point-Based Algorithm

In contrast to the grid-based algorithm, the point-based stay region extraction is designed to exploit the maximum spatial accuracy possible. To extract individuals' whereabouts [including their stationary stay locations (to infer activity types) and their moving pass-by locations (to infer travel path and road usage)] from phone records, Jiang et al. (2013) employed a method inspired by Hariharan and Toyama (2004) that was originally designed for processing GPS traces.

GPS data are recorded with a high frequency so that they can be treated as continuous trajectories. Unlike GPS data, cell phone data are perceived with indefinite gaps in space and time. Furthermore, the locational accuracy of cell phone data is lower than that of pinpointed GPS traces, depending on the technology (Renso et al. 2008). On the basis of these differences, Jiang et al. (2013) tailored the algorithms for the CDR data by using the spatial and temporal approach described in Sections 5.3.2.1 and 5.3.2.2.

5.3.2.1 Spatial Dimension

Let sequence $D_i = (d_i(1), d_i(2), d_i(3), \ldots, d_i(n_i))$ be the observed data for a given anonymous user i,

where

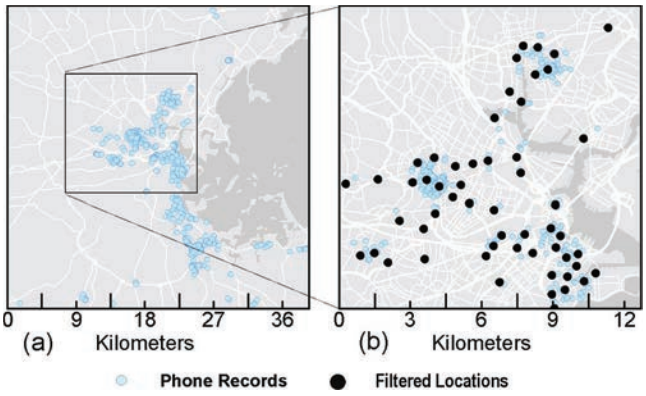
$d_i(k) = (t(k), x(k), y(k))1$ for $k = 1, \ldots, n_i$;
$t(k)$ = time observation;
$x(k)$ = longitude; and
$y(k)$ = latitude of the kth observation of user i.

First, points $d_i(k)$ that are spatially within the roaming distance of 300 meters to their subsequent observations are extracted, say, $d_i(k+1), d_i(k+2), \ldots, d_i(k+m)$.

To reduce the jumps in the location sequence of the cell phone data, it is assumed that $d_i(k), \ldots, d_i(k+m)$ are observed when user i is at a specific location. That location corresponds to the medoid (Med) of the set of locations $(x_i(k), y_i(k))1, \ldots, (x_i(k+m), y_i(k+\mathrm{m}))1$, and it is denoted by $\mathrm{Med}((x_i(k), y_i(k))1, \ldots, (x_i(k+m), y_i(k+m))1)$.

This treatment respects the time order, at first, to ignore noisy jumps in the estimated location. At the next step, the treatment disregards time ordering to apply Hariharan and Toyama's (2004) agglomerative clustering algorithm, which consolidates points that are close in space but may be far apart in time. The points to be consolidated together form a cluster whose diameter is

[1]For more details about the algorithm, readers are referred to Zheng et al. (2010).



Source: Jiang et al. 2013.

Figure 5-2. Filtered locations from phone records with a 300-meter threshold: (a) raw phone records and (b) filtered locations. Two months of data from an anonymous user were used in the filtering.

required to be no more than a certain threshold. Again, the observation is moved to the location of the new medoid of the clusters, as shown in Figure 5-2.

5.3.2.2 Temporal Dimension

The top part of Figure 5-3 presents the frequency distribution and the bottom part the cumulative distribution function of stay duration for the extracted filtered locations identified in Figure 5-2. The stay duration criterion is imposed on the filtered data, and the stay locations whose duration exceeds a certain temporal threshold are extracted. The temporal threshold was set at 10 minutes. The cyan vertical bars in Figure 5-3 show the positions of the 10-minute stay duration in these two distributions.

In the example discussed, 31 distinct stay locations were extracted from the 1,776 phone records in the 2-month period of an anonymous user represented by the red points in Figure 5-4. The points represented by black circles in Figure 5-4 are pass-by points at which lengthy stays were not observed.

It is possible that the user stayed in some of these pass-by locations as well as in other locations that were not observed. In these cases, information about time and location is totally or partially latent, as they are not observed in the cell phone records. However, all the stay locations frequently visited by the user ought to be extracted from the cell phone data, especially if the observation period is long enough, as was the case with this example. Therefore, the pass-by locations were filtered out and the stays were assumed to be true trip origins or destinations, between which trips were made.

5.4 Stay Extraction Results

5.4.1 Individual Example

This section discusses the results of applying the algorithms to the data from the anonymous student user. As discussed earlier, the value of the student user data set was that it functioned as a validation step. The detailed trace data were combined with prompted recall information not available in the CDR data. Therefore, this step provided information critical to understanding whether or not the algorithms performed a reasonable job in measuring stays and pass-by locations.

Figure 5-5 presents the stay extraction results obtained using the grid-based algorithm presented in Section 5.3.1. Figure 5-6 presents the extracted stay results obtained using the

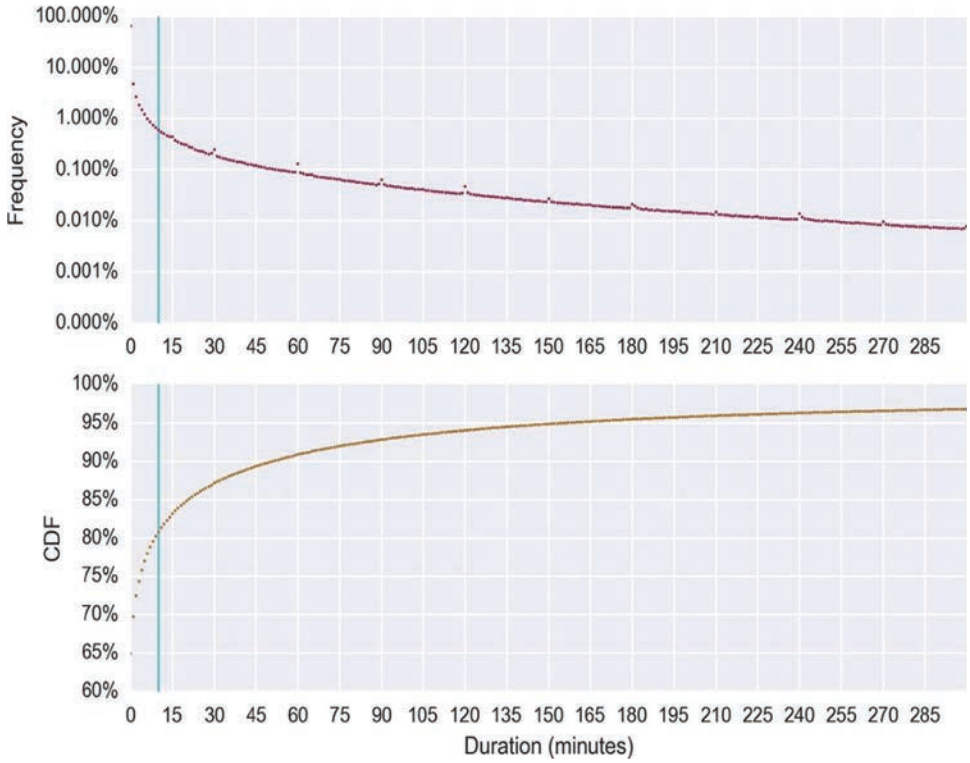
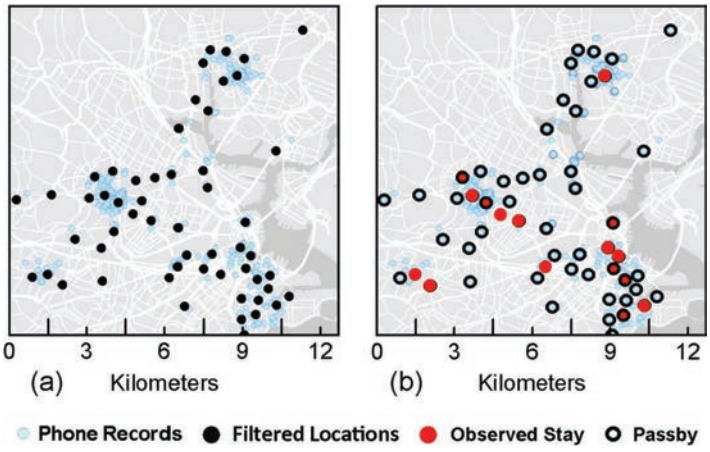


Figure 5-3. Pattern of stay durations: (top) frequency distribution of stay durations and (bottom) cumulative distribution of stay durations. Data represent stay durations for all filtered locations in 2 months of data from Boston.



Source: Jiang et al. 2013.

Figure 5-4. Inference of stays and pass-by areas by using a 10-minute threshold: (a) filtered locations and (b) stays and pass-by areas. Two months of data from an anonymous user were filtered with a 10-minute threshold.

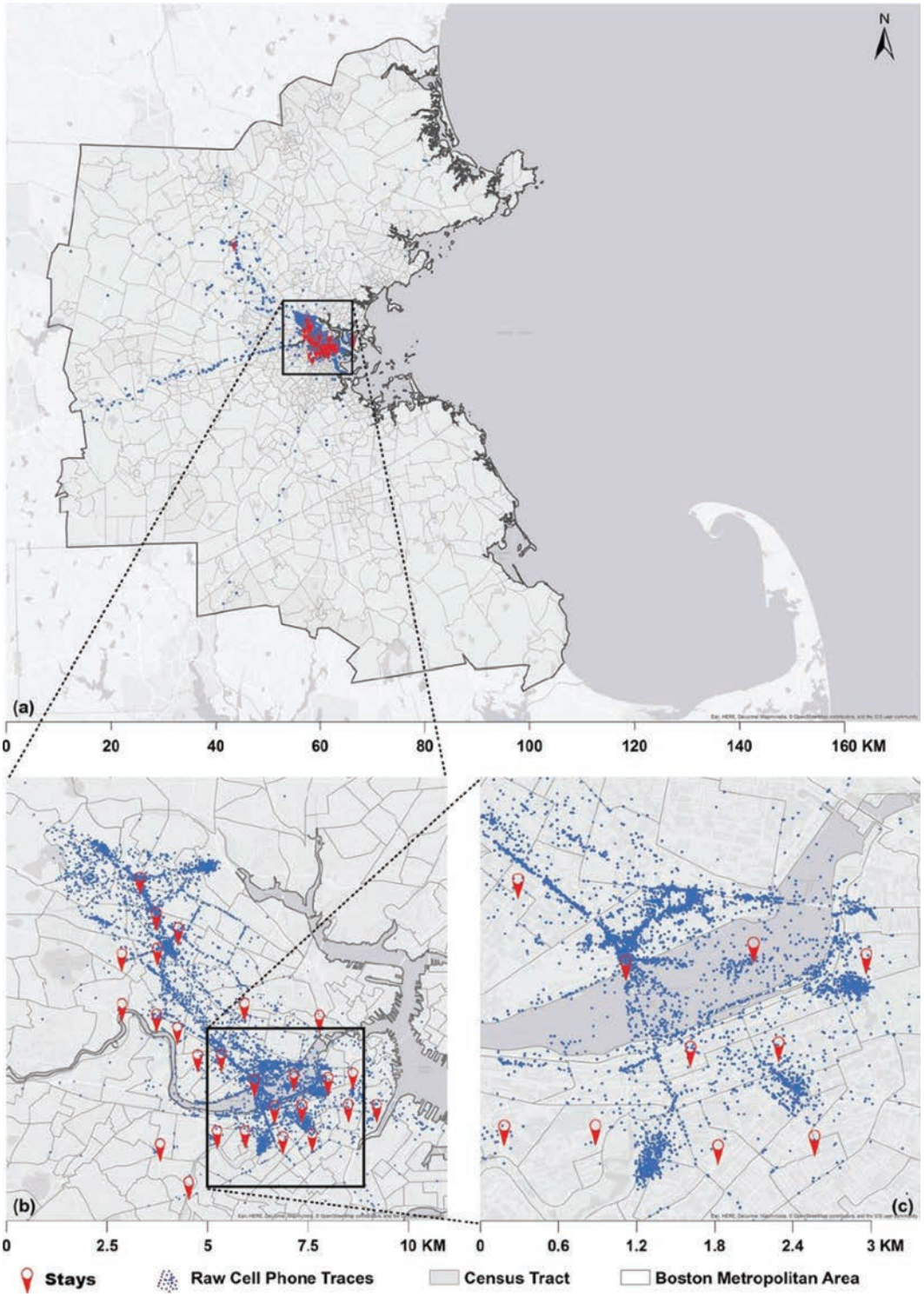


Figure 5-5. Stay locations extracted by using grid-based algorithm (blue points = user's raw cell phone data; red bulbs = stays extracted by using grid algorithm and anonymous user data. Grid-based algorithm used 18 months of data from an anonymous user.

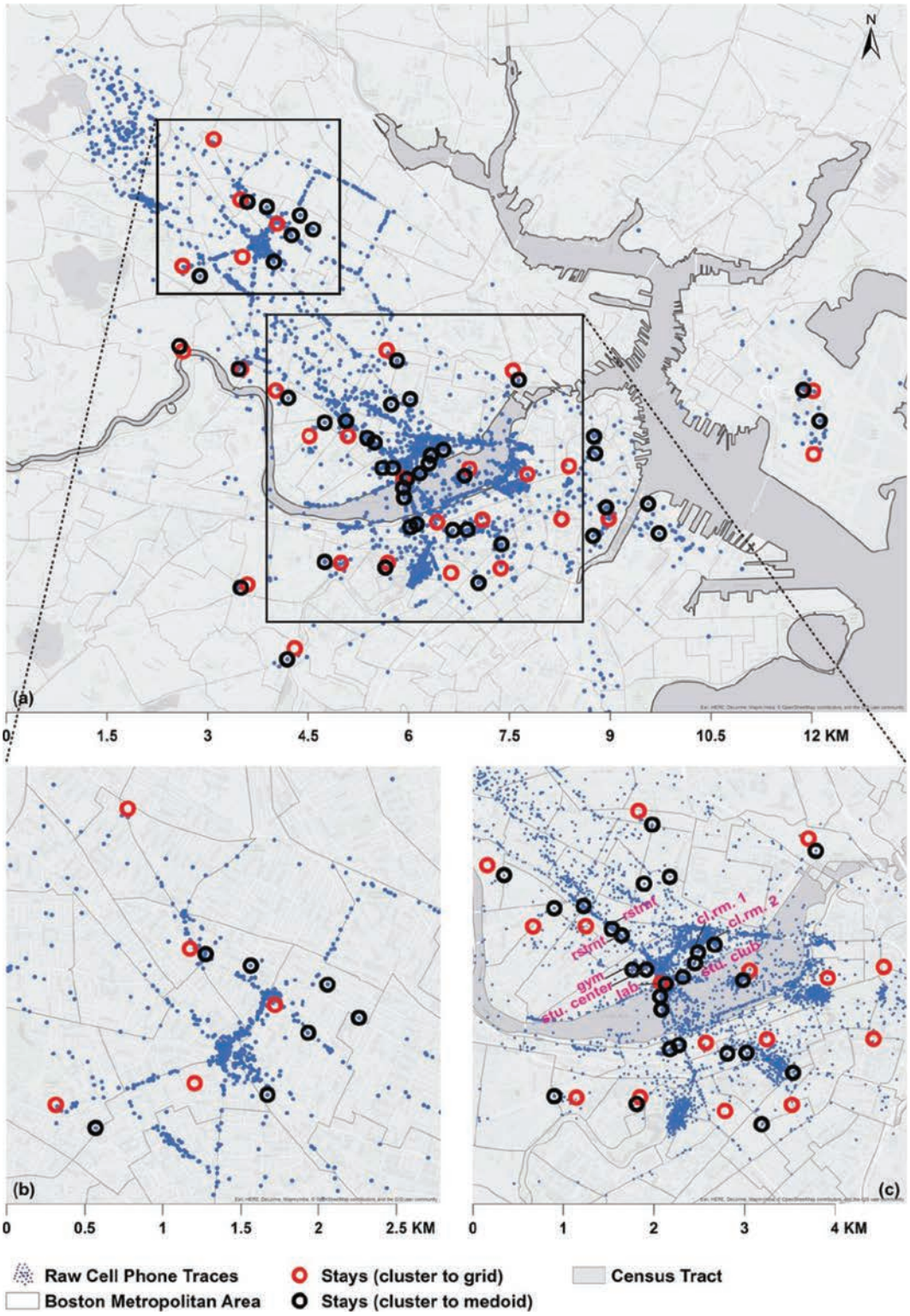


Figure 5-6. Effect of grid- and point-based algorithms on stay locations (blue points = raw phone records; circles = filtered locations). Eighteen months of data from an anonymous user were used in the filtering.

point-based algorithm presented in Section 5.3.2 and compares the results with those derived using the grid-based algorithm. In both cases, the noise in the raw cell phone data was removed and locations were reduced to a few anchor points where the individual was estimated to have conducted activities.

- The comparison in Figure 5-6 of the stay extraction results from the grid-based algorithm and the point-based algorithm suggests that the two algorithms have different advantages. In most cases, the stays extracted by these two algorithms were close to each other.
- However, the stays extracted from the point-based algorithm are always on top of existing raw cell phone records, and thus more sensitive and relevant when referring to local spatial context. In contrast, the stay results from the grid-based algorithm are regionalized into locations that are not necessarily near the exact locations the user visited.
- One of the advantages of the grid-based algorithm is that it is faster. However, when information with high spatial resolution is aggregated into a coarser resolution, some local details are lost.
- The grid-based algorithm also has advantages in terms of privacy protection when compared with the point-based algorithm. Given that the point-based algorithm uses the agglomerative clustering method to preserve much of the spatial information in the raw data, it keeps the original spatial resolution of the data but is costly in terms of computing speed.
- For purposes of travel demand estimation, grid-based stay extraction may be good enough, given that data from the extracted stay points are aggregated into zones such as Census tracts or traffic analysis zones (TAZs).
- For other studies, such as inferring population density at the building or block level, a point-based stay extraction algorithm that will provide higher spatial accuracy may be more appropriate.

5.4.2 Regional Perspective

This section uses data from the same sample day presented in Figure 5-4 to demonstrate the results of extracted pass-by locations and stays in the Boston region.

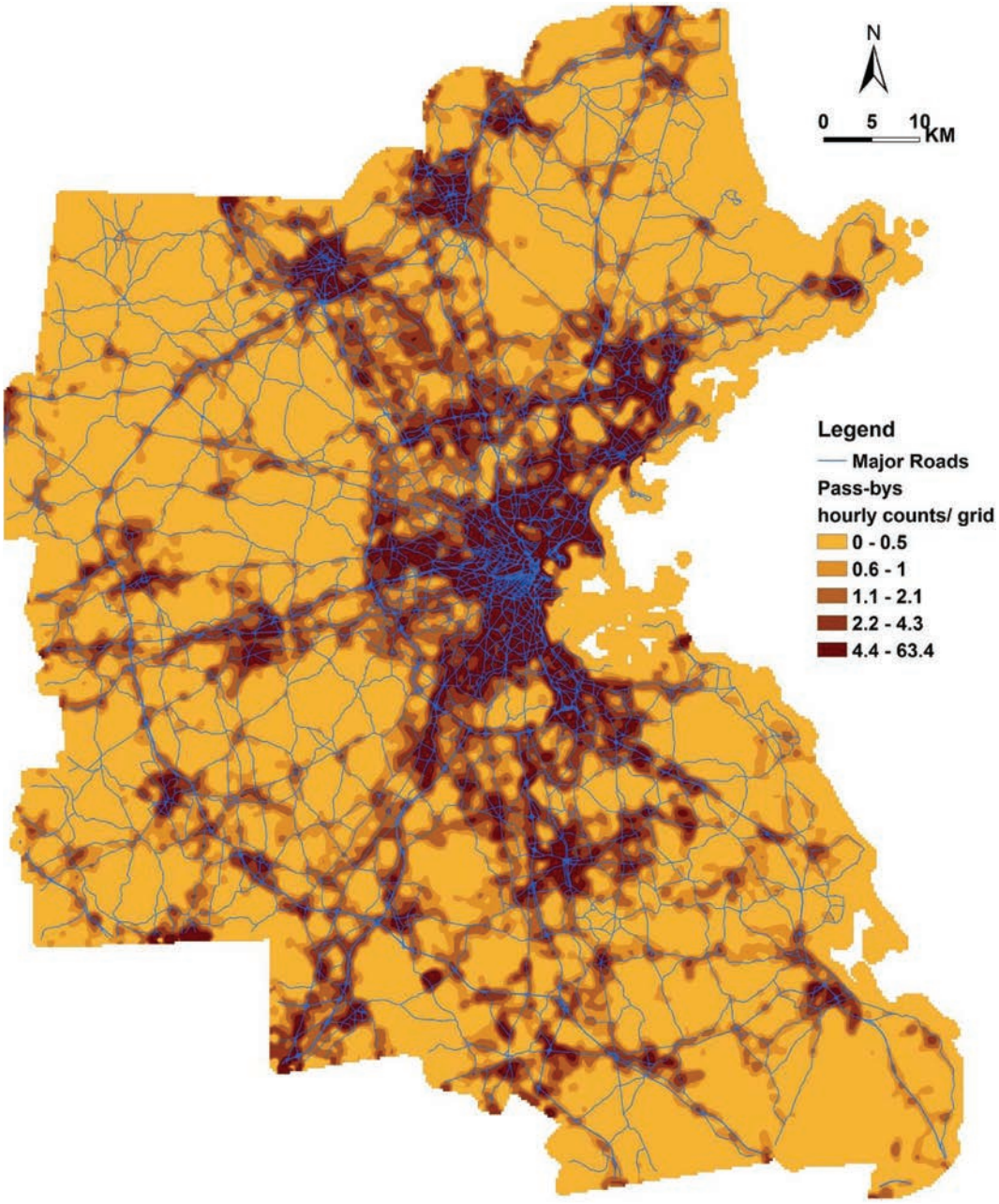
Figure 5-7 shows the spatial distribution of the kernel density estimation of pass-by filtered locations with a temporal threshold of less than 10 minutes, as discussed in Section 5.3.1. The major roads in the region are presented in blue to show their spatial relationship to the pass-by locations extracted from the CDR data.

As the road network in the downtown area is denser, it is harder to see a clear relationship, other than the high pass-by density in the downtown area as compared with the suburbs. However, Figure 5-7 shows a clearer spatial correlation of the major roads and the pass-by locations along major transportation corridors, especially in the outer areas of the metropolitan region. Presumably, those cell phone traces correspond to cell phones that were being used along those roads.

Figure 5-8 presents the spatial distribution of the kernel density estimation of the stays, which represent filtered anchor locations with a duration of 10 minutes or more in the CDR data for the same sample day. It shows a higher density of stays in the center of the region, while it also highlights several subcenters in the suburbs.

5.4.3 Stay Duration by Arrival Time

To validate the extracted stays in terms of the temporal dimension, the research team looked further at stay duration by arrival time observed both in the travel survey data and the synthesized CDR data for the Boston region.

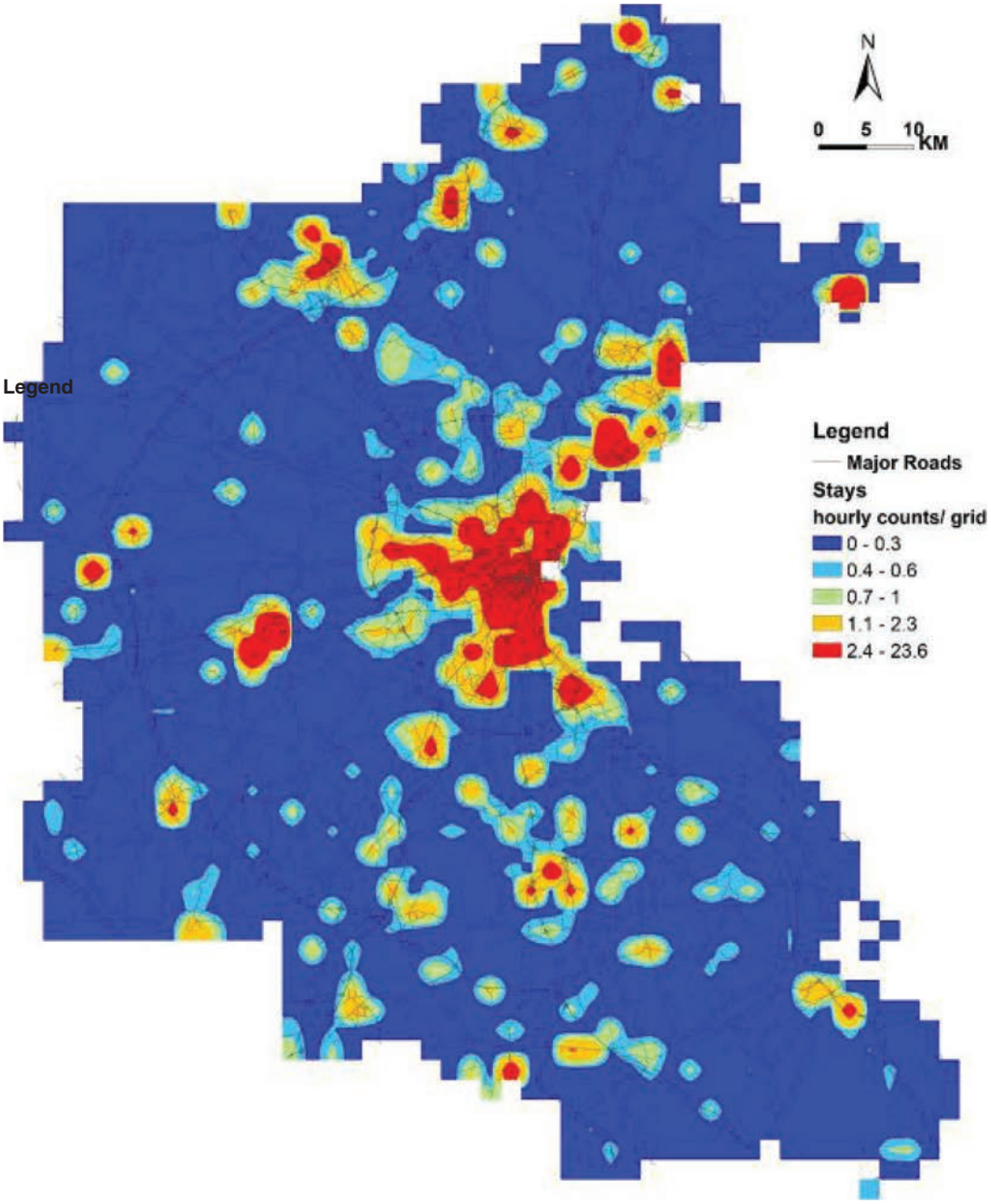


Source: Jiang et al. (2013)

Figure 5-7. Pass-by locations and their spatial distribution in the region. CDR data for a 2-month sample period in Boston (same as in Figure 4-7) were used to show the kernel density estimation of the pass-by locations.

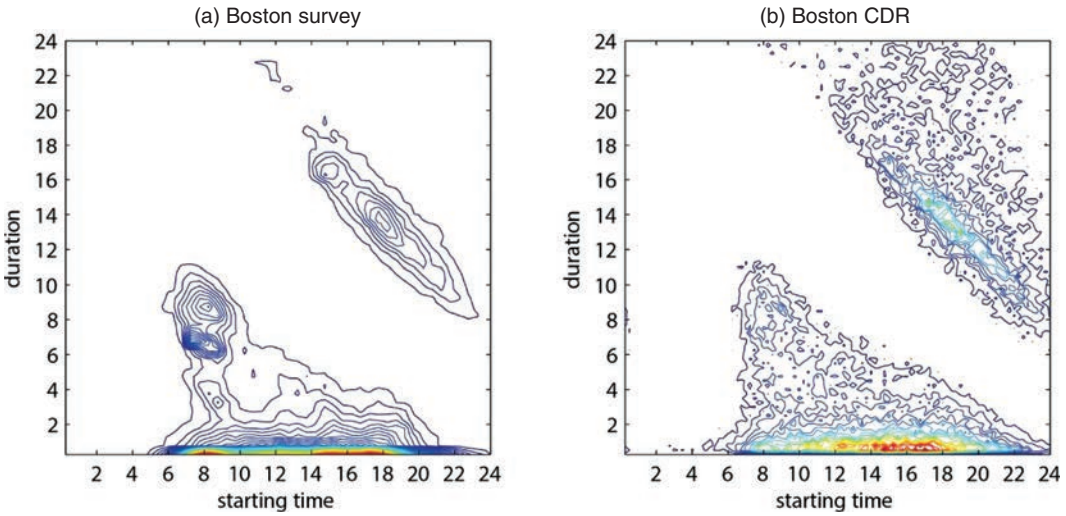
Figure 5-9, which was presented in an earlier published study of the authors' research group (Widhalm et al. 2015), demonstrates the validity of the stays extracted from the CDR data in the temporal dimension. Figure 5-9*a* shows the stay duration by arrival time derived from the 2011 Massachusetts Travel Survey (MTS) data, while Figure 5-9*b* shows the distribution derived from the 2-month CDR data.

In the aggregate, the two figures show similar temporal patterns for activity start and activity duration. There is a concentration of 6- to 8-hour stays that start around 8 a.m. and most likely



Source: Jiang et al. (2013)

Figure 5-8. Stay locations and their spatial distribution in the region. CDR data for a 2-month sample period in Boston (same as in Figure 4-7) were used to show the kernel density estimation of the pass-by locations.



Source: Widhalm et al. 2015.

Figure 5-9. Stay duration (hours) in (a) MTS and (b) CDR data.

correspond to work activities. There is also a concentration of 10- to 16-hour stays starting between 4 p.m. and midnight that probably reflect home stay activities.

These patterns suggest that the methods of extracting stays presented earlier are reliable and can give robust estimates of stay duration. These duration estimates are important for further categorizing activity types as "home," "work," and "other" and analyzing O-D matrices to provide estimates of travel demand.

5.5 Mapping Stay Locations to Zones

5.5.1 Creating and Storing Geographic Data

To estimate zone-to-zone O-D matrices, it is necessary to assign extracted stay locations to Census tracts or another type of zone that is defined. A relational database was used to store Census information for the study area in a standard format. A Postgres program along with an open-source spatial extension PostGIS was used to store and manipulate Census data and other geographic data, such as road network data. Given the current cost of computing resources, these systems provide adequate performance for storing static GIS and Census data. They have convenient, mature interfaces for easy access. In addition, this database was also used to store aggregated results from the various analyses so that they could be made available to interactive web application programming interfaces (APIs) and visualization platforms.

5.5.2 Aggregating Stay Points to Zones

Polygons of Census tracts or TAZs and demographic information associated with them were stored in a relational database. However, it is computationally inefficient to perform point-in-polygon calculations for each user or call record in a CDR data set. To dramatically speed up these computations, the research team rasterized polygons into small pixel grids in which each pixel value is a unique identifier for the Census tract covering that pixel. This raster was then used as a look-up table for converting the latitude and longitude of CDR data into Census tract IDs. The rasterization introduced some error along the borders of tracts, but these errors were minimized by making pixel sizes much smaller than the resolution of the location estimates

of calls between 10 meters and 100 meters. By using this method, the stay points were easily converted into zones such as Census tracts. The method is critical, given that it supports further in-depth analyses of cell phone data for providing estimates of travel demand.

5.6 Summary

This technical chapter discusses the details of how daily trajectories are extracted from noisy and massive cell phone CDR data. The discussion starts with the motivation for this approach and focuses on how the noise from these CDR data is removed to arrive at meaningful data for transportation analysis. The value of the CDR data lies in their analysis to make key inferences regarding the locations and the activities of the respondent carrying a cell phone device throughout the day.

The chapter first reviews the steps needed to identify activity types at detected stay locations where individuals spend time to conduct their activities. The temporal patterns related to each activity, including the start times and the duration of individual activities, are also examined.

The chapter describes how grid-based and point-based algorithms are used to identify and extract stay points and stay regions by analyzing traces, locations, and the time spent at each location. Examples of stay extraction results are presented with data from an individual user. The density of stay locations and the duration of activities by time of arrival are presented. Finally, the results are aggregated to the zone level to facilitate O-D comparisons.

CHAPTER 6

Measuring Individual Activities: Home, Work, "Other"

6.1 Roadmap to the Chapter

This report has so far discussed raw cell phone data, ways to remove noise from call detail record (CDR) data, and methods for extracting meaningful stay points that reflect locations where individual activities are anchored. To derive reliable origin–destination (O-D) trips by purpose, it is important to identify activity types that correspond to "home," "work," and "other" stay locations.

Practitioners traditionally use household survey data to establish respondents' locations of home, work, and school and to study educational, recreational, shopping, and personal business purposes in detail. This chapter first reviews the analysis steps needed to identify activity types at the inferred home, work, and "other" stay locations detected by the CDR data.

Sample expansion is key for practitioners who develop detailed weights by comparing survey and population totals for selected household characteristics and market segments. This chapter presents the filtering of the CDR sample to remove observations with infrequent cell phone use. The expansion of active cell phone users to the metropolitan population is also discussed.

Taken together, the inference and sample expansion methods for CDR data discussed in this chapter provide the building blocks for developing estimates of total trip making and O-D person-trip tables at a regional level.

Finally, the chapter compares the expanded home and work trips produced and attracted on the basis of CDR data versus the Census Transportation Planning Products (CTPP) journey-to-work data. The Boston, Massachusetts, region is used as a case study to make these comparisons.

6.2 Activity Inference

6.2.1 Goal and Approach

It is well documented in the transportation literature that trips are induced by the need or desire to engage in activities (Manheim 1979, Pinjari and Bhat 2011). Therefore, an understanding of patterns and types of activities is crucial in deriving estimates of travel flows and travel demand.

Recent studies from various cities (Alexander et al. 2015, Colak et al. 2015, Toole et al. 2015) have used cell phone data at city scale and with low costs to enhance knowledge about human mobility and methods of estimating O-D trip tables. It has been demonstrated that human mobility patterns are characterized by regularity, with frequent returns to previously visited locations (González et al. 2008; Song et al. 2010a, 2010b; Schneider et al. 2013; Jiang et al. 2013; Hasan et al. 2013). Because of this predictability, stay activities for users' most-visited locations

can reasonable be inferred from observations of cell phone records made over multiple days. For each user, the stay extraction process detailed in Chapter 5 results in a time stamp and duration for each observed visit to a stay location. With trajectories of stay points without the noise of raw data, the next step is to infer contextual information about each location. There are two approaches to infer activity types.

The first approach depends only on cell phone data and estimates activity types on the basis of circadian rhythms and regularities exhibited in human mobility. Alexander et al. (2015) and Colak et al. (2015) improved on methods introduced by Wang et al. (2012) and Iqbal et al. (2014) by using visitation frequency and temporal data to infer contextual information such as a location's function or trip purpose.

The second approach incorporates additional land use information and data on points of interest in addition to the individual trips extracted from cell phone data to infer activity types in detailed categories such as home, work, recreation/leisure, shopping, and other. For example, Jiang et al. (2013) proposed to infer activity types on the basis of dependencies among daily mobility motifs, temporal information about trips, and data on land use and points of interest. Widhalm et al. (2015) demonstrated the estimation using a relational Markov network with Vienna, Austria, and Boston as examples. They found that the inferred activity clusters were stable across days. Widhalm et al. (2015) also pointed out limitations of the approach in areas with mixed land use with regard to inferring detailed activities.

This section focuses on the first approach for three reasons:

- Its simplicity for application and limited requirements for external data make the analysis approach functional and standard.
- Detailed land use and point-of-interest data may not be available consistently throughout the country. Therefore, it is unclear whether commercial vendors can use the second approach consistently.
- The Boston region is highly urbanized and has several zones with mixed land use. Breaking down the actual activity at the trip end for areas with mixed land use is tricky and may introduce errors.

This section describes the assumptions and methods used in the research team's approach to assigning an activity type of home, work, or "other" to each user's stay locations and validates the number of trips produced and attracted. In Chapter 7, the distribution of trips by purpose and by time of day is discussed and the results from CDR models are compared with those from traditional survey summaries and regional model outputs. In Chapter 8, travel flows are examined and O-D trips are compared by trip purpose and time of day with results from the Boston Metropolitan Planning Organization's travel demand model.

6.2.2 Algorithms

6.2.2.1 Inferring Home Location

Each user's home location was identified as the stay location that had the most visits on weekends and on weekday nights as defined by a time parameter specific to the local context. This parameter represents the time window(s) during which users are expected to spend a substantial amount of time at home. For the study area, the period between 7 p.m. of a given day and 8 a.m. of the next weekday was defined as a weekday night.

In addition to inferring trip purpose, the home stay location of each user was also used to filter out users with too few data points. Another important function of the home location is that it provided control totals for data expansion from the sample of cell phone users to the population in the study area.

6.2.2.2 Inferring Work Location

Inferring a user's work location involves considerably greater uncertainty than inferring his or her home location. Two methods of inferring a user's work location, each of which is based on different assumptions, are discussed below.

Conservative Model. The assumption of this model is based on the rationale and historical evidence (Levinson and Kumar 1994, Schafer 2000) that, for a given frequency of visits, longer-distance trips are more likely to be work trips than are shorter-distance trips, which are more likely to be for nonwork purposes such as going to a nearby grocery store.

A work location is identified as the stay not previously labeled as home to which the user travels the maximum total distance from home $\max(d * n)$, where n is the total number of visits to a given stay on weekdays between, for example, 8 a.m. and 7 p.m. and d is the straight-line distance between the home stay location and the given stay as calculated by plane approximation.

If the user visits the identified work stay less than once per week on average (i.e., eight times in total during the observation period of 8 weeks), or if the distance is less than 0.5 km ($d < 0.5$), then the activity of the stay region is identified as "other" rather than as work. In effect, not all users are assigned a work stay, in recognition that not all users commute to a job. These classification assumptions serve to avoid falsely identifying a location as work that either is not visited frequently enough to be a work location or is close enough to a user's home that it could reflect signal noise rather than a distinct work location.

Relaxed Model. The second approach relaxes the condition for labeling as work. A user's work location is defined as the stay point other than home that a user visits most often during the daytime on a weekday between the hours of 8 a.m. and 7 p.m. Because many individuals do not work, the work location is left blank if the candidate location is not visited more than once per week or if the location is less than 500 meters from the home location.

Neither approach to inferring work location distinguishes between work and school or university locations. The work location must actually be considered the location for the most common mandatory activity. Of the two methods, the more conservative method may be better suited to modeling work activity because it minimizes the error of falsely assigning the label "work" to nonwork locations.

6.2.2.3 Inferring "Other" Locations

All remaining stay locations that were not identified as either home or work locations were designated as "other." Future research can expand the designation "other" to reflect activity types such as school, shopping, recreation, and social. To further distinguish activities at this level of detail, additional contextual information such as detailed land use data will most likely be required. The research team used "other" to represent all nonhome and nonwork activities.

The team acknowledges that under these simple assumptions, a user's true home and work locations might be misidentified, along with their corresponding trip purposes. For example, a school activity might be misidentified as a work activity if it satisfies the conditions assumed by the work location identification model. However, the comparisons with Census data presented in the model validation in Section 6.3 suggest that the procedure offers good estimates of the distribution of home and work locations and home–work flows in the study region.

These assumptions are also related to the length of the observation period and the spatial resolution of this CDR data set. It may therefore be necessary to adjust the criteria used for applications of this method to other data sets.

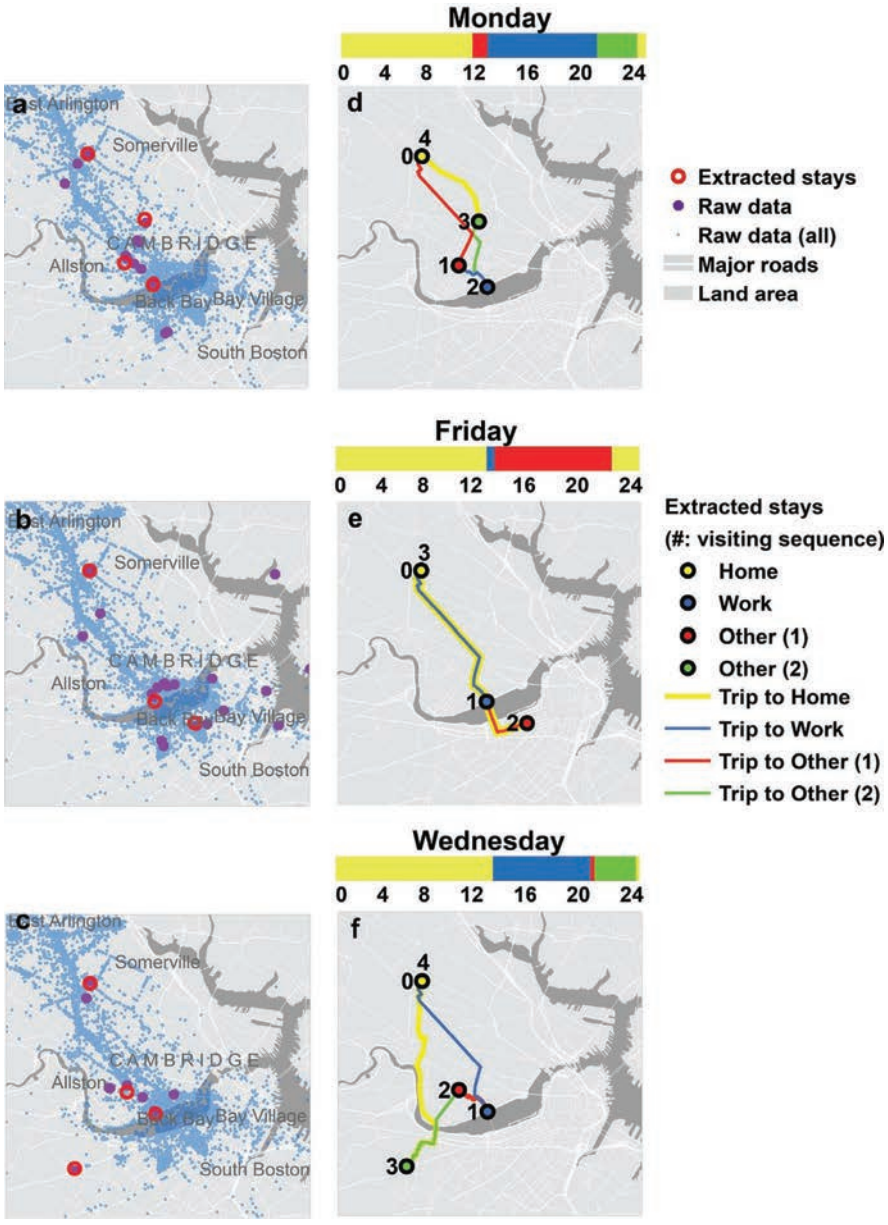
The following sections illustrate the results in space and time by implementing the activity inference algorithms that have been discussed here.

6.2.3 An Individual Example

Figure 6-1 shows the results of analysis of 3 days' worth of data by the student who voluntarily donated his self-selected cell phone data, collected over a period of 18 months, to the Massachusetts Institute of Technology HuMNet Lab for research purposes.

The spatial distribution of the raw data is shown in Figure 4-11. Figure 6-1 shows the sequence that translates the raw data into the inferred activity types for this individual. The extracted stays were developed in three steps:

- The raw cell phone data over the course of 18 months are shown as blue points in Figure 6-1*a*;
- The raw data of each selected day are the purple dots in Figure 6-1*b*; and
- The extracted stays for the day are shown as red circles in Figure 6-1*c*.



Source: Jiang et al. 2016.

Figure 6-1. Inference of student's activities on basis of 3 days of data.

The inference of activity types at the extracted stays are also shown:

- Home is represented by the yellow-faced circles in Figure 6-1, *d–f*;
- Work is represented by the blue-faced circles in Figure 6-1, *d–f*; and
- "Other" is represented by the red-faced circles in Figure 6-1, *d–f*, and the green-faced circles in Figure 6-1, *d* and *f*.

The number next to each stay location represents the visitation sequence within each day. Trips from one stay location to another are color-coded with the same color as the activity type at the destination. For example, in Figure 6-1*d*, the trip from home to "other 1" is in red, the trip from "other 1" to work is in blue, the trip from work to "other 2" is in green, and the trip from "other 2" to home is in yellow.

Above Figure 6-1, *d*, *e*, and *f*, there is a time bar for a 24-hour period. The color represents the inferred activity type while the length of the bar shows the inferred duration of the activity. The methods used to infer trip departure time are discussed in detail in Chapter 7.

6.2.4 Sample Filtering and Expansion

6.2.4.1 Sample Filtering

By definition, cell phone CDR data will not provide a full picture of some users' travel patterns. In particular, users who do not use their cell phone often for calls, texts, or Internet data access will yield too few events to allow for meaningful extraction of their stay locations and travel patterns.

To address this question, the research team filtered out observations with fewer than eight visits over the 2-month period to home stays. This filter corresponds to less than one visit per week on average to designated home stay locations. This filter served the additional purpose of ensuring, within a reasonable degree of certainty, that the designated stay was the user's home, a key assumption in the team's method of expanding users to population.

This filtering process by definition excludes visitors for whom a home location is not observed in the data. Future research could look at extracting visitor trips from CDR data by using an assumption other than home location to expand these trips.

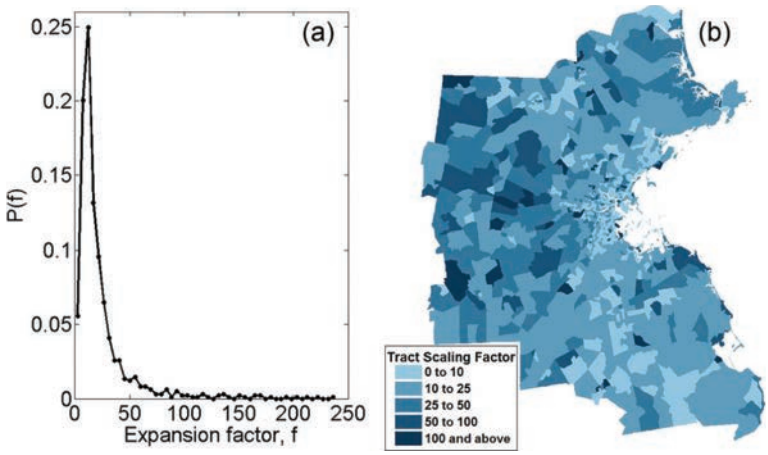
Following the application of this filter, 335,795 users remained in the Boston CDR data set. This sample size is an order of magnitude larger than that of most household travel surveys and could also increase, given a longer period of observation.

6.2.4.2 Sample Expansion

To expand the filtered sample of cell phone users to the total population of the study region, the number of home stays was aggregated to Census tracts in the Boston metropolitan area. An expansion factor was then calculated for each tract as the ratio of the 2010 Census population and the number of residents identified in the CDR data. There were a few Census tracts with fewer than 10 CDR residents. For those tracts, the expansion factor was set to zero to ensure that users who might not be representative of a given Census tract were not overweighted.

Figure 6-2 shows the distribution of the expansion factor values. The values of the first, second, and third quartiles of the expansion factors were 9.4, 14.2, and 25.1, respectively. Figure 6-2 also illustrates the spatial distribution of the expansion factors. The research team's analysis suggests that the tracts in the suburban western portion of the study area tend to be more heavily weighted than the core central area, which is better represented in the sample.

The availability and analysis of cell phone CDR data for a period greater than 2 months would most likely require lower expansion factors and result in better spatial distribution of users. The



Source: Alexander et al. 2015.

Figure 6-2. Expansion factors for Census tracts: (a) probability distribution of expansion factors for Census tracts and (b) spatial distribution of expansion factors at Census tract level.

analogy with traditional surveys is the increase in sample size and the focus on geographic and socioeconomic market segments to improve the representativeness of the sample.

6.2.4.3 Analogies with Household Surveys

There are some interesting analogies that are worth noting when the cell phone expansion factors are compared with the sampling weights in a traditional household survey.

First, the motivation behind the sampling weights in traditional surveys is the difference in making contact and the willingness to participate in a survey. Both of these steps in traditional surveys require a correction through the development of sample weights that expand the sample to be more representative of the regional population.

- These survey sampling weights reflect the under- and over-representation of certain geographic and socioeconomic market segments in the survey. Implicit in sample weighting is the need to adjust the representation by members of these market segments to avoid a model that under- or overpredicts travel in the region.
- In the case of the cell phone data, the need for expansion factors similar to the sampling weights is directly linked to the market penetration and use of cell phones during a typical day. Younger, more educated, and more technology-savvy users of cell phones are likely to provide more traces of their daily routines through their increased use of calls, text messages, and Internet data access when visiting websites or receiving passive signals from apps.

Second, the cell phone expansion weights are smaller in magnitude compared with the sampling weights of a traditional household survey. This is an expected result, given that a sampling rate of 1% in a traditional survey would correspond to an average expansion factor of about 100, with higher values for difficult-to-reach geographic and socioeconomic market segments.

Third, one expects to find differences due to geography and socioeconomics for both traditional surveys and cell phone use.

- In traditional surveys, large households and households that include respondents who are younger, rely less on phone land lines, have lower incomes, and live in urban areas are likely

to have a lower response rate compared with older, suburban respondents who are more likely to be contacted and to respond to a survey.
- However, market penetration is higher and the usage of cell phones more extensive among younger cohorts of the population. Data and analyses based on a cell phone sample are likely to reflect the travel behavior and habits of some of the hard-to-reach segments of the household survey.

6.3 Validation

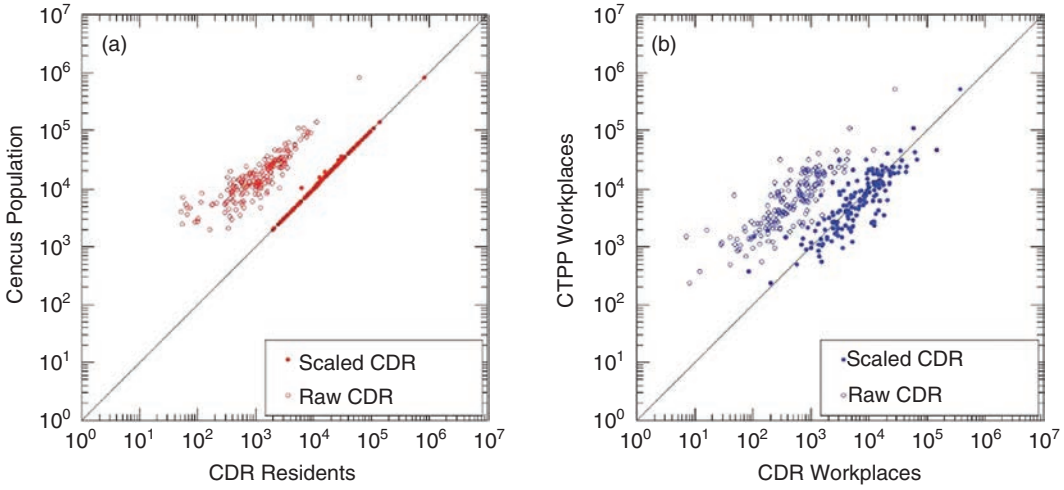
6.3.1 Productions and Attractions

Accurate extraction of users' stays and proper expansion to the regional population are critical to trip generation and estimates of total travel in a region. The regularity of human behavior (González et al. 2008; Song et al. 2010a, 2010b; Jiang et al. 2013) enabled the research team to infer users' home stays and, where applicable, their work stay locations from the CDR data.

The spatial distribution of home and work locations, when aggregated to the 164 cities and towns in the study area (MassGIS, 2014), looks reasonable. Chapters 7 and 8 discuss in more detail the effect of geographic aggregation on how the results compare with traditional data and models.

Figure 6-3*a* compares home locations by town on the basis of 2010 Census data and raw and expanded CDR data. As expected, given that the Census tract population was used to expand the data, the number of residents in each town is almost identical to the estimates from the expanded CDR data. The raw CDR home location data are shown as hollow red circles; the expanded CDR home location data are shown as solid red bullets.

The slope of a best-fit line through the raw CDR data is also close to 1, which strongly suggests that the overall distribution of CDR users in the raw data is fairly representative and that a simple factoring method is appropriate for expanding the cell phone users to the total population and its distribution across the region.



Source: Alexander et al. 2015.

Figure 6-3. Comparison of scaled CDR residents with Census data by town: (a) residents estimated with CDR data versus 2010 Census population by town before and after population expansion and (b) workers estimated with CDR data versus 2013 CTPP workers by town before and after population expansion.

Figure 6-3*b* also compares work locations aggregated at the town level. As with the raw CDR data on the home end, the distribution of raw workplaces is fairly consistent with the 2006–2010 CTPP. The raw CDR work location data are shown as hollow blue circles; the expanded CDR work location data are shown as solid blue bullets.

The data slope is again approximately 1, and the sample expansion method adjusts well for the differences in magnitude across towns. This strong correlation is noteworthy, considering that each user's home and work locations were expanded on the basis of their home location only.

6.3.2 Commuting Flows in Space

Comparisons of the CDR and the CTPP data sets were also made by using town pair flows. Figure 6-4 shows the CDR and CTPP home-to-work flows for all of the intratown and intertown pairs, with correlations of 0.99 and 0.95, respectively.

Figure 6-4 strongly suggests that the validation of town pairs that have many trips is better than that of pairs with few trips, especially those with fewer than about 500 daily trips. This trend is most likely due to the scarcity of data for the smaller markets.

Figure 6-4 also shows the trip-length distribution of daily home-to-work flows derived from the CDR data and the 2006–2010 CTPP data. The results show that the estimated home-to-work flows from the CDR data are close to those reported in the CTPP data at a town level.

Figure 6-4, *c* and *d*, illustrate spatially the distribution of home-to-work flows for key markets (intertract pairs with greater than 1,000 daily trips) for the CDR and Census data, respectively. These patterns suggest that the CDR data capture patterns similar to those of the CTPP commuting data, with the majority of flows directed in and out of Boston as well as a few shorter-distance markets in suburban towns.

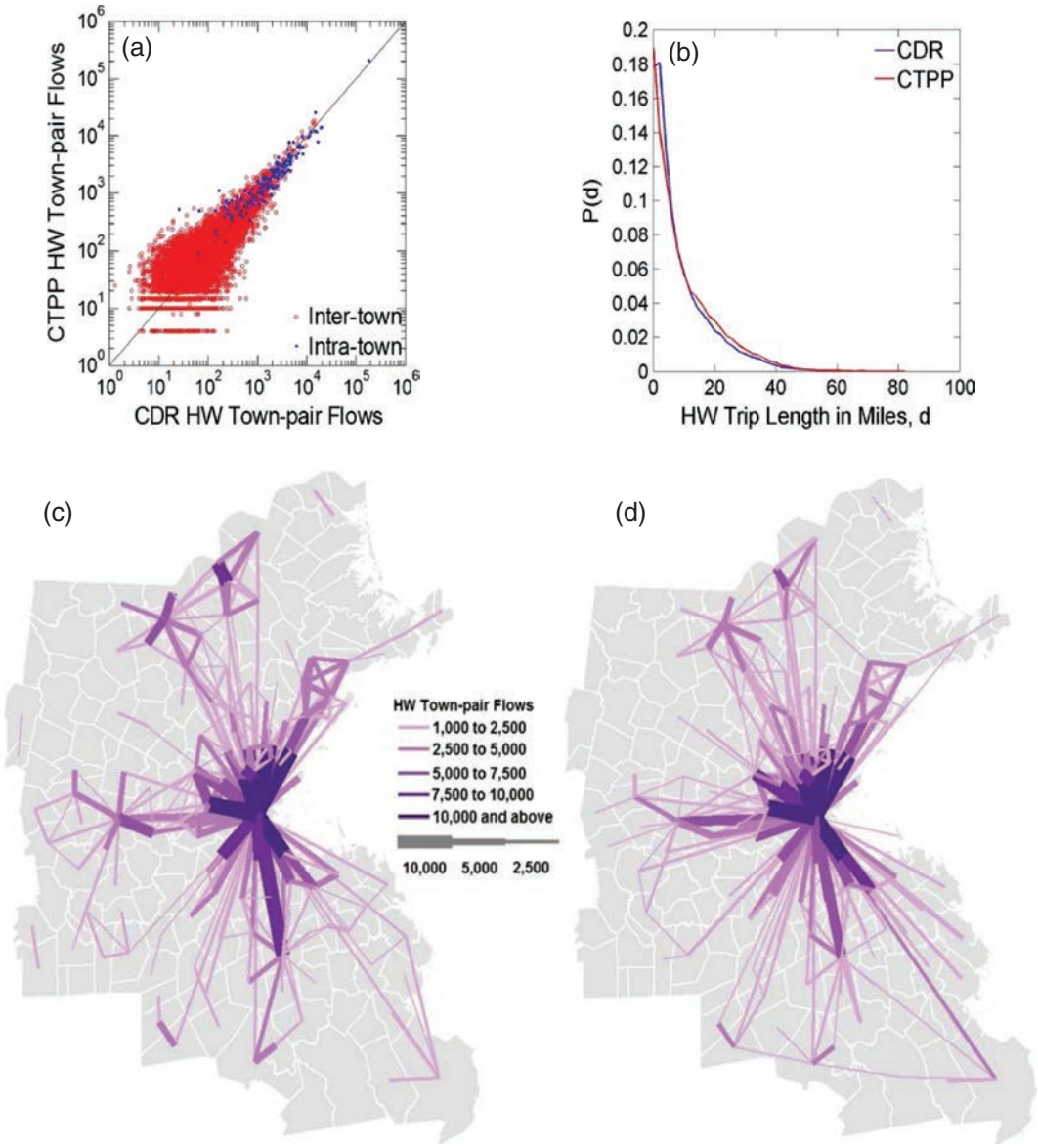
6.4 Summary

The definitions of activities, their location, and the duration spent at each stop are key elements of any trip-based or activity-based model. This chapter describes the research methods used to extract three locations—home, work, and "other"—from cell phone CDR data. The steps needed to detect and identify these stay locations are reviewed and the different algorithms used for this purpose are discussed.

The chapter presents how the sample was filtered by removing observations for which cell phone use during the period of observation was infrequent. The population-based expansion of the data from active cell phone users to regional totals is discussed, and the expanded home and work trips produced and attracted from CDR data are compared with those from the CTPP journey-to-work data for the Boston region.

The home, work, and "other" locations are key designations for practitioners who traditionally use household survey data to establish respondents' locations of home, work, and school. Surveys are also analyzed to describe in great detail nonwork travel related to educational, recreational, shopping, and personal business.

The sample expansion presents a key consideration for practitioners. In traditional approaches, the distribution of sample observations is compared with population totals for market segments reflecting household characteristics. The expansion weights typically account for differences in household size, number of workers, and number of autos per household.



Source: Alexander et al. 2015.

Figure 6-4. Work trips: (a) travel flows, (b) trip lengths, and O-D trip patterns for (c) CDR data and (d) Census data.

The CDR data, however, are anonymized and do not include socioeconomic characteristics. The sample expansion used a simpler approach that compares the total population in each Census tract with the number of cell phone users who live each tract.

Taken together, the inference and sample expansion methods discussed in this chapter provide the building blocks for developing estimates of total trip making and O-D person-trip tables at a regional level. These estimates are refined further in Chapter 7 by focusing on trip tables by purpose and time of day.

CHAPTER 7

Trips by Purpose and Time of Day

7.1 Roadmap to the Chapter

Chapter 5 discussed how respondents' trajectories were analyzed and what assumptions were made to infer respondents' stay locations from triangulated call detail record (CDR) data. Chapter 6 further discussed how respondents' individual activities and inferred home, work and "other" locations were identified and outlined how the sample of cell phone users was expanded to regional population totals by using Census tracts as the sampling unit.

A key consideration for practitioners is whether the total number of trips matches reasonably well with traditional modeling approaches and data. Another key consideration is whether the patterns of trips by purpose and by time of day are similar to the patterns in traditional survey data sources.

Chapter 7 introduces the notion of what constitutes ground truth in Section 7.2. The chapter then outlines methods for producing CDR travel estimates that are comparable to survey data and model outputs. How inferences were made is described:

- First, a departure time was assigned to each stay. This allowed the researchers to group the stays during a typical day to four periods: a.m. peak, midday, p.m. peak, and rest of the day (Section 7.3).
- Second, purpose and time-of-day inferences for each activity were combined to develop estimates of travel flows in a format comparable to the trip distribution outputs (Section 7.4). Home-based work (HBW) trips reflect travel between the home and work locations. Non-home-based (NHB) trips include stay points other than home. Home-based other (HBO) trips reflect travel between a user's home and locations other than work.

The research team compared the inferred travel patterns in the CDR data with those in the 2009 National Household Travel Survey (NHTS) and two regional surveys from Boston, Massachusetts. Travel patterns and the researchers' comparisons are discussed in the following sections:

- Time-of-day patterns for a.m. peak, midday, p.m. peak, and rest of the day are discussed in Section 7.5;
- Patterns of activity duration are shown in Section 7.6;
- Comparisons of travel estimates from CDR and survey data by the three purposes and the four times of day are made in Section 7.7; and
- Comparisons at the Census tract and town pair levels are used to compare CDR travel-to-work estimates with journey-to-work data in Section 7.8.

7.2 Concept of Ground Truth

This chapter presents a first comparison of the CDR-derived estimates of travel by purpose and by time of day with traditional surveys and Census Transportation Planning Products (CTPP)

Ground Truth

Ground truth is a single point of reference that is often difficult to establish in many scientific and professional disciplines. In travel demand modeling, the reliance on surveys from a sample of the population to estimate models; the use of traffic count, transit ridership, and Census data to validate models; and the use of different analysis methods make it difficult to establish a single source as a unique point of reference that reflects ground truth.

journey-to-work estimates. These comparisons highlight the promise and the challenge of using and understanding CDR data for various planning and modeling purposes.

A key issue that does not have a clear or straightforward answer is which of these sources constitutes ground truth. This chapter compares CDR-derived results with those of household surveys, established Census estimates of commute travel, and well-understood model outputs. However, the different nature of each data source and the fact that each source reflects a sample of observations that has strengths, weaknesses, assumptions, and errors embedded in it must be recognized.

A related question includes the assumptions made and inferences drawn when each data source is being analyzed. The following weaknesses of analysis approaches reflect different assumptions about what drives travel and how travel is inferred from data:

- **Weighting of household surveys**. A small sample is collected and weighted on the basis of regional distribution of socioeconomic characteristics. The implicit assumptions are that the determinants of travel are properly reflected in the segments used in the sampling plan and that enough observations are collected in each cell to properly assess travel within each segment.
- **Weaknesses in model development**. In both trip-based and activity-based models, errors are likely to propagate throughout the model components. These errors may reflect limited data for certain market segments, errors or omissions of important variables in model specification, and linkages between models that are not properly accounted for during the analysis.
- **Census data.** Journey-to-work travel flows offer probably the strongest ground truth data source for the daily commute market. However, both CTPP and American Community Survey (ACS) data also represent a sample of a region's households. Weaknesses of these data include absenteeism, the reporting of the primary work location only, and lower sampling rates in smaller geographical areas.
- **CDR data assumptions**. This promising new source of data benefits from a large sample size, the ability to observe the same cell phone device over a long period, and the advantage of making inferences about activities using repeated observations. The value of CDR data is also constrained by the following issues:
  - Passive or active use of the phone is needed to record travel,
  - There is uncertainty in stay locations that are inferred,
  - The number of inferred trip purposes is small,
  - Trips by members of the same household are not linked, and
  - Lack of socioeconomic data prevents analysis by market segment.

In evaluating the findings in this chapter and in the remainder of the report, it is important to keep in perspective the nuances of each data source and analysis method and the lack of absolute and definitive ground truth estimates.

7.3 Modeling Departure Time

This section describes in detail how time of day was incorporated into the CDR data analysis. Daily trips were estimated from filtered users by analyzing consecutive observations at different stay points during a given time window. The process begins by defining an effective day as a period between 3 a.m. on Day 1 and 3 a.m. on the following day. This definition is consistent with the approach used in household travel diary surveys.

7.3.1 Concept

The analysis thus far has discussed a user who traveled between two observed stay points and whose activity at each stay was inferred. However, the user's precise departure time is not known because the time stamp and duration associated with each stay reflect the observed time of phone usage rather than the true arrival time and duration of each stay.

To account for this uncertainty, probability density functions can be used to infer the hour of departure for a trip. This concept was implemented by assigning a departure time on the basis of the conditional probability that a user departed between the time he or she was last observed at the origin stay location and the time he or she was first observed at the destination stay location.

The conditional probability function for departure time can either be derived from surveys such as the NHTS or can be estimated empirically by using the observed call frequencies of all users over the course of the day. The research team used the NHTS data to derive departure time for each stay.

7.3.2 Method in Detail

7.3.2.1 Algorithm

A simplifying assumption was made that a user must start and end each 24-hour period at home. If a user is not observed in the CDR data to be at his or her home stay location for the first (or last) record of the 24-hour period, then the first (or last) trip is completed by beginning (or ending) at a home stay.

A trip is made between two consecutive stays $(i, i+1)$ that occur within a 24-hour period that begins and ends at 3 am. The first and last trips are assumed to occur at a point within the range of [3 a.m., $t_i + 1$] ([$t_i + \delta_i$, 3 a.m.]),

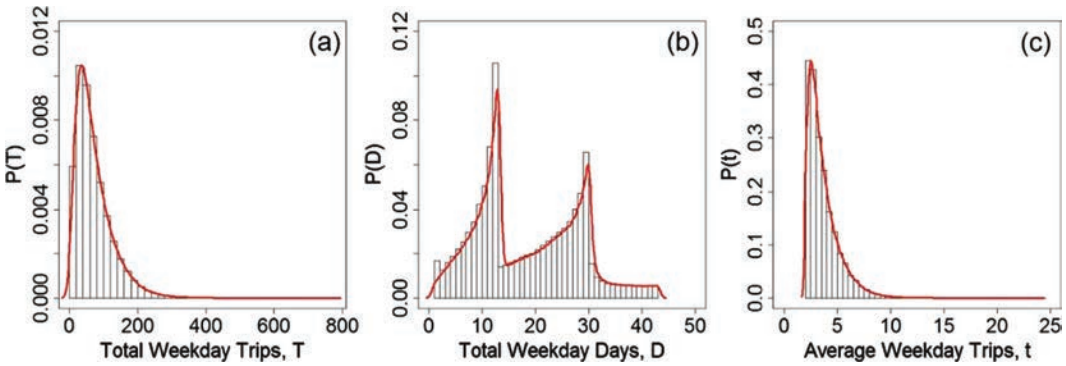
where

t_i = observed arrival time of the current stay (i),
δ_i = observed duration at stay (i), and
$t_i + 1$ is the observed arrival time at the next stay $(i+1)$.

A key concept is the time window within which a user's trip occurs. In particular, a trip occurs at a point in time spanned by the range $[t_i + \delta_i, t_i + 1]$, where t_i, δ_i, and $t_i + 1$ are defined in the same manner as above. The departure hour is generated within this time window, with an empirically derived conditional probability of hourly departure.

7.3.2.2 Empirical Conditional Probability

The 2009 NHTS data were analyzed to derive the conditional probability for hourly departure that corresponds to the day and trip purpose combinations inferred from the analysis of the cell phone sample. The NHTS data were filtered to include respondents who resided in large metropolitan areas of 3 million or more. The research team focused on temporal travel patterns in major U.S. cities that are comparable to Boston because focusing on the Boston Metropolitan Statistical Area alone would yield a small sample. These departure time data were used to generate six



Source: CDR Data for the Boston Region; Alexander et al. 2015.

Note: P(T) = share of total weekday trips; P(D) = share of total weekday days; P(t) = share of average weekday trips.

Figure 7-1. Patterns of weekday observations and user trips in CDR data for Boston region: (a) distribution of total weekday trips per user during 2-month period of cell phone data collection, (b) distribution of weekdays on which each user was observed during the 2-month period, and (c) distribution of average daily weekday trips per user.

hourly distributions for weekdays, weekends, and each of the three trip purposes. The patterns of weekday observations in the cell phone CDR data are illustrated in Figure 7-1.

- The distribution of total weekday trips per user over the course of 2 months is shown in Figure 7-1*a* with first, second, and third quartiles of 33, 58, and 96 trips respectively.
- The number of days during which a user was observed is shown in Figure 7-1*b*. This graph clearly shows the reindexing of anonymous user IDs in the raw CDR data at approximately the 12th day and the 30th day of observation as shown in the two peaks of the distribution.
- Reindexing was carried out on a regular basis to maintain user anonymity. Despite this reindexing, each user was observed for a sufficiently large number of days, with first, second, and third quartiles of 11, 17, and 21 days, respectively.
- The average number of weekday trips per user was derived by dividing each user's total weekday trips by his or her total weekdays (Figure 7-1*c*). Although the distribution has a long tail, the first, second, and third quartiles correspond to 2.6, 3.2, and 4.3 average trips per weekday, respectively.

This analysis suggests that during the 2-month period, individuals' cell phone use and travel behavior were repeatedly observed, despite the anonymizing process that effectively breaks the sample into two subsamples. The average number of weekday trips observed during this period demonstrates that the vast majority of users has a reasonably small number of daily trips, with a median of slightly more than three trips per day.

7.4 Modeling Person-Trips

The first key question, from a practitioner's perspective, is how the inferences about stay locations, activities, and time of day are combined to construct a user's trips at the origin–destination (O-D) level. Although the assumptions differ from those for a typical household survey, the cell phone CDR sample still needs to be normalized to reflect a typical weekday's travel, filtered to remove observations with incomplete data, and expanded to represent the population of a region.

A follow-up question with implications important to practitioners is how the cell phone–derived estimates of travel by purpose and by time of day compare with known and more familiar estimates of travel. A range of comparisons with traditional estimates from regional household surveys was made and is discussed in Sections 7.5 through 7.8 to provide insights into the robustness of the cell phone estimates.

Assigning a purpose and departure time to every stay inferred from the cell phone sample allows O-D trips made by a user to be constructed for any given day. As is the case with household survey data, the O-D trips were allocated to a traffic analysis zone or district for analysis purposes. The case study used Census tracts as the unit of analysis, which resulted in a database with a vector of trips between Census tracts in the Boston region for each user in the CDR data set.

7.4.1 Average-Day Normalization

The average number of trips made by each user during a given time window was calculated by dividing the number of trips counted by the number of days that each individual user was observed in the cell phone database.

7.4.2 Filtering Users

It is known that the daily usage of mobile phones within the population varies considerably. There are cases in which users do not make enough calls, send enough texts, or use enough data to correctly infer their movements and travel patterns during that day. As a result, users who did not have sufficient records in the CDR data were dropped.

The unavailability of travel information resulting from infrequent cell phone use by an individual is an inherent weakness of the cell phone data for purposes of travel analyses. Fewer calls and texts and lower data use do not necessarily correspond to less travel. Such patterns may be more correlated with users' socioeconomic characteristics, their familiarity with technology, or their need to stay connected.

These questions, however, cannot be addressed unless a dedicated sample of cell phone users is tracked and surveyed to provide a means of linking cell phone use to travel patterns, but also to socioeconomic characteristics, familiarity with technology, or their need to stay connected.

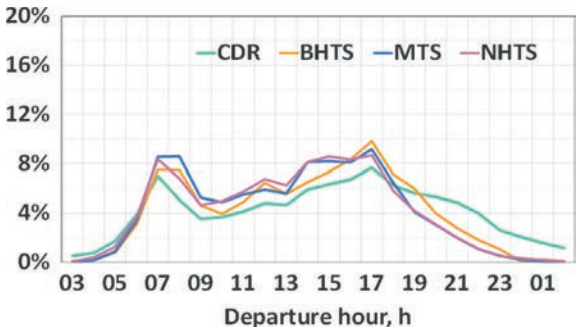
Given that the trips generated from the CDR will eventually be assigned to the transportation network, it is important to estimate the total number of trips taken and the distribution of trips across the region correctly. A study by Toole et al. (2015) found that filtering out users who made fewer than 2.5 trips per day still left a large sample size of active users and resulted in valid estimates of trip tables and O-D matrices. Subsequent sections show the comparisons made after data with fewer than 2.5 trips per day were filtered out.

7.4.3 Trip Expansion

While a trip represents an observation of movement of at least one person between two locations, these trips come only from a sample of individuals and need to be expanded to represent the regional population. To obtain the average daily O-D trips, the researchers multiplied each user's trips by the expansion factors described in Section 6.2.4. The population in each user's home Census tract was used, and the number of days from which a user's trips were constructed was controlled for.

A simplification was made for users who were assigned a work stay. For those users, weekday trips were constructed only on those days on which the user was observed at his or her work stay, to capture representative weekday travel by commuters. Each user's average daily trips were then aggregated into O-D trip matrices for weekdays and weekend days, differentiating by trip purpose and hour of departure.

As is the case with respondents in traditional surveys, the ratio of cell phone users to the population was not uniform within the region. Unlike traditional travel surveys, in which respondents provide travel data for 1 or 2 recent days, the cell phone method has the advantage of capturing many days per user and includes more variation in each user's daily travel behavior.



Source: Alexander et al. 2015.

Figure 7-2. Departure time patterns for all trips (CDR = CDR Model 1; BHTS = 1991 Boston Household Travel Survey; MTS = 2011 Massachusetts Travel Survey; NHTS = 2009 National Household Travel Survey).

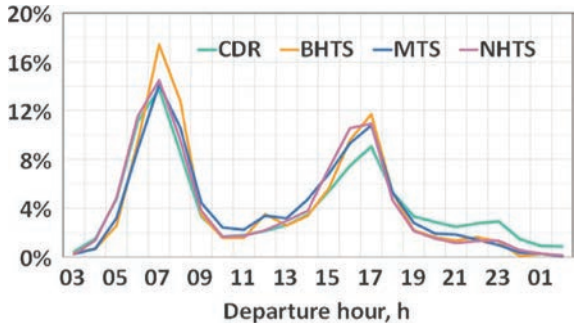
Given the much larger size of the cell phone user sample versus a traditional survey sample, the CDR method also has the advantage of smaller expansion weights than a traditional survey. In Boston, the majority of these expansion factors were found to be less than 10, although expansion factors can be larger in places with a lower level of cell phone market penetration.

7.5 Time-of-Day Patterns

The robustness of the approach to developing estimates of travel by time of day is reflected in two key comparisons. First, the distribution of trips by time of day was compared with that of the NHTS and the two regional Boston surveys. The distributions of trips for each of the three purposes were examined in more detail.

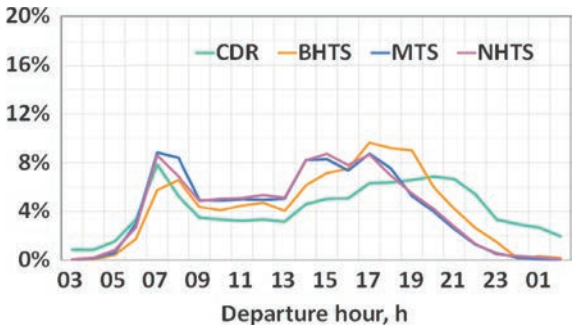
Figure 7-2 illustrates the time-of-day patterns for average weekday trips. The shapes of the cell phone–derived distributions and those from the three surveys are broadly comparable. This suggests that the departure times imputed in Section 7.3 are relatively robust. It should be noted that the CDR data have a greater share of trips starting around 8 p.m. and a lower share of trips between 8 a.m. and 5 p.m.

Figure 7-3 shows the time-of-day distributions for work-related travel. Most transportation planning applications focus on trips in the morning and evening peak periods, when congestion is most prevalent and imposes the greatest demands on the transportation infrastructure. There



Source: Alexander et al. 2015.

Figure 7-3. Departure time patterns for HBW trips.



Source: Alexander et al. 2015.

Figure 7-4. Departure times for HBO trips.

is a close match between the cell phone–derived patterns and each of the three surveys. The four lines track closely with two distinct a.m. and p.m. peaks.

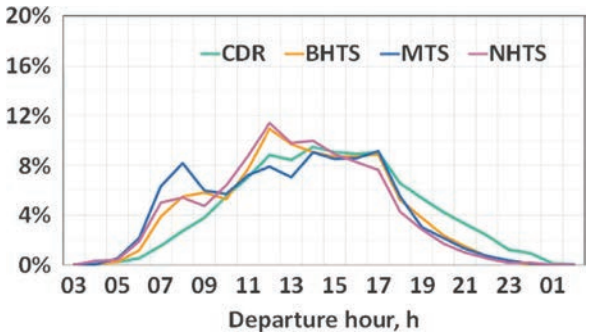
Figure 7-4 shows the time-of-departure patterns for HBO trips, and Figure 7-5 the patterns for NHB trips. The shape of the curves are similar for both purposes. However, there are consistently more CDR trips in the early evening and late night hours as compared with the three surveys. This pattern may reflect a mismatch between a lower frequency of cell phone use during work hours and higher trip-making throughout the day. It may also highlight the ability of CDR data to capture early evening and late-night trips not typically reported in surveys during a typical day.

7.6 Activity Duration Patterns

A comparison of the duration of activities corresponds roughly to the trip-length distribution comparisons of traditional modeling approaches. The departure time for each stay having been modeled as home, work, or other, the temporal distribution of stay durations in the CDR and the travel survey data were checked.

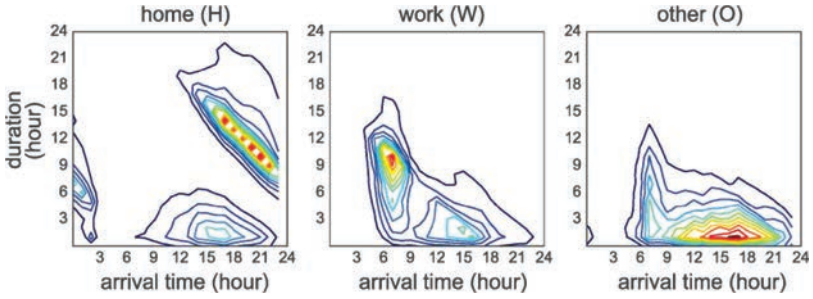
It should be noted that, in the analysis of the CDR data, it was assumed that the arrival time at the current location was equivalent to the modeled departure time from the previous location. Although this simplification addresses a weakness in the CDR data, it is a reasonable assumption in most cases of urban travel. Given that the temporal resolution is in 1-hour increments, this assumption can be interpreted as a user arriving at the current location within the same hour that he or she departed from the previous location.

Figure 7-6 shows the distribution of stay duration by arrival time and activity type modeled from the Boston CDR data. These data were compared with the distribution derived from the 2011 MTS

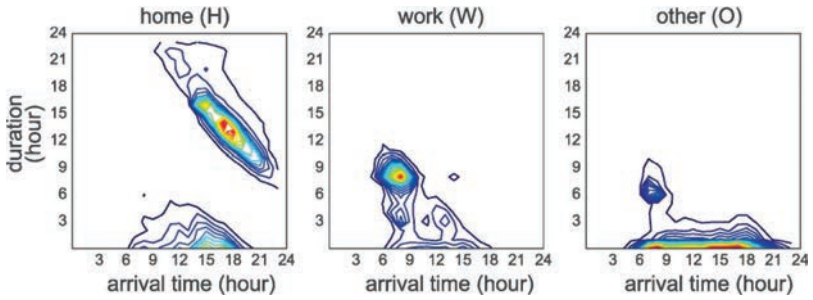


Source: Alexander et al. 2015.

Figure 7-5. Departure times for NHB trips.



(a) Boston CDR Model 1



(b) Boston Travel Survey Data

Figure 7-6. Patterns of arrival time and trip duration.

travel survey data for the Boston region. In general, the patterns of activity stay duration for home, work, and "other" locations at 1-hour intervals for the two data sets were comparable.

The modeled arrival times and stay durations at home locations are shown on the left-hand side of Figure 7-6. The CDR data in Figure 7-6*a* suggest that arrivals started between 3 and 6 p.m. and continued late into the evening. The duration of home stays ranged between 8 and 15 hours. The survey data suggest arrivals starting at about the same time but ending earlier around 10 p.m. The home stay durations in the survey data are similar, ranging mostly from 9 to 16 hours.

The analysis of work locations is shown in the middle of Figure 7-6. The CDR data in Figure 7-6*a* suggest that most workers arrive at their work locations between 6 and 9 a.m. and that most of them stay at work between 7 and 10 hours. The survey data in Figure 7-6*b* suggest similar patterns of arrivals at work between 6 and 9 a.m. and staying at work for 7 to 10 hours. The CDR data also suggest more spread out arrival times outside the a.m. peak period and more dispersed durations as compared with the survey data.

There are more differences when arrivals at and stay durations in locations classified as "other" are considered, as shown on the right side of Figure 7-6. The CDR data in Figure 7-6*a* suggest that visits to other locations happened mostly between 3 and 6 p.m. Most stay durations were less than 3 hours long. In contrast, the survey data suggest a more concentrated pattern of trips to other locations in the a.m. and p.m. peak periods. The durations were generally less than 2 hours long, with the exception of a concentration of stays that were about 6 hours long. As with the work activities, the CDR data have more spread-out arrival times and, especially, durations than the survey data.

7.7 Daily Trip-Making Patterns

The comparisons of total trips, trips by purpose, and trips by time of day represent a critical test from a practitioner's point of view, given that they help assess the robustness of trip-making estimates produced by the CDR data.

Table 7-1. Total daily trips by purpose.

	Total Daily Trips			
Variable	**HBW**	**HBO**	**NHB**	**Total**
CDR trips (millions)	2.81	7.84	4.73	15.38
MTS trips (millions)	2.14	8.99	7.18	18.31
Share of CDR trips by purpose (%)	18	51	31	100
Share of MTS trips by purpose (%)	12	49	39	100
Tract pair correlation	0.3	0.64	0.58	0.58
Town pair correlation	0.96	0.97	0.98	0.98

Source: Alexander et al. 2015.

Note: CDR data for the Boston region; 2011 Massachusetts Travel Survey (MTS).

Table 7-1 summarizes estimates of total daily trips by purpose. CDR trip estimates are compared with the Massachusetts Travel Survey (MTS) data which are weighted and expanded to the regional population estimated from the 2006–2010 ACS:

- The MTS reports about 19% more daily trips than were observed in the analysis of the cell phone data (Table 7-1).
- The trip rate implicit in the MTS is about 4.24 trips per person per day, compared with 3.5 trips per person per day in the CDR data.
- Although these estimates are comparable to the average of four daily person-trips reported in the FHWA Validation Manual (Cambridge Systematics, Inc. 2010), they point to fewer trips on average in the CDR data and more trips on average in the MTS.

Table 7-1 also summarizes the Pearson correlations of the spatial distribution of the daily CDR and MTS trips at the tract pair and town pair level. The correlation coefficients of the trip matrices are significantly better when the data are aggregated to the 164 study area cities and towns, which suggests that the value of the CDR data is greater when the data are aggregated.

7.7.1 Trips by Purpose

The patterns of the relative shares of trips by purpose shown in Table 7-1 reflect the CDR data and traditional survey data:

- HBO trips account for roughly half of all trips and are similar in the two data sources.
- The share of NHB trips is lower in the CDR model than in the MTS (31% versus 39%, respectively).
- The share of HBW trips is higher in the CDR model than in the MTS (18% versus 12%, respectively).

It is likely that these different patterns reflect the effect of the 10-minute criterion used in analyzing the CDR data. The heuristic rule of 10 minutes will miss, by definition, some of the short-duration intermediate stops made on the way to and from work and will not recognize them as true activities. As a result, this criterion will artificially increase the number of HBW trips in the CDR model while reducing the nonwork trip estimates.

In traditional surveys where all daily trips are listed, a portion of the stops lasting less than 10 minutes will correspond to a true activity. A traditional model based on these survey data would then create an HBO trip and an NHB trip that take into account this short-duration activity. In contrast, the CDR model would create an HBW trip from home to work, given that the short-duration intermediate stop is not taken into account as a true activity.

Table 7-2. Total daily trips by time of day.

	Total Daily Trips				
	A.M. Peak	Midday	P.M. Peak	Rest of Day	Total
CDR trips (millions)	2.46	4.12	4.15	4.65	15.38
MTS trips (millions)	3.99	6.24	6.06	2.31	18.6
Share of CDR trips by time of day (%)	16	27	27	30	100
Share of MTS trips by time of day (%)	21	34	33	12	100
Tract pair correlation	0.42	0.65	0.54	0.4	0.58
Town pair correlation	0.97	0.98	0.97	0.96	0.98

Source: Alexander et al. 2015.

Note: CDR data for the Boston region, 2011 MTS.

Finally, the correlation coefficients in Table 7-1 suggest that aggregating trips from the tract to the town level yields the greatest improvement for HBW trips. This may reflect the role of tract size, especially in the smaller zones in downtown Boston, where many of the morning commute trips end.

7.7.2 Trips by Time of Day

Table 7-2 shows the same databases as Table 7-1, analyzed by time of day. The main difference is the lower share of CDR model trips in both of the peak periods and during the midday period compared to the much larger share of CDR trips in the rest of the day.

It is again reasonable to postulate that the 10-minute heuristic rule reduces the number of activities that correspond to short-duration intermediate stops. Given that this is more likely to happen during daytime travel, it effectively generates fewer CDR trips during the peak periods and the midday.

The reverse pattern is true during the rest of the day, when the CDR data pick up a much larger share of trips than typical surveys do. In this regard, evening travel may be underreported in traditional surveys. Furthermore, short-duration stops are less likely to happen in the evening and early morning hours, which reduces the effect of the 10-minute heuristic rule on the CDR trip estimates.

7.8 Commuter Flows

Beyond the big-picture estimates of total travel and travel flows by purpose, practitioners often need to focus on corridor-level comparisons and analyses, especially for work-related travel. This section investigates the accuracy of CDR data at different levels of spatial aggregation by focusing on commuter flows, which predominantly occur during peak periods.

Commuting trips represent a key travel market and source of daily recurring roadway congestion. The accurate representation of these trips is an important step in validating trips estimated with the CDR data. The research team compared flows between respondents' home and work locations as reported in the 2006–2010 CTPP journey-to-work data. Commuting flows link home and work locations and are not affected by residents' complex daily trip chains, which may include intermediate stops on the way to work or NHB work trips to or from locations other than home.

Table 7-3. Commuting flows from cell phone data and CTPP.

Source	Home-to-Work Trips (millions)	Commuting Flow		Average Trip Length (miles)
		Intertract (%)	Intertown (%)	
CDR	2.11	94	68	9.67
CTPP	2.10	90	68	10.72

Source: Alexander et al. 2015.

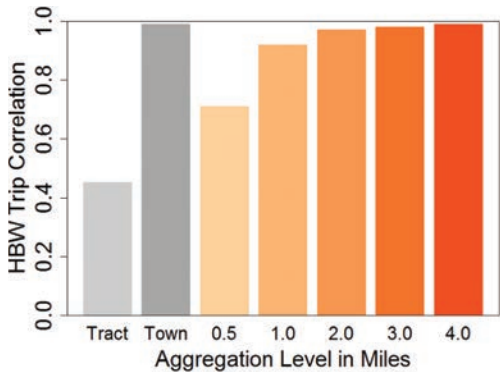
Note: CDR data for the Boston region; 2010 CTPP.

Table 7-3 summarizes the comparison of CDR and 2006–2010 CTPP commuting flows:

- The estimates of the total number of work trips are comparable.
- The percentages of intertract and intertown flows are similar, which suggests a consistency at the spatial level.
- The average trip length is lower in the cell phone data but generally comparable with the CTPP journey-to-work data, which suggests similar distributions of commuting flows.
- The correlation between CDR and CTPP home-to-work tract-to-tract flows has a low value of 0.45.
- The correlation grows to 0.99 when town-to-town O-D travel flows are analyzed, which suggests a higher level of consistency when trips are aggregated at the town level.

Another way to look at the effect of spatial aggregation is offered by Figure 7-7. Buffers of different sizes are drawn around each origin and destination tract to evaluate the effect of spatial aggregation. As the average size of the spatial unit increases, the correlation between CDR and CTPP commuting flows increases as well. The biggest improvement in correlation happens when a half-mile buffer is added to the Census tracts and when the buffer increases from 0.5 miles to 1 mile, as shown in Figure 7-7. Increasing the size of the buffer beyond 1 mile improves the match, but at a diminishing rate.

In effect, using a half-mile buffer aggregates the small, dense tracts that are mostly in the city center and results in a notable improvement in accuracy. In the absence of meaningful districts or communities to which to aggregate, this method can inform suitable distance thresholds for trip clustering to overcome limitations of sparse data or spatial inaccuracy. Another option is



Source: Alexander et al. 2015.

Note: CDR data for the Boston region; 2010 CTPP.

Figure 7-7. Impact of spatial aggregation on CDR and CTPP correlation.

to aggregate CDR data to commonly used geographic units such as the traffic analysis district, which was developed following the 2010 Census in support of the CTPP.

7.9 Summary

This chapter outlines the methods used to analyze CDR data to produce travel estimates that were compared with traditional surveys and can be further compared with traditional model outputs. These comparisons highlight the promise and the challenge of understanding and using CDR data for various planning and modeling purposes.

The following inferences were made in analyzing the CDR data and comparing the CDR estimates with survey data:

- A departure time was assigned to each stay location, and these times were grouped into four periods: a.m. and p.m. peaks, midday, and rest of the day.
- The stay locations of each activity were analyzed to assign the trip purpose: HBW, HBO, or NHB.
- The CDR trip tables were summarized by day of the week, trip purpose, and time of day.
- The CDR data were compared with the survey travel patterns in the 2009 NHTS, the 2011 MTS, and the 1991 BHTS.
- CDR commuter flows were also compared with the 2010 CTPP journey-to-work data by using aggregations at the town pair and the Census tract levels.

The first broad issue discussed is identifying the source that constitutes ground truth. Although the CDR results were compared with traditional surveys and Census estimates of commute travel, the research team recognizes the assumptions that are present and the inferences that need to be made in every data source and model.

The findings discussed in this chapter can be summarized as follows:

- The time-of-day patterns suggest great similarity in the CDR and survey data on work trips. The differences between HBO and NHB trips suggest more CDR trips during the rest of the day than what is reported in surveys.
- The analyses of arrival times and trip durations also suggest a close correspondence between CDR and survey data on travel to work. Trips to home and other nonwork locations differed, with the CDR data suggesting more trips later in the day and a greater variability in trip durations as compared with the survey data.
- The comparison of the share of trips by purpose yielded mixed results. The CDR data produced a higher share of HBW trips, a similar share of HBO trips, and a lower share of NHB trips as compared with the survey data.
- The comparison of trips by time of day was also mixed. The CDR data produced a higher share of trips in both peak periods and in the midday and a much higher mix of trips during the rest of the day.
- Finally, the commuter flows for the CDR and journey-to-work data matched. As was expected, the aggregation of the CDR data to the town level produced a much better correlation than that at the Census tract level.

CHAPTER 8

Model Comparison: Origin–Destination Trips

8.1 Roadmap to the Chapter

Chapter 7 provided the cornerstone for the estimation of origin–destination (O-D) trip matrices using call detail record (CDR) data by identifying activity types for the stay locations home, work, and "other." The expanded estimates of home and work activity were then compared with Census Transportation Planning Products (CTPP) journey-to-work travel data.

In Chapter 7, CDR data were compared with data from household surveys and model outputs for trips by purpose and time of day. From a practitioner's perspective, these comparisons are vital to assessing how CDR data can be used to support planning decisions or to enhance a regional travel demand model with up-to-date data.

Chapter 8 discusses how cell phone CDR data were used to develop trip tables and compares these trip tables with those from the Boston, Massachusetts, household travel surveys and the Boston regional travel demand model.

Two methods were used in the analysis of the raw CDR data collected over 2 months in 2010 and presented as CDR Models 1 and 2. The O-D person-trip comparisons focused on

- Trip tables at the regional, city, and town levels;
- Travel by purpose, including home-based work (HBW), home-based other (HBO), and non-home-based (NHB) trips; and
- Trip tables by time of day (a.m. peak, midday, p.m. peak, and evening/night).

The O-D flows estimated from the raw CDR data were also compared with up to six of the following sources of person-trip tables:

- 2009 National Household Travel Survey (NHTS), 2011 Massachusetts Travel Survey (MTS), and 1991 Boston Household Travel Survey (BHTS);
- 2007 and 2010 Boston regional models developed by the Central Transportation Planning Staff (CTPS); and
- 2015 third-party CDR estimates provided by a data vendor.

8.2 Data Sources and Model Definition

8.2.1 Surveys

The following national and local household travel surveys were summarized for the analysis:

- **2009 NHTS.** The 2009 NHTS was used to infer departure times for the CDR Model 1, discussed in Section 7.3. Summaries developed using the NHTS included the average distribution of

departure times and the distribution of trip purposes by time of day. The NHTS data were also used as a benchmark for comparisons with CDR estimates.

- **2011 MTS.** The 2011 MTS is the most recent local travel survey in Massachusetts and includes the Boston metropolitan area. It contains data on more than 153,000 trips made by nearly 33,000 individuals (Massachusetts Department of Transportation 2012). This survey was expanded to match population estimates from the 2006–2010 American Community Survey (ACS). The ACS data for the Boston metropolitan area were compared with CDR trip table estimates of O-D trip matrices by purpose and by time of day.
- **1991 BHTS.** The 1991 BHTS is an earlier survey covering the Boston region. It contains information on 39,300 trips made by almost 3,800 households (Boston Metropolitan Planning Organization 1991). This survey was the input for the 2010 Boston Region Metropolitan Planning Organization (MPO) travel demand model (Central Transportation Planning Staff 2013).

8.2.2 MPO Models

8.2.2.1 2010 Boston Region MPO Travel Demand Model

The Boston CTPS provided a set of model results for comparison with the CDR-generated O-D trip tables summarized in the next section. The modeled area encompasses 164 cities and towns, including 101 cities and towns in the Boston MPO area and 63 other communities (Figure 8-1).

The MPO travel demand model follows the traditional four-step modeling framework that includes trip generation, trip distribution, mode choice, and trip assignment. The model provides estimates of present and future average weekday transit ridership and highway traffic. It uses regional socioeconomic data, transportation networks, and multimodal levels of service. The spatial unit of analysis is a traffic analysis zone (TAZ). The modeled area is divided into 2,727 internal TAZs that can be aggregated into 164 cities and towns (Figure 8-1).

The trip purposes summarized from the Boston Region MPO model include

- HBW trips, which combine work and work-related trips;
- HBO trips, which include home-based personal business, social–recreational trips, and pick-up and drop-off trips; and
- NHB trips, which combine all NHB trips, regardless of trip purpose.

The trip generation component considered daily trips for an average weekday and used a.m. peak, p.m. peak, and an off-peak period for other times of day.

Mode choice models were applied after the trip distribution step. Trips from the peak and off-peak periods were split into four periods: a.m. peak (6–9 a.m.), midday (9 a.m.–3 p.m.), p.m. peak (3–6 p.m.), and nighttime (6 p.m.–6 a.m.). It should be noted that the end of the p.m. peak period was 6 p.m., whereas the CDR analysis used an end time of 7 p.m. Correspondingly, the start of the nighttime period of the MPO model was an hour earlier, at 6 p.m. rather than 7 p.m.

The model results provided by the CTPS included trip tables for all specified modes and each of the four periods. For comparison purposes, trips were combined into total person-trips by period, accounting for the following modes:

- Walking access transit trips,
- Driving access transit trips,
- Single-occupancy vehicles and person-trips,

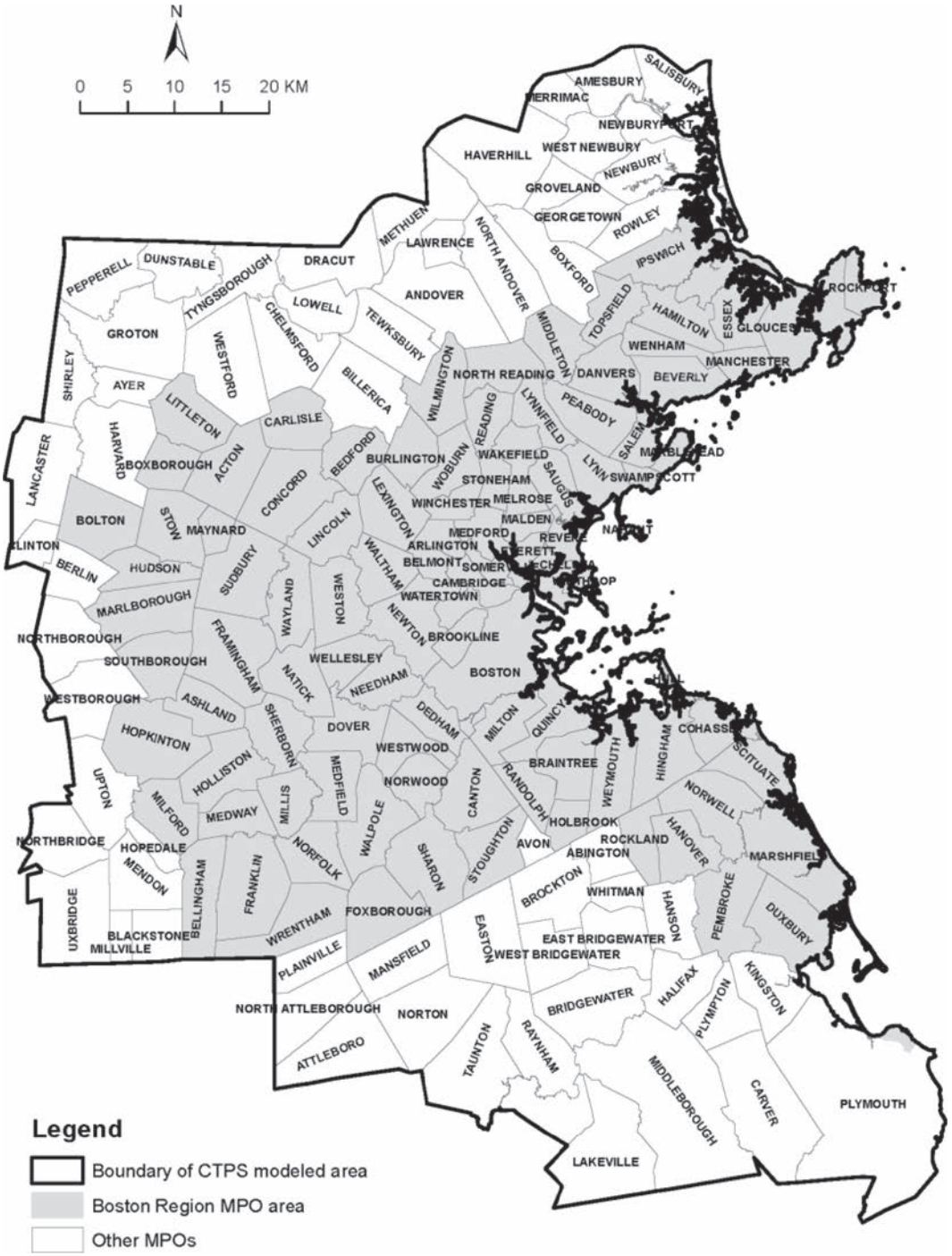


Figure 8-1. Overview of the Boston metropolitan area.

- High-occupancy vehicles (HOVs) with two, three, or more persons that were converted to equivalent HOV person-trips, and
- Walk-only person-trips.

The Boston model also considers internal–external trips and external–external trips. For comparison purposes, only internal–internal trips were examined to reflect travel to and from zones within the Boston region.

8.2.2.2 2007 Boston Region MPO Travel Demand Model

The research team obtained a published version of the MPO model results for the Boston region (Central Transportation Planning Staff 2008). The online report summarizes the model outputs by purpose, time of day, and mode. For consistent comparisons with the CDR trip tables and the household surveys, the team combined the trip purposes to account for HBW trips, HBO trips that included home-based school trips, and NHB trips by time period. The total number of person-trips was compared at an aggregate level, given that no TAZ pair or town pair O-D trips were available.

8.2.3 CDR Trip Tables

8.2.3.1 CDR Models 1 and 2

The data inputs, modeling assumptions, and analysis frameworks of the two CDR models are reported in Chapters 4 to 7. These two models use the same raw CDR data that correspond to 2 million cell phone subscribers in the Boston area during a 2-month period in 2010. The cell phone users' trips were expanded on the basis of population derived from the corresponding 2010 Census data.

- **Work/other identification.** As discussed in Sections 6.2.1 and 6.2.2, different methods can be used to infer locations and the activities that are associated with them. CDR Model 1 uses a conservative assumption in labeling a stay point as a work activity.

  Specifically, CDR Model 1 identifies a cell phone user's work location as the stay to which the user travels the maximum total distance from home, defined as the distance from home multiplied by the weekday visitation frequency during the daytime.

  CDR Model 2 relaxes this condition. Specifically, it identifies a user's work location as the second most frequently visited stay location other than home that a user visits on weekdays.

  Both CDR Models 1 and 2 leave the work location blank if the candidate location is not visited more than once per week or if the location is less than 500 meters from the home location.
- **Departure time modeling.** As outlined in Section 7.3, CDR Model 1 employs the distribution of hourly travel departure frequency derived from the 2009 NHTS for each of the trip purposes.

  CDR Model 2 departure times do not depend on a survey and reflect the empirical hourly phone usage activity in the raw CDR data. The weakness of this approach is that it does not differentiate by trip purpose.

Both CDR models produce O-D trips by Census tract pair for an average weekday and weekend day for three trip purposes and across four times of day. Given that the trips were expanded to the regional population on the basis of cell phone user observations alone, for which only the home location can be identified, the trips that were included account solely for area residents and do not include any visitors.

8.2.3.2 Proprietary CDR Results: CDR Model 3

CDR Model 3 is a proprietary model estimated by CDR data vendor AirSage. The model uses 3 months of 2015 cell phone data expanded on the basis of the 2010 Census population data for the Boston region. The estimation methods and procedures are proprietary and are not known to the research team. This set of O-D estimates was analyzed for evaluation purposes.

It should be noted that these vendor data are different from the 2010 raw CDR data used to demonstrate the concept of CDR analysis. Although the vendor data were weighted to the same 2010 population totals, the usage of cell phones for calls, text messages, and data increased between 2010 and 2015, yielding a richer data set with more data points related to locations, daily activities, and travel.

The results of CDR Model 3 also include O-D person-trips for an average weekday and weekend day by purpose and by time of day at the Census tract level. For consistency in the comparisons, weekday estimates were included in this study. The evening and early morning periods were combined into one period between 7 p.m. and 6 a.m.

It should also be noted that the vendor data differentiate between trips made by Boston residents and visitors. To make comparisons consistent across all data sources and the MPO models, only trips made by residents were used in this study.

8.3 Comparisons at the Regional Level

To compare the results of the three CDR models with those of the travel behavior surveys and the regional MPO models, the research team first aggregated tract-level O-D person-trips by purpose and by time of day to the metropolitan level.

8.3.1 Total Person-Trips

Table 8-1 presents total daily trips and the average trips per person for the Boston region on the basis of the 2011 MTS, the regional MPO models, and the three versions of the CDR data analysis. The total travel estimates across these sources range from just less than 13 million trips to almost 19.5 million trips. These results underscore the difficulty of establishing a ground truth estimate against which these estimates can be compared.

- The 2010 Boston MPO model had the lowest number of total daily person-trips, with 12.92 million trips. The CDR Model 3 (third-party) estimate of 19.42 million total daily trips was the highest, followed by the 2011 MTS with 18.31 million trips.
- In contrast, the results of CDR Models 1 and 2 were similar to those of the 2007 Boston MPO model:
  - The results of CDR Models 1 and 2 are similar, with total daily person-trips of 15.36 million and 15.70 million, respectively;
  - Both of these estimates are slightly larger than but comparable to the 2007 Boston MPO model estimate of 14.23 million daily trips; and
  - These comparisons suggest a reasonably close correspondence in total trip-making patterns in the region.

Table 8-1. Total daily person-trips.

	Person-Trips (millions)	
Estimation Source	**Total Daily**	**Individual Daily Average**
2011 MTS (Massachusetts Department of Transportation 2012)	18.31	4.11
2010 Boston MPO model (2010 CTPS travel demand model)	12.92	2.90
2007 Boston MPO model (2007 CTPS travel demand model)	14.23	3.20
CDR Model 1 (cell phone model using 2010 raw cell phone data)	15.36	3.45
CDR Model 2 (cell phone model using 2010 raw cell phone data)	15.70	3.52
CDR Model 3 (third-party 2015 cell phone data processed by data provider)	19.42	4.36

Table 8-2. Daily weekday person-trips by purpose.

	Person-Trips (%) by Purpose			
Estimation Source	**HBW**	**HBO**	**NHB**	**Total**
2009 NHTS (Federal Highway Administration)	13	55	32	100
2011 MTS (Massachusetts Department of Transportation 2012)	12	49	39	100
1991 BHTS (Boston MPO 1991)	20	48	32	100
2010 Boston MPO model (2010 CTPS travel demand model)	23	55	22	100
2007 Boston MPO model (2007 CTPS travel demand model)	20	49	31	100
CDR Model 1 (cell phone model using 2010 raw cell phone data)	18	51	31	100
CDR Model 2 (cell phone model using 2010 raw cell phone data)	27	42	31	100
CDR Model 3 (third-party 2015 cell phone data processed by data provider)	18	49	33	100

Table 8-1 also provides estimates of individual daily average trips per person in the Boston region. As expected, CDR Models 1 and 2 and the 2007 Boston MPO model had similar results, with averages ranging from 3.2 to 3.5 daily trips per person.

These estimates are lower than the approximately 4 daily trips per person provided as guidance in Table 5.4 of *NCHRP Report 716: Travel Demand Forecasting: Parameters and Techniques* (Cambridge Systematics, Inc. et al. 2012).[1] The NCHRP report suggests that average trip rates in households of different sizes vary in a rather narrow range from 3.7 to 4.1 trips per person. The 2011 MTS, with an average of 4.11 trips per person (Table 8-1), is the data source closest to the trip rate estimates in *NCHRP Report 716*.

8.3.2 Person-Trips by Purpose

Table 8-2 presents the share of person-trips by purpose estimated by the three surveys, the two Boston regional models, and the three versions of CDR data analyses. Although the 2009 NHTS covers the exact same study area as the other surveys and models, these data were included to evaluate the share of daily person-trips by purpose.

The distribution of trips by purpose suggests that the relative incidence of work, nonwork, and non-home-based trips generally falls within the range provided by other sources and *NCHRP Report 716* (Cambridge Systematics, Inc. et al. 2012). Therefore, the inferences made by CDR Model 1 about the home, work, and other activity locations and the relative incidence of inferred trip purposes appear to be reasonable.

8.3.2.1 HBW Trips

The analysis of the HBW trips shows a rather wide range of estimates across the various data sources (Table 8-2), specifically:

- The share of work trips as a percentage of total daily trips ranged from a low of 12% to a high of 27%:
  - The 2011 MTS and the 2009 NHTS had the smallest HBW shares (12% and 13%, respectively).

[1]It should be noted that *NCHRP Report 716* differentiates trip rates by household size and by income category. Average trip rates vary considerably by income, with higher trip rates corresponding to households with higher incomes.

  - CDR Model 2 and the 2010 Boston regional model had the largest HBW shares (27% and 23%, respectively).
  - CDR Model 2 used a relaxed definition for the work location. Its high share (27%) of HBW trips suggests that the definition may be too broad compared with the more conservative definition used in CDR Model 1.
- The other four sources of data show a much tighter range, with work trips representing between 18% and 20% of total daily travel.
  - CDR Model 1, which used a conservative method to label work activity, had an 18% share of HBW trips. The estimate for CDR Model 3, which used vendor-processed CDR data, was the same.
  - The close match between CDR Model 1 and the vendor data set suggests that the methods used to differentiate between trips were broadly similar.
  - The 2007 Boston MPO model and the 1991 Boston survey provided similar estimates of the HBW share.
- Table 5.8 of *NCHRP Report 716: Travel Demand Forecasting: Parameters and Techniques* shows the range of the share of HBW trips for cities of different sizes (Cambridge Systematics, Inc. et al. 2012).
  - The estimates in Table 5.8 of *NCHRP Report 716* came from *NCHRP Report 187* (Sosslau et al. 1978), *NCHRP Report 365* (Martin and McGuckin 1998), and the 2009 NHTS and ranged from 14% to 25% for cities comparable to the Boston metropolitan area that had a population between half a million and 3 million residents.
  - The share of HBW trips was highest in the 1978 NCHRP report (25%) and lowest in the 2009 NHTS data (14%).
  - The trend in lower shares for HBW trips reflects, to some extent, the better reporting of shorter NHB trips in more recent survey efforts.
- In summary, a share of 18% to 20% of work trips is within the range provided by *NCHRP Report 716*, suggesting that the estimates from CDR Model 1 and CDR Model 3 (the vendor data set) are reasonable (Cambridge Systematics, Inc. et al. 2012).

8.3.2.2 HBO Trips

The share of HBO trips across all eight sources fell within a relatively narrower range, with a low of 42% and a high of 55% (Table 8-2).

- The CDR Model 2 approach resulted in the lowest share of HBO trips (42%), which reflects the much higher share of HBW trips obtained under this method.
- The 2010 Boston MPO model and the 2009 NHTS had the highest share of HBO trips, at 55% of total daily travel.
- The share of HBO trips in the other five sources of data ranged between 48% and 51%, a much tighter range. The 51% share for CDR Model 1 and 49% share for CDR Model 3 (the vendor data set) were similar to the shares for the 2007 Boston MPO model and the 1991 BHTS and 2011 MTS survey.
- Table 5.8 of *NCHRP Report 716: Travel Demand Forecasting: Parameters and Techniques* suggests that the share of HBO trips for urban areas comparable to Boston is 54% to 56% of total daily trips. This estimate has stayed relative stable over the years (Cambridge Systematics, Inc. et al. 2012).
- In summary, the CDR Model 1 estimate of HBO trips as 51% of daily trips is comparable to, but a little lower than, the guidance in *NCHRP Report 716* on HBO travel accounting for 56% of total daily trips.

8.3.2.3 NHB Trips

Reflecting the variation in HBW and HBO trip estimates, the NHB share of trips also ranged a lot between a low of 22% and a high of 39% across the eight sources of data (Table 8-2).

- The two outliers are the 2010 Boston MPO model, with the lowest NHB share of 22%, and the 2011 MTS, with the highest NHB share of 39%. Both of these shares fall outside the range of NHB travel observed by planners across different regions and can be considered as outliers.
- The range of NHB trips among the other sources of data is narrow, ranging between 31% and 33%.
- The estimates from CDR Model 1, CDR Model 2, and CDR Model 3 (the vendor data set) are similar to those from the 2007 Boston MPO model, the 1991 BHTS, and the 2009 NHTS.
- Table 5.8 of *NCHRP Report 716: Travel Demand Forecasting: Parameters and Techniques* suggests that the share of NHB trips for urban areas similar to the Boston region is 30% of total daily trips (Cambridge Systematics, Inc. et al. 2012).
  - These NHB estimates have changed over time, from 21% in 1978 to 22% in 1998 to 30% in 2009 for regions comparable to the Boston area.
  - The trend in a higher share for NHB trips reflects, to some extent, the better reporting of shorter NHB trips in more recent survey efforts.
- In summary, the shares of total daily travel constituted by NHB trips for CDR Models 1 and 2 (31% each) and for CDR Model 3, the vendor data set (33%), are consistent with the guidance in *NCHRP Report 716* (Cambridge Systematics, Inc. et al. 2012).

8.3.3 Person-Trips by Time of Day

The analysis of travel patterns was also extended to evaluate travel by time of day. Table 8-3 presents the relative shares of average weekday trips by time of day derived from the same sources of CDR data, survey data, and regional models reported on above. The comparison of these eight

Table 8-3. Share of daily weekday person-trips by time of day.

	Share (%)				
Estimation Source	A.M. Peak (6–9 a.m.)	Midday (9 a.m.–3 p.m.	P.M. Peak (3–7 p.m.)	Rest of Day (7p.m.–6 a.m.)	Total
2009 NHTS (Federal Highway Administration)	19	37	31	13	100
2011 MTS (Massachusetts Department of Transportation 2012)	21	34	33	12	100
1991 BHTS (Boston MPO 1991)	18	32	33	17	100
2010 Boston MPO model (2010 CTPS travel demand model)[a]	11	51	21	17	100
2007 Boston MPO model (2007 CTPS travel demand model)[a]	16	34	28	22	100
CDR Model 1 (cell phone model using 2010 raw cell phone data)	16	27	27	30	100
CDR Model 2 (cell phone model using 2010 raw cell phone data)	17	36	27	20	100
CDR Model 3 (third-party 2015 cell phone data processed by data provider)[b]	20	36	27	18	100

Note: Detail may not add to total because of rounding.
[a] The total number of trips for the p.m. period (PM) and rest-of-day period (RD) was adjusted from the original periods of the MPO model, which used 3 to 6 p.m. as the p.m. peak (PM*) and 6 p.m. to 6 a.m. as the rest of the day (RD*). RD = RD*/12 × 11. PM = total − AM − MD − RD, where AM = a.m. peak and MD = midday.
[b] Data were scaled to the 2010 Census population.

98 Cell Phone Location Data for Travel Behavior Analysis

sources highlights some new patterns that are specific to differences in time of day and different from trip purposes. The a.m. and p.m. peak periods are discussed separately and then compared with the midday and rest-of-day travel patterns.

8.3.3.1 Peak Period Traffic

- The 2010 Boston MPO model appears to underrepresent both a.m. and p.m. peak period travel as compared with all other surveys and models. For purposes of this discussion, it can be considered as an outlier.
- Although there are some differences in the a.m. peak share of trips across the other data sources, the range of these differences is narrow—between 16% and 21%.
- The same pattern holds true for the p.m. peak share of trips, which ranged from a low of 27% to a high of 33%.
- CDR Model 1 had a.m. and p.m. peaking characteristics similar to those of the 2007 Boston MPO model. The same was true of CDR Model 2, which was considered less reliable because its share of HBW trips was much higher than expected.
- The three survey data sources (2009 NHTS, 2011 MTS, and 1991 BHTS) and CDR Model 3 (the vendor CDR data) shared many similarities. All of these estimates had consistently pronounced peaking patterns in both the a.m. and p.m. peaks.
- A comparison of these findings with the time-of-day guidance in the Travel Model Validation Manual (Cambridge Systematics, Inc. 2010) is not as clear, given the different definition of the time periods. However,
  - The Validation Manual's 2001 NHTS estimate of a 12% share of daily trips between 7 and 9 a.m. is broadly consistent with the 19% share for the 3-hour a.m. peak in the 2009 NHTS summary.
  - Similarly, the Validation Manual's 24% share of trips between 3 and 6 p.m. is broadly consistent with the 31% share for the 4-hour p.m. peak in the 2009 NHTS.
- In summary, CDR Model 1 had comparable but lower shares of trips in both peak periods as compared with the other sources of data but was similar to the 2007 Boston MPO model. CDR Model 3 (vendor-provided data set) was close to the 2011 MTS in the a.m. peak but lower in the p.m. peak.

8.3.3.2 Midday and Rest-of-Day Traffic

As expected, the results for the midday period between 9 a.m. and 3 p.m. and the rest of the day between 7 p.m. and 6 a.m. were, to a large extent, mirror images of the patterns observed for the two peak periods.

- The 2010 Boston MPO model can again be treated as an outlier, given that it had 51% of all daily trips during the midday—a much higher percentage than any other data source.
- The midday share of trips ranged between 32% and 37% of all daily trips, with one key exception: at 27%, CDR Model 1's share of midday trips was much lower than that of all the other data sources. CDR Model 3's estimate of 36% was close to that of the other data sources.
- The distribution of trips during the rest of the day showed a somewhat different pattern that is worth discussing:
  - The 2009 NHTS and the 2011 MTS had low shares of 13% and 12% of midday trips, respectively, compared with the 17% and more observed in the other data sources.
  - However, the much larger share of late evening and early morning trips under CDR Model 1 should be noted. This is a pronounced difference when compared with any of the other sources of data.
  - The similarity of the CDR Model 2 and the 2007 Boston MPO model shares in each of the four periods should also be noted. CDR Model 2 was also consistent with the 1991 BHTS and CDR Model 3.

- In summary, CDR Model 1 had the highest share of rest-of-day trips. The CDR Model 3 (vendor-provided data) estimate fell within the range of the rest-of-day estimates for the other data sources.

Overall, the time-of-day patterns evaluated were not as clear as the patterns observed in the other evaluations, in part because there were considerable differences between the estimates provided by the Boston models and the regional surveys.

- CDR Model 1 had comparable but lower shares of trips for both peak periods as compared with the other sources of data but was similar to the 2007 Boston MPO model. CDR Model 1 had a 27% share of midday trips, which was much lower compared with all other data sources. It also had the largest share by far of rest-of-day trips.
- CDR Model 3 (vendor-provided data) was close to the three survey estimates for the a.m. peak but lower in its p.m. peak estimates. This CDR source was close to the other data sources for midday trips and fell within the range of the rest-of-day estimates of the other data sources.
- On balance, the vendor-provided CDR estimates were more comparable to the other data sources in terms of the time-of-day distribution of trips.

8.3.4 Comparisons at the City and Town Level

To examine the spatial distribution of the CDR model results, the research team compared O-D person-trips in each CDR model with those in the 2010 Boston MPO model. The tract pair level CDR results were aggregated to the city and town level, and the same process was repeated for the Boston model by aggregating TAZ pair results. The comparisons of total trips, trips by purpose, and trips by time of day suggested the following:

- At the city pair and town pair levels, O-D person-trips estimated from the CDR models were highly correlated with the MPO model for total trips, across the three trip purposes, and across the four times of day.
- The intratown O-D person-trips that reflect travel between zones within the same city or town showed a higher degree of correlation between each CDR model and the MPO model.
- The intertown O-D person-trips reflecting travel between different cities and towns showed a lower correlation that was still satisfactory.
- Table 8-4 shows the comparisons between each CDR model and the Boston MPO model results for all trips, intratown travel, and intertown trips.
  - The correlation for all trips ranged between 0.96 and 0.98;
  - The correlation for intratown trips was high—above 0.98;
  - The correlation for intertown trips was lower, ranging from 0.90 to 0.94; and
  - Despite small differences in correlation across the CDR methods, CDR Model 1 had a slightly higher correlation for each of the three types of trips.

8.3.5 Daily O-D Person-Trips

Figure 8-2 shows the comparison of daily O-D person-trips for weekday travel for each of the three CDR models and the 2010 Boston MPO model. The horizontal *x*-axis represents observations in each CDR model, while the vertical *y*-axis represents observations from the 2010 Boston MPO model. The results shown in Figure 8-2, *a* and *b*, are consistent with Table 8-4 and provide a qualitative assessment of how well the CDR data match the model.

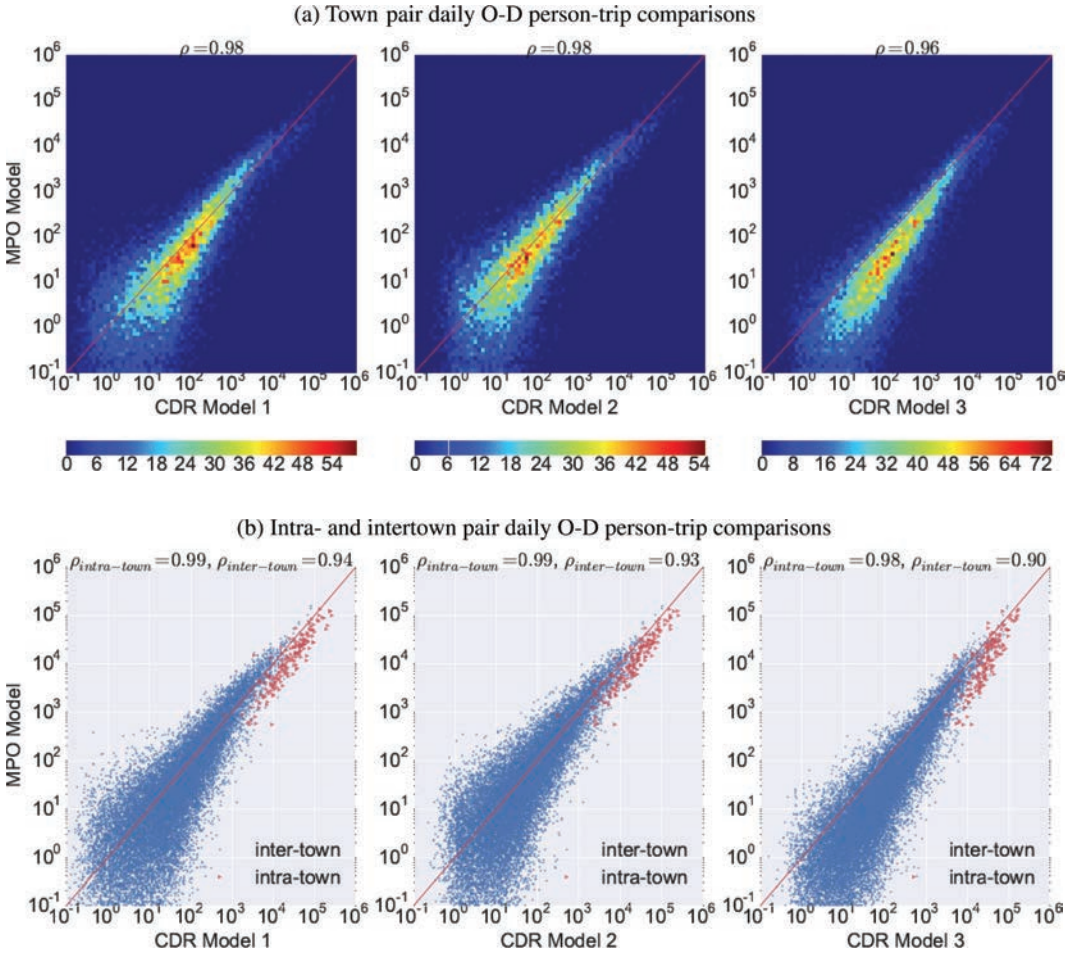
Figure 8-2*a* shows that CDR Model 1 has more observations close to the 45° line, suggesting a better match with the MPO model. Figure 8-2*b* differentiates between the intratown O-D pairs, shown as red dots, and the intertown O-D pairs, shown as blue dots. Again, the horizontal *x*-axis

Table 8-4. Correlation of person-trips.

Estimation Source	Total (Pearson correlation coefficient)
All Pairs	
CDR Model 1	0.98
CDR Model 2	0.98
CDR Model 3	0.96
Intratown	
CDR Model 1	0.99
CDR Model	0.99
CDR Model 3	0.98
Intertown	
CDR Model 1	0.94
CDR Model 2	0.93
CDR Model 3	0.90

Source: 2010 Boston MPO model and CDR Models 1–3.

Note: CDR Model 1 = cell phone model using 2010 raw cell phone data; CDR Model 2 = cell phone model using 2010 raw cell phone data; and CDR Model 3 = third-party 2015 cell phone data processed by a data provider. CDR Models 1–3 were compared with the 2010 Boston MPO model as the baseline.



Source: 2010 Boston MPO model and CDR Models 1, 2, and 3.

Figure 8-2. Comparison of O-D person-trips by geography.

represents observations in each of the CDR models while the vertical *y*-axis represents observations from the 2010 Boston MPO model.

These three comparisons suggest that CDR Models 1 and 2 had a better correspondence with the MPO model than CDR Model 3 (the vendor product) for the intratown pairs shown with the blue dots. The results are less clear for the intertown pairs, although again CDR Models 1 and 2 have a better correspondence with the MPO model.

8.3.6 O-D Person-Trips by Purpose

Table 8-5 extends the analysis of correlation patterns by focusing on HBW, HBO, and NHB trips. Each of the three CDR models was compared with the 2010 MPO model at the city pair and town pair levels.

- The intratown O-D person-trips showed a high correlation coefficient of 0.96 or greater when each CDR model was compared with the MPO model.
- The intertown O-D person-trips showed lower correlation coefficients, with values ranging between 0.76 and 0.93.
  - Among HBW trips, the lowest correlation is observed for CDR Model 2, a result consistent with the analysis of total work trips.
  - For HBO trips, the lowest degree of correlation is offered by CDR Model 3, the vendor-provided data set.
  - For NHB trips, the lowest correlation is again present when CDR Model 3 is compared with the 2010 Boston MPO model.
  - These patterns are consistent with the differences in the mix of trip purposes between CDR methods and the Boston model shown in Table 8-2.

Figure 8-3 shows how the results from the three CDR models compare with the 2010 Boston MPO model when the data are differentiated by trip purpose. The horizontal *x*-axis represents

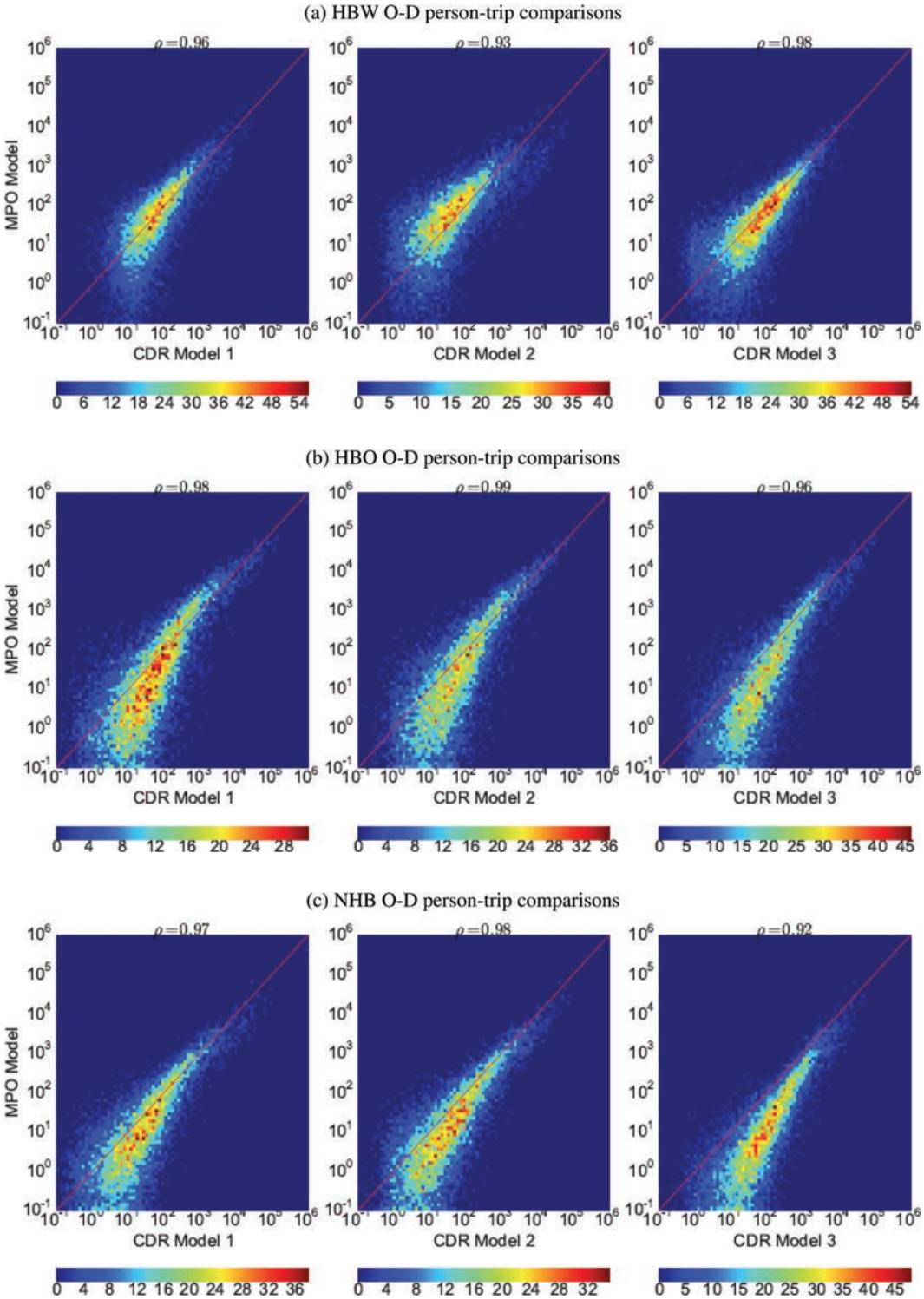
Table 8-5. Correlation of person-trips by purpose.

	Pearson Correlation Coefficient by Trip Purpose		
Estimation Source	**HBW**	**HBO**	**NHB**
All Pairs			
CDR Model 1	0.96	0.98	0.97
CDR Model 2	0.93	0.99	0.98
CDR Model 3	0.98	0.96	0.92
Intratown			
CDR Model 1	0.99	0.99	0.99
CDR Model 2	0.97	0.99	0.99
CDR Model 3	1.00	0.98	0.96
Intertown			
CDR Model 1	0.89	0.93	0.88
CDR Model 2	0.80	0.92	0.88
CDR Model 3	0.90	0.89	0.76

Source: 2010 Boston MPO model and CDR Models 1–3.

Note: CDR Model 1 = cell phone model using 2010 raw cell phone data; CDR Model 2 = cell phone model using 2010 raw cell phone data; and CDR Model 3 = third-party 2015 cell phone data processed by a data provider. CDR Models 1–3 were compared with the 2010 Boston MPO model as the baseline.

102 Cell Phone Location Data for Travel Behavior Analysis



Source: 2010 Boston MPO model and CDR Models 1, 2, and 3.

Figure 8-3. Comparison of O-D person-trips by purpose.

the CDR models, while the vertical *y*-axis represents observations from the 2010 Boston MPO model. The patterns in these graphs provide a qualitative way to evaluate the degree of match with the 2010 Boston MPO model:

- The first row of figures refers to HBW trips and shows that CDR Models 1 and 3 have more observations close to the 45° line and therefore offer a better match with the 2010 Boston MPO model than does CDR Model 2.
- The second row of figures refers to HBO trips and shows that CDR Models 1 and 2 have more observations close to the 45° line.
- The third row of figures refers to NHB trips and shows that CDR Models 1 and 2 have more observations close to the 45° line and therefore offer a better match with the 2010 Boston MPO model.

Figure 8-4, *a–c*, shows similar comparisons by trip purpose but further differentiates between intratown O-D pairs, shown in red, and intertown pairs, shown in blue. The horizontal *x*-axis represents the CDR models, while the vertical *y*-axis represents observations from the 2010 Boston MPO model.

- Both the intertown and the intratown HBW trips show a better match between CDR Model 3 and the 2010 Boston MPO model.
- The intertown HBO trips show that CDR Models 1 and 2 have more observations close to the 45° line. The intratown comparisons suggest that CDR Model 2 has a better match with the 2010 Boston MPO model.
- The intertown NHB trips show a closer match between CDR Models 1 and 2 and the 2010 Boston MPO model. The intratown trips do not have a good match, although CDR Models 1 and 2 are again more similar to the 2010 Boston MPO model.

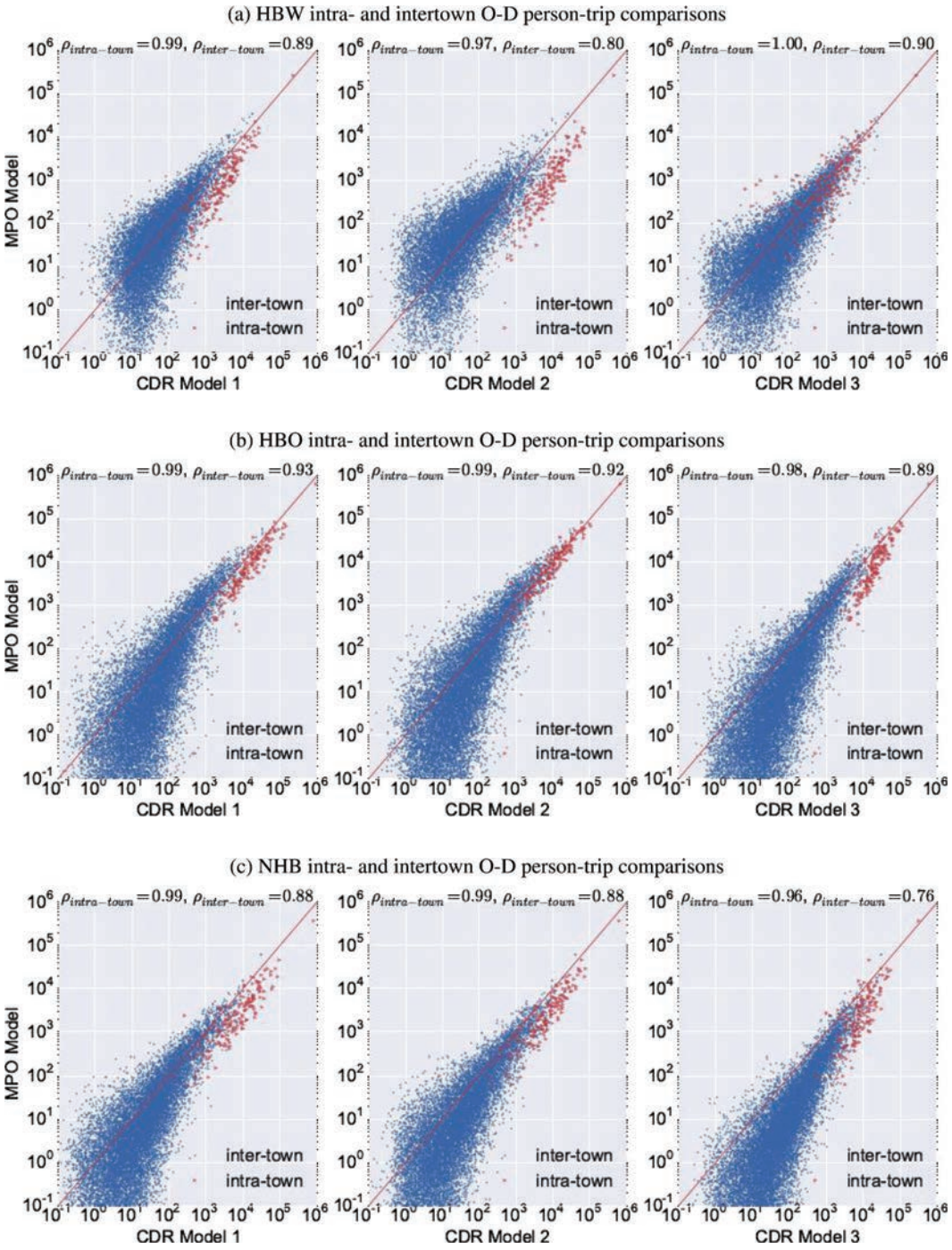
8.3.7 O-D Person-Trips by Time of Day

This section presents a similar analysis of correlation patterns that focus on trips by time of day. The analysis is repeated for each of the four periods and compares the three CDR models and the 2010 Boston MPO model at the city pair and town pair levels for an average weekday.

The correlation patterns in Table 8-6 for all trips, intratown trips, and intertown O-D person-trips can be summarized as follows:

- The correlation coefficients for all O-D person-trips are high, with values ranging between 0.95 and 0.98.
- The intratown O-D person-trips show high correlation coefficients of 0.98 or greater when each CDR model is compared with the 2010 Boston MPO model.
- The intertown O-D person-trips show somewhat lower correlation coefficients, with values ranging between 0.89 and 0.96.
  - Among a.m. and p.m. peak trips, the highest correlation is observed for CDR Model 1, a result consistent with the analysis of work trips.
  - The same pattern applies to midday and rest-of-day trips, with CDR Model 1 and the 2010 Boston MPO model reflecting similar distributions.
  - These results suggest the uniformly better ability of CDR Model 1 to replicate the 2010 Boston MPO model across all time periods.
  - In contrast, the vendor-provided data in CDR Model 3 provide a lower degree of correspondence with the 2010 Boston MPO model.

104 Cell Phone Location Data for Travel Behavior Analysis



Source: 2010 Boston MPO model and CDR Models 1–3.

Figure 8-4. Comparison of inter- and intratown O-D person-trips by purpose.

Table 8-6. Correlation of person-trips by time of day

	Pearson Correlation Coefficient by Time of Day			
Estimation Source	**A.M. Peak**	**Midday**	**P.M. Peak**	**Rest of Day**
All Pairs				
CDR Model 1	0.98	0.98	0.98	0.98
CDR Model 2	0.97	0.98	0.98	0.98
CDR Model 3	0.95	0.96	0.97	0.96
Intratown				
CDR Model 1	0.99	0.99	0.99	0.99
CDR Model 2	0.99	0.99	0.99	0.99
CDR Model 3	0.98	0.98	0.98	0.98
Intertown				
CDR Model 1	0.95	0.94	0.96	0.95
CDR Model 2	0.92	0.93	0.94	0.92
CDR Model 3	0.93	0.89	0.93	0.91

Source: 2010 Boston MPO model and CDR Models 1–3.

Note: CDR Model 1 = cell phone model using 2010 raw cell phone data; CDR Model 2 = cell phone model using 2010 raw cell phone data; and CDR Model 3 = third-party 2015 cell phone data processed by a data provider. CDR Models 1–3 were compared with the 2010 Boston MPO model as the baseline. For CDR models, PM = 3 to 7 p.m.; RD = 7 p.m. to 6 a.m. For 2010 Boston MPO model, PM = 3 to 6 p.m.; RD = 6 p.m. to 6 a.m.

Figure 8-5 shows the comparison between the CDR models on the vertical axis and the 2010 Boston MPO model in the horizontal axis for all city pair and town pair O-D trips broken out by time period. These qualitative comparisons further underscore the strong correspondence between the results from CDR Models 1 and 2 and those from the 2010 Boston MPO model for a.m. peak, p.m. peak, and midday travel.

Figure 8-6 further differentiates O-D trips by time of day for intratown pairs, shown in red, and intertown pairs, shown in blue. Similar to the trends seen in the comparisons by purpose, the intertown pairs are more widely distributed than the intratown pairs, which are more tightly distributed. The horizontal *x*-axis represents the CDR models, while the vertical *y*-axis represents observations from the 2010 Boston MPO model.

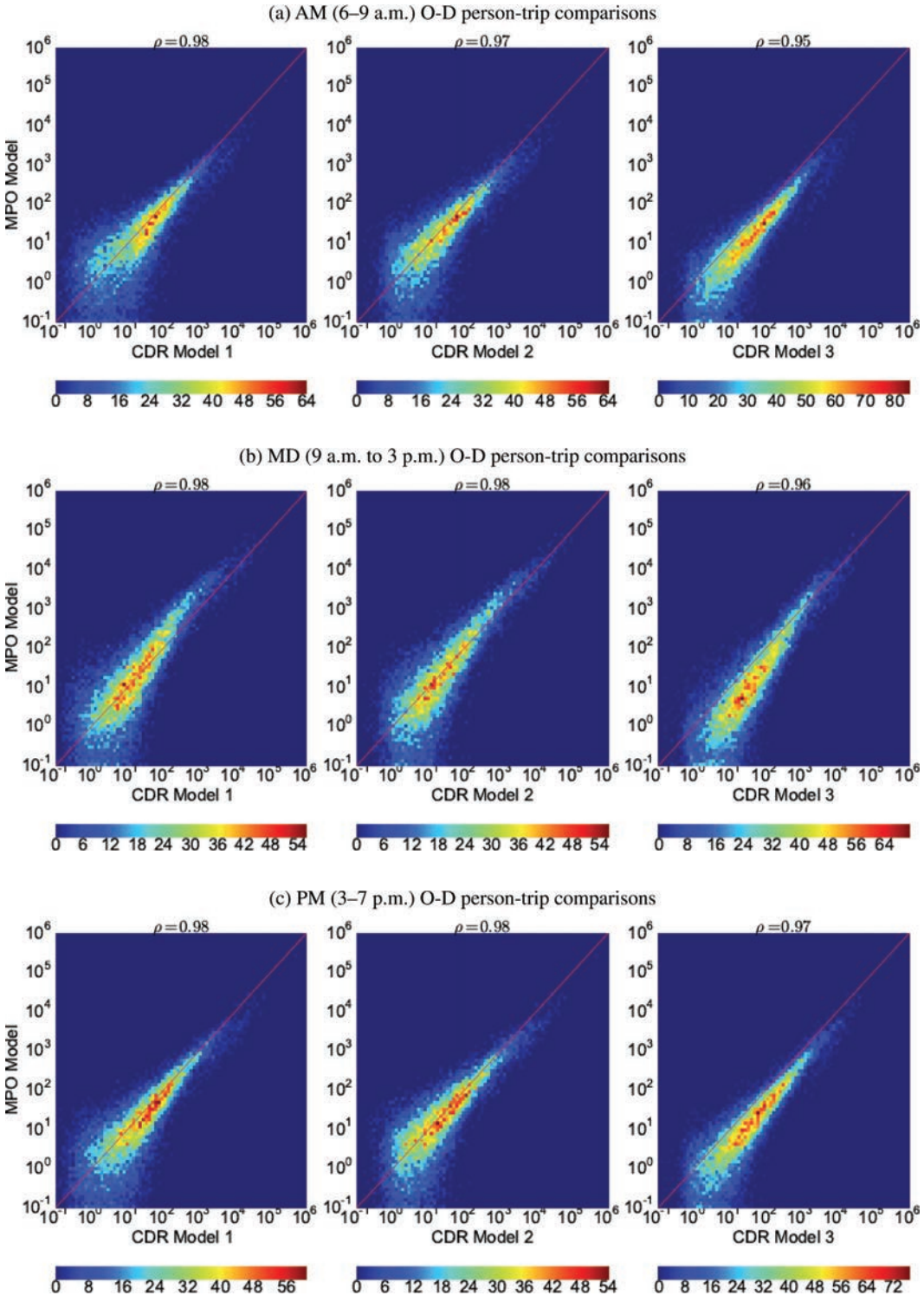
- The intertown a.m. peak trips show a better match between CDR Model 1 and the 2010 Boston MPO model. The intratown a.m. peak observations are clustered but are also concentrated below the 45° line for all CDR models.
- The intertown midday trips show more observations close to the 45° line for CDR Models 1 and 2. The same pattern applies to intratown comparisons.
- The intertown PM peak trips show a close match between CDR Models 1 and 2 and the 2010 Boston MPO model. The intratown trips do not have as good a match and are again consistently below the 45° line. CDR Models 1 and 2 are more similar to the 2010 Boston MPO model.

8.4 Summary

This chapter goes to the heart of the comparisons between CDR-derived travel estimates and traditional measures of travel, including traditional household surveys and model outputs. The findings of these comparisons are briefly summarized below.

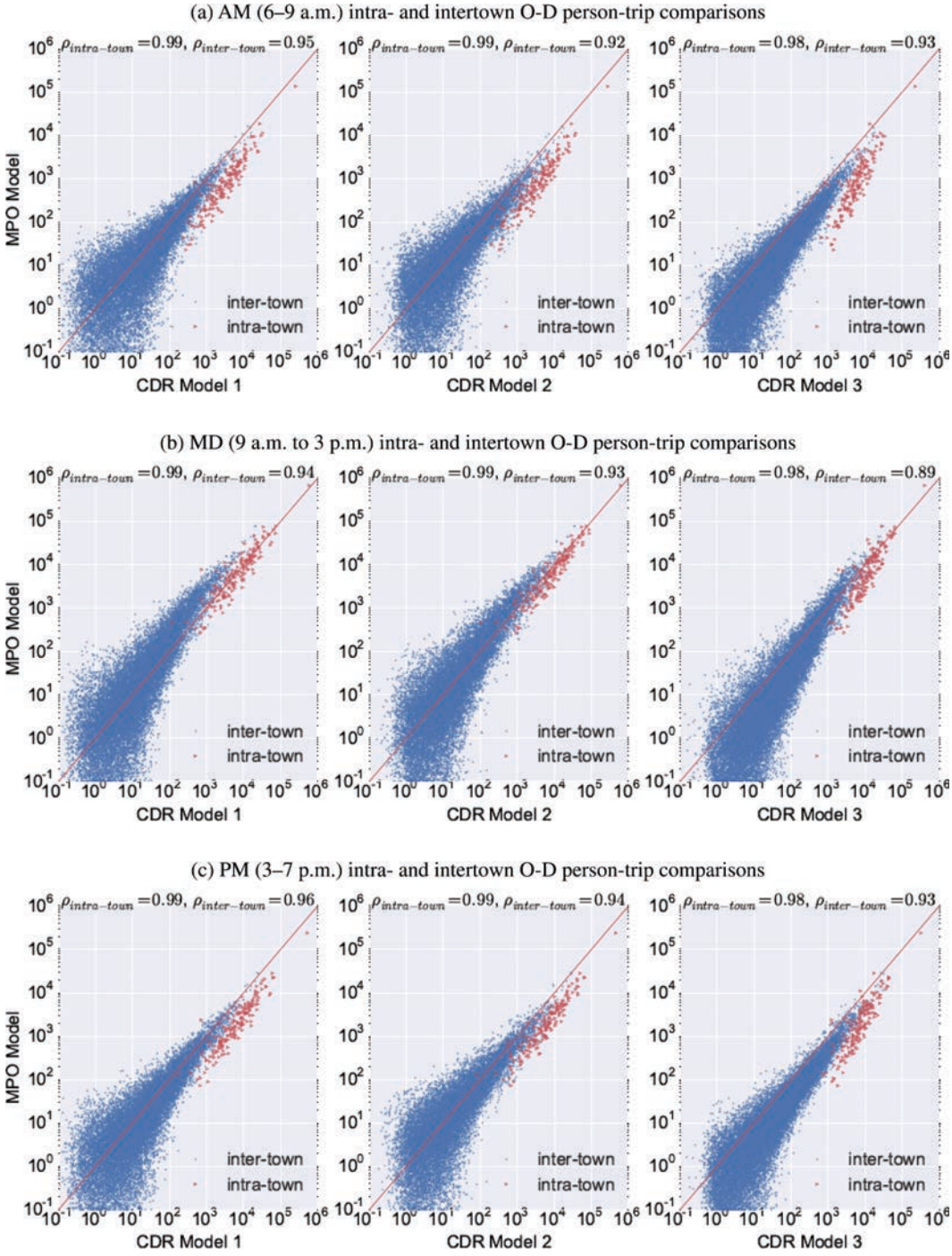
The total volume of trip making and the distribution of trips by purpose were evaluated by comparing the three CDR models, one a vendor product, with three regional surveys and two

106 Cell Phone Location Data for Travel Behavior Analysis



Source: 2010 Boston MPO model and CDR Models 1–3.

Figure 8-5. Comparison of O-D person-trips by time of day.



Source: 2010 Boston MPO model and CDR Models 1–3.

Figure 8-6. Comparison of intra- and intertown O-D person-trips by time of day.

versions of the Boston MPO model. These results were also compared with the guidance in *NCHRP Report 716* where applicable (Cambridge Systematics, Inc. et al. 2012):

- CDR Models 1 and 2 and the 2007 Boston MPO model were similar, producing 3.2 to 3.5 daily trips per person. These estimates are comparable to but a little lower than the approximately four daily trips per person reported in *NCHRP Report 716*.
- CDR Model 1 and CDR Model 3 (vendor-provided data) showed a share of 18% to 20% of work trips, which is within the range provided by *NCHRP Report 716*.
- CDR Model 1 estimated HBO trips as 51% of daily trips; this estimate is comparable to, but a little lower than, the guidance in *NCHRP Report 716*.
- The CDR Model 1 and CDR Model 3 (vendor-provided data) shares of NHB trips (31% and 33%, respectively) are consistent with the guidance in *NCHRP Report 716*.

The time-of-day patterns are less clear, in part because there were considerable differences in the estimates provided by the regional surveys and models. On balance, Model 3's vendor-provided CDR estimates were more comparable to those of the other data sources in terms of the distribution of trips by time of day:

- CDR Model 1 had comparable but lower shares of trips for both peak periods. Its 27% share of midday trips was much lower than that of all other data sources, and it also had the largest share of rest-of-day trips by far.
- CDR Model 3's (vendor-provided data) estimates for the a.m. peak were close to those of the three surveys, but its p.m. peak estimate was lower. This model's results for midday trips were also close to those of the other data sources and fell within the rest-of-day range of estimates.

The analysis of O-D trips at different levels of geographic detail was carried out for total trips, trips by purpose, and trips by time of day:

- At the city pair and town pair level, O-D person-trips estimated from the CDR models correlated highly with the 2010 Boston MPO model for total trips, trips by purpose, and trips by time of day.
- The intratown O-D person-trips that reflect travel between zones within the same city or town showed a higher degree of correlation than the intertown trips that reflect travel between different cities and towns.
- CDR Models 1 and 2 had a better correspondence with the 2010 Boston MPO model for both the intratown and intertown pairs.

The detailed analysis of O-D person-trips by purpose and time of day suggested the following:

- CDR Model 1 and CDR Model 3 (vendor-provided data) offered a better match with the 2010 Boston MPO model for HBW trips. CDR Models 1 and 2 showed a closer match for HBO trips and for NHB trips.
- For the a.m. and p.m. peak periods, the highest correlation was observed for CDR Model 1; this result was consistent with the analysis of work trips. The same pattern applied to midday and rest-of-day trips, with CDR Model 1 matching the 2010 Boston MPO model more closely.
- These results suggest that CDR Model 1 can replicate the results of the 2010 model across purposes and time periods better than the other CDR approaches.

These comparisons provide a basis for evaluating the ability of CDR data to emulate the results obtained by the analysis of traditional surveys and regional models. As discussed in this chapter, it is necessary to be aware of the lack of definitive ground truth when these comparisons are carried out. It is also necessary to remain aware of the assumptions and inferences embedded in each data source and made for each type of analysis.

The value of these comparisons lies in their transparency, in that they can serve as a benchmark for practitioners assessing the value of CDR data for different purposes. Additional comparisons such as trip-length distributions and screenline comparisons can be carried out to provide more insight into the value of CDR data. Alternatively, different assumptions, such as the duration threshold used for defining a stay in CDR data, can be tested and evaluated.

The next chapter summarizes key considerations about potential uses of CDR data and provides guidelines for practitioners of planning and modeling. The chapter focuses on the questions practitioners typically ask about the properties of data and models to shed more light on the potential value of cell phone CDR locational data.

CHAPTER 9

Guidelines for Practitioners

9.1 Roadmap to the Chapter

Chapters 4 through 8 highlighted three-way comparisons between survey data, traditional models, and call detail record (CDR)-derived estimates to help bridge the gap between research and practice. This chapter summarizes key considerations about the potential uses of CDR data and provides guidelines for practitioners of planning and modeling.

Specifically, the chapter focuses on the questions practitioners typically ask about data and models to shed more light on the potential value of cell phone CDR locational data. To extract the maximum value from cell phone data for planning and modeling purposes, practitioners can consider the following general principles:

- Be aware of the underlying assumptions that are made to process cell phone data to determine locations and to infer activities and purposes. Vendors are likely to make inferences that are similar to the method documented in this report to process CDR data and develop origin–destination (O-D) tables.
- Recognize that results from traditional surveys and models are also built on different sets of assumptions. Although transportation practitioners are more familiar with these methods, traditional survey and model results do not necessarily provide a true ground truth baseline either.
- Expect that with the increase in the quantity and quality of cell phone data and the use of powerful machine learning algorithms, vendors will improve their analytical methods to analyze locational data to infer travel patterns. As new products are developed, practitioners need to continue asking typical questions about the underlying data and assumptions used.
- Appreciate the uncertainty underlying both CDR estimates and traditional measures of travel patterns. Although it is not easy to quantify uncertainty, practitioners should be open to using ranges of estimates for both new and traditional data sources.

The practitioner guidelines presented here are grouped into three categories:

- Administrative considerations (Section 9.2);
- Data considerations, juxtaposing CDR data with traditional data sources (Section 9.3); and
- Modeling considerations and the potential of CDR data to support different model components (Section 9.4).

The advent of technology and how the availability of more and better-quality CDR data, coupled with new research in locational data, may create a new generation of enhanced CDR data products is discussed in Section 9.5. The chapter concludes with a discussion of examples of recent literature that benefit from richer and more detailed data sources and the use of advanced analytics.

9.2 Administrative Considerations

The specification, purchase, and licensing of CDR data is a complex transaction with cost and legal implications. The terms of payment need to be considered in the financial planning undertaken by public agencies. The cost of such data may also exceed the financial limits set for incidental purchases.

For example, a state department of transportation (DOT) may need to disclose plans for similar data purchases in advance as part of its State Planning and Research Work Programs. Similarly, a metropolitan planning organization (MPO) may need to disclose plans for purchasing CDR data in its Unified Planning Work Program. In the event that funds, contracts, or grants from other parties are used to finance such data purchases, then a pay-when-paid approach should be negotiated with the data vendor.

This section summarizes the financial, legal, schedule, technical, and communication considerations that practitioners are likely to face.

9.2.1 Financial Considerations

Agencies need to consider the financial implications of acquiring processed CDR data from third parties in their financial planning. Questions include the timing of the payments, the terms of the payments, and the delivery of available data to the agency.

Agency staff need to develop detailed specifications for the data request. The parameters that may affect the quality and cost of the CDR data set and that need to be defined by the agency include the following:

- Size of the data set;
- Spatial coverage;
- Duration of the observation period;
- Desired geographic level of detail;
- Types of travel, including purposes and resident-versus-visitor markets;
- Temporal detail for time-of-day analyses;
- Time period during which the data may be used;
- Ability to refresh the data for one or more future years;
- Single-use versus multiple-use data purchase; and
- Ability to share the data with agency partners.

9.2.2 Legal Considerations

Privacy considerations and proprietary methods have introduced complications in the use and dissemination of CDR data. It is safe to assume that, once purchased or licensed for a specific purpose, CDR data may not be used for any other purpose by a public agency.[1] Furthermore, it is also reasonable to assume that contract language will regulate how these CDR data can be shared with agency partners, including consulting firms working for a public agency.

Agency staff will probably need to address the following issues in a legal document such as a data-sharing agreement:

- The potential uses of the CDR data,
- The parties by whom these CDR data may be used or with whom they may be shared, and
- The products that can be derived from the CDR data.

[1] It is unlikely that the processed CDR O-D data can be purchased outright. The right to use the data will likely be licensed to public agencies for specific agreed-upon uses.

If the public agency does not have the signatory authority to enter into legal agreements, then legal staff of the signatory agency will need to enter into a legally binding agreement. An example of an MPO administered by a separate legal entity is the Greater Buffalo Niagara Regional Transportation Council, which is administered by the Niagara Frontier Transportation Authority.

Agency staff should also recognize and address in a legal document the inherent potential conflicts between the confidential nature of CDR data and any Freedom of Information Act or other state legislation to which the public agency may be subjected.

9.2.3 Schedule Considerations

The time required to obtain the CDR data may be longer than the agency's experience with traditional travel and survey data sources, owing to the financial and legal requirements that were discussed. As part of the agency's financial plans, agency staff should consider the potential contract negotiation delays in data acquisition.

Agency staff should also recognize that the negotiations are likely to affect the project schedule. Care should be taken to adjust the period of performance of contracts by a third party, including other agencies, agency partners, and consulting firms that are expected to use the CDR data for project-related analyses.

9.2.4 Technical Considerations

CDR data are different from data from household travel surveys and other data sources used to support planning analyses and model development. In this respect, CDR data have different strengths and weaknesses as compared with traditional data sources and may not conform exactly to a practitioner's expectations or needs.

The data considerations described in Section 9.3 and the modeling and analysis considerations discussed in Section 9.4 are accompanied by checklists of the types of questions that agency staff should think about and discuss when they are considering the purchase of CDR data.

Agency staff should review and clearly communicate to the data vendor the agency's data and analysis needs for planning and modeling purposes. Such a discussion helps the agency and the vendor to compare and contrast existing sources of data and model outputs with the corresponding properties of the CDR-derived travel data. This discussion will also help ensure that the vendor data product is specified on the basis of the agency's needs and that the CDR travel data provide value by supporting the agency's planning and modeling applications.

9.2.5 Communications Considerations

Given the complex nature of CDR data, it is also critical that agencies develop a clear and well-defined document that can be available to legislative bodies, news agencies, and informed citizens and that addresses the following considerations:

- Statutory language that allows and regulates the CDR data purchase;
- Description of the types of CDR data that are being collected and the purpose(s) of collecting and using these data;
- The kinds of planning questions that these data are expected to answer;
- The steps that have been taken to anonymize data and preserve privacy;
- The spatial and temporal resolution of the data; and
- Practices related to data access, retrieval, archiving, and deletion.

The design and availability of this document during the preliminary stages of project development will ensure that any privacy issues are dealt with in a comprehensive and transparent manner.

9.3 Data Considerations

Chapters 6 through 8 discuss how CDR data and the O-D trip tables derived from them are similar to and different from travel survey results and model outputs. This section summarizes the key data features of CDR data to highlight their strengths and weaknesses.

In addition to the 2010 CDR data that were used in this research effort and the 2015 vendor CDR data, the discussion in this section includes GPS logger and Bluetooth device data collection technologies that were used in a recent Transportation Model Improvement Program report (Hard et al. 2016). The discussion also addresses the features and potential of the smartphone app data collection option.

GPS logger data for personal or truck travel rely on the GPS functionality of a logger carried by a respondent as part of a household survey or embedded in a commercial vehicle as part of a freight survey. These GPS data are often combined with a follow-up telephone survey that obtains additional information and context about individual stops and activities. This method of enhancing the GPS data creates an integrated data source that benefits from both the contextual information and the detailed location data provided by GPS technology.

Bluetooth-based data passively capture and identify the location of vehicles or devices by employing Bluetooth readers on corridors and locations of interest. Little contextual or socioeconomic information is known about the user or the owner of the device. These data provide value in cases in which the analyst is interested in counts of vehicles along a facility. These data cannot be integrated with other sources to provide either socioeconomic or contextual information.

Smartphone apps represent a new wave of technology that offers a new option for integrating passive and active collection of locational data to infer travel. Respondents who agree to participate in the survey provide socioeconomic data and can also identify locations that they visit often. They agree to be passively monitored by the smartphone app, which traces their daily travel through the GPS tracking option embedded in their cell phones.

Respondents are asked to actively validate their travel through a prompted recall method in which they provide information about individual activities. A more sophisticated approach includes embedded machine learning software that makes inferences about the day's activities and presents them to respondents, asking them to verify the inferences.

Chapter 3 discussed how the various elements of travel obtained from traditional surveys are similar to and different from travel estimates obtained from the analysis of CDR data. This comparison clarified the strengths and weaknesses of CDR data as compared with traditional survey data. Table 3-3 is the key table, repeated here as Table 9-1, to highlight the contrast between traditional survey data and CDR data for key variables that are critical to planning and modeling analyses.

Table 9-2 extends this comparison by highlighting how specific data properties differ across four technology options: CDR data, GPS loggers, smartphone apps, and Bluetooth devices. The entries in Table 9-2 correspond to elemental properties of data and include

- Raw versus processed data records,
- Spatial and temporal resolution of each data source,
- Level of technology used, and

114 Cell Phone Location Data for Travel Behavior Analysis

Table 9-1. Travel elements in traditional surveys and CDR data.

Variable of Interest	Travel Data from Traditional Surveys	Travel Data Based on Cell Phone Use
Total daily travel	Self-reported in survey diaries. Travel may be underreported. Prompted recall offers an improvement.	Passive cell signals over days may offer more robust metrics than surveys. Unit is device-trips rather than person-trips. Quality depends on CDR data density.
Time of travel	Self-reported in survey diaries. Times may be inaccurate and incomplete.	Accurate time stamps. Need to infer activity and link it to the time stamp versus en route travel.
Stops versus activities	Self-reported in survey diaries. Detailed log of stops and activities. Good detail on all travel purposes.	Need to infer stops, activities, segments. Nonwork purposes are difficult to infer.
Location of activities	Self-reported in survey diaries. Smart geocoding needed to match. Prompted recall offers an improvement.	Difficult to infer the location of activities. A challenge in mixed land use areas.
Travel purpose	Self-reported in survey diaries. Prompted recall offers an improvement.	Home and work locations are inferred. Poor inference on nonhome and nonwork.
Joint travel	Self-reported in survey diaries. Risk of underreporting. Prompted recall offers an improvement.	Not feasible to record or capture.
Mode of travel	Self-reported in survey diaries. Good detail by tour and segment. Walk and bike trips may be underreported.	Not readily inferred.
Route assignment	Not usually captured in surveys.	Depends on trace data and algorithm.
Tour generation	Self-reported in detail in a survey. Analysis by using heuristics and rules.	Data products do not include chains. Only aggregate trips are sold.

Source: Cambridge Systematics, Inc.

- Contextual information available in each data source to
  - Differentiate between commercial and passenger travel,
  - Identify activities and travel purposes, and
  - Use socioeconomic information to expand the sample.

The data elements shown in Tables 9-1 and 9-2 can serve as a data checklist to help agency staff identify the specific features of CDR data, their ability to provide the required information, and their relative value compared with traditional surveys, GPS logger data, smartphone surveys, and Bluetooth data.

The remainder of this section focuses on and briefly discusses each of the 11 data properties in Table 9-2. Key findings are summarized and specific recommendations are made for each individual data entry.

Table 9-2. Properties of different locational-based data sources.

Data Property	CDR Data	Personal GPS-Derived Data	Smartphone Survey	Custom Bluetooth Data
CDR data in raw form	Raw data likely not available due to privacy concerns.		Raw data are available to data analysts.	
Processed CDR data available to analyst	Processing method is not known to analyst.		Method can be shared with the analyst.	Limited data processing is possible.
Zonal size and spatial resolution	Low spatial accuracy. Zone size and number of zones affect pricing.	Spatial accuracy greater than CDR data.	Spatial accuracy similar to personal GPS data.	Data can be used to support corridor traffic analysis.
External zones and external stations	External travel may be obtained.		Depends on survey methodology and participant travel.	Yes, but depends on survey locations.
Trip purpose	Activities and purposes are inferred. Three purposes are available: HBW, HBO, and NHB.		Detailed trip purposes through prompted recall.	Not possible.
Socioeconomics	Not available.		Available.	Not available.
Technology	Advances in technology will yield more accurate data. More frequent data points. Greater spatial accuracy.	Standardized technology. Potential to improve pulse rates versus battery life.		Standardized technology.
Time periods and temporal resolution	Depends on cell utilization and interaction with network.	Depends on level of interaction with network.	Very detailed resolution.	Possible to summarize data by time of day.
Commercial and passenger travel	Not possible to differentiate between vehicle classes.	Able to differentiate between vehicle classes.		Not possible to differentiate.
Expansion of sample	Expansion is driven by population and geography. No socioeconomic or market segment data. Vendor-driven methods are used.		Customized expansion by socioeconomics and geographic detail.	Expansion can be made to vehicle counts.
Path traces	Unreliable path traces. Infrequent transactions. Low spatial accuracy.	Unreliable traces for slow data transaction rate.	Very reliable path traces.	Not possible.

Source: Cambridge Systematics, Inc.

9.3.1 Raw CDR Data

This report benefited from access to the raw, disaggregate, and anonymized 2010 CDR data that were used for research purposes. However, current thinking and legal privacy considerations make it unlikely that raw cell phone CDR data will be available to practitioners in the future.

This is a key consideration for practitioners who are accustomed to traditional analysis methods that allow testing and experimentation with household travel survey data. This natural part of

Raw CDR data will not be available because of privacy considerations.

the model estimation and discovery process is anchored in the behavioral paradigm approach but is not feasible with CDR data.

Agency staff will be able to specify their customized data requirements to CDR data vendors. However, access to processed CDR estimates instead of raw CDR data most likely will not allow them to

- Test for themselves the sensitivity of different assumptions about the sampling of cell phone devices,
- Use different criteria to select a CDR sample for estimation,
- Weigh and expand the CDR sample to different control totals or account for the presence of different service providers in the marketplace;
- Infer stops and activities from CDR traces; or
- Construct CDR trip tables by purpose and by time of day in response to different assumptions.

Agency staff can evaluate the processed CDR data indirectly by discussing with the data vendor the properties of CDR data for these specific assumptions. Table 9.2 can be used as a guide to clarify the processed and aggregated CDR data and to help address the strengths and weaknesses of CDR data.

9.3.2 Processed CDR Data

Innovative ideas are more likely to be embraced and adopted in practice after a high level of collaboration and vetting of ideas between academia, the industry, and planning agencies. Academic research to harness data from disruptive cell phone technology can be tested and verified by the industry and can then be more easily adopted by agencies as a proven method and product.

Although business realities may prevent the sharing of proprietary methods by vendors, greater transparency of the black box will increase the value of CDR data to practitioners and planning agencies. Recognition of the strengths and weaknesses of CDR data, the analysis methods used, and underlying assumptions will increase the industry's confidence in cell phone data products.

At the time of development of this report, only one data vendor was using CDR data to infer stops, activities, and a limited number of travel purposes before expanding the processed CDR sample to generate O-D trip tables. As discussed in Chapters 4 to 8, the vendor's processed and aggregated 2015 trip tables are broadly comparable to the measures derived from the research team's analysis of the raw 2010 CDR data. This suggests that CDR data vendors are most likely using methods and assumptions that are consistent with the Boston case study research.

Vendors provide aggregate trip tables from processed CDR data.

A key drawback for practitioners is that they need to rely on processed CDR data without the benefit of having access to the underlying raw data or the methods used to analyze them. As a result, practitioners implicitly have to accept the methods used for processing, expanding, and interpreting the raw data. Given that vendors may not provide enough methodological details because of proprietary considerations, practitioners need to ask specific questions about the CDR product to better understand its properties.

Agency staff routinely ask data-related questions and challenge assumptions and methods used in traditional trip-based or advanced activity-based models. Agency staff and practitioners can better understand elements of the CDR black box by asking questions about

- The spatial and temporal accuracy of the CDR data to ensure that the data resolution is appropriate for the agency's project needs;

- The population of cell phone users in the CDR sample and its representativeness of the population at large;
- The incidence of different market segments in the sample of cell phone users and the method used to expand the sample; and
- The methods used to detect home and work locations, stops, and activities that are then used to infer daily travel by purpose and time of day.

Agency staff will develop more confidence in a vendor's analyses and products if they are convinced that the processed CDR data generate results that are broadly consistent and comparable to the outcomes of traditional data sources and modeling methods. These questions can be addressed directly and definitively if

- The underlying raw CDR data are available for analysis by the agency,
- The assumptions and methods used by a CDR vendor are disclosed to a greater extent than is available today, or
- Practitioners, academics, and vendors collaborate to analyze and test different assumptions by using raw CDR data to gain greater confidence in the final product and the methods used.

If such options are not realistic, agency staff and practitioners need to use indirect ways to assess the quality of the O-D trip tables by engaging the vendor in a discussion about the underlying specific assumptions and methods that drive the final product. The entries in Table 9-2 can serve as a data checklist with the types of questions typically asked during traditional data design in advance of model development, validation, and application.

9.3.3 Zonal Size and Spatial Resolution

CDR data offer a wealth of spatial and temporal information, but the uncertainty about stay locations and activities has practical implications for the accuracy of travel data. Location inferences are made by triangulating between cell towers, which results in different degrees of spatial accuracy. In the case of the 2010 raw CDR data, the accuracy was as low as 300 meters.

Practitioners are familiar with traditional surveys in which locations are provided by respondents but reporting errors may include missing an activity or not providing an accurate location for an activity. Data collection with GPS loggers or smartphones, aided by prompted recall and verification, provides full and direct reporting of locations compared with CDR data, for which locations need to be inferred.

CDR spatial resolution may effectively preclude analysis at the traffic analysis zone (TAZ) level and require aggregation of existing TAZs. CDR-based trip tables at a more aggregate geographic level are likely to provide results closer to those of traditional surveys and models than are trip tables at the TAZ level. This is not an unexpected finding, given that it is also true of models developed by using travel surveys. Modeled trips at an O-D level are often aggregated at a district level to compare them with other data sources.

Locations in CDR data are inferred and may not be accurate.

Agency staff need to decide on geographic coverage and the desired geographic detail, given the spatial accuracy of CDR data. Although CDR data may be purchased at a TAZ level, the data will need to be aggregated for most practical applications. An example of district aggregation is the traffic analysis district (TAD) which was developed after the 2010 Census in support of the Census Transportation Planning Products (CTPP). TADs are aggregates of select TAZs. When TAZs are delineated for a given area, TAD boundaries follow the outermost boundaries of the TAZs they are intended to encompass, are contiguous, and do not extend into other areas.

Agencies can specify to the vendor the desired level of aggregation on the basis of CDR spatial properties, activity centers in the region, and the existence of other sources of data at comparable levels of aggregation. Given that vendors work with point-level data, they can provide customized aggregations of zonal data based on a model's TAZs, Census geographic boundaries, or an agency's custom-built layers. When zonal data are being aggregated, it is recommended that a nesting structure within commonly used geographic layers be preserved for comparability with other data sources. Finally, the cost of the CDR data may depend not only on coverage but also on the number of zones and level of geographic detail.

CDR trip tables are more robust at a more aggregate geography.

On balance, the research team believes that district-level zones similar to the TAD system are preferred to traditional urban area TAZs. The spatial accuracy of up to 300 meters that is present in the 2010 research CDR data prevented accurate analysis at a typical urban TAZ level. The use of the TAD system provides other data sources with which the CDR data can be compared.

Agency staff evaluating cell phone–derived data need to weigh the effect of less-accurate activity location data in relation to benefits such as lower costs and larger sample sizes. Questions to discuss with the vendors include

- The cost of the CDR data purchase as a function of the model coverage, the size of individual zones, and the use of the CDR data for specific analyses;
- The spatial accuracy of the CDR data to determine whether the model's TAZ system or a more aggregate level needs to be used; and
- The definition of a robust district-level zone system for CDR data that is consistent with the geography of Census data and other local databases.

9.3.4 External Zones and Stations

CDR-derived O-D data are processed at the zonal level for the region under study. Zones available as trip ends in the CDR data will typically not include external stations that are not represented as a model zone. Some minor additional processing will be needed to designate zones that correspond to current external stations.

Capturing travel by nonresidents can be of high value.

CDR data are available for a broader geographical area that includes zones both inside and outside of the model region. Agencies that need estimates for internal–external traffic and external–external traffic that passes through their region can define and purchase data for zones outside the regional model boundary.

For agencies interested in traffic originating or destined outside their region from both a modeling and economic policy perspective,

- CDR data provide a valuable tool for understanding total travel made by visitors whose inferred home address is outside the study region;
- CDR nonresident data can augment current methods of collecting visitor data and may be packaged as part of the regional CDR data request; and
- Both passenger and commercial vehicles will be captured in these estimates, but distinguishing the two segments will not be possible.

9.3.5 Trip Purpose

A key difference between traditional and CDR data and methods is the ability to infer activities and trip purposes. In traditional surveys, a sample of individuals and household members record their daily activities at a great level of detail. Survey data are then analyzed to infer travel at a similar level of activity and purpose detail for the entire population in a region.

In contrast, CDR data are limited in the detail they offer for activities and purposes. As described in Chapters 5 and 6, the analysis of CDR data relies on heuristic rules and algorithms to infer home and work locations. The most frequently observed location for a cell device during the nighttime hours is assumed to be the home end, while the most frequently inferred CDR trip end during the daytime is assumed to be the place of work.

However, CDR data are much weaker when it comes to nonwork travel, given that they cannot distinguish between different types of nonwork activities. CDR-derived trip purposes are defined by the land use activities at each trip end. Three trip purposes include home-based work (HBW), home-based other (HBO), and non-home-based (NHB) trips. Analysts need to accept that CDR data do not include detailed travel purposes and that HBO and NHB trips cannot be expanded to a wider range of travel purposes.[2]

CDR trip purposes are limited and they are inferred.

Agency staff need to decide the following when considering a CDR data purchase:

- Is a data set with three trip purposes adequate for the agency's planning needs?
  - Traditional models offer a much more detailed set of trip purposes.
  - Activity-based models further link trips together into tours to better represent an individual's and a household's daily travel.
- Are the relative magnitudes of CDR travel by purpose reasonable?
  - The home–home database must be small (almost zero).
  - The home–work and work–home matrices must be roughly comparable to the journey-to-work data.
  - The percentage of work and nonwork travel should be roughly comparable to that of past regional surveys and models.

9.3.6 Socioeconomic Data

A key strength of CDR data is the large sample of locational data points in comparison to the small sample of traditional diary surveys that offer more depth and detail. A key weakness of CDR data is the lack of socioeconomic information that would provide the context for the travel patterns observed for the large sample of a region's residents. By definition, this weakness limits an agency's understanding of differences in the travel behavior across market segments in the region.

Agency staff need to accept this key weakness of CDR data and its corresponding effects on sample expansion, market segmentation, household travel patterns, and overall resolution of the underlying data. These effects, which are discussed in other sections, include the following:

Lack of socioeconomic data limits the value of CDR travel patterns.

- The unit of the analysis is the cell phone device instead of the individual.
- Although two or more devices belong to the same household, their trips are not connected, and household interactions are not taken into account.
- The sample expansion methods for CDR data are simpler and are based on the population of cell phone subscribers at the home location.
- The models and travel patterns inferred from CDR data cannot reflect differences in travel behavior by market segment.

9.3.7 Technology of CDR Data

As described in Chapters 3 through 5, CDR data currently include transmissions that correspond to telephone calls made or received, incoming and outgoing text messages, and access

[2] Academic research that is under way aims to address this weakness, which is more difficult to overcome in urban areas with many mixed land use parcels.

to the web. Smartphone devices now include active and passive data transmissions, such as podcast or music downloads, e-mail updates, data on maps and directions, and the use of various apps.

Inference of trip ends depends on cell phone technologies captured.

The 2010 raw CDR research data set used in this analysis and the 2015 aggregate cell phone data set purchased from a vendor include time stamp and location for every instance of phone use in the service network. This includes information about location every time a phone call is made or received, a text message is sent or received, or data are accessed on the device. Agency staff should confirm that the CDR data used by a vendor include at least the level of active and passive transmissions reflecting calls, texts, and Internet data access as was used in the 2010 CDR case study. The inclusion in a CDR sample of smartphones that use the 4G-LTE spectrum will increase the frequency of device sightings and will provide analysts and machine learning algorithms with more data with which to make inferences about travel.

Capturing more passive signals improves the data for travel analysis.

The quality and representativeness of CDR data also depend on the cell phone service providers from whom the CDR data are obtained, their market share in a region, and the data plans that they offer, such as unlimited data transmission. A CDR data set that uses data from multiple vendors, records all types of transmissions, and uses the latest technology will yield a richer set of signals that are transmitted, recorded, and processed.

Agency staff need to be comfortable with the coverage of the cell phone data, the sample of CDR data in the region, the quality of the CDR data used, and the representativeness of the cell phone sample for travel by the region's residents. The relevant questions to discuss with vendors to address these objectives include the following:

- Who are the cell phone service providers in each region?
- What is the market share of each provider?
- Which cell phone service providers are included in the vendor's sample?
- Are there distinct market segments and parts of the region where specific cell phone service providers have a greater presence?
- Are there important differences in the socioeconomic and usage profile of the markets served by each cell phone service provider?
- Are calls, texts, and Internet data access recorded as part of the captured transmissions?
- Do these transmissions account for both active and passive signals?
- What cell phone technologies are captured in the vendor's CDR data?
- Are the more frequent 4G-LTE technology signals used by new smartphones captured in the vendor data?

From a practitioner's perspective, the ideal CDR data product would be based on a sample from two or more cell phone service providers that accounts for the majority of cell phone users and is representative of the region's population. The CDR data should account for active and passive transmissions of calls, text messages, and Internet access and should reflect the prevailing technology used on the cell phones of most subscribers.

A product that meets these conditions would benefit from the higher quality and greater quantity of cell phone signals. A sample of cell phone users that is a representative sample of the population would result in a representative and rich sample of locational data that would allow the analyst to draw better travel inferences.

9.3.8 Time Periods and Temporal Resolution

The 2010 CDR data processed and presented in this case study allowed the inference of a large number of trip ends at a high level of temporal resolution. These CDR data can be grouped in

different ways without any major limitation on the number of time periods or the duration of each time period.

The location and timing of the inferred CDR trip ends are processed, aggregated, and expanded by the vendor to preserve the confidentiality of individual subscribers. Although existing products may use a default time period that is not suitable to an agency's needs, there is no practical limitation in grouping the data using different time period definitions.

The definition of the number and duration of time periods is flexible.

Agency staff can benefit in the following ways from the scale and detail of the time-of-day information that is provided by CDR data:

- Flexibility in defining the time periods best suited to an agency's purposes allows time-of-day summaries at the desired level of temporal resolution.
- Although mode-specific information is not available, time-of-day profiles reflect fluctuations in the observed total demand for regional travel.
- The large sample size of CDR data allows for detailed time-of-day comparisons, as follows:
  - Peak and shoulder peak period traffic during a specific weekday;
  - Late night and early morning traffic patterns not captured accurately by traditional models;
  - Peaking patterns by day of the week and during the weekend; and
  - Changes in peaking patterns observed by time of the year, as a result of weather or other special events, and in response to recurring or incident congestion.

9.3.9 Commercial and Passenger Travel

In recent years, there has been increased emphasis on commercial travel and freight flows in urban areas and regional corridors and at the state level. Sources of commercial traffic data used in freight and truck analyses include commodity flow surveys, networks and data from the freight analysis framework, and GPS trace data from the American Transportation Research Institute.

In the case of CDR data, there is no identifying information to classify a cell phone device as being used for personal or commercial use. It is therefore not possible to classify the inferred trip end activities as serving a passenger or commercial purpose. As a result, CDR data provided as part of a vendor data set account for total cell use and reflect total travel in a region.

Trip ends are inferred from all sampled devices.

Agency staff need to focus on alternative data sources to assess freight and truck traffic in a region. These sources of data include

- Customized traditional surveys of truck drivers, establishments, rail and truck companies, and freight forwarders;
- GPS data extracted from devices installed on trucks to track their movements; and
- Smartphone surveys that target truck drivers and are used to measure commercial trip ends and O-D truck flows.

9.3.10 Expansion of the CDR Sample

Traditional sample weighting and expansion techniques use detailed approaches by market segment that can account for household size, vehicle ownership, number of workers in a household, and geography. The rationale is that residents' inherent propensity to travel and their travel patterns differ across these market segments. Therefore, it is important to account for these differences by developing distinct weights during sample expansion.

The expansion of CDR data is more simplistic by definition. Given that CDR data do not include socioeconomic data or contextual information, the analyst cannot differentiate devices

CDR expansion relies on the population of cell service subscribers at the home location.

across market segments. Therefore, the population of cell phone subscribers within the zone identified as one's home location is used as a simpler measure of expected CDR use and travel activity.

Practitioners recognize that this approach does not explicitly account for differences in cell phone use and travel by each market segment in the CDR sample. Population-based weights are also not satisfactory for trips such as long-distance travel where the home and/or the work location are outside the model's coverage.

Agency staff should

- Inquire about the details of the weighting method used to expand the CDR data to verify that a simpler, population-based expansion method is acceptable for their intended use of the CDR sample;
- Ask which cell service provider data are included in the sample and what is the market share of the service provider in the region; and
- Discuss whether subscribers of this service provider represent a random sample of the population or whether this provider serves market segments with distinct socioeconomic characteristics and cell phone usage profiles.

9.3.11 Travel Times and Path Traces

Travel time along paths can be determined by CDR data typically discarded as part of the processing of O-D data. Travel times and detailed actual path traces along each O-D pair may be available for specific cell phone devices. However, these detailed data are not typically included in CDR data, given that they are masked during the aggregation and expansion of the origin and destination trip end data.

Agency staff who are interested in detailed O-D trip times and path travel times can discuss the option of obtaining such data as part of the data specification:

- Travel time estimates and O-D path traces are not reliable for shorter distances, where location errors affect their accuracy; and
- GPS data provide an alternative source of data that is more likely to provide the desired level of travel time and speed detail by using a sample of vehicles in the network.

9.4 Modeling Considerations

Regional transportation models vary in size, scope, and complexity across state DOTs and regional MPOs, which have different experiences and track records with the design, collection, and analysis of travel data for planning and modeling purposes. In this context, CDR data have been evaluated by agencies as a potential source of data that could support different aspects of

- Model estimation,
- Model validation,
- Model updates for intermediate years between releases of Census data or between years with a major regional survey data collection,
- Corridor studies and microsimulation analyses,
- Special generator studies,
- Assessment of visitor markets, and
- Estimates of long-distance travel markets.

Table 9-3 outlines each of these modeling options and discusses how each option can benefit from CDR data. The table also compares the value of CDR data with similar nontraditional data

Table 9-3. Modeling applications supported by CDR, GPS, smartphone survey, and Bluetooth data.

Type of Model	CDR Data	Personal GPS Derived Data	Smartphone Survey	Custom Bluetooth Data
Estimation of regional models	No socioeconomic data. No detailed activity, purpose, mode and tour data. Spatial resolution can vary.		All data needed to develop detailed regional travel demand models are captured.	na
Validation of regional models	Aggregate validation for trip generation, trip distribution.	Aggregate validation for trip generation, trip distribution, and possibly highway assignment.	Detailed validation for all aspects of regional travel demand models.	Validation for small corridors or locations with traffic counts.
Model updates	Documentation of changes in travel patterns. Measurement of changes in total travel flows. Identification of changes in travel flows by time of day.		High costs for frequent large-scale data collection. Refresh is feasible with small sample & key travel markets.	Estimates of changes in corridor-level traffic at different points in time can be used.
Corridor and traffic impact studies	Data at the corridor level. Spatial resolution may not be sufficient.	Spatial resolution possible. Important to capture trip start and end locations.	High cost of survey data for corridor-level studies.	Estimates of traffic counts by time of day.
Microsimulation studies	Precise temporal resolution is an additional concern.	GPS data are better suited than CDR data.	High cost of survey data for microsimulation studies.	Data can be used to support traffic analysis.
Special generator studies	Applicable to special generators and special events. Trip generation and trip-length estimates. Airports, universities, malls, and sports arenas.		Data on socioeconomics, mode used, and trip purpose. High single-purpose survey cost.	Traffic counts at special generators can be made.
Visitor models	Long-term study of aggregate movements of visitors to a region and within a region.	Study of aggregate movement of visitors to and within a region.	Difficult to target visitors.	na
Long-distance models	CDR and GPS offer suitable data sources. Ability to monitor visitors and residents.		Long-distance travel underrepresented in typical surveys. Long observation period is needed.	na

Source: Cambridge Systematics, Inc.

Note: na = not applicable.

that can be obtained from GPS loggers, smartphone apps, and Bluetooth devices. The detailed discussion of modeling options can help agency staff prioritize their plans to use CDR and other nontraditional data to support, augment, or replace one or more model components. The modeling components in Table 9-3 are discussed in separate sections, each of which summarizes key observations and provides practitioners with specific recommendations for individual modeling options.

9.4.1 Estimation of Regional Models

Traditional and activity-based regional models have been estimated and validated with detailed travel data from household travel surveys that also include person-level socioeconomic data. Traditional disaggregate models often account for travel patterns by market segment, while activity-based models provide more detail in accounting for tours and reflecting intrahousehold travel interactions.

As discussed in Chapters 4 through 8, outputs of regional models, summaries from household travel surveys, and inferred O-D flows from CDR data are broadly comparable, especially at an aggregate level. However, CDR data require different analysis approaches. Although they benefit from a much larger sample size and accurate temporal resolution, CDR data are used to infer up to three trip purposes and have a lower spatial resolution for shorter trips. CDR data also do not take into account the effect of socioeconomic characteristics on daily travel, given that they are anonymized because of privacy considerations. As a result, CDR-based models provide considerably less context and depth than traditional and activity-based models.

CDR data cannot be used to estimate models at the traditional level of detail and resolution.

Agency staff need to recognize that CDR data cannot be used to estimate models at the same level of detail and resolution as traditional or activity-based models. The weaknesses of CDR data for model estimation can be summarized as follows:

- Only work and nonwork travel purposes are inferred, in comparison with the more detailed purposes obtained in traditional and activity-based models.
- Locations of activities are inferred by using heuristic rules and are subject to error, especially for shorter trips and for activities of short duration.
- Socioeconomic data at the individual level are not available and cannot be used to estimate models that differentiate between market segments.
- Estimates of regional travel obtained from CDR data are provided at the trip level instead of the tour or activity level.
- Passenger and freight-related travel estimates cannot be distinguished, given that the unit of analysis is the cell phone device.

Exploratory research is being conducted to generate synthetic populations using marketing data sets and to assign cell phone O-D data to these populations. In addition, academic research that is now under way focuses on inferring more travel purposes by using a new set of heuristic rules to infer purposes in more detail.

9.4.2 Validation of Regional Models

Practitioners are always interested in independent sources of data that can be used to validate the outputs of travel demand models. Models are typically estimated and calibrated with detailed travel survey data. These estimates are validated by comparison with independent sources of travel flows such as the CTPP journey-to-work data, American Community Survey data and other Census estimates, and link-level measures such as highway traffic counts, transit ridership estimates, and O-D travel times.

CDR-derived travel and trip table estimates provide an alternative source of data for validating individual model components of trip generation and distribution as well as estimates of travel by time of day. However, these CDR-derived estimates are not available by market segment such as income, auto availability, or household size. CDR validation data are also only available at the trip level and do not provide the ability to connect trips into tours. CDR data also do not provide information on modes used.

CDR data offer an additional source for validation of model components.

Agency staff can evaluate the degree to which processed CDR-derived estimates and O-D trip tables can provide a valuable source of validation data that can be used independently to

- Compare total trips produced and attracted, including both passenger and freight, without the ability to further differentiate by mode;
- Provide estimates of O-D flows for home-based work, home-based other, and non-home-based travel purposes combined;
- Generate trip-length distributions for the three CDR inferred purposes that could be used to calibrate trip distribution models;
- Capture a greater level of temporal detail and generate O-D trip tables to validate models by time of day; and
- Quantify external–internal travel and through travel to support model validation in addition to the typical internal travel in a region.

9.4.3 Model Updates

Regional planning agencies do not update travel demand models often, primarily for reasons of data availability and cost considerations. Practical constraints include the need for updated socioeconomic and Census data, revised highway and transit networks, and up-to-date traffic counts and transit ridership data.

Agencies that update or revalidate their models for an intermediate year rely on their existing model structure and use updated data to fine tune their model for new base year conditions. The motivation to update regional models is more easily justified in cases of significant socioeconomic changes, the introduction of a new mode, major changes to highway or transit services, and new technologies that bring about major changes in travel behavior.

CDR data can provide information on changes in travel patterns in intermediate years.

CDR data can provide agency staff with updated O-D trip matrices to support more frequent model updates that do not require a major data collection effort. The framework of using CDR data for updating a model for an intermediate year would require the following steps:

- CDR data can be purchased for the original base model year and compared with model outputs to ensure that the two data sets provide broadly comparable travel behavior metrics. Comparisons may include:
  - Trip rates by geography,
  - Temporal distribution of trips,
  - Trip-length distributions, and
  - Share of total trips by purpose.
- CDR data can be purchased for an intermediate year, assuming that the same method and underlying source of data are used. This intermediate year data set can be compared with the original base-year CDR data to quantify the magnitude and reasonableness of the observed change in travel flows.
- The percentage of growth or decrease measured by the two CDR data sets can be applied to the results of the original base-year model to generate travel estimates and metrics for the intermediate year.

- In addition to the CDR data, intermediate year traffic counts, transit ridership data, and socioeconomic data can be used to paint a complete picture of the updated base year travel patterns.

9.4.4 Corridor, Traffic Impact, and Microsimulation Studies

CDR data can provide input data for broadly defined corridors or parts of an urban area.

CDR data can support corridor-level studies in cases of large, long, and well-defined corridors or parts of an urban area to allow meaningful estimates of travel within, to, and from the corridor or area boundaries. The format of CDR data limits their suitability for a detailed traffic assessment or microsimulation studies, which require a greater level of detail.

CDR data can be used to generate synthetic O-D tables that can be further adjusted by using matrix adjustment methods to match base-year traffic counts. This approach may be a preferred alternative to generating subarea models. Analysts can then assign these synthetic O-D tables to the network to support corridor studies.

Agency staff are aware of the following issues that affect corridor-level studies:

- CDR data do not distinguish between passenger and commercial vehicle travel;
- Lack of socioeconomic data does not allow analysts to assign economic benefits to different market segments; and
- The spatial inaccuracies of CDR data may affect the validity of the analysis in cases in which the corridor is narrowly defined or in which there are multiple competing facilities in proximity.

With regard to supporting traffic microsimulation studies and models, the requirements for detailed geographic and time period data are even greater, and path trace information becomes even more important. In such cases, the use of detailed local traffic counts, GPS data, and specifically focused travel surveys provide the desired level of resolution, which is higher than that provided by CDR data.

9.4.5 Special Generator and Special Event Studies

Special generators cause a lot of passenger and commercial activity that is not always captured well by regional models. Typical special generators include airports, large malls, parks, ballparks or stadiums that have recurring sports events, ports with commercial activity, and universities. Typically, special surveys are conducted to capture activity at these locations. Models based on these surveys are relatively straightforward and capture travel by using broad aggregate measures.

Special generator and special event studies can benefit from CDR data.

Such an aggregate modeling framework is suitable for CDR data. Data purchases can be structured to include only those trips for which the special generator is either an origin or a destination. Metrics such as trip rates, trip-length distributions, and temporal distributions can be obtained using CDR data. In addition, CDR data can be used to study other aspects of special generator travel, including

- Seasonal variation in travel;
- Before-and-after studies in cases in which the modes and the level of service to reach the special generator have changed;
- Visitor travel to the special generator, for the purpose of understanding the effect of such locations in attracting out-of-town visitors; and
- Contribution of a special generator to local travel, especially during peak hours.

Agency staff need to specify the CDR data purchase, especially with respect to zone definition, and provide input to sample expansion:

- The zone system needs to be outlined carefully. When specifying the special generator, agency staff need to account for the spatial inaccuracy of the CDR data and choose an appropriate buffer around the generator.
- Agency staff must discuss with the vendor the weighting method used. The preferred way would be to develop weights for the entire database and then carve out the portion of the trip table for the special generator study. Agencies can also use independent counts to scale the CDR data further.

Special event facilities generate activity over a few days each year and include concert halls, conference centers, and stadium facilities. CDR data can be suitable for special events too, but three points of note must be considered by agency staff:

- Special events often attract nontypical crowds, and O-D patterns at the same location may vary across events. A sufficiently large sample within a broad time horizon may be needed to better capture average special event activity.
- The temporal dimension of travel is determined by the time of year when the special event is being held. Again, obtaining CDR data for a broad time horizon will help mitigate any skewed travel patterns.
- Finally, some special events may attract visitors from long distances. Establishing detailed external locations is vital in capturing special event travel patterns.

9.4.6 Visitor Models

Visitor travel patterns can be assessed with CDR data.

The ability to distinguish between residents and visitors is a strength of CDR data. Understanding and quantifying travel by visitors whose home address is outside the study region is a valuable tool for urban areas from the perspective of modeling, planning, and economic policy. Agency staff can use CDR data to augment their current methods of collecting visitor travel data, especially if they are not interested in classifying or otherwise segmenting the visitor travel market.

Capturing visitor data has traditionally been challenging within the context of travel demand models. Agencies typically intercept visitors and conduct specialized surveys at hotels, airports, rail stations, and other ports of entry. These surveys face numerous challenges:

- Visitor surveys are typically costly and are often limited in scale, given that regions have multiple ports of entry and visitors have many places to stay—a pattern exacerbated by new services like Airbnb.
- Visitor profiles and their travel patterns tend to vary by time of the year. The scheduling of conferences, festivals, and sporting events affects travel and may require conducting a survey that spans several months and results in an expensive and time-consuming effort.

Agency staff can use CDR data to study visitor travel patterns in a region by using the following framework:

- The analyst defines the area of interest for studying visitor movements.
- The data vendor identifies cell phone devices that travel within this region, but whose home location is outside the area.
- The data vendor provides trip tables for these visitors over a specified time frame that could span several months or a whole year.

9.4.7 Long-Distance Models

Long-distance travel is infrequent and is not adequately represented in most traditional surveys, which capture resident travel within a region over a 1- to 5-day observation period. As with

CDR data can support long-distance travel models.

visitor models, CDR data can provide valuable information for modeling long-distance travel by a region's residents.

Agencies typically model long-distance travel by using broad classification schemes to identify differences in total trip making, destinations, and modes used within their population. Given the lack of socioeconomic information in CDR data, only a more aggregate long-distance model is feasible without the benefit of additional market segmentation.

The considerations for agency staff are similar to those they have with visitor models:

- The area of interest for studying long-distance movements needs to be defined.
- Criteria such as "resident travels 50 miles beyond the study region" are defined to eliminate travel to the outer reaches of the study region that may happen on a more regular basis.
- The time frame during which long-distance travel needs to be examined is defined and could span several months or a whole year.
- The data vendor identifies residents' cell phone devices that travel outside the specified geographic region during the specified time frame.

9.5 Future Research Directions

The collection and analysis of CDR locational data represent dynamic areas of data and research that can benefit transportation planning and modeling. Technological advancements, the changing patterns of cell phone use by segments of the population, and strong academic research efforts to harvest the value of locational data will continue to change the properties and potential value of CDR data.

The interest in locational data is reflected in analytical methods and machine learning approaches aimed at using CDR locational data for transportation planning and various other purposes. The following have recently been seen: the emergence of new data collection methods that use smartphone devices, ongoing research to better infer travel purposes on the basis of land use data, and the fusion of and quilting together of different data sources.

9.5.1 Dynamic Nature of Cell Phone Data

The collection and analysis of CDR locational data represent dynamic areas of data and research that can benefit transportation planning and modeling. Technological advancements, changing patterns of cell phone use by segments of the population, and strong academic research efforts to harvest the value of locational data will continue to change the properties and value of CDR data.

This report focused on the properties and the analysis of cell phone CDR records from 2010 that were available to the research team and 2015 data provided by a vendor. The raw nature of the 2010 CDR data allowed the research team to analyze these data in detail and compare them with traditional transportation surveys and travel demand models for the Boston area. It is therefore important to put in context the 2010 CDR data compared with today's cell phone technology and changing use of cell phone devices.

The 2010 CDR data represent a period of less-intense use of cell phones for calls, text messaging, and Internet data access as compared with today. The increase in cell phone use over the past 7 years has resulted in an increase in the density of CDR signals and the amount of trace information. The availability and analysis of more and better-quality locational data during a typical day can produce more robust results about daily activities and travel.

Use of cell phones is increasing and yields more locational data.

The cost of cell phone service has also been decreasing over time, especially for text and data, with packages now often including unlimited data transmissions. Cell phones are becoming more of a necessity for many households, including those in lower-income market segments. Use of cell phones to access data has increased to the point that, for many users, their cell phone provides the principal means of accessing the Internet.

In November 2016, the Pew Research Center reported that 77% of American adults owned a smartphone of some kind, up from 35% in the spring of 2011 (Pew Research Center 2017). The market penetration of smartphones is likely to increase further and thus increase the size of CDR databases and data records as compared with the research case study discussed in this report. Furthermore, as both active and passive data transmissions from smartphones increase, the cell phone data records available also will increase accordingly. As the density of CDR signals and locational data increases for every device, the ability to determine trip end locations and time of activities could also improve.

GPS-enabled smartphones yield higher resolution data.

In addition, all smartphone devices include a GPS location service. When this feature is enabled, the location of the device can be more precisely determined, thus improving the quality of locational information and providing CDR vendors with an additional source for mining cell phone location data.

- A 2015 Pew Research Center report on cell phone use titled *The Smartphone Difference* notes that smartphones already help users navigate their environment with real-time directions provided by the inbuilt GPS system.
- As many as two-thirds of smartphone owners use their phone at least occasionally for turn-by-turn navigation while driving, with 31% saying that they do so "frequently."
- One in four respondents uses his or her phone at least occasionally to get public transit information, with 10% doing so "frequently."
- While Tables 9-2 and 9-3 make a distinction between data obtained from cell phones and data obtained from personal GPS devices, this distinction will become more blurred with GPS-enabled smartphones.
- As more GPS-enabled devices are used during traveling, traces from CDR data will continue to increase and will provide more data points that should help improve an analyst's ability to determine trip ends.

The technology used to collect cell phone data records has been changing in a manner consistent with changes in the technology serving cell phones. In 2010, wireless industry data transmission standards were 1G and 2G transmissions. The industry standard today is 4G/LTE service, while many providers are advertising faster, enhanced service. Although it is unlikely that raw CDRs using these technologies will become available, practitioners should ask whether a vendor's product with processed CDR data takes advantage of the latest technology. This becomes even more important if the new technology commands a large share of the cell phone marketplace.

Monitoring of changes in technology and marketplace is needed for a representative sample.

In addition to the technological methods used, the sources of CDR records available from vendors may be changing. The wireless and cell phone provider market is extremely competitive and is changing rapidly. Wireless providers who are used by a CDR vendor should be disclosed to ensure that the CDR data being processed are representative of the market for cell phones in the area for which the processed CDRs are being obtained.

The research presented in this report indicates the potential to develop credible and useful travel information on O-D patterns from voluminous cell phone data. Much of that information has to be inferred and is limited in nature; the limitations include three travel purposes linked to home, work, and "other" locations; the lack of socioeconomic information; and the lack of information on the travel modes used. However, technological changes suggest that a greater amount of CDR data with higher resolution will be available in the future. With continuing research in

inferring activities and trips by purpose, the increased volume and quality of CDR data make it easier to address data gaps.

9.5.2 Status of Relevant Research in Locational Data

The growing penetration of cell phones among different age cohorts, the increased use of cell phones for Internet data access, and the improvements in the technology embedded in smartphones provide opportunities to leverage locational data for use by transportation planners, modelers, and decision makers. The following sections briefly comment on the current status of some of the most relevant research activities in locational data and highlight their potential promise to improve the industry's best practices in data, planning, and modeling methods.

9.5.2.1 Quilting of Diverse Data Sources

The cell phone is a truly disruptive technology that has significantly changed the way people communicate, get information, and travel today. Cell phones are increasingly being used to hail cabs, build transit itineraries, plan travel by auto, and use bike-sharing modes. The question for planners and modelers is how to leverage these data to study accessibility, improve mobility, and provide a better quality of life.

The strengths and weaknesses of new and traditional data sources suggest the need for continuing research to integrate data from diverse data sources and transportation user apps with socioeconomic and land use data to provide a complete picture of travel. Such a quilting of data sources must be conducted with a data transfer protocol that is mindful of user privacy considerations.

9.5.2.2 Smartphone Apps for Data Collection

A recent trend in survey data collection is the use of cell phones as the means for collecting data at the individual and household levels. The owner of a cell phone device is recruited and agrees that his or her travel will be monitored over a given period along with that of other members of the household.

The methods used to validate the day's travel differ across the different companies that have invested in this technology. Prompted recall methods are used to validate a specific activity or to verify a day's travel patterns. Respondents are asked to edit the responses inferred or left blank by the software to provide a full day's worth of travel information.

Some of the methods and products rely on sophisticated machine learning algorithms to benefit from respondent entries and improve the data collected (Ghorpade et al. 2015).

9.5.2.3 Machine Learning Concepts

Recent research has used CDR data and machine learning algorithms to annotate user activities and fill gaps to reveal temporal activity profiles and transitioning between activities. This research focused on activity patterns; the socioeconomic, land use, and mode-specific information that is needed to provide the complete picture of travel behavior is still missing. Therefore, future research should consider evaluating and expanding these machine learning techniques to determine how these CDR data can provide a complete picture of travel behavior (Yin et al. 2016).

9.5.2.4 Consumer Data as Inputs to Models

An example of ongoing research that aims to bridge the gap between traditional survey data and passive cell phone data is offered by a model developed for the Asheville region of North Carolina (Kressner et al. 2016). This research presents an approach to overcoming the limitations of passive location data. In addition to traditional National Household Travel Survey data and travel time data, this method uses consumer data that include much of the household and

individual socioeconomic information used in travel demand modeling. It builds a tour-based model with passive data by using a person-based discrete event simulation framework. A comparison of assignment results and average link error with the trip-based model for Asheville showed the results of this innovative approach to be promising.

9.5.2.5 Streamlined Collection of Survey Data

Localities, regions, and states currently use third-party tools to collect travel surveys. While these tools provide information that is relevant to planners and modelers, they sometimes require installation of an additional app that may require more than one touch by the user and thus potentially affect response rates.

An alternative option is the use of inbuilt Google Maps location data that are available to each smartphone user. While Google Maps is installed by default in Android phones, iPhone devices require that Google Maps be installed as an independent app. Given the market penetration of Google Maps, it is reasonable to expect that it would be part of an iPhone user's app environment.

Given this ubiquity and its editing features, this tool can supplement travel surveys. However, research is needed on issues of sample design, response rates, cell phone operation system biases, and the effect on users who do not use smartphones. The promise of this technology is that it reduces costs for data collection, given that it already uses an existing app and, more importantly, provides information on long-distance trips. The challenge is addressing biases resulting from lower smartphone penetration among the elderly and poor.

9.5.2.6 Machine Learning Versus Econometric Modeling

One session at the 2017 annual meeting of the Transportation Research Board compared the benefits of traditional econometric methods with machine learning approaches. This session, titled "Machine Learning Is from Venus, Econometric Modeling Is from Mars: Two Different Travel Forecasting Perspectives," highlighted many of the themes discussed in the literature and the practitioner considerations mentioned in this report. The discussion was moderated by David Ory, and the panel members included Josephine Kressner, Joel Freedman, Alexei Pozdnoukhov, and Eric Miller.

Alexei Pozdnoukhov argued in favor of machine learning techniques. He reviewed a paper by Eric Miller presenting a tour-based mode choice model and critiqued its econometric approach. In his summary Pozdnoukhov pointed out that the model is applied and validated to the same data set; it does not apply outside the data sample; and it is probably more complicated than needed.

Eric Miller argued in favor of econometric approaches. He reviewed a paper by Alexei Pozdnoukhov that describes a way to supplement traditional household survey data with cell phone data. The approach uses input–output hidden Markov models to infer travelers' activity patterns from their CDR records. In his summary, Miller pointed out that the model is also based on random utility; it does not use socioeconomic attributes; and it does not take into account household interactions. On the basis of the validation of activity durations by purpose and across space and of the assignment of trips to the network, the overall model performance seems very good.

A key point during the panel discussion was that although socioeconomic characteristics may not be essential to match counts or predict future traffic, they are critical in understanding traveler behavior and addressing policy questions on subject areas such as toll roads, transit travel, and managed lanes.

9.5.2.7 The Promises of Big Data and Small Data for Travel Behavior

In a 2016 review paper, Cynthia Chen and her colleagues discussed the potential for collaboration and sharing of cross-discipline ideas between transportation researchers who focus on

models of travel behavior and the computer scientists and physicists who use big data to address human mobility patterns (Chen et al. 2016).

Chen et al. pointed out the "tension" between the traditional behavioral approach that aims to formulate and represent causality in travel behavior and the analysis of passive CDR data that aims to identify mobility patterns. The discussion cautions about the potential risk of ecological fallacy in CDR data, a consideration that was also present in aggregate travel demand models. The risk of ecological fallacy is that although a model mechanism can predict average and total regional travel conditions, it does not necessarily perform well at the individual level without an approach rooted in the behavioral paradigm and a conceptual framework that explains travel behavior at the individual traveler level.

The conclusions and cautionary notes in Chen et al. reflect the line of thinking that is discussed in this report. On the one hand, the CDR locational data have tremendous potential to provide travel-related data and estimates at the regional, state, and national levels. On the other hand, the potential of CDR data can only be enhanced by adopting a travel behavior perspective and building on the collective experience of the field while opening up the black box to the community to reveal the assumptions underlying the data.

9.6 Epilogue

This guidebook for transportation practitioners concludes with a few thoughts about the properties of cell phone data and the types of behavioral modeling approaches that will help shape best practices in the future.

Practitioners evaluating their policy needs, data options, and modeling tools should be guided by the following principles:

- Be aware of the underlying assumptions made to process cell phone data to determine locations and to infer activities and purposes.
- Recognize that results from traditional surveys and models are also built on different sets of assumptions and that ground truth is tough to establish.
- Expect that increases in the quantity of CDR data, improvements in signal and CDR data quality, and the use of machine learning algorithms will improve methods for analyzing locational data and inferring travel patterns.
- Appreciate the uncertainty underlying both CDR estimates and the traditional data and measures of travel patterns.
- Use as a guide in your evaluation the conceptual framework that is based on the behavioral paradigm examining individuals' travel behavior.

The field's collective experience, academic research over the years, and the collaborative approach linking research to practice have helped refine the data design process, spawned new and more sophisticated methods, and increased the understanding of travel behavior and its drivers. As new data and methods are introduced, the interpretation of their value and uses through a behavioral framework lens will help improve the state of the art in this profession and community.

A transparent approach to the strengths and weaknesses of CDR and traditional data sources, along with a thoughtful approach that integrates new and old data and methods, will help refine the state of best practice. The approaches that will emerge may be different than those used today but will continue to have a behavioral foundation and will leverage the strengths of different data sources. This will help improve the field's ability to make inferences about travel and will result in analysis methods that best harness the value of new and traditional data sources to interpret, quantify, and forecast travel behavior and mobility patterns.

Glossary

The objective of this appendix is to document and explain in a stand-alone section some of the key new concepts related to the cell phone data and their analysis. The objective is to remove the jargon and use nontechnical terms to describe the essence of each term and its relevance to the study.

active and passive CDR data. Active use of a cell phone device includes making a call, sending a message, or visiting a website; passive use includes receiving a call, receiving a message, or being pinged by an app running in the background. Both active and passive signals trigger the capture of CDR data from a single cell phone.

advanced positioning. Advanced positioning techniques offer a finer spatial resolution than tower-based methods and include triangulation, transmission delay from multiple base stations, and other techniques of identifying the location of phones anywhere in a cell. Additional infrastructure such as monitoring devices along freeway segments and technology such as built-in GPS receivers in phones can provide accurate phone location.

call detail record (CDR). Call detail records are automatically collected by service providers for billing purposes. Each record contains time-stamped coordinates of anonymized customers when they use their phone in the cellular network either actively or passively.

CDR location data. The compiled data set of CDR records has locations for each cell phone device. Locations are inferred either by observing the cellular tower through which the phone is connected or by triangulation with nearby towers.

cell phone tower. Network operators often record the location of mobile phones in terms of the cell tower to which the phones are currently connected, in part because of privacy considerations. In this case, users' traces are represented by time-ordered sequences of cell tower IDs. If the geographic locations of cell towers are known, the latitude and longitude coordinates of the tower are used.

clustering method. A second method used to aggregate spatial points. The clustering method recognizes the uncertainty in location estimates. Although a device remains at the same location, it may be assigned to multiple neighboring location estimates. Clustering-based approaches allow points to be aggregated by using zones with arbitrary shapes. These methods use distance or travel time thresholds as inputs.

device oscillation. This is the phenomenon in which, although a cell phone is stationary, a radio signal from multiple nearby towers may reach the device. In such a case, a stationary device may appear to move between multiple towers, which creates the appearance of movement.

134 Cell Phone Location Data for Travel Behavior Analysis

grid method. One method of aggregating spatial points is by imposing a grid over the space and aggregating points that fall within each grid cell. This method depends on the layout of the grid, including the size and shape of the grid cell.

grid-based stay extraction. This process identifies stay regions by using a grid-based clustering method to cluster stay points. A study area is divided into rectangular cells; all stay points are mapped to the appropriate cell; and an iterative method is used to merge each unlabeled cell with the maximum stay points and its unlabeled neighbors to create a stay region.

inter-event time. The inter-event time is the time gap between successive uses of a cell phone for calls, text messages, and Internet data access. Frequent cell phone use results in shorter inter-event times and thus provides more location data points and a richer data set on travel patterns. Infrequent usage of a cell phone provides a more limited set of travel information with fewer available location data.

location of cell phone device. The location of a device is recorded when a phone call, text message, or data request is registered by carriers for billing, network performance, and legal purposes.

medoid. A medoid represents a cluster and is similar in concept to means or centroids. A medoid is a member of the data set and it minimizes the distance between points that belong in a cluster and a point designated as the center of that cluster. The term is used in data clustering algorithms in computer science. A set of medoids is first chosen at random and the distances to the other points are computed. Data are clustered according to the medoid to which they are most similar and the medoid set is optimized in an iterative process.

point-based algorithm. This process identifies stay regions by using a point-based method that exploits the maximum spatial accuracy possible. This method is similar to those originally designed for processing GPS traces. It is tailored to process cell phone data, which have lower locational accuracy and gaps in space and time, to extract individuals' whereabouts.

positioning. The position of a device is recorded by network operators when a user communicates with the network. Positioning data describe users' locations only when an event occurs. Every time a user initiates a network connection event (voice call, text message, or data access), the cellular network operator needs to know the user's location to determine the cell tower to channel this event. Cellular network operators do not maintain positions of users at all times in order to improve network performance, save bandwidth, and protect users' privacy.

spatial resolution. Unlike the accuracy of continuous GPS traces, the spatial resolution of cell phone data depends on the location, number, and density of cell towers; the land use and activity density; and the size of the zones used in the analysis.

spatial resolution–tower-based CDR data. The spatial resolution of these data is determined by the density of cell towers, which varies from a few hundred meters in metropolitan areas to a few kilometers in rural regions. If a geographic area or zone that includes one or more cell towers is used, phone activity routed through a tower will result in a record with the zone location. The spatial resolution of location records greatly depends on the size of these zones.

stay point. A stay or stay point corresponds to an activity location where the device and the individual user are engaged in an activity.

stay region. Different stay points identified in a user's several trajectories may refer to the same location, but the coordinates of these stay points are unlikely to be exactly the same. A grid-based clustering method is used to cluster stay points to obtain stay regions.

stays and stay extraction. The identification of stays is important to identifying devices of users who engage in activities at a location instead of simply passing by the location while traveling along a trajectory.

temporal resolution. Temporal resolution is the precision of a measurement with respect to time. The frequency of cell phone use for events such as calls, text messages, and Internet data access; the daily patterns of cell phone use; and the distribution of cell phone use over a typical day are important elements of temporal resolution. The quality of temporal resolution depends on the mechanism that triggers what is recorded. Originally, each record corresponded to calls made by cell phone users. In recent years, a record is generated each time an activity is performed on the cell phone, including calling, texting, and Internet browsing, resulting in a finer temporal resolution.

tower-based CDR data. When a cell phone device connects to cellular networks for a call, message, or data transmission, a cellular tower ID is recorded with a time stamp. The tower ID identifies the cellular tower to which the device connected when its user made a call, sent a text message, or used data.

triangulated CDR data. As technology advances and cell phones are used more frequently, the location of a device can be pinpointed more accurately while it connects to operators' service networks. A time stamp records the connection of a cell phone to the network. Longitude- and latitude-pinpointed coordinate pairs are estimated with a reported accuracy of 200 to 300 meters.

triangulation. The location of a cell phone device at any given time can be approximated by triangulating the signals sent from two or more cell towers to a phone device. In more densely populated urban areas, where towers are closely spaced, the location of a cell phone can be determined more accurately.

uncertainty in location estimates. In areas with a single cell tower, the location of the cell phone is approximate, cannot be triangulated, and falls within a radius. This is often the case in rural areas. In dense urban areas with a higher density of cell towers, multiple signals are sent to the device and its location can be approximated. Triangulation can be more accurate but may also result in false signals that suggest movement even when the device is stationary.

References

Aharony, N., Pan, W., Ip, C., Khayal, I., and Pentland, A. (2011). Social fMRI: Investigating and shaping social mechanisms in the real world. *Pervasive and Mobile Computing*, 7, 643–659. ISBN 9781605588438. https://doi.org/10.1016/j.pmcj.2011.09.004.

Ahas, R., Aasa, A., Silm, S., and Tiru, M. (2010). Daily rhythms of suburban commuters' movements in the Tallinn metropolitan area: Case study with mobile positioning data. *Transportation Research Part C: Emerging Technologies, 18*(1), 45–54. https://doi.org/10.1016/j.trc.2009.04.011.

Ahmed, N., and Miller, H. J. (2007). Time–space transformations of geographic space for exploring, analyzing and visualizing transportation systems. *Journal of Transport Geography, 15*(1), 2–17. https://doi.org/10.1016/j.jtrangeo.2005.11.004.

Alexander, L. P. (2015). *Cell phone location data for travel behavior analysis*. S.M. Thesis in Transportation, Massachusetts Institute of Technology, Cambridge. http://dspace.mit.edu/handle/1721.1/99592.

Alexander, L., Jiang, S., Murga, M., and González, M. C. (2015). Origin–destination trips by purpose and time of day inferred from mobile phone data. *Transportation Research Part C: Emerging Technologies*, 58, Part B: 240–250. ISSN 0968-090X. https://doi.org/10.1016/j.trc.2015.02.018.

Barabási, A.-L. (2005). The origin of bursts and heavy tails in human dynamics. *Nature, 435*(7039), 207–211. https://doi.org/10.1038/nature03459.

Bayir, M. A., Demirbas, M., and Eagle, N. (2010). Mobility profiler: A framework for discovering mobility profiles of cell phone users. *Pervasive and Mobile Computing, 6*(4), 435–454. https://doi.org/10.1016/j.pmcj.2010.01.003.

Beimborn, E., and Kennedy, R. (1996). Inside the black box: Making transportation models work for livable communities. Citizens for a Better Environment and the Environmental Defense Fund. https://www4.uwm.edu/cuts/blackbox/blackbox.pdf.

Belik, V., Geisel, T., and Brockmann, D. (2011). Natural human mobility patterns and spatial spread of infectious diseases. *Physical Review X, 1*(1), 011001. https://doi.org/10.1103/PhysRevX.1.011001.

Bell, M. G. H. (1991). The estimation of origin–destination matrices by constrained generalized least squares. *Transportation Research Part B: Methodological, 25*(1), 13–22. https://doi.org/10.1016/0191-2615(91)90010-G.

Ben-Akiva, M., and Lerman, S. (1985). *Discrete choice analysis*. Boston: MIT Press.

Bhat, C. R., and Koppelman, F. S. (1999). Activity-based modeling of travel demand. In R. W. Hall (Ed.), *The handbook of transportation science* (pp. 35–61). Norwell, Mass., Kluwer Academic Publishers. https://doi.org/10.1007/978-1-4615-5203-1_3.

Boston Metropolitan Planning Organization. (1991). Boston Household Travel Survey. http://www.surveyarchive.org/Boston/Boston_91.zip.

Bowman, J. L., and Ben-Akiva, M. (2001). Activity-based disaggregate travel demand model system with activity schedules. *Transportation Research Part A: Policy and Practice*, *35*(1), 1–28. http://www.sciencedirect.com/science/?article/pii/S0965856499000439. https://doi.org/10.1016/S0965-8564(99)00043-9.

Bricka, S., and Bhat, C. R. (2006). A comparative analysis of Global Positioning System–based and travel survey–based data. *Transportation Research Record, 1972*, 9–20. https://doi.org/10.3141/1972-04.

Brockmann, D., Hufnagel, L., and Geisel, T. (2006). The scaling laws of human travel. *Nature, 439*(7075), 462–465. https://doi.org/10.1038/nature04292.

Bureau of Transportation Statistics. (2017). *National transportation statistics 2017*. Washington, D.C.: U.S. Department of Transportation. https://www.bts.gov/sites/bts.dot.gov/files/docs/browse-statistical-products-and-data/national-transportation-statistics/217651/ntsentire2017q4.pdf.

Caceres, N., Wideberg, J. P., and Benitez, F. G. (2008). Review of traffic data estimations extracted from cellular networks. *IET Intelligent Transport Systems, 2*(3), 179–192. https://doi.org/10.1049/iet-its:20080003.

Calabrese, F., Di Lorenzo, G., Liu, L., and Ratti, C. (2011a). Estimating origin–destination flows using mobile phone location data. *IEEE Pervasive Computing, 10*(4), 36–44. https://doi.org/10.1109/MPRV.2011.41.

Calabrese, F., Smoreda, Z., Blondel, V. D., and Ratti, C. (2011b). Interplay between telecommunications and face-to-face interactions: A study using mobile phone data. *PLOS ONE, 6*(7), e20814. https://doi.org/10.1371/journal.pone.0020814.

Calabrese, F., Diao, M., Di Lorenzo, G., Ferreira, J., Jr., and Ratti, C. (2013). Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. *Transportation Research Part C: Emerging Technologies, 26*, 301–313. https://doi.org/10.1016/j.trc.2012.09.009.

Calabrese, F., Pereira, F. C., Di Lorenzo, G., Liu, L., and Ratti, C. (2010). The geography of taste: Analyzing cell phone mobility and social events. In P. Floréen, A. Krüger, and M. Spasojevic (eds.), *Pervasive Computing: Proceedings of the 8th International Conference on Pervasive Computing*. Helsinki, Finland, May 17–20. https://doi.org/10.1007/978-3-642-12654-3_2.

Cambridge Systematics, Inc. (2010). *Model validation and reasonableness checking manual, Second Edition*. Washington, D.C.: Travel Model Improvement Program, Federal Highway Administration, U.S. Department of Transportation. https://www.fhwa.dot.gov/planning/tmip/publications/other_reports/validation_and_reasonableness_2010/fhwahep10042.pdf.

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

Candia, J., González, M. C., Wang, P., Schoenharl, T., Madey, G., and Barabási, A.-L. (2008). Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical, 41*(22), 224015. http://iopscience.iop.org/article/10.1088/1751-8113/41/22/224015/meta.

Cascetta, E. (1984). Estimation of trip matrices from traffic counts and survey data: A generalized least squares estimator. *Transportation Research Part B: Methodological, 18*(4–5), 289–299. https://doi.org/10.1016/0191-2615(84)90012-2.

Castiglione, J., Bradley, M., and Gliebe, J. (2015). *SHRP 2 Report S2-C46-RR-1: Activity-based travel demand models: A primer*. Washington, D.C.: Transportation Research Board, http://www.trb.org/Main/Blurbs/170963.aspx.

Central Transportation Planning Staff. (2013). *Methodology and assumptions of Central Transportation Planning Staff regional travel demand modeling*. http://www.ctps.org/Drupal/data/pdf/about/mpo/recert_2014/CTPS_GLE_Modeling_Method_20130416.pdf.

Central Transportation Planning Staff. (2008). *Regional travel demand modeling methodology and assumptions*. http://studylib.net/doc/13042279/central-transportation-planning-staff-regional-travel-dem.

Chapin, F. S. (1974). *Human activity patterns in the city: things people do in time and in space*. New York: Wiley. http://www.getcited.org/pub/101488314.

Chen, C., Ma, J., Susilo, Y., Liu, Y., and Wang, M. (2016). The promises of big data and small data for travel behavior (aka human mobility) analysis. *Transportation Research Part C: Emerging Technologies, 68*, 285–299. https://doi.org/10.1016/j.trc.2016.04.005.

Colak, S., Alexander, L. P., Alvim, B. G., Mehndiratta, S. R., and González, M. C. (2015). Analyzing cell phone location data for urban travel: Current methods, limitations and opportunities. *Transportation Research Record*, *2526*, 126–135.

Cottrill, C. D., Pereira, F. C., Zhao, F., Dias, I. F., Lim, H. B., Ben-Akiva, M., and Zegras, P. C. (2013). Future mobility survey: Experience in developing a smartphone-based travel survey in Singapore. *Transportation Research Record*, *2354*, 59–67. http://dx.doi.org/10.3141/2354-07.

Couronné, T., Smoreda, Z., and Olteanu, A.-M. (2011). Chatty mobiles: Individual mobility and communication patterns. NetMob, Boston. Analysis of Mobile Phone Datasets and Networks, Oct. 10–11, 2011, MIT, Cambridge, Mass.

Daganzo, C. F. (1980). Optimal sampling strategies for statistical models with discrete dependent variables. *Transportation Science, 14*(4), 324–345. https://doi.org/10.1287/trsc.14.4.324.

de Montjoye, Y.-A., Hidalgo, C. A., Verleysen, M., and Blondel, V. D. (2013). Unique in the crowd: The privacy bounds of human mobility. *Nature Scientific Reports*, *3*, 1376. http://www.nature.com/srep/2013/130325/srep01376/full/srep01376.html.

de Montjoye, Y.-A., Smoreda, Z., Trinquart, R., Ziemlicki, C., and Blondel, V. D. (2014). D4D-Senegal: The second mobile phone data for development challenge. http://arxiv.org/abs/1407.4885.

Dimitriou, H. T., and Gakenheimer, R. A. (Eds.). (2011). *Urban transport in the developing world: A handbook of policy and practice*. Edward Elgar Publishing. https://doi.org/10.4337/9781849808392.

Eagle, N., and Pentland, A. S. (2009). Eigenbehaviors: Identifying structure in routine. *Behavioral Ecology and Sociobiology*, *63*, 1057–1066. https://doi.org/10.1007/s00265-009-0739-0.

Experian Simmons. (2011). *The 2011 mobile consumer report*. http://www.experian.com/assets/simmons-research/white-papers/experian-simmons-2011-mobile-consumer-report.pdf.

Federal Highway Administration. (2009). National Household Travel Survey. U.S. Department of Transportation. http://nhts.ornl.gov/download.shtml.

Federal Highway Administration. (2013). CTPP 2006–2010 Census tract flows. U.S. Department of Transportation. http://www.fhwa.dot.gov/planning/census_issues/ctpp/data_products/2006-2010_tract_flows/index.cfm.

Ghorpade, A., Pereira, F. C., Zhao, F., Zegras, C., and Ben-Akiva, M. (2015). An integrated stop-mode detection algorithm for real-world smartphone-based travel survey. Presented at 94th Annual Meeting of the Transportation Research Board, Washington D.C.

González, M. C., Hidalgo, C. A., and Barabási, A.-L. (2008). Understanding individual human mobility patterns. *Nature, 453*(7196), 779–782. http://www.ncbi.nlm.nih.gov/pubmed/18528393.

GSMA. (2011). *European mobile industry observatory 2011*. Technical report. http://www.gsma.com/publicpolicy/wp-content/uploads/2012/04/emofullwebfinal.pdf.

Hägerstrand, T. (1989). Reflections on "what about people in regional science?" *Papers in Regional Science, 66*(1), 1–6. http://onlinelibrary.wiley.com/doi/10.1111/j.1435-5597.1989.tb01166.x/abstract.

Hard, E., Chigoy, B. Songchitruksa, P., Farnsworth, S., Borchardt, D., and Green, L. (2016). *Synopsis of new methods and technologies to collect origin–destination (O-D) data*. FHWA-HEP-16-083. FHWA, U.S. Department of Transportation.

Hariharan, R., and Toyama, K. (2004). Project Lachesis: Parsing and modeling location histories. In M. J. Egenhofer, C. Freksa, and H. J. Miller (Eds.), *Geographic information science* (pp. 106–124). Springer.

Hasan, S., Schneider, C. M., Ukkusuri, S. V., and González, M. C. (2013). Spatiotemporal patterns of urban human mobility. *Journal of Statistical Physics, 151*(1–2), 304–318. https://doi.org/10.1007/s10955-012-0645-0.

Hazelton, M. L. (2003). Some comments on origin–destination matrix estimation. *Transportation Research Part A: Policy and Practice, 37*(10), 811–822. https://doi.org/10.1016/S0965-8564(03)00044-2.

Hidalgo, C. A. (2006). Conditions for the emergence of scaling in the inter-event time of uncorrelated and seasonal systems. *Physica A: Statistical Mechanics and its Applications 369*(2), 877–883. https://doi.org/10.1016/j.physa.2005.12.035.

Huntsinger, L. F., and Donnelly, R. (2014). Reconciliation of regional travel model and passive device tracking data. Presented at 93rd Annual Meeting of the Transportation Research Board, Washington, D.C. http://docs.trb.org/prp/14-1058.pdf.

IBM. (2014). The four V's of big data. IBM Big Data and Analytics Hub. http://www.ibmbigdatahub.com/infographic/four-vs-big-data.

International Energy Agency. (2015). *Energy and climate change: World energy outlook special report*. https://www.iea.org/publications/freepublications/publication/WEO2015SpecialReportonEnergyandClimateChange.pdf.

International Telecommunication Union (2014). The world in 2014. *ICT facts and figures*. http://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2014-e.pdf.

Iovan, C., Olteanu-Raimond, A.-M., Couronné, T., and Smoreda, Z. (2013). Moving and calling: Mobile phone data quality measurements and spatiotemporal uncertainty in human mobility studies. In D. Vandenbroucke, B. Bucher, and J. Crompvoets (Eds.), *Geographic information science at the heart of Europe* (pp. 247–265). Springer.

Iqbal, M. S., Choudhury, C. F., Wang, P., and González, M. C. (2014). Development of origin–destination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies, 40*, 63–74. https://doi.org/10.1016/j.trc.2014.01.002.

Janelle, D. G. (2012). Space-adjusting technologies and the social ecologies of place: Review and research agenda. *International Journal of Geographical Information Science*, *26*(12), 2239–2251. https://doi.org/10.1080/13658816.2012.713958.

Jiang, S. (2015). *Deciphering human activities in complex urban systems—Mining big data for sustainable urban future*. Ph.D. dissertation. Massachusetts Institute of Technology, Cambridge.

Jiang, S., Ferreira, J., Jr., and González, M. C. (2012a). Clustering daily patterns of human activities in the city. *Data Mining and Knowledge Discovery*, *25*(3), 478–510. https://doi.org/10.1007/s10618-012-0264-z.

Jiang, S., Ferreira, J., Jr., and González, M. C. (2012b). Discovering urban spatial-temporal structure from human activity patterns. In *UrbComp '12: Proceedings of the ACM SIGKDD International Workshop on Urban Computing*, 95–102. https://doi.org/10.1145/2346496.2346512.

Jiang, S., Fiore, G. A., Yang, Y., Ferreira, J., Frazzoli, E., and González, M. C. (2013). A review of urban computing for mobile phone traces: Current methods, challenges and opportunities. In *UrbComp '13: Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*, *2*, 1–2, p. 9. https://doi.org/10.1145/2505821.2505828.

Jiang, S., Yang, Y., Gupta, S., Veneziano, D., Athavale, S., and González, M. C. (2016). The TimeGeo modeling framework for urban mobility without travel surveys. *Proceedings of the National Academy of Sciences of the United States of America, 113*(37), E5370–E5378. https://doi.org/10.1073/pnas.1524261113.

Karsai, M., Kaski, K., Barabási, A.-L., and Kertész, J. (2012). Universal features of correlated bursty behaviour. *Scientific Reports*, *2*, Article number 397. doi:10.1038/srep00397.

Kosta, E., Graux, H., and Dumortier, J. (2014). Collection and storage of personal data: A critical view on current practices in the transportation sector. In B. Preneel and D. Ikonomou (Eds.), *Privacy technologies and policy.* Lecture Notes in Computer Science, 8319 (pp. 157–176). Berlin, Heidelberg: Springer-Verlag. https://doi.org/10.1007/978-3-642-54069-1_10.

Kressner J. D., Macfarlane, G.S., Huntsinger, L., and Donnelly, R. (2016). Using passive data to build an agile tour-based model: A case study in Ashville. Presented at 6th Transportation Research Board Conference on Innovations in Travel Modeling (ITM), May 1–4, Denver, Colo.

Krumm, J., and Horvitz, E. (2006). Predestination: Inferring destinations from partial trajectories. In P. Dourish and A. Friday (Eds.), *UbiComp 2006: Ubiquitous computing. Lecture Notes in Computer Science, 4206* (pp. 243–260). Berlin, Heidelberg: Springer-Verlag. https://link.springer.com/chapter/10.1007/11853565_15.

Lane, N. D., Miluzzo, E., Lu, H., Peebles, D., Choudhury, T., and Campbell, A. T. (2010). A survey of mobile phone sensing. *IEEE Communications Magazine, 48*(9), 140–150. http://ieeexplore.ieee.org/document/5560598/.

Laney, D. (2001). 3D data management: Controlling data volume, velocity, and variety. *Application delivery strategies.* META Group Inc. https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

Lee, J.-K., and Hou, J. C. (2006). Modeling steady-state and transient behaviors of user mobility: Formulation, analysis, and application. *MobiHoc 2006: Proceedings of the 7th ACM International Symposium on Mobile ad hoc Networking and Computing* (pp. 85–96). https://doi.org/10.1145/1132905.1132915.

Levinson, D. M., and Kumar, A. (1994). The rational locator: Why travel times have remained stable. *Journal of the American Planning Association, 60*(3), 319–332. https://doi.org/10.1080/01944369408975590.

Lo, H. P., Zhang, N., and Lam, W. H. K. (1996). Estimation of an origin–destination matrix with random link choice proportions: A statistical approach. *Transportation Research Part B: Methodological, 30*(4), 309–324. https://doi.org/10.1016/0191-2615(95)00036-4.

Lu, C.-C., Zhou, X., and Zhang, K. (2013). Dynamic origin–destination demand flow estimation under congested traffic conditions. *Transportation Research Part C: Emerging Technologies, 34*, 16–37. https://doi.org/10.1016/j.trc.2013.05.006.

Lu, X., Bengtsson, L., and Holme, P. (2012). Predictability of population displacement after the 2010 Haiti earthquake. *Proceedings of the National Academy of Sciences of the United States of America, 109*(29), 11576–11581. https://doi.org/10.1073/pnas.1203882109.

Lynch, K. (1976). *What time is this place?* Cambridge, Mass.: MIT Press.

Maher, M. J. (1983). Inferences on trip matrices from observations on link volumes: A Bayesian statistical approach. *Transportation Research Part B: Methodological, 17*(6), 435–447. https://doi.org/10.1016/0191-2615(83)90030-9.

Malmgren, R. D., Stouffer, D. B., Motter, A. E., and Amaral, L. A. N. (2008). A Poissonian explanation for heavy tails in e-mail communication. *Proceedings of the National Academy of Sciences of the United States of America, 105*(47), 18153–18158. https://doi.org/10.1073/pnas.0800332105.

Manheim, M. L. (1979). *Fundamentals of transportation systems analysis*. Cambridge, Mass.: MIT Press.

Mao, G., Fidan, B., and Anderson, B. D. O. (2007). Wireless sensor network localization techniques. *Computer Networks, 51*(10), 2529–2553. https://doi.org/10.1016/j.comnet.2006.11.018.

Martin, W. A., and McGuckin, N. A. (1998). *NCHRP Report 365: Travel estimation techniques for urban planning*. Washington, D.C.: TRB, National Research Council.

Massachusetts Department of Transportation. (2012). *Massachusetts Travel Survey 2010–2011*. http://www.massdot.state.ma.us/planning/Main/MapsDataandReports/Reports/TravelSurvey.aspx.

MassGIS. (2014). MassGIS Data: Community boundaries (towns). http://www.mass.gov/anf/research-and-tech/it-serv-and-support/application-serv/office-of-geographic-information-massgis/datalayers/towns.html.

McNally, M. G. (2008). The four-step model. In D. A. Hensher and K. J. Button (Eds.), *Handbook of transportation modeling* (pp. 35–36). Amsterdam, London: Elsevier.

NuStats. (2002). *2000–2001 California Statewide Household Travel Survey*. Final Report. Sacramento: California Department of Transportation.

Ortúzar, J., and Willumsen, L. G. (2011). *Modelling transport, fourth edition*. Chichester, UK: Wiley. https://doi.org/10.1002/9781119993308.

Pew Research Center. (2017). Mobile factsheet. http://www.pewinternet.org/fact-sheet/mobile/.

Pew Research Center. (2015). *The smartphone difference*. http://www.pewinternet.org/2015/04/01/us-smartphone-use-in-2015.

Pinjari, A. R., and Bhat, C. R. (2011). Activity-based travel demand analysis. In A. de Palma, R. Lindsey, E. Quinet, and R. Vickerman (Eds.), *A handbook of transport economics*. https://www.elgaronline.com/view/9781847202031.00017.xml.

Ranjan, G., Zang, H., Zhang, Z.-L., and Bolot, J. (2012). Are call detail records biased for sampling human mobility? *Mobile Computing and Communications Review, 16*(3), 33–44. https://doi.org/10.1145/2412096.2412101.

Reddy, S., Mun, M., Burke, J., Estrin, D., Hansen, M., and Srivastava, M. (2010). Using mobile phones to determine transportation modes. *ACM Transactions on Sensor Networks, 6*(2), Article No. 13. https://doi.org/10.1145/1689239.1689243.

Renso, C., Puntoni, S., and Frentzos, E. (2008). Wireless network data sources: tracking and synthesizing trajectories. In F. Giannotti and D. Pedreschi (Eds.), *Mobility, data mining and privacy*. Berlin, Heidelberg: Springer-Verlag. https://doi.org/10.1007/978-3-540-75177-9_4.

Richardson, A. J., Meyburg, A. H., and Ampt, E. S. (1995). *Survey methods for transport planning.* Eucalyptus Press.

Rose, G. (2006). Mobile phones as traffic probes: Practices, prospects and issues. *Transport Reviews, 26*(3), 275–291. https://doi.org/10.1080/01441640500361108.

Schafer, A. (2000). Regularities in travel demand: An international perspective. *Journal of Transportation and Statistics, 3*(3), 1–31.

Schneider, C. M., Belik, V., Couronné, T., Smoreda, Z., and González, M. C. (2013). Unravelling daily human mobility motifs. *Journal of the Royal Society, Interface, 10*(84), 20130246. https://doi.org/10.1098/rsif.2013.0246.

Schrank, D., Eisele, B., Lomax, T., and Bak, J. (2015). *2015 Urban mobility scorecard*. College Station, Tex.: Texas A&M Transportation Institute and INRIX. https://mobility.tamu.edu/ums/report/.

Schrank, D., Eisele, B., and Lomax, T. (2012). *2012 Urban mobility report*. College Station, Tex.: Texas A&M Transportation Institute https://static.tti.tamu.edu/tti.tamu.edu/documents/ums/archive/mobility-report-2012-wappx.pdf.

Scott, B. (2015). Big data with volume, velocity, variety, veracity and value. *Insights.* Sage Sustainable Electronics. http://www.sagese.com/resources/volume-velocity-value.

Smith, M. E. (1979). Design of small-sample home-interview travel surveys. *Transportation Research Record, 701*, 29–35.

Song, C., Koren, T., Wang, P., and Barabási, A.-L. (2010a). Modelling the scaling properties of human mobility. *Nature Physics, 6*(10), 818–823. https://doi.org/10.1038/nphys1760.

Song, C., Qu, Z., Blumm, N., and Barabási, A.-L. (2010b). Limits of predictability in human mobility. *Science, 327*(5968), 1018–1021. https://doi.org/10.1126/science.1177170.

Sosslau, A. B., Hassam, A. B., Carter, M. M., and Wickstrom, G. V. (1978). *NCHRP Report 187: Quick-response urban travel estimation techniques and transferable parameters: User's guide*. Washington, D.C.: TRB, National Research Council.

Spiess, H. (1987). A maximum likelihood model for estimating origin–destination matrices. *Transportation Research Part B: Methodological, 21*(5), 395–412. https://doi.org/10.1016/0191-2615(87)90037-3.

Stopher, P. R., and Greaves, S. P. (2007). Household travel surveys: Where are we going? *Transportation Research Part A: Policy and Practice, 41*(5), 367–381. https://doi.org/10.1016/j.tra.2006.09.005.

Toole, J. L. (2015). *Putting big data in its place: Understanding cities and human mobility with new data sources*. Ph.D. dissertation. Massachusetts Institute of Technology, Cambridge, Mass.

Toole, J. L., Colak, S., Sturt, B., Alexander, L. P., Evsukoff, A., and González, M. C. (2015). The path most traveled: Travel demand estimation using big data resources. *Transportation Research Part C: Emerging Technologies. 58*(Part B), 162–177

United Nations. (2014). World urbanization prospects: The 2014 revision. Technical report. http://esa.un.org/unpd/wup/Highlights/WUP2014-Highlights.pdf.

U.S. Energy Information Administration. (2015). How much energy is consumed in the world by each sector? http://www.eia.gov/tools/faqs/faq.cfm?id=447&t=1.

Van Zuylen, H. J., and Willumsen, L. G. (1980). The most likely trip matrix estimated from traffic counts. *Transportation Research Part B: Methodological, 14*(3), 281–293. https://doi.org/10.1016/0191-2615(80)90008-9.

Vázquez, A., Oliveira, J. G., Dezsö, Z., Goh, K.-I., Kondor, I., and Barabási, A.-L. (2006). Modeling bursts and heavy tails in human dynamics. *Physical Review E, 73*(3), 036127. https://doi.org/10.1103/PhysRevE.73.036127.

Wang, P., Hunter, T., Bayen, A. M., Schechtner, K., and González, M. C. (2012). Understanding road usage patterns in urban areas. *Scientific Reports, 2*(1), 1001. https://doi.org/10.1038/srep01001.

Wesolowski, A., Eagle, N., Tatem, A. J., Smith, D. L., Noor, A. M., Snow, R. W., and Buckee, C. O. (2012). Quantifying the impact of human mobility on malaria. *Science, 338*(6104), 267–270. https://doi.org/10.1126/science.1223467.

Widhalm, P., Yang, Y., Ulm, M., Athavale, S., and González, M. C. (2015). Discovering urban activity patterns in cell phone data. *Transportation, 42*(4), 597–623. https://doi.org/10.1007/s11116-015-9598-x.

Yang, H., Sasaki, T., Iida, Y., and Asakura, Y. (1992). Estimation of origin–destination matrices from link traffic counts on congested networks. *Transportation Research Part B: Methodological, 26*(6), 417–434. https://doi.org/10.1016/0191-2615(92)90008-K.

Ye, Y., Zheng, Y. Chen, Y., Feng, J., and Xie, X. (2009). Mining individual life pattern based on location history. *MDM'09. 10th International Conference on Mobile Data Management: Systems, Services and Middleware, 2009.* http://ieeexplore.ieee.org/document/5088915/.

Ygnace, J.-L., Benguigui, C., Delannoy, V., Remy, J.-G., Auclair, P., Bosseboeuf, J.-L., Schwab, N., and da Fonseca, V. (2001). *Travel time/speed estimates on the French Rhone corridor network using cellular phones as probes.* Final Report of the SERTI V Program, INRETS, Lyon, France.

Yin, M., Sheehan, M., Feygin, S., Paiement, J. F., and Pozdnoukhov, A. (2016). A generative model of urban activities from cellular data. *IEEE Transactions in ITS,* September. http://faculty.ce.berkeley.edu/pozdnukhov/papers/IEEE_ITS_cellular_abm.pdf.

Zang, H., Baccelli, F., and Bolot, J. (2010). Bayesian inference for localization in cellular networks. *2010 Proceedings IEEE INFOCOM*. http://ieeexplore.ieee.org/document/5462018/.

Zhao, Y. (2000). Mobile phone location determination and its impact on intelligent transportation system. *IEEE Transactions on Intelligent Transportation Systems, 1*(1), 55–64. https://doi.org/10.1109/6979.869021.

Zheng, V. W., Zheng, Y., Xie, X. and Yang, Q. (2010). Collaborative location and activity recommendations with GPS history data. In *Proceedings of the 19th International Conference on World Wide Web*, 1029–1038. ACM. https://doi.org/10.1145/1772690.1772795.

Zheng, Y. (2015). Trajectory data mining: An overview. *ACM Transactions on Intelligent Systems and Technology, 6*(3), Article 29. https://doi.org/10.1145/2743025.

Zmud, J., and Wolf, J. (2003). Identifying the correlates of trip misreporting: Results from the California Statewide Household Travel Survey GPS Study. Presented at Moving Through Nets: The Physical and Social Dimensions of Travel: 10th International Conference on Travel Behaviour Research. https://www.researchgate.net/publication/242144239_Identifying_the_Correlates_of_Trip_Misreporting_-_Results_from_the_California_Statewide_Household_Travel_Survey_GPS_Study.

Additional Resources

Ahas, R., Aasa, A., Mark, Ü., Pae, T., and Kull, A. (2007). Seasonal tourism spaces in Estonia: Case study with mobile positioning data. *Tourism Management, 28*(3), 898–910. https://doi.org/10.1016/j.tourman.2006.05.010.

Axhausen, K. W., Löchl, M., Schlich, R., Buhl, T., and Widmer, P. (2007). Fatigue in long-duration travel diaries. *Transportation, 34*(2), 143–160. https://doi.org/10.1007/s11116-006-9106-4.

Bagrow, J. P., Wang, D., and Barabási, A.-L. (2011). Collective response of human populations to large-scale emergencies. *PLOS ONE, 6*(3), e17680. https://doi.org/10.1371/journal.pone.0017680.

Becker, R., Cáceres, R., Hanson, K., Isaacman, S., Loh, J. M., Martonosi, M., Rowland, J., Urbanek, S., Varshavsky, A., and Volinsky, C. (2013). Human mobility characterization from cellular network data. *Communications of the ACM, 56*(1), 74–82. https://doi.org/10.1145/2398356.2398375.

Bekhor S., Cohen, Y., and Solomon, C. (2013). Evaluating long-distance travel patterns in Israel by tracking cellular phone positions. *Journal of Advanced Transportation*, *47*(4), 435–446. (First published online Feb. 2011.) https://doi.org/10.1002/atr.170.

Bekhor S., Hirsh, M., Nimre, S., and Feldman, I. (2015). Identifying Spatial and Temporal Congestion Characteristics Using Passive Mobile Phone Data. Presented at 87th Annual Meeting of the Transportation Research Board, Washington, D.C.

Bengtsson, L., Lu, X., Thorson, A., Garfield, R., and von Schreeb, J. V. (2011). Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: A post-earthquake geospatial study in Haiti. *PLOS Medicine, 8*(8), 1–9. https://doi.org/10.1371/journal.pmed.1001083.

Carrion C., Pereira, F. C., Ball, R., Zhao, F., Kim, Y., Zheng, N., Zegras, P. C., and Ben-Akiva, M. E. (2014). Evaluating FMS: A preliminary comparison with a traditional travel survey. Presented at 93rd Annual Meeting of the Transportation Research Board, Washington, D.C.

Chen, C., Gong, H., Lawson, C., and Bialostozky, E. (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: Policy and Practice, 44*(10), 830–840. https://doi.org/10.1016/j.tra.2010.08.004.

Federal Highway Administration and Federal Transit Administration. (2013). *Status of the nation's highways, bridges, and transit: Conditions & performance*. U.S. Department of Transportation. https://www.fhwa.dot.gov/policy/2013cpr/pdfs/cp2013.pdf.

Girardin, F., Vaccari, A., Gerber, A., Biderman, A., and Ratti, C. (2009). Towards estimating the presence of visitors from the aggregate mobile phone network activity they generate. https://pdfs.semanticscholar.org/0801/bc25c0fc5a86ae40c3aecbc5314e3ca3cc66.pdf.

Gong, H., Chen, C., Bialostozky, E., and Lawson, C. T. (2012). A GPS/GIS method for travel mode detection in New York City. *Computers, Environment and Urban Systems, 36*(2), 131–139. https://doi.org/10.1016/j.compenvurbsys.2011.05.003.

Gur, Y. J., Bekhor, S., Solomon, C., and Kheifits, L. (2009). Intercity person trip tables for nationwide transportation planning in Israel obtained from massive cell phone data. *Transportation Research Record, 2121*, 145–151. https://doi.org/10.3141/2121-16.

Hartgen, D. T., and San Jose, E. (2009). Costs and trip rates of recent household travel surveys. http://www.hartgengroup.net/Projects/National/USA/household_travel_summary/2009-11-11_Final_Report_Revised.pdf.

Hu, P. S., and Reuscher, T. R. (2004). *Summary of travel trends: 2001 National Household Travel Survey*. FHWA, U.S. Department of Transportation. http://nhts.ornl.gov/2001/pub/stt.pdf.

Isaacman, S., Becker, R. A., Cáceres, R., Kobourov, S. G., Martonosi, M., Rowland, J., and Varshavsky, A. (2011). Ranges of human mobility in Los Angeles and New York. In K. Lyons, J. Hightower, and E. M. Huang (Eds.), *Pervasive computing. Pervasive 2011. Lecture Notes in Computer Science*, *6696*. Berlin, Heidelberg: Springer. https://www.researchgate.net/publication/221036879_Ranges_of_Human_Mobility_in_Los_Angeles_and_New_York.

Isaacman, S., Becker, R. A., Cáceres, R., Kobourov, S. G., Martonosi, M., Rowland, J., and Varshavsky, A. (2011). Identifying important places in people's lives from cellular network data. In K. Lyons, J. Hightower, and E. M. Huang (Eds.), *Pervasive computing. Pervasive 2011. Lecture Notes in Computer Science*, *6696*. Berlin, Heidelberg: Springer. https://link.springer.com/chapter/10.1007%2F978-3-642-21726-5_9.

Jiang, S., Ferreira, J., Jr., and González, M. C. (2017). Activity-based human mobility patterns inferred from mobile phone data: A case study of Singapore. *IEEE Transactions on Big Data, 3*(2), 208–219.

Milone, R. (2014). Initial analysis of AirSage O-D cellular data for the TPB modeled area. Presentation to the Travel Forecasting Subcommittee, National Capitol Region Transportation Planning Board and Metropolitan Washington Council of Governments, July 18. http://www1.mwcog.org/uploads/committee-documents/ZV1YW1Zc20140718142637.pdf.

Rwanda hits 55pc mobile phone penetration rate. (2012). *Africa Review.* http://www.africareview.com/Business—Finance/Rwanda-mobile-phone-penetration-rate/-/979184/1713912/-/format/xhtml/-/dr1a8kz/-/index.html.

Sagl, G., Loidl, M., and Beinat, E. (2012). A visual analytics approach for extracting spatio-temporal urban mobility information from mobile network traffic. *ISPRS International Journal of Geo-Information, 1*(3), 256–271. https://doi.org/10.3390/ijgi1030256.

Sevtsuk, A., and Ratti, C. (2010). Does urban mobility have a daily routine? Learning from the aggregate data of mobile networks. *Journal of Urban Technology, 17*(1), 41–60. https://doi.org/10.1080/10630731003597322.

Sun, J. B., Yuan, J., Wang, Y., Si, H. B., and Shan, X. M. (2011). Exploring space–time structure of human mobility in urban space. *Physica A, 390*(5), 929–942. https://doi.org/10.1016/j.physa.2010.10.033.

Tarasov, A., Kling, F., and Pozdnoukhov, A. (2013). Prediction of user location using the radiation model and check-ins. *Proceedings of the 2nd ACM SIGKD International Workshop on Urban Computing*, Article No. 8, New York. https://doi.org/10.1145/2505821.2505833.

Traag, V.A., Browet, A., Calabrese, F., and Morlot, F. (2011). Social event detection in massive mobile phone data using probabilistic location inference. *IEEE Third International Conference on Privacy, Security, Risk, and Trust, and IEEE Third International Conference on Social Computing* (pp. 625–628). http://ieeexplore.ieee.org/document/6113183/.

Wang, M., Chen, C., and Ma, J. (2015). On making more efficient location prediction. Presented at 94th Annual Meeting of the Transportation Research Board, Washington, D.C.

Wang, M., Chen, C., and Ma, J. (2015). Time-of-day dependence of location variability: An application of passively-generated mobile phone dataset. Presented at 94th Annual Meeting of the Transportation Research Board, Washington, D.C.

Wang, P., González, M. C., Hidalgo, C. A., and Barabási, A.-L. (2009). Understanding the spreading patterns of mobile phone viruses. *Science, 324*(5930), 1071–1076. https://doi.org/10.1126/science.1167053.

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
FAST	Fixing America's Surface Transportation Act (2015)
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
TDC	Transit Development Corporation
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

TRANSPORTATION RESEARCH BOARD
500 Fifth Street, NW
Washington, DC 20001

ADDRESS SERVICE REQUESTED

NON-PROFIT ORG.
U.S. POSTAGE
PAID
COLUMBIA, MD
PERMIT NO. 88

ISBN 978-0-309-39035-4
90000
9 780309 390354

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

The nation turns to the National Academies of Sciences, Engineering, and Medicine for independent, objective advice on issues that affect people's lives worldwide.
www.national-academies.org