Cell Phone Location Data for Travel Behavior Analysis

by

Lauren P. Alexander

B.S., University of Texas at Austin (2009)

Submitted to the Department of Civil and Environmental Engineering
in partial fulfillment of the requirements for the degree of

Master of Science in Transportation

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2015

Author ..... Signature redacted

Department of Civil and Environmental Engineering
May 21, 2015

Certified by ..... Signature redacted

Marta C. González
Associate Professor of Civil and Environmental Engineering
Thesis Supervisor

Accepted by ..... Signature redacted

Heidi M. Nepf
Donald and Martha Harleman Professor of Civil and Environmental Engineering
Chair, Departmental Committee for Graduate Students

Cell Phone Location Data for Travel Behavior Analysis

by

Lauren P. Alexander

Submitted to the Department of Civil and Environmental Engineering
on May 21, 2015, in partial fulfillment of the
requirements for the degree of
Master of Science in Transportation

Abstract

Mobile phone technology generates vast amounts of data at low costs all over the world. This rich data provides digital traces when and where individuals travel, improving our ability to understand, model, and predict human mobility. Especially in this era of rapid urbanization, mobile phone data presents exciting new opportunities to plan transportation infrastructure and services that meet the mobility needs and challenges associated with increasing travel demand. But to realize these benefits, methods must be developed to utilize and integrate this data into existing urban and transportation modeling frameworks.

In this thesis, we draw on techniques from the transportation engineering and urban computing communities to estimate travel demand and infrastructure usage. The methods we present utilize call detail records (CDRs) from mobile phones in conjunction with geospatial data, census records, and surveys, to generate representative origin-destination matrices, route trips through road networks, and evaluate traffic congestion. Moreover, we implement these algorithms in a flexible, modular, and computationally efficient software system. This platform provides an end-to-end solution that integrates raw, massive data to generate estimates of travel demand and infrastructure performance in any city, and produces interactive visualizations to effectively communicate these results. Finally, we demonstrate an application of these data and methods to evaluate the impact of ride-sharing on urban traffic.

Using these approaches, we generate travel demand estimates analogous to many of the outputs of conventional travel demand models, demonstrating the potential of mobile phone data as a low cost option for transportation planning. We hope this work will serve as unified and comprehensive guide to integrating new big data resources into transportation modeling practices.

Thesis Supervisor: Marta C. González
Title: Associate Professor of Civil and Environmental Engineering

Acknowledgments

First and foremost, I want to thank my advisor, Marta, for her support and contagious enthusiasm. She introduced me to a field of interdisciplinary research, which I find fascinating and meaningful. I truly appreciate the flexibility she has afforded me to explore this new world, while being there to patiently guide me along the way.

To all of my HuMNet friends, thank you for the countless conversations and laughs. Being surrounded every day by such an intelligent, passionate, and fun(ny) group of people is something I will never forget and can only hope to find again. And to my other officemates, classmates, and professors, thank you for helping to make my time at MIT enjoyable and enriching.

To my SDG colleagues, thank you for sparking my interest in transportation. Beginning with the *red bus/blue bus* lesson on day 1, you taught me, inspired me, and supported me on my path to MIT. To my Bridj colleagues, thank you for the opportunity to apply theory in practice this summer to move real people from point A to B; it was a truly rewarding experience that cemented my desire to work at the intersection of transportation and data science.

To my sisters, thank you for always being there for me, and especially for your silliness. To Drew, thank you for your endless support despite taking the brunt of my stress these past two years. And to all of my friends and family, I would be lost without you in my life. Despite distance and busy schedules, know that you are always on my mind. And from the bottom of my heart, thank you.

But most of all, thank you to my parents for the abundance of encouragement, advice, and unconditional love. I am lucky to have two amazing role models who continue to inspire me personally and professionally. You've both shown me that working hard and enjoying what you do can go hand-in-hand, and I will forever strive for this goal.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview and motivation

According the United Nations Population Fund (UNFPA), 2008 marked the first year in which the majority of the world's population lived in cities. Rapid urbanization places enormous strain on already burdened transportation infrastructure critical to providing residents with access to places, people, and goods. Delays and poor levels of service resulting from such congestion waste time and money and exacerbate harmful vehicle emissions.

Effectively moving people and goods–the fundamental task of transportation planners, modelers, and engineers–is increasingly challenging in this era of rapid population growth in cities. Meanwhile, transportation services and infrastructure effect economic growth and quality of life within cities. Given the direct and varied impacts it has on society, the transportation industry attracts planners, engineers, and economists to address these complex challenges.

Interdisciplinary approaches are key to understanding and modeling human mobility patterns and future mobility needs. The economists' principles of supply, demand, and pricing, along with the planners' concepts of transportation system dynamics, underpin much of the framework for modeling human mobility. Combining broad and varied expertise, cities can adopt strategies to plan more efficient, sustainable, and equitable transportation systems.

Travel demand models are essential for managing existing transportation systems and planning for future development. Demand estimates output from such models are relied upon for transportation plans, environmental impact studies, and infrastructure investment and prioritization decisions [11]. Travel demand models widely-used in industry fall into two main categories: the traditional four-step or trip-based models, and the newer activity-based or schedule-based models.

But despite the sophistication of these models, they require quality input data for development, calibration, and validation. Accordingly, a large amount of time and money are spent on data collection. Data demands include detailed data on transportation networks, capacities, and levels of service, as well as behavioral data collected from surveys. In addition to sociodemographic information, household travel surveys provide travel-activity diaries detailing specific trips and travel characteristics of the respondent. Because they are expensive and intrusive, such surveys typically describe just one recent day, limiting their ability to capture irregular and/or leisure activities.

In contrast to survey data, ubiquitous mobile computing, namely the pervasive use of cellular phones, has generated a wealth of data that can be analyzed to understand and improve urban infrastructure systems. The penetration of these devices is astounding with six billion mobile phones nearly tripling the number of internet users. Penetration rates of over 100% are routinely found in the developed word, e.g. 104% in the United States and 128% in Europe[1] and rates are of over 85%[2] are observed developing contexts. These devices and the applications that run on them passively record social, mobility, and a variety of other behaviors of their users with extremely high spatial and temporal resolution.

However, before the benefits of this massive, passive data can be realized within the transportation domain, methods must be developed and assessed with respect to their applicability and limitations. In particular, mobile phone data has the potential

[1]GSMA European Mobile Industry Observatory 2011 http://www.gsma.com/publicpolicy/wp-content/uploads/2012/04/emofullwebfinal.pdf

[2]ITU. (2013) ICT Facts and Figures http://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2013-e.pdf

to complement or substitute for household travel surveys. However, despite the fact that it can be gathered more frequently and economically, mobile phone data lacks information about a respondent (e.g. age or income) or his/her trip (e.g. purpose or mode) [70, 81, 44]. Furthermore, mobile phone data contains traces of a user at approximated locations when his/her phone communicates with a cell phone tower, providing an inexact and incomplete picture of daily trip-making. Accordingly, much research has focused on developing methods to extract meaningful information about human mobility from mobile phone traces.

Adding to the existing body of work using mobile phone data, we present methodology to go from raw mobile phone data to road usage, analogous to the trip generation, trip distribution, mode choice, and traffic assignments procedures of traditional travel demand models. By paralleling the framework commonly used by transportation planners and modelers, we are able to compare and contrast these methods and results with traditional survey data and models. Moreover, we present a flexible, modular, and computationally efficient software system to integrate these algorithms and visualize results in any city for which mobile phone data is available. Lastly, we present an application that utilizes all of these methods to evaluate the impact of ridesharing on urban traffic.

The gamut of travel demand estimation using big data resources is presented through the methods, validation, implementation, and applications described in this thesis. Demonstrating the validity of mobile phone data as an end-to-end solution for travel demand estimation will hopefully support the incorporation of new big data resources into transportation demand modeling approaches.

1.2 Literature review

Researchers across different domains approach travel demand modeling in different ways. Here, we present an overview of bodies of work in two communities, in particular: transportation engineering and urban computing. The transportation engineering community typically uses models to derive demand and behavior based on socioeco-

nomic, land use, and transportation characteristics. The urban computing community brings together statistical models, data mining, and machine learning techniques to extract mobility patterns from large amounts of data. As automatically collected data is becoming more widespread, the boundary between these communities is diminishing.

1.2.1 Transportation engineering

Travel demand models are widely used for transportation policy, planning, and engineering applications in order to estimate infrastructure capacity requirements, financial and social viability, and environmental impacts of proposed transportation projects. The fundamental task of these models is to adequately model the travel decision process, a problem simplified by aggregating decisions and decision-makers in space (e.g. dividing a study area into zones) and time (e.g. discrete time periods).

Travel demand models were first developed in the US in response to post-war development and economic growth, with the first comprehensive application being the Chicago Area Transportation Study in the 1950s. This study implemented a four-step model, modeling estimation procedures in four sequential steps: trip generation, trip distribution, mode choice, and trip assignment. Federal legislation introduced in the 1960s requiring urban transportation planning institutionalized the four-step model. In the 1970s additional legislation called for improved models, with particular emphasis on multimodal and environmental planning, leading to the development and integration of more sophisticated demand and assignment methods into four-step models. Growing recognition of the limitations of the four-step modeling approach in the late 1970s and 1980s led to Travel Model Improvement Program[3], which has worked towards advancing modeling capabilities and supporting transportation professionals since the early 1990s. Efforts in the last few decades can be characterized as improving state-of-the-practice conventional models while further developing state-of-the-art methodologies [57].

Approaches to modeling travel demand fall into three main categories–trip-based,

[3]http://www.fhwa.dot.gov/planning/tmip/

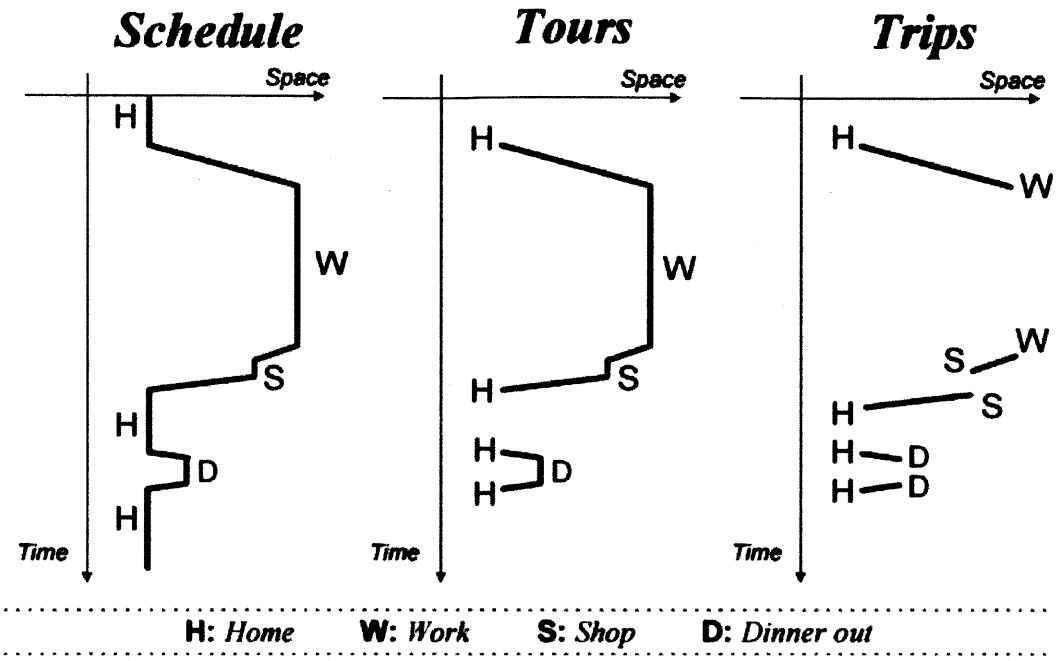

Figure 1-1: Schedule-based, tour-based, and trip-based travel demand modeling frameworks. Trip-based models represent unlinked trips, tour-based models chains these trips into tours, and schedule-based models schedule these tours. Source: `http://ocw.mit.edu/courses/civil-and-environmental-engineering/1-201j-transportation-systems-analysis-demand-and-economics-fall-2008/lecture-notes/MIT1_201JF08_lec05.pdf`

tour-based, and schedule-based, as summarized in Figure 1-1. Trip-based models represent unlinked trips, tour-based models chains these trips into tours, and schedule-based models schedule these tours. The four-step model is an example of a trip-based approach, while newer activity-based models use the schedule-based approach.

Four-step models–still the most widely used model in practice–are developed, updated, and applied by many metropolitan planning organizations (MPOs) and planning agencies across the US and abroad. In contrast to microscopic agent-based models, this framework aggregates travelers and trips at the level of traffic analysis zones (TAZs) rather than simulating travel behavior of individuals. Each of the four steps are described in more detail below:

1. *Trip generation*: This step determines trip origins and destinations based on the distribution of households, employment, and land use. More specifically, trip productions are estimated based on household characteristics such as size, income,

car ownership, density, and accessibility. Similarly, trip attractions are estimated using land use, employment by sector, and accessibility data. The models in this step typically use regression [37, 48, 56, 62], cross classification [68], and growth factor analysis.

2. *Trip distribution*: This step estimates the distribution of trips as a function of the generalized cost of travel between origins and destinations. Trip distribution uses the origins and destinations estimated in the first step as marginal totals from which to estimate the elements of a trip matrix. Various aggregate models of trip distribution have been proposed [63, 82, 90, 91, 92, 95]. Among all these efforts, the gravity model, which assumes that the number of movements between an OD pair decays with their distance or cost, is the most widely used [95, 28, 33]. When an empirical OD matrix is available from survey data, for example, a method called iterative proportional fitting (IPF) method can be used [75]. This procedure adjusts matrix values from the input (or seed) matrix, in order to match the input row and column totals (or marginals).

3. *Mode choice*: This step computes the share of origin-destination (OD) trips that use each available transportation mode. This step is dominated by discrete choice models [16, 15, 59, 27], which model choices between discrete alternatives using an economic utility-maximization framework. Here, a decision maker selects the alternative (in this case mode) with the highest utility among all the alternatives in the choice set. The utility of an alternative is modeled as a function of the characteristics of traveler (i.e. socioeconomic characteristics) and the characteristics of each travel mode (i.e. levels of service such as time and cost). Logit models, which can take on a nesting structure to model hierarchy in the decision-making process, are most commonly used to compute mode shares.

4. *Trip assignment*: This step allocates origin-destination trips of a given mode to a particular path or route. Route choice can be modeled as a deterministic choice (i.e. shortest path or minimum generalized cost), or a stochastic choice (i.e. discrete choice using a logit model). Moreover, modeling approaches can use non-equilibrium, heuristic assignment methods (i.e. all-or-nothing or incremental traffic assignment)

or equilibrium methods (i.e. user optimal or system optimal). Lastly, traffic assignment can be dynamic (DTA) which uses an equilibrium approach at fine-grain temporal intervals (i.e. real-time or pseudo real-time) [58], or static assignment, which assumes fixed-demand typically at the interval of an hour. All of these assignment methods incorporate the relationship between volume, capacity, and travel time using volume-delay functions such as the Bureau of Public Roads (BPR), conical congestion function, and Akcelik flow delay function.

As we use iterative traffic assignment (ITA) due its computational advantages in subsequent chapters, we present more detail on the limitations of this traffic assignment heuristic. ITA is a static, non-equilibrium method which assigns batches (e.g. 40%, 30%, 20%, 10%) of trips serially and updates costs between increments, an improvement over all-or-nothing assignment. However, it does not represent the optimal traffic assignment outcome. The smaller the increments, the closer this method is to user equilibrium algorithms, and the closer the solution is to the Wardrop principles [88], or Nash Equilibria, where in the final traffic conditions, no driver has an incentive to change their route as all driver paths minimize their total travel time.

Despite its wide-use, the four-step model has several limitations, including:

1. demand is modeled for trips rather than activities, which governs trip-making in reality; and

2. trips are the unit of analysis, therefore interdependence of trips cannot be captured; and

3. aggregating trips by discrete spatial, temporal, and demographic characteristics introduces errors; and

4. the sequential nature of the four-step procedure does not enable interdependence of these choices.

These limitations led to the development of activity-based models, which use a schedule-based approach illustrated in Figure 1-1. Under this method, travel demand is derived from demand for activities rather than trips themselves, sequences of trips

are modeled as interdependent tours, and activity and travel scheduling constraints impact travel decisions [7, 18, 69, 83]. Activity-based models replace the trip generation and distribution of four-step models, instead modeling the number, purpose, and sequence of tours. First, the model predicts activity patterns, including a primary tour type and the number and purpose of secondary tours. Next, the model estimates the timing, destination, and mode of primary tours, followed by that of the secondary tours. Despite its benefits, activity-based models have much larger choice sets and are therefore computationally burdensome, and are still unable to completely represent schedules and constraints.

Both four-step and activity-based models rely heavily on survey data for development, calibration, validation. These models combine meticulous methods of statistical sampling in local [31, 76] and national household travel surveys [81, 70] to process and infer trip information between areas of a city. Travel surveys are typically administered by state or regional planning organizations and are integrated with public data such as census tracts and the demographic characteristics of their residents, made available by city, state, and federal agencies. While the surveys that provide the empirical foundation for these models offer a combination of highly detailed travel logs for carefully selected representative population samples, they are expensive to administer and participate in. As a result, the time between surveys range from 5 to 10 years in even the most developed cities.

The estimates these surveys and models produce are critically important for understanding the use of transportation infrastructure and planning for its future[86, 79, 54, 51, 43, 42, 41, 52, 25, 13]. As data becomes better and more widely available, models continue to improve, and computational resources increase, they are increasingly useful and accurate tools for transportation planning applications. Modeling methods are increasingly moving from aggregate to disaggregate models and from static to dynamic procedures, as with more detailed agent-based, microsimulation models. Such trends also support more detailed representation of behavior, capturing more heterogeneous populations and preferences, and complex, interdependent trip-making.

Given the complexity of travel demand estimation procedures, several widely-used commercial software packages exist to implement these models. The vast majority of travel demand models used in practice are implemented in TransCAD, Cube, or Emme. These three software platforms consist of standard GIS capabilities as well as built in functions supporting travel demand forecasting, including four-step and activity-based modeling procedures. These proprietary software platforms are updated to incorporate state-of-the-art methods and use menu-based graphical user interfaces (GUIs) for ease of use, but are expensive and are considered by some to be *black boxes*, with their inner workings unknown to most users.

1.2.2 Urban computing

The interdisciplinary field of urban computing has emerged in recent years, developing computational methods focused on supporting livable, efficient, and sustainable cities for generations to come [47]. Central to this research are estimates of where, when, and how people move within a city and use its facilities. Here, we focus on methods of urban computing as they relate to human mobility patterns.

Given the heterogeneity of urban populations, as well as the immense number of activities and spatial and temporal options in which to perform to these activities, mobility estimates have proved difficult to attain. Moreover, stochasticity is present not only in individuals' choices of locations, times, and activities, but also their travel modes, routes, and trip sequences used to perform these activities. Despite this complexity, however, researchers have found that human mobility can in fact be characterized by regularity and preferential attachment, enabling the development of models to predict mobility patterns [78, 77, 38, 40, 20].

At the same time, technological advances such as increased storage capacity and cloud computing has made it possible to capture petabytes of data from individuals worldwide, from internet usage, credit card transactions, GPS-equipped vehicles, and transit smart cards. These data streams produce massive amounts of time-stamped location data saved in real-time [30, 21, 34]. However, these data pale in comparison to that produced by mobile phones around the globe. Given the frequency of use and

penetration of these ubiquitous devices, they serve as effective sensors of our daily movement.

Mobile phones provide digital footprints of our whereabouts anytime we send a text, make a phone call, or browse the web. Moreover, even when we aren't interacting with our phones, they periodically communicate with cellular network access points and towers. And given the ability to store such data at decreasing costs year after year, such data is increasingly being collected by mobile phone providers. This rich source of data presents new opportunities to understand, infer, and predict human behavior [35]. In developing countries, which often lack reliable data resources such as local and national surveys, mobile phones are a particularly promising source of mobility information. Especially in these contexts, mobility data from mobile phones could be crucial for modeling epidemic spreading, disaster and emergency evacuation response, and effective resource allocation.

Data generated by the pervasive use of cellular phones has offered insights into abstract characteristics of human mobility patterns. Recent work has found that individuals are predictable, unique, and slow to explore new places [38, 20, 32, 78, 77, 24, 23]. The availability of similar data nearly anywhere in the world has facilitated comparative studies that show many of these properties hold across the globe despite differences in culture, socioeconomic variables, and geography. The benefits of this data have been realized in various contexts such as daily mobility motifs [73, 74], disease spreading [12, 89] and population movement [53]. While these works have laid an important foundation, there still is a need to integrate these data into transportation planning frameworks.

Individual survey tracking and stay extraction [8], OD-estimation and validation [22, 60, 87, 46], traffic speed estimation [9, 94], and activity modeling [64, 67] have all been explored using new massive, passively collected data. However, these studies generally present alternatives for only a few steps in traditional four-step or activity based models for estimating travel demand or fail to compare outputs to travel demand estimates from other sources. Moreover, many methods offered to date lack portability from one city to many with minimal additional data collection or calibra-

tion required.

1.3 Outline

The remaining chapters of this thesis present methods and applications using mobile phone data for travel demand estimation.

Chapter 2 presents methods, results, and validation of estimates of origin-destination trips by purpose and time of day using mobile phone data. The results of this chapter are analogous to outputs of the first two steps of four-step travel demand models: trip generation and trip distribution. Chapter 3 follows with methods, results, and validation of estimates of vehicle trips and road usage, analogous to the last two steps of four-step travel demand models: mode choice and traffic assignment. The methods described in both chapters are applied to mobile phone data in Boston, demonstrating the applicability and validity of these methods compared with local and national survey data.

Chapter 4 presents an overview of a portable, efficient software platform built to implement the methods presented in Chapter 2 and 3. This system enables researchers to import mobile phone data sets to produce trip matrices and road usage in any city. Moreover, the platform visualizes these outputs to effectively communicate mobility patterns to planners, stakeholders, and decision-makers. The platform is an alternative to expensive proprietary transportation software packages and built specifically to handle massive mobile phone data sets and additional, open-source data. Results are presented for five cities in the US, Latin America, and Europe, demonstrating the flexibility and extensibility of the platform.

Chapter 5 presents an application of the methods developed in previous chapters. Here, we evaluate the impact of ridesharing on urban traffic. Assuming hypothetical adoption rates of ridesharing, we estimate the number of rideshare vehicles and change in total, network-wide vehicles. Here, we use travel demand estimated in Chapter 2, and modify methods in Chapter 3 to measure the impact of rideshare service on urban congestion in Boston.

Each chapter begins with an introduction, follows with descriptions of data and methods, and concludes with a discussion of results and conclusions. Although each chapter can stand-alone, they reference and build upon ideas and methods covered in previous chapters. Accordingly, each chapter provides context useful for understanding subsequent methods and applications. Finally, we summarize the over-arching results, limitations, and applications of this work in Chapter 6. Appendix A provides pseudo-code describing the algorithms presented in Chapters 2 and 3 as they are implemented in Chapter 4.

Chapter 2

Inferring origin-destination trips by purpose and time of day

2.1 Introduction

The ubiquity of cell phones, along with rapid advancement in mobile technology, has made them increasingly effective sensors of our daily whereabouts [49]. Call detail records (CDRs) from mobile phones contain time-stamped coordinates of anonymized customers, thereby providing rich spatiotemporal information about human mobility patterns. Since CDRs are automatically collected by cell phone carriers for billing purposes, this data can be gathered more frequently and economically than travel survey data collected once (or twice) a decade for transportation planning purposes. Additionally, mobile phone data offers digital footprints at a scale and resolution that may not be captured by surveys that typically record one day of travel diaries per household.

Despite these advantages, mobile phone data lacks information typically available from travel surveys about a respondent (e.g. age or income) or his/her trip (e.g. purpose or mode) [70, 81, 44]. Furthermore, CDRs contain traces of a user at approximated locations when his/her phone communicates with a cell phone tower, providing an inexact and incomplete picture of daily trip-making. Accordingly, much research has focused on developing methods to extract meaningful information about

human mobility from mobile phone traces as well as understanding its limitations.

It has been demonstrated that CDR data can be used to infer origin-destination (OD) trips using microsimulation and limited traffic count data [46]. At the level of the individual, daily trip chains/trajectories constructed from mobile phone data are consistent with household surveys [47, 73]. Further, road usage inferred from the CDR data has been validated against GPS speed data [87] and highway assignment results from a travel demand model [45].

There is still work to be done to explore the usage of phone data to generate trip distributions of different modes, purposes, and times of day. As a step in that direction, this research proposes a methodology to extract OD trips by purpose and time of day from CDR data. This segmentation captures distinct trip-making patterns pertinent for transportation planning applications. Moreover, other than CDR data, the techniques presented in this paper rely only upon nationally-available survey data to allow transferability of the methodology to other study areas in the US.

Extensive research has been conducted into OD estimation, as these trips provide the basis for transportation feasibility and impact studies. Conventional OD estimation approaches rely on surveys and/or travel demand models to provide trip matrices. Often, such trip matrices are calibrated or updated using traffic counts and estimation techniques such as maximum likelihood, generalized least squares, and optimization [79, 25, 13, 93]. This research provides a realistic, cost-effective alternative to these traditional OD data sources and estimation approaches. By presenting a systematic and replicable procedure to extract data relevant to the transportation community, we hope this work will help to facilitate the use of mobile phone data in practice.

In this chapter, we demonstrate methods to analyze mobile phone records for the Boston metropolitan area. In Section 2.2 and Section 2.3, we present an overview of the data and the methods developed to produce OD trips by purpose and time of day. In Section 2.4, we summarize and validate our results against independent data sources for the study area, including the US Census and household travel surveys. Based on these findings, we conclude with a discussion of the limitations and

applications of CDR data in the context of transportation planning and modeling.

2.2 Data Description

The studied dataset contains more than 8 billion anonymized mobile phone records (from several carriers) from roughly 2 million users in the Boston metropolitan area over a period of two months in the Spring of 2010. Although the CDR data spans 60 days, the data provider reindexed the anonymous user IDs for most of the users after the 17th day of the dataset. Effectively, we observe some users for at most 17 days, some users for at most 43 days, and still others for up to 60 days.

Each record contains an anonymous user ID, longitude, latitude, and timestamp at the instance of a phone call or other types of phone communication (such as sending SMS, etc.). The coordinates of the records are estimated by service providers based on a standard triangulation algorithm, with an accuracy of about 200 to 300 meters. In typical mobile phone data sets, locations are represented by cell towers rather than triangulated coordinates and therefore have a lower spatial resolution [77, 87]; however, the method proposed here holds for such cases, as demonstrated in Chapter 4.

2.3 Data Processing

2.3.1 Stay Extraction

The first step to reliably infer activities and trips from CDR data is to filter out noise resulting from (1) tower-to-tower call balancing performed by the mobile service provider, creating the appearance of false movements, and (2) inexact signal triangulation. Furthermore, we wish to distinguish users' stationary stay locations (when/where users engage in an activity) from their moving pass-by locations (when/where users are en-route to activities). To do so, we develop a method based in the work of Hariharan and Toyama [39] for processing GPS traces. The spatial and temporal filtering methods are discussed below and illustrated in Figure 2-1.

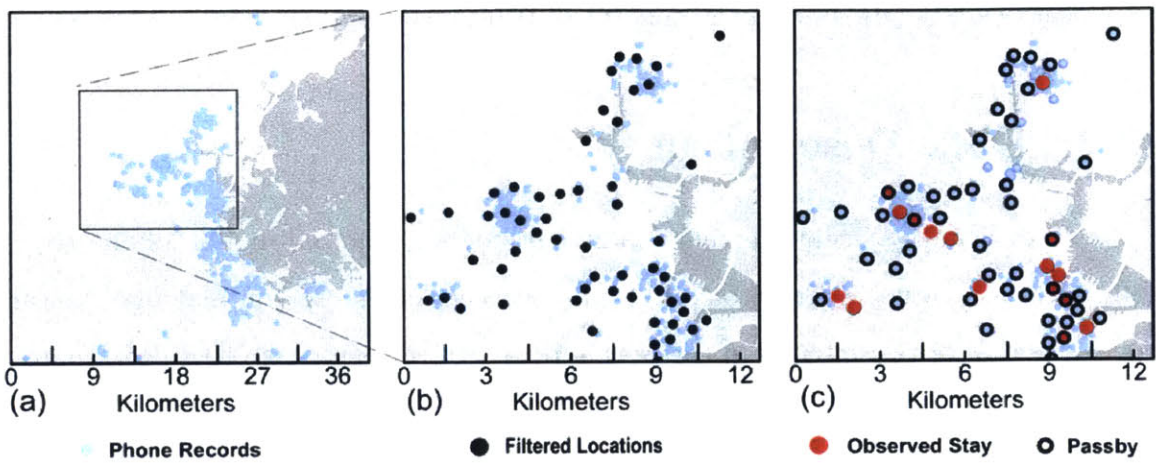

Figure 2-1: Extracting stay and pass-by areas from the phone data for an anonymous user in the 2-month period

Let sequence $D_i = (d_i(1), d_i(2), d_i(3), ..., d_i(n_i))$ be the observed data for a given anonymous user i, where $d_i(k) = (t(k), x(k), y(k))'$ for $k = 1, ..., n_i$, and $t(k), x(k)$, and $y(k)$ are the time, longitude, and latitude of the k-th observation of user i. First, we extract points $d_i(k)$ that are spatially close (i.e. within roaming distance of 300 meters) to their subsequent observations, say, $d_i(k+1), d_i(k+2), ..., d_i(k+m)$. To reduce the jumps in the location sequence of the mobile phone data, we assume that $d_i(k), ..., d_i(k+m)$ are observed when user i is at a specific location, i.e., the medoid of the set of locations $(x_i(k), y_i(k))', ..., (x_i(k+m), y_i(k+m))'$, which is denoted by

$$Med((x_i(k), y_i(k))', ..., (x_i(k+m), y_i(k+m))').$$

This treatment respects the time order at first, to ignore noisy jumps in estimated location, but then disregards time ordering to apply the *agglomerative clustering algorithm* [39] to consolidate 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 required to be no more than a certain threshold (set as 500 meters). Again we modify the observation locations to the corresponding medoids of the clusters (see Figures 2-1a and 2-1b).

Next, we impose the time duration criterion on the clean data, and extract the

stay locations whose durations exceed a certain threshold (set as 10 minutes). In the example presented in the figure we extract 31 distinct stay locations from the 1,776 phone records in the two-month period of the exhibited anonymous user (see Figure 2-1c). The rest of the points are called pass-by points, at which we don't observe any lengthy stays. Note that it is possible that the user stays in some of these pass-by locations as well as locations that we don't observe. In these cases, information about time and location is totally or partially latent to us as we don't observe it from the phone records. However, all the stay locations frequently visited by the user ought to be extracted from the mobile phone data, if the observation period is long enough. As such, the pass-bys are filtered out and the stays are assumed to be trip origins or destinations, between which trips are made. Analysis of the pass-by points is out of the scope of the present work, in which we focus on simple trip chains with origins and destinations labeled as: home, work, or other.

2.3.2 Activity Inference

Trips are induced by the need or desire to engage in activities [65] and therefore understanding patterns and types of activities is crucial in estimating travel demand. It has been demonstrated that human mobility patterns are characterized by regularity with frequent returns to previously visited locations [78, 77, 40]. Due to this predictability, we are able to reasonably infer stay activities for users' most visited locations (i.e. home and work).

Accordingly, our first task is to label the stay regions in order to assign trip purpose. For each user, the stay extraction process detailed above results in a timestamp and duration for each observed visit to a stay location. For this study, we assign an activity type of either *home*, *work*, or *other* to each users' stay locations. Future research can expand the *other* designation to activity types such as school, shopping, recreation, social, etc., using land use information.

Each user's *home* location is identified as the stay with the most visits on (i) weekends and (ii) weekdays between 7 pm and 8 am, representing the time windows in which we expect users to spend substantial amounts of their time at home. In

addition to inferring trip purpose, the *home* stay location of each user is used to filter out users with too few data points and expand the data from phone users to study area population, as summarized in Section 2.3.3.

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 8 am and 7 pm and d is the distance between the latitude-longitude coordinates of the *home* stay and the given stay using plane approximation. This assumption is based on the rationale and historical evidence [50, 72] that for a given frequency of visits, longer distance trips are more likely to be work trips than shorter distance trips, which are more likely to be for non-work purposes (i.e. to the nearby grocery store).

If the user visits the identified *work* stay less than 8 times ($n<8$; once a week, on average) or the distance is less than 0.5 km ($d<0.5$), then the activity of the stay region is identified as *other* rather than *work*. In effect, not all users are assigned a *work* stay, accounting for the fact that not all users commute to a job. Subsequently, all the remaining stay locations not identified as *home* or *work* are designated as *other*. These classification assumptions serve to avoid falsely identifying a location as work that is either not visited frequently enough or close enough to a user's home that it could reflect signal noise rather than a distinct location.

We acknowledge that under these simple assumptions we may misidentify users' *true* home and work locations and, by extension, their trip purposes. However, based on comparisons with census data (presented below) this procedure give us very good estimates of the distribution of home and work locations and home-work flows in our study region. Note that these assumptions are related to the duration and spatial resolution of this dataset, and it may be necessary to adjust them for applications of other datasets.

2.3.3 Filtering and Expansion

For users with too few stay locations, the CDR data may not fully represent their travel patterns. Accordingly, users with fewer than 8 (one per week, on average) visits

to designated *home* stays are filtered out. This filter serves the additional purpose of ensuring with a reasonable degree of certainty that the designated stay is the user's home, a key assumption in our method of upscaling users to population. Note that this filtering process necessarily excludes visitors, for whom a home location is not observed in the studied dataset. Future research could look at extracting visitor trips from CDR data using an assumption other than home location to upscale these trips.

After this filtering, 335,795 users remain in the Boston CDR dataset. This sample size is an order of magnitude larger than in most household travel surveys, and should increase given longer periods of observation. To upscale these users to total population of the study region, the number of *home* stays were aggregated to the 974 Census tracts in the study 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. For the 10 Census tracts with fewer than 10 CDR residents, the scaling factor is set to 0 to ensure that we don't overweight users that are not representative of a given Census tract. The 1st, 2nd, and 3rd quartiles of the expansion factors are 9.4, 14.2, and 25.1, respectively, as illustrated by the tight probability distribution of expansion factors in Figure 2-2a. The spatial distribution illustrated in Figure 2-2b suggests that the tracts in the western portion of the study area tend to be more heavily weighted. CDR data for a period greater than 60 days would likely have lower expansion factors and an improved spatial distribution of users, however, we show that already this limited data set gives reasonable results.

2.3.4 Trip Estimation

With stays for each user designated by activity type and expansion factors to upscale users to population, average daily origin-destination trips can be constructed by time of day and purpose—home-based work (HBW), home-based other (HBO), and non-home based (NHB). This segmentation allows us to capture distinct trip-making patterns and is consistent with segmentation in the trip distribution stage of trip-based travel demand models.

Since the timestamp and duration associated with each stay reflect the *observed*

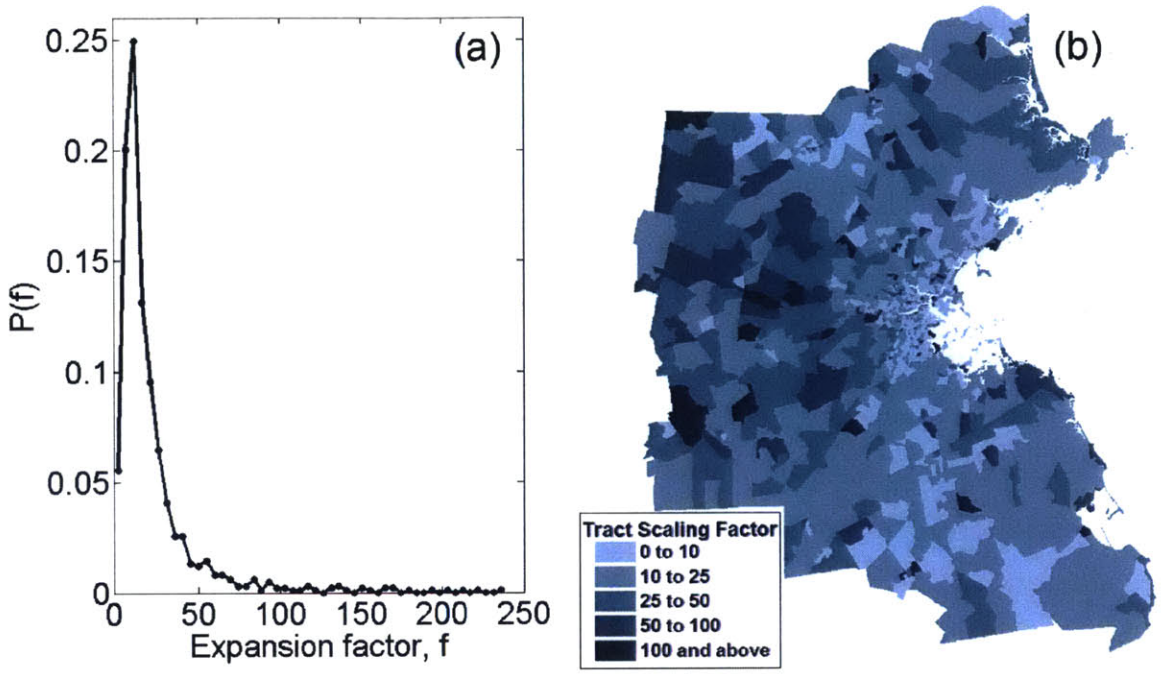

Figure 2-2: (a) Probability distribution of Census tract expansion factors. (b) Thematic map showing the spatial distribution of Census tract expansion factors.

(based on phone usage) rather than *true* arrival time and duration of a user, we infer trip departure hour using probability density functions to account for this uncertainty. The publicly-available 2009 National Household Travel Survey (NHTS) [84], filtered for respondents residing in a consolidated metropolitan statistical area (CMSA) or MSA with populations greater than or equal to 3 million, is a reasonable source as it approximates temporal travel patterns of major US cities comparable to Boston, while allowing for transferability of this methodology to other US cities. Using this departure time data, we generate six hourly distributions for weekdays and weekends and the following trip purposes: HBW, HBO, and NHB.

For each user, it is assumed that a trip is made between two consecutive stays $(i, i+1)$ occurring within a 24-hour period beginning and ending at 3am. The trip occurs at a point in time spanned by the range $[s_i+\delta_i, s_{i+1}]$, where s is the observed arrival time and δ is the observed duration of a stay. The departure hour is randomly generated in this time window using the NHTS distribution that corresponds to the day (weekday, weekend) and the trip purpose identified from the origin and destination stay activities

(HBW, HBO, NHB).

Furthermore, it is presumed that a user starts and ends each 24-hour period at home such that if a user is not recorded at his/her *home* stay for the first (last) record of the 24-hour period, his/her first (last) trip begins (ends) at his/her *home* stay. The first (last) trips are assumed to occur at point in time spanned by the range $[3AM, s_{i+1}]$ ($[s_i + \delta_i, 3AM]$), where s is the observed arrival time and δ is the observed duration of a stay. As before, the departure hour is randomly generated in this window using the NHTS distribution that corresponds to the day (weekday, weekend) and the trip purpose based on the destination (origin) stay activity (HBW, HBO).

Through this process, we construct trips on all days we observe each user. The frequency of weekday observations per user is illustrated in Figure 2-3. The distribution of total weekday trips per user is shown in Figure 2-3a, with first, second, and third quartiles of 33, 58, and 96 trips, respectively. The reindexing of anonymous user IDs mentioned previously in Section 2.2 is evident in the two peaks of the distribution of the number of weekday days we observe each user, as seen in Figure 2-3b. Despite this reindexing, we achieve a sufficiently large number of observation days per person, with first, second, and third quartiles of 11, 17, and 21 days, respectively. Dividing each user's total weekday trips by his/her total weekday days, we get the distribution of average weekday trips shown in Figure 2-3c. The distribution has a long tail, however, the first, second, and third quartiles are 2.6, 3.2, and 4.3 average trips per weekday, respectively, demonstrating that the vast majority of users have a reasonably small number of daily trips.

In order to obtain average daily OD trips, each user's trips are multiplied by the expansion factors described in Section 2.3.3 for the user's *home* Census tract and divided by the number of days from which we constructed the user's trips. For users assigned a *work* stay, weekday trips are only constructed on days in which the user is observed at his/her *work* stay to ensure we capture representative weekdays of commuters. Unlike traditional travel surveys which ask a respondent details about one or a few recent days, this method has the advantage of capturing many days per

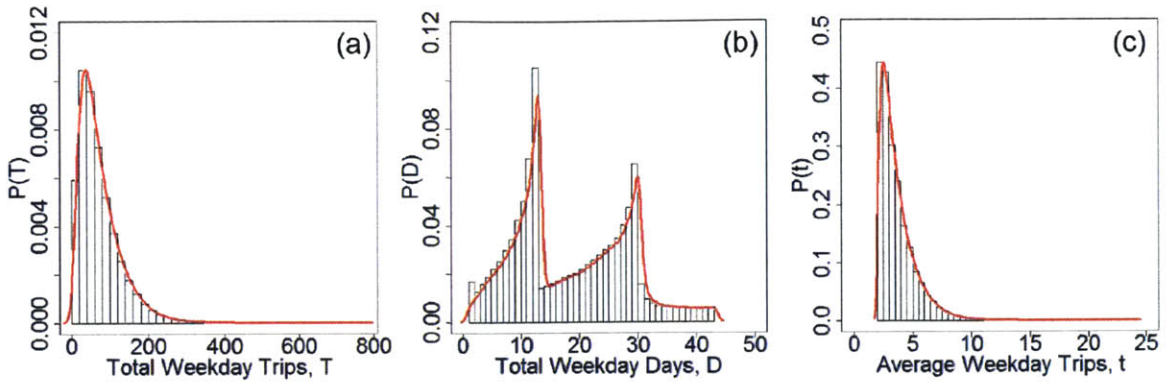

Figure 2-3: Frequency of weekday observations per user. (a) Probability distribution of total weekday trips per user. (b) Probability distribution of total weekday days per user. (c) Probability distribution of average weekday trips per user.

user and thus variations in his/her daily travel behavior. Lastly, each user's average daily trips are aggregated into Census tract pair trip matrices by day type (weekday, weekend), purpose (HBW, HBO, NHB), and hour of departure.

2.4 Results and Validation

2.4.1 Productions and Attractions

Accurately extracting and upscaling users' stays is crucial to trip generation. Due to the regularity of human behavior [78, 77, 40], we are able to infer users' *home* and (if applicable) *work* stay locations from CDR data. For this dataset, we find that we can reasonably represent the spatial distribution of home and work locations when aggregated to the 164 study area cities and towns [55]. Refer to Section 2.4.2 below for more information on the impact of aggregation level on accuracy. Figure 2-4a shows a comparison of home locations by town from 2010 Census data and the raw and upscaled CDR data.

As we would expect since tract population was used to upscale the data, the number of residents in each town is almost identical to that of the upscaled CDR data. However, the slope of a best-fit line through the raw CDR data is close to 1, which speaks to the fact that the overall distribution of raw CDR users is fairly representative and a simple factoring method is in fact appropriate to expand the phone

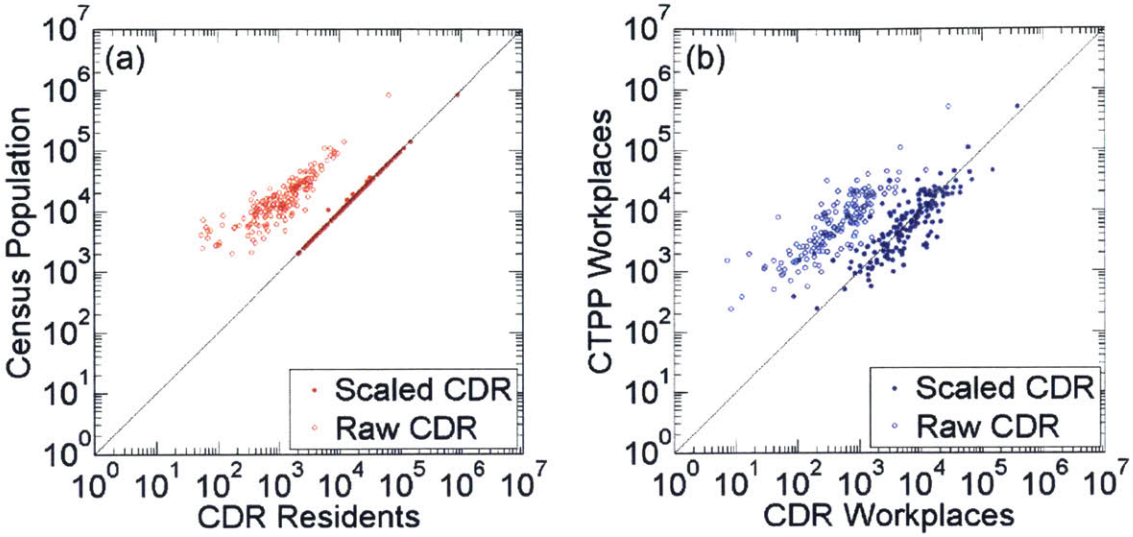

Figure 2-4: (a) CDR residents vs. 2010 Census population by town before and after population expansion. (b) CDR vs. Census Transportation Planning Products (CTPP) [85] workers by town before and after population expansion.

users to population. Similarly, Figure 2-4b shows a comparison of work locations aggregated by town. As with the raw CDR data on the home-end, the distribution of raw workplaces is fairly consistent with the 2006-2010 Census Transportation Planning Products (CTPP) [85] data (slope approximately 1), and the upscaling method adjusts well for the difference in magnitude. This strong correlation is noteworthy considering that each users' *home* and *work* locations were scaled based on their *home* location only.

2.4.2 Trip Distribution

With the establishment of reasonable distributions of trip productions and attractions, we next validate the distribution of trips using two local surveys. The 1991 Boston Household Travel Survey (BHTS) contains information on 39,300 trips made by 3,737 households [17], while the 2010/2011 Massachusetts Travel Survey (MHTS) contains data on 153,099 trips made by 32,739 people [61]. We find that the CDR trips compare well with trips from these data sources by time of day and purpose. Figure 2-5 illustrates the distributions of hourly departure times for (a) HBW, (b) HBO,

Source	HBW	HBO	NHB	Morning 6a-9a	Mid-day 9a-3p	Evening 3p-7p	Rest-of-day 7p-6a
CDR	18%	51%	31%	16%	27%	27%	30%
BHTS	20%	48%	32%	18%	32%	33%	17%
MHTS	12%	49%	39%	21%	34%	33%	12%
NHTS	14%	55%	30%	19%	37%	31%	13%

Table 2.1: Average weekday trip shares by purpose and period from CDR data, 1991 Boston Household Travel Survey (BHTS) [17], the 2010/2011 Massachusetts Travel Survey (MHTS) [61] and the 2009 National Household Travel Survey (NHTS) [84].

(c) NHB, and (d) total average weekday trips. Note that we also benchmark against the NHTS departure time distributions, which were used to infer departure time for the CDR trips. Accordingly, differences between each of the hourly NHTS and CDR distributions reflect the observed arrival and duration times of CDR stays.

Most notably, there are consistently more CDR trips in the late night hours than that of the surveys. While this may be due to a slight mismatch between the frequency of calling and trip-making throughout the day, it may also highlight an advantage of CDR data to capture late night trips not typically reported in survey responses of an average day. Regardless, most transportation planning applications focus on trips in the morning and evening peak periods, when congestion is most prevalent, and for which we compare well. Similar trends are evident for average weekday trip shares segmented by key time periods, as presented in Table 2.1.

Furthermore, the relative share of average weekday trips for each trip purpose is comparable for the CDR and survey data. Table 2.1 shows that the shares of HBW, HBO, and NHB CDR trips are within the ranges of trip purpose shares across all three surveys. This again suggests that our inferences of *home*, *work*, and *other* activities, as well as their relative prevalence in the data set, seem reasonable.

To draw comparisons on the magnitude of daily CDR trips, we MHTS data, which includes weights to expand respondents to population estimated from the 2006-2010 American Community Survey [61]. Table 2.2 shows a comparison of average weekday

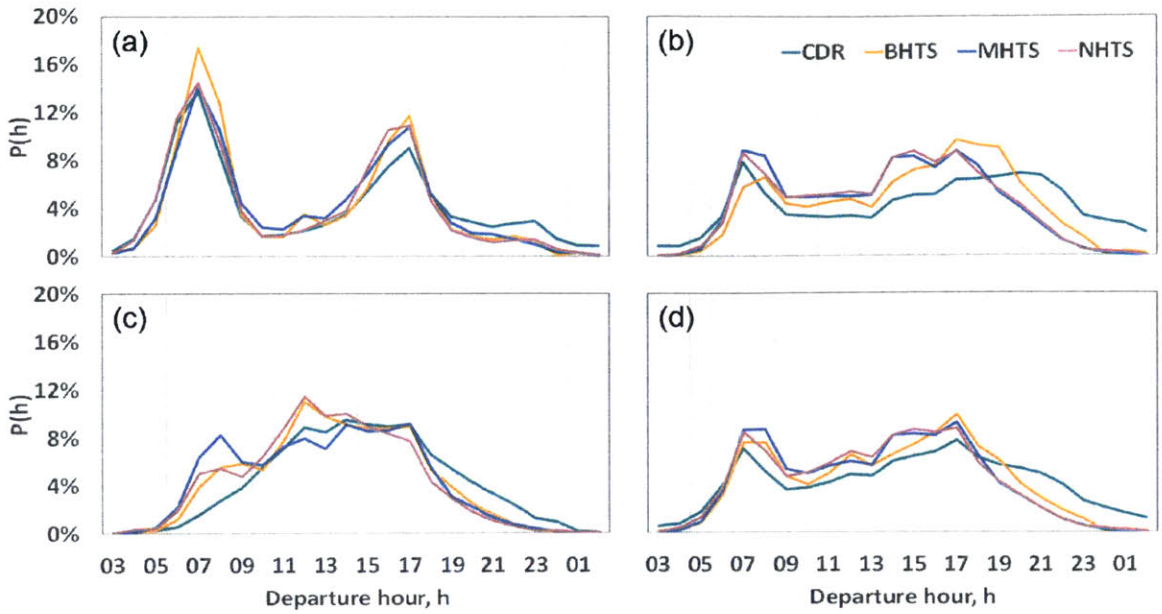

Figure 2-5: Distribution of average weekday hourly departure time from CDR data, 1991 Boston Household Travel Survey (BHTS) [17], the 2010/2011 Massachusetts Travel Survey (MHTS) [61], and 2009 National Household Travel Survey (NHTS) [84] for (a) Home-based Work Trips, (b) Home-based Other Trips, (c) Non-home Based Trips, and (d) All Trips.

trips by purpose and period of the day for the CDR trips and weighted MHTS trips. The survey reports more daily trips than we observe in the CDR data, with most of the difference coming from the NHB trip segment. Still, the total CDR and MHTS trips imply reasonable numbers of average weekday trips per person–3.50 and 4.24, respectively.

Lastly, Table 2.2 presents a comparison of the spatial distribution of daily CDR and MHTS trips at the tract-pair and town-pair level. The correlation coefficients of the trip matrices improve significantly with aggregation to the 164 study area cities and towns. In particular, the HBW and AM correlations at the tract-pair level see the largest improvement. This may be indicative of the role of the size of tracts, which are considerably smaller in downtown Boston where many of the morning commute trips end. We discuss the relationship between aggregation level and correlation in more detail in Section 2.4.3 below.

	HBW	HBO	NHB	AM 6a-9a	MD 9a-3p	PM 3p-7p	RD 7p-6a	Total
CDR Trips (in Millions)	2.81	7.84	4.73	2.46	4.12	4.15	4.65	15.37
MHTS Trips (in Millions)	2.14	8.99	7.18	3.99	6.24	6.06	2.31	18.61
Tract-pair Correlation	0.30	0.64	0.58	0.42	0.65	0.54	0.40	0.58
Town-pair Correlation	0.96	0.97	0.98	0.97	0.98	0.97	0.96	0.98

Table 2.2: Average daily trips by purpose and period from CDR data and the 2010/2011 Massachusetts Travel Survey (MHTS) [61], as well as the correlation coefficients of CDR and MHTS tract-pair and town-pair trips.

Source	Daily HBW Trips, Millions	Inter-tract Share, %	Inter-town Share, %	Average Trip Length, Miles
CDR	2.11	94	68	9.67
Census	2.10	90	68	10.72

Table 2.3: Comparison of average weekday HW CDR and 2006-2010 CTPP [85] flows.

2.4.3 Home-Work Flows

Commuting trips represent a key travel market and source of daily roadway congestion, and accurately representing these trips is an important step in validating trips estimated from CDR data. Accordingly, we next compare with flows between people's home and work locations, as reported by the 2006-2010 Census Transportation Planning Products (CTPP) [85]. Distinct from the average daily HBW trips compared in Section 2.4.2, these flows simply link home and work, ignoring that people's daily trip chains may in fact include work trips to/from locations other than home.

Table 2.3 summarizes statistics that support the comparison of CDR and CTPP home-work (HW) flows. In addition to the total magnitude of trips, the similarities between the percentages of inter-tract and inter-town flows and average trip length give a high-level indication that the distributions of HW flows are similar.

At the flow level, we find that the correlation between CDR and CTPP HW tract-to-tract and town-to-town flows is 0.45 and 0.99, respectively, indicating that the level

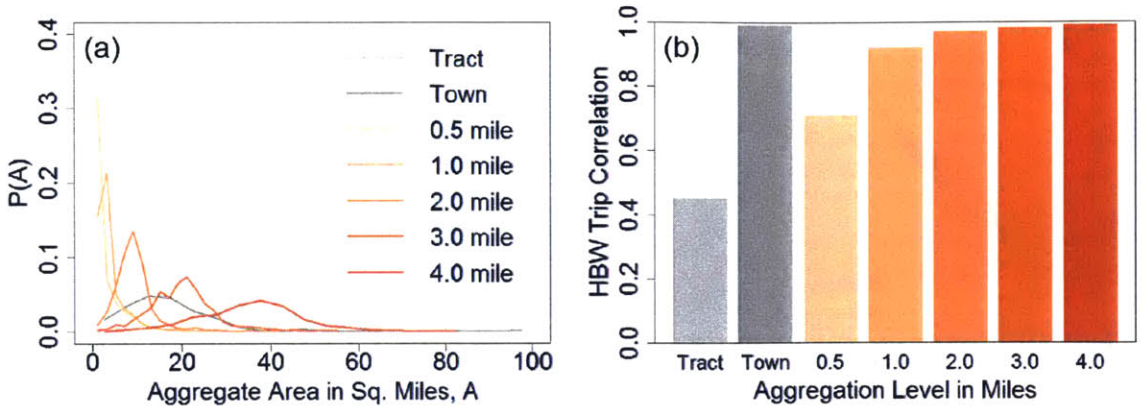

Figure 2-6: (a) Probability density distributions of aggregation area size by designated areas (tract or towns) and variable buffers. (b) Correlation between HW CDR and 2006-2010 CTPP [85] flows corresponding to the aggregation levels in (a).

of aggregation of trips has a significant impact on accuracy. We demonstrate that as we gradually increase average aggregation size using variably-sized buffers around each origin and destinationtTract (Figure 2-6a), the correlation between CDR and CTPP HW trips increases as well (Figure 2-6b). We find that using small aggregation buffers has the most significant impacts on correlation, while having minimal influence on average aggregation size (as illustrated by the fact that the distribution for the 0.5 mile buffer obscures that of the tract-level aggregation in Figure 2-6b). In effect, using a 0.5 mile buffer aggregates the small, dense tracts (i.e. 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 can inform suitable distance thresholds for trip clustering to overcome limitations of sparse data and/or spatial inaccuracy.

We further investigate comparisons of the data sets using town-pairs flows. Figure 2-7a shows the CDR and CTPP HW flows for all of the intra-town and inter-town pairs, which have correlations of 0.99 and 0.95, respectively. It is evident from Figure 2-7a that town pairs with many trips validate better than those pairs with few trips, especially those with fewer than about 500 daily trips. This trend is likely due to sparsity in data for these smaller markets. Figure 2-7b and Figure 2-7c illustrate spatially the HW flow distribution for key markets (inter-tract pairs with greater than $1,000$ daily trips) for the CDR and Census data, respectively. Inspecting the

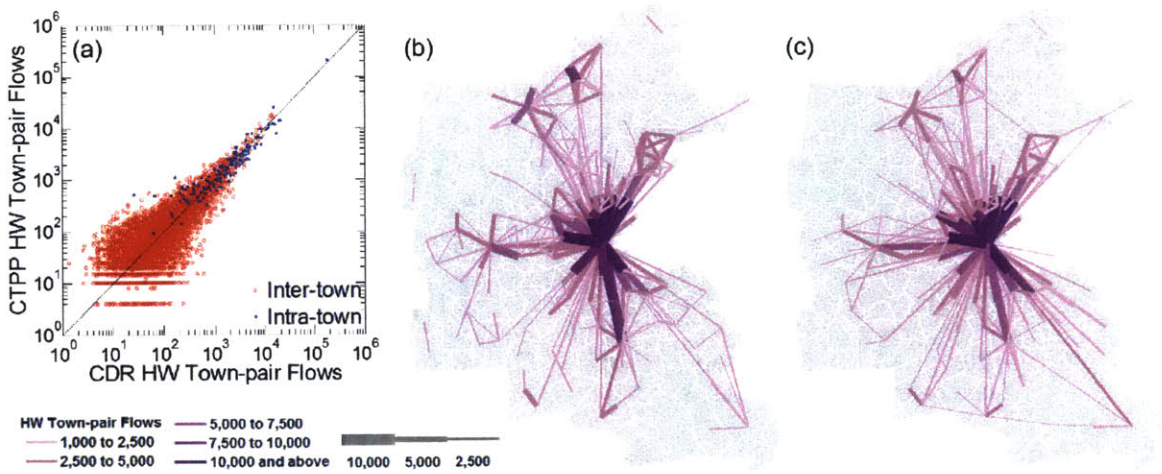

Figure 2-7: (a) Intra-town and inter-town pair daily HW CDR flows and 2006-2010 CTPP [85] flows. (b) Spatial distribution of daily inter-town HW CDR flows (>1,000). (c) Spatial distribution of daily inter-town HW 2006-2010 CTPP [85] flows (>1,000).

figure, it is evident that the CDR data captures very similar patterns to that 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 the suburban towns.

2.5 Conclusions

In this chapter, we detailed steps necessary to extract average daily origin-destination trips by purpose and time of day from mobile phone call detail records (CDRs). The proposed techniques were applied to CDRs in the Boston metropolitan area and validated against local and national surveys. The methods are transferable to other study areas and could be reproducible by researchers and practitioners using mobile phone and census data.

Emphasizing the importance of data preprocessing, much of the methods serve to filter out noise and extract accurate travel patterns representative of the study area. While this processing reduces the immensity of the CDR data, we are left with a sample size that is an order of magnitude larger than most household travel surveys. Further, we observe many days per user, allowing us to capture variation in daily behavior, including weekends, not typically reported household travel surveys.

We find that the size of the areas used to aggregate trips is a very important factor in how well the CDR and survey data compare. We observe significantly higher trip correlation when aggregating origins and destinations to 164 cities and towns rather than the 974 Census tracts in the study area. This improvement in accuracy is seemingly an effect of aggregating small Census tracts (i.e. in the city center), for which CDR data may not have a sufficiently-large sample size or the necessary spatial accuracy. In general, aggregating trip origins and destinations to areas greater than 1 square mile produces agreement with survey data. As mobile phone providers collect more dense data such as GPS traces or wifi access points, spatial and temporal data sparcity will decrease, and accordingly, aggregation size can decrease relative to a given level of precision. Although we can reasonably represent average daily activity and trip patterns with CDRs, data limitations preclude its use in applications requiring richer data such as real-time, dynamic OD estimation.

Aggregating to towns results in similar distributions of upscaled home and work locations inferred from the CDR data and the home- and workplace-based tabulations from the 2006-2010 US Census Transportation Planning Package (CTPP) [85]. Additionally, our inferred distributions of trips by hour of the day and purpose are comparable with the 1991 Boston Household Travel Survey [17], 2010/2011 Massachusetts Travel Survey [61], and the 2009 National Household Travel Survey [84] (filtered for trips in MSAs and CSAs with populations greater than 3 million). Finally, the spatial distribution of home-work flows is highly correlated with that of the CTPP, a well-established nation-wide source for tract-to-tract commuting data.

In validating OD trips by purpose and time of day, we demonstrate that CDR data can be effectively used to represent distinct mobility patterns across market segments typically relevant to transportation planning applications. In particular, CDR data can be used to augment or complement traditional survey data, which provides detailed information about a respondent and his/her trip but is more costly and onerous to collect. Transportation models rely heavily on survey data for inputs, calibration, and validation, and CDR data can be a valuable new resource. Furthermore, the outputs of our proposed methodology are analogous to the outputs of the trip gener-

ation and distribution steps of traditional four-step travel demand models. In areas where public transportation is significant, OD matrices developed from CDRs can be post-processed to obtain mode-specific trip tables, equivalent to the mode split step. As such, CDR data can be very useful for planning applications and/or study areas where running such a model is either not feasible or not necessary.

Chapter 3

Estimating vehicle trips and road usage

3.1 Introduction

Understanding demand for transportation infrastructure is crucial for designing accessible, equitable, and sustainable cities. In the US, in particular, most of this demand utilizes private vehicles to move around cities, using a network of highways and local roads. Especially on weekday mornings and evenings when commuters travel to and from work at similar times, many roads become congested as the number of vehicles using a roads exceed its capacity. As a result, travelers lose time and money to traffic delays, and vehicle emissions worsen.

To understand the demand for and impacts of a new transportation facility, transportation planners have typically relied upon travel demand models to estimate vehicle demand and traffic patterns. In particular, with person trips estimated from the first two steps of a four-step travel demand model, the third step estimates mode choice, generating OD matrices for vehicle trips based on the relative attractiveness of all competing modes. The fourth step then allocates vehicle trips to a road network, estimating the route of every vehicle and road segment level of service characteristics such as volume and travel time.

In this chapter, we build on the methods presented in Chapter 2 to estimate vehicle

trips and road usage from OD person trips inferred from mobile phone data. Our approach offers a simplistic approximation of the outputs of a mode choice model using OD commuting shares from the Census to convert person trips into vehicle trips. We then implement a computationally efficient traffic assignment algorithm to route vehicle trips in a road network. By comparing the results of these methods with survey data, we demonstrate the ability of these approximations to represent vehicle trip and road usage patterns in Boston.

3.2 Data

3.2.1 OD trips inferred from mobile phone data

As in Chapter 2, we utilize a mobile phone CDR dataset for the Boston metropolitain area. A summary of the dataset and methodology to extract OD trips is presented in this section, but refer to Chapter 2 for a detailed description.

CDRs in Boston are first converted into clustered locations or *stay points* at which users engage in activities for an observed duration. These locations are inferred to be home, work, or other depending on observation frequency, day of week, and time of day, and represent a user's origins and destinations. Next, we construct trips between two consecutive stay points in a day. Since the arrival time and duration at these locations reflect the *observed* (based on phone usage) rather than *true* arrival time and duration, we probabilistically infer departure hour h using the NHTS survey data on trips in major US cities.

For each user u, we generate trip matrices t_{ij} by summing the number of trips from origin Census tract i to destination tract j. By dividing these trips by the number of days n on which we observed the user, we compute average daily transition probabilities—the probabilities that a user makes a trip between any origin and destination pair ij on an average weekday. Lastly, user trips are multiplied by expansion factors w based on the population of a user's home Census tract. Summing across all individuals', we compute average daily trip matrices T_{ij}, as summarized in Equation

3.1.

$$T_{ij}(h) = \sum_{u=1}^{U} t_{ij}(u,h)/n(u) * w(u) \tag{3.1}$$

3.2.2 OD vehicle trips from the Massachusetts Household Travel Survey

In Chapter 2, average daily tract-pair trips from the Massachusetts Household Travel Survey (MHTS) were compared with the OD trips inferred from CDR data. In this chapter, we now compare outputs of our vehicle trip estimation with vehicle trips reported by the MHTS. Additionally, we can apply our traffic assignment algorithms to the MHTS OD vehicle trips matrices and compare traffic conditions and travel times with that of the OD trips and road usage inferred from CDR data. Further, the MHTS explicitly asked respondents for the travel time of trips, which can be compared with that of the MHTS trips routed using our traffic assignment algorithms to assess the accuracy of this method.

3.2.3 GIS/Survey

- Road network: For traffic simulation, we use a GIS shapefile from the local transportation authority containing road characteristics such as speed limits, road capacities, number of lanes, and classifications [26].

- Census tracts: CDR trips are aggregated to the spatial resolution of 974 study area Census tracts, which contain roughly 5000 residents each. We use a GIS shapefile as well as population estimates from the American Community Survey to expand observed users to total population [1, 2].

- Communities: CDR trips are aggregated to 174 areas (163 study area towns and 11 Boston neighborhoods), referred to as communities in this chapter. Note that in Chapter 2, the aggregation level that we referred to as town consisted of 164 areas–163 towns and Boston. We divided Boston into 11 neighborhoods for

this analysis to calculate mode share at a finer resolution since mode shares vary across Boston depending on transit access. We used a GIS shapefile of town boundaries developed by MassGIS to map Census tracts to these communities, but further split Boston into neighborhoods using local knowledge of neighborhood boundaries [55].

- Census commuting trips: The 2006-2010 Census Transportation Planning Products (CTPP) Part 3 provides commuting characteristics between Census tract pairs [85]. This nationally-available dataset provides tabulations across 16 different travel modes, which we use to infer mode shares.

3.3 Vehicle trip estimation

Capturing spatial variation in mode shares is essential to estimating reasonable distributions of vehicle trip patterns from the person trips inferred from CDR data. To infer travel mode, we use the CTPP commuting data aggregated from Census tract pairs ij to community pairs IJ in order to minimize the effects of matrix sparsity and small sampling size. We compute (i) drive-alone and taxi mode share d_{IJ}, (ii) carpool mode share c_{IJ}, and (iii) non-auto mode share o_{IJ}. These mode shares are mutually exclusive and exhaustive, such that $\sum(d_{IJ} + c_{IJ} + o_{IJ}) = 1$. Community pairs with sampling issues, including few total trips or zero auto trips, are assigned average mode shares depending on their geography: Urban-Urban, Suburban-Urban, or Suburban-Suburban. Communities lying within Boston's I-95 highway ring are designated as Urban, and all other communities are designated Suburban.

For every tract-pair and hour, we use the community-pair mode shares to compute the number of total vehicles V_{ij} (Equation 3.4). Total vehicles are comprised of single-occupancy vehicles $v_{d,ij}(h)$ computed by multiplying tract-pair ij person trips by the drive-alone mode share of the corresponding community-pair IJ (Equation 3.2). Similarly, carpool vehicles $v_{c,ij}(h)$ are computed using the carpool mode share divided by the average vehicle occupancy of study area carpools p ($p = 2.18$ in Boston; Equation 3.3).

$$v_{d,ij}(h) = T_{ij}(h) * d_{IJ} \tag{3.2}$$

$$v_{c,ij}(h) = T_{ij}(h) * c_{IJ}/p \tag{3.3}$$

$$V_{ij}(h) = v_{d,ij}(h) + v_{c,ij}(h) \tag{3.4}$$

Next, we estimate peak hourly vehicle trips, as is conventionally done in travel demand models before static traffic assignment is performed. To do so, we compute peak hourly factors as the ratio of the maximum number of vehicle trips in a time period to the total number of vehicle trips in the time period. For the AM period (6 to 9 am), we compute peak hourly factors of 0.438 and 0.419 for the CDR and MHTS vehicle trips, respectively. Similarly, for the PM period (3 to 7 pm), we calculate peak hourly factors of 0.284 and 0.288 for CDR and MHTS vehicle trips, respectively. By comparison, if trips are evenly distributed across the 3-hour AM and 4-hour PM periods, the peak hourly factors would be 0.333 and 0.250, respectively. Lastly, we can compute the peak hourly vehicle trips in a given time period by summing the vehicle trips for each OD pair across all hours in the period and multiplying by the corresponding peak hourly factor.

3.4 Traffic assignment

On most city roads, free-flow speeds are rarely achieved due to congestion. As a result, traffic patterns may significantly change the time costs associated with using a particular route. In conventional four-step travel demand models, vehicle trips are allocated to road networks using a traffic assignment algorithm that captures the impact of congestion on travel time and route choice.

In this chapter, we distribute trips on the roadway network using Incremental Traffic Assignment (ITA) [3, 63]. Our ITA algorithm assigns trips in a series of increments and updates the costs of edges in the network based on the number of vehicles that were previously assigned to that road between increments. For example,

the first increment assigns 40% of trips for each pair assuming each driver experiences free-flow speeds. The travel time cost associated with every road segment is then adjusted based on how many drivers were assigned to that road and the total number of cars a road can accommodate in unit time. The next 30% of drivers are then routed in the updated conditions. This process is repeated until all users have been assigned a route.

Although this incremental approach allows us to capture the impact of congestion on travel times of each subsequent batch, once a driver has been assigned a route it does not change. Consequently the approach does not converge to Wardrop's equilibrium even for very small increment sizes. Despite its shortcomings, however, ITA is attractive for its ease of implementation. A more detailed description of ITA and other static and dynamic traffic assignment methods are discussed in detail in the literature review in Section 1.2.1.

For traffic assignment, we also utilize centroid connectors to distribute trips to road segments. This method, frequently used in travel demand models, assumes that trips begin and end at the geographic centroid of their origin and destination zones. For every zone, one or more links is added from each centroid to a nearby road segment intersection. In effect, OD demand is routed along a road network first by a centroid connector from the origin centroid and finally by a centroid connector to the destination centroid. Centroid connectors can therefore be thought of as proxy links for local roads that feed into the more major roads represented in a road network, but that are not themselves represented. Accordingly, we route tract-pair trips using centroid connectors that are given low speeds (10 mph; to add time representative of that needed to get in and out of a neighborhood, for example) and unlimited capacity (so that congestion is not introduced during assignment). Note that intra-tract trips (trips beginning and ending within the same tract) are not routed under this method.

Relating travel performance to traffic conditions has been a long standing problem in transportation. Many different characterizations exist, ranging from conical volume-delay functions to more complex approaches [19, 80, 4]. One of the most simplistic and common metrics used in determining the travel time associated with a

specific flow level is the ratio between the number of cars actually using a road (volume) and its maximum flow capacity (volume-over-capacity or V/C). At low V/C, drivers enjoy large spaces between cars and can safely travel at free-flow speeds. As roads become congested and V/C increases, drivers are forced to slow down to insure they have adequate time to react. Based on the volume-over-capacity (V/C) for each road, costs are updated according to Eq. 3.5. The Bureau of Public Roads's (BPR) default guidelines use $\alpha = 0.15,\ \beta = 4$[1].

$$t_{current} = t_{freeflow} \cdot (1 + \alpha(V/C)^{\beta}) \tag{3.5}$$

As often done in traffic assignment modeling, we modify the default coefficients of the BPR function in order to better represent local traffic patterns. By comparing the reported travel time from the MHTS survey with that of MHTS vehicle trips assigned using ITA, we select $\alpha = 0.85$ (compared with the default value of 0.15), but maintain the default value for β (4). $\alpha = 0.85$ is in line with transportation literature, which typically increases the value of α for highways and major roads. Lastly, we underestimate total travel time using ITA without applying time penalties for intersection and traffic light stops and queues. To account for such delays, we add two minutes to all travel times.

Although simplistic, total travel times after these adjustments correspond well to the travel times reported in the MHTS survey, as shown in Figure 3-1. Figure 3-1a and Figure3-1b compare the distributions of AM and PM peak travel times for the MHTS tract-pair vehicle trips estimated using both ITA and User Equilibrium (UE) assignment methods with the average travel time reported by respondents in the MHTS survey. Some of the differences between the reported and assigned MHTS trips may be due to imperfect and/or biased recollection. For example, the peaks at 30, 45, and 60 minutes are likely to be caused by the fact that many people approximate and report travel times at these key 15-minute increments. However, differences are also due to traffic assignment itself, which merely approximates average traffic flow

[1] Travel Demand Modeling with TransCAD 5.0, User's Guide (Caliper., 2008).

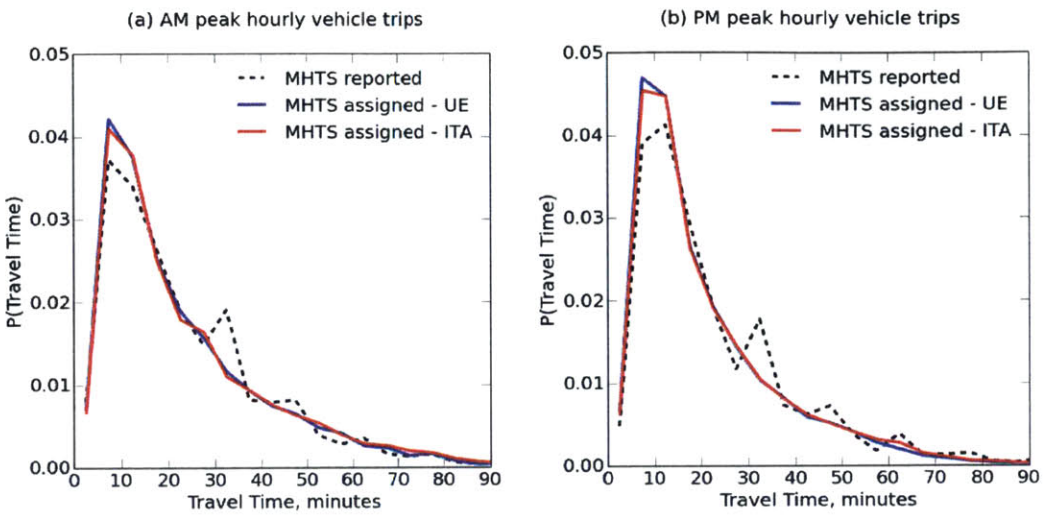

Figure 3-1: Travel time (in minutes) distribution of assigned CDR and MHTS trips, as well as average travel time as reported in the MHTS survey for (a) AM and (b) PM peak hourly vehicle trips.

for static hour-long periods. We see, however, that despite ITA not approaching an optimal equilibrium solution, it produces results very similar to the UE method. Therefore, whether or not these observed differences are due to response bias or assignment error, in general, traffic assignment–and ITA in particular–is a suitable approximation of road usage.

3.5 Results and validation

In order to validate both our methods of converting person trips to peak hourly vehicle trips and allocating these vehicle trips to road segments, we compare with that of the Massachusetts Household Travel Survey (MHTS). We first compare and contrast aggregate statistics between the CDR and survey data, indicative of the validity of our methods at a high level. We then go a step deeper, evaluating travel time distributions for tract-pair OD trips. Lastly, we look at the properties of road segments themselves, which provide insights into congestion levels resulting from our traffic assignment methodology.

	AM 6a-9a	MD 9a-3p	PM 3p-7p	RD 7p-6a	Total
CDR trips (in Millions)	1.85	2.99	3.07	3.44	11.36
MHTS trips (in Millions)	1.79	3.18	2.90	1.23	9.09
Community-pair correlation	0.87	0.88	0.87	0.81	0.89

Table 3.1: Average daily vehicle trips by period from CDR data and the 2010/2011 Massachusetts Travel Survey (MHTS) [61], as well as the correlation coefficients of CDR and MHTS community-pair trips.

3.5.1 Vehicle trips

Among other information, MHTS respondents, who are associated with expansion factors in order to be representative of the entire population, reported the departure hour, mode, and travel time of recent trips. For this analysis, we aggregate expanded vehicle trips to the same time periods used in Chapter 2: morning (AM, 6a-9a), mid-day (MD, 9a-7p), evening (PM, 3p-7p), and rest-of-day (RD, 7p-6p). Table 3.1 summarizes the total number of vehicle trips for both the CDR and MHTS datasets. The number of vehicle trips in each time period is very similar across the two data sets, except for the rest-of-day period, a trend previously illustrated by the person trip totals in Table 2.2. Moreover, the community-pair correlations between the two data sets are high. Given the agreement in magnitude and distribution, our method to convert person trips to vehicle trips appears reasonable on the aggregate.

The magnitude of total vehicle trips in Table 3.1, however, is noteworthy when you recall the relative magnitudes of total person trips in Table 2.2. Total person trips from the survey are greater than that of the CDR trips, whereas the opposite is true with respect to vehicle trips. 11.36 million vehicle trips implies 0.74 vehicle trips per person trips from the CDR data, while 9.09 million vehicle trips implies just 0.49 vehicle trips per person from the MHTS data.

Figure 3-2a and Figure 3-2b illustrate the correlation between MHTS and CDR vehicle trips spatially for the AM and PM periods, respectively. We see that for community-pairs with higher vehicle trips, the correlation between the CDR and

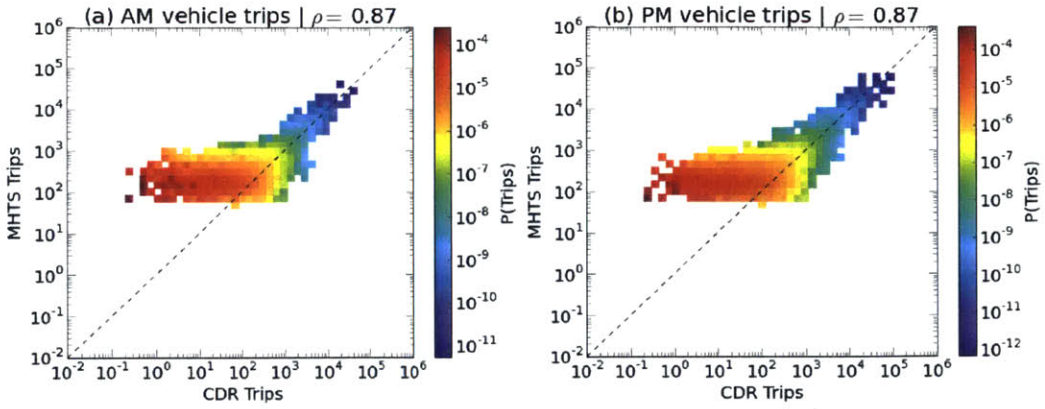

Figure 3-2: 2D histogram of community-pair MHTS and CDR vehicle trips in the (a) morning (AM, 6a-9a) and (b) evening (PM, 3p-7p) peak periods.

survey trips are very similar. Moreover, we see that CDR trips are distributed over many more community-pairs, with many having less than 10 trips. For the MHTS survey, however, the smaller sample size and larger expansion factors results in a much sparser matrix, with all community-pairs having over 10 trips. Despite these differences, for community-pairs with at least 100 trips a day, the two sources have similar vehicle trip matrices.

3.5.2 Road usage

With general agreement between the vehicle trip distributions, we next evaluate the traffic patterns resulting from allocating these trips to a road network. We are able to estimate average OD travel times for the peak AM and PM vehicle trips from the MHTS survey using our traffic assignment algorithm. Comparisons between traffic patterns estimated from the CDR data and that from the MHTS survey data are impacted by differences in magnitude and distribution of trips (as estimated in Chapter 2), in addition to errors due to our methods of approximating vehicle trips and simulating traffic conditions presented in this chapter.

Table 3.2 summarizes the aggregate results from assigning tract-pair AM and PM peak hourly vehicle trips. Note that the vehicle trips listed here are inter-tract only, since intra-tract trips are not assigned to the road network (as described in Section 3.4). The magnitude of peak hourly inter-tract trips are similar for the MHTS and

	Trips (thousands)		Time (minutes)		Distance (miles)	
	AM	PM	AM	PM	AM	PM
MHTS	646.4	717.8	24.3	21.2	8.8	7.8
CDR	687.5	694.9	29.1	22.1	10.1	8.2

Table 3.2: Inter-tract vehicle trips, average travel time, and average distance for the AM and PM peak hours from CDR data and the 2010/2011 Massachusetts Travel Survey (MHTS) [61].

CDR data, as are the average travel time and distances.

Next, we compare the travel times of assigned MHTS and CDR trips at the tract-pair level. Figure 3-3a and Figure 3-3b illustrate the tract-pair travel time correlation across the two data sets in 2D-histograms with bin sizes of two minutes. The correlations between OD travel times for the AM and PM peaks are very high, at 0.96 and 0.98, respectively. Considering that the tract-pair correlations between person trips presented in Table 2.2 are considerably lower, this suggests that assigning trips reduces some of the noise present in the tract-pair trip matrices. For example, the origin tract of a trip may or may not reflect the true origin due to spatial inaccuracy and noise in the CDR data, but given that the true origin is likely to be in a nearby tract otherwise, the travel path and by extension travel time will be similar regardless of spatial errors. Accordingly, the correlation between MHTS and CDR tract-pair travel times is higher than that of the tract-pair trips themselves.

Delving deeper, we next compare the results of our vehicle estimation and traffic assignment at the level of roads. Figure 3-4a and Figure 3-4b illustrate the correlation between MHTS and CDR road segment volumes. The AM and PM peak hourly correlations of 0.92 and 0.91, respectively, again indicate reasonable results, with road segments serving high volumes of trips (e.g. highways) having the highest correlation. In contrast, there is much more variability across road segments serving less demand.

While the MHTS data set does not have trips with weights less than 34.8, the methods we use to extract trips from the CDR dataset allows for trips with magnitudes less than 1. Accordingly, the CDR trip matrix is far less sparse than that of

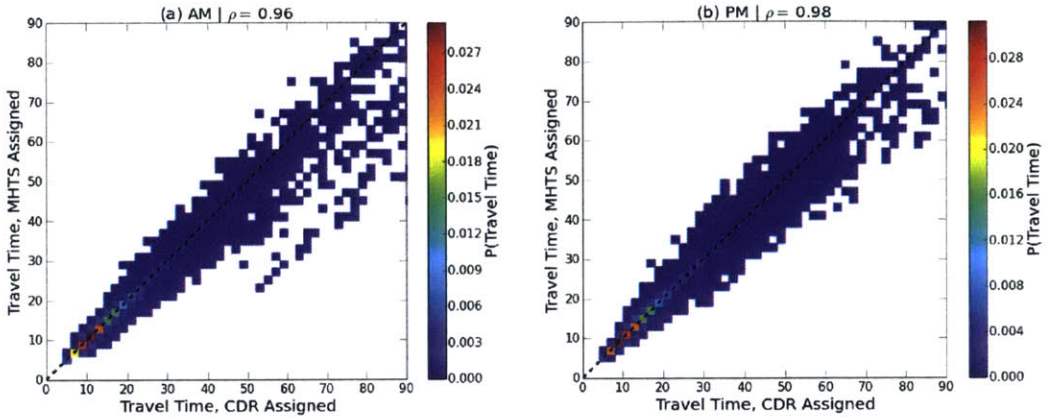

Figure 3-3: Distribution of average tract-pair travel time (in minutes) of assigned CDR and MHTS trips for (a) AM and (b) PM peak hourly vehicle trips.

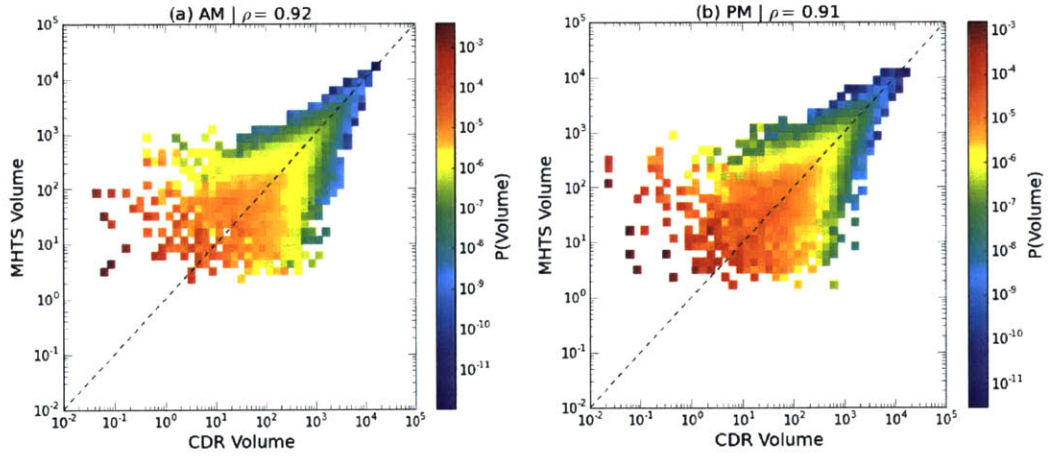

Figure 3-4: Distribution of road segment volumes of assigned CDR and MHTS trips for (a) AM and (b) PM peak hourly vehicle trips.

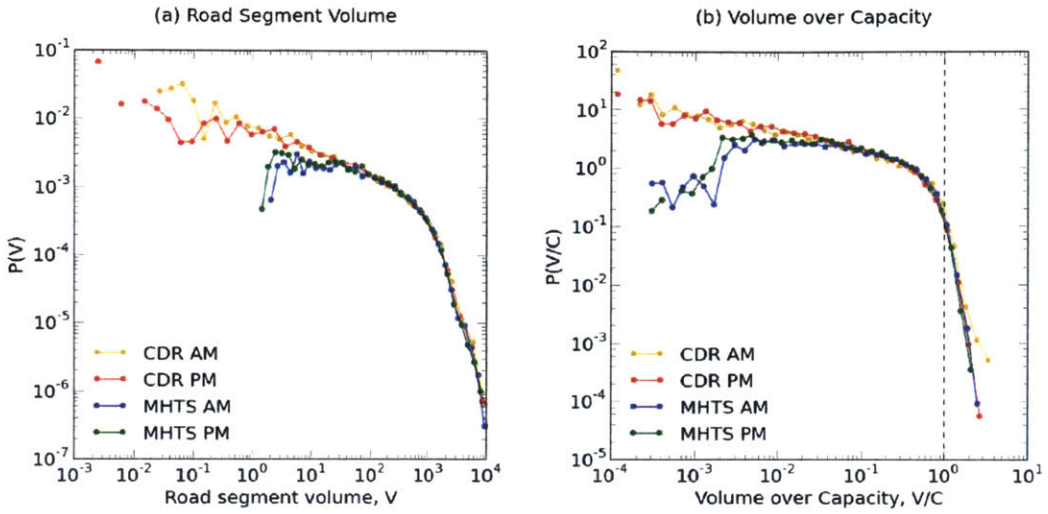

Figure 3-5: (a) Road Segment Volume and (b) Volume over Capacity ratio for MHTS and CDR hourly vehicle trips in the AM and PM peaks.

the MHTS survey, resulting in some road segments having very small CDR vehicle volumes as evidenced in Figure 3-4. Figure 3-5a further illustrates this trend. The probability distribution for road segments having volumes greater than about 10 are very similar across AM and PM peak hours and CDR and MHTS trips. However, for road segments carrying less than about 10 trips, the probability distributions show that very small CDR volumes are observed for many roads, while MHTS volumes are never less than 1. A similar trend is illustrated in Figure 3-5b, with many more road segments having very small volume-over-capacity V/C ratios (e.g. very uncongested). For MHTS and CDR vehicle trips in both the AM and PM peaks, the majority of road segments are uncongested (V/C < 1), and we observe the worst congestion for the AM CDR trips, with a small number of road segments having (V/C > 2).

3.6 Conclusions

Using approaches analogous to the mode choice and trip assignment steps of conventional travel demand models, we demonstrate methods to convert OD trips inferred from mobile phone data to vehicle trips and allocate these vehicle trips to a road network. These methods are validated against vehicle trip patterns and travel times

reported by the Massachusetts Household Travel Survey (MHTS).

Although simplistic, the mode choice approximation allows us to represent the distribution of vehicle trips reasonably well. Future work should consider developing more sophisticated methods for modeling mode choice that incorporate characteristics of trip origins and destinations, such as transit accessibility, auto ownership, and parking costs, as well as characteristics of the trip itself, such as distance and time of day.

Using ITA with a modified BPR volume delay function to perform traffic assignment, we are able to match reported travel times from the MHTS survey reasonably well. Moreover, despite ITA not reaching an optimal routing solution, it produces traffic patterns very similar to a more sophisticated algorithm that converges to equilibrium. With reasonable estimates of road usage and traffic congestion, we can better plan infrastructure, services, and strategies to help mitigate urban traffic and pollution.

Combined with the methods in Chapter 2 to estimate OD person trips, the methods presented in this chapter enable us to perform all four steps of traditional travel demand estimation. These methods therefore provide an alternative to attaining and running proprietary transportation software suites and travel demand models. With these benefits in mind, we hope this reseach helps support the use of mobile phone data for transporation planning applications.

Chapter 4

Integrating travel demand algorithms and big data sources into a portable software platform

4.1 Introduction

The rise of ubiquitous mobile computing has led to a dramatic increase in new, *big data* resources that capture the movement of vehicles and people in near real time and promise solutions to some of these deficiencies. With these new opportunities, however, come new challenges of estimation, integration, and validation with existing models. While these data are available nearly instantaneously and provide large, long running, samples at low cost, they often lack important contextual demographic information due to privacy reasons, lack resolution to infer choices of mode, and have their own noise and biases that must be accounted for. Despite these issues, their use for urban and transportation planning has the potential to radically decrease the time in-between updated surveys, increase survey coverage, and reduce data acquisition costs. In order to realize these benefits, a number of challenges must be overcome to integrate new data sources into traditional modeling and estimation tools.

Here we fill this gap with a modular, efficient computational system that performs

many aspects of travel demand estimation billions of geo-tagged data points as an input. We review and integrate new and existing algorithms to produce validated origin-destination matrices and road usage patterns. We begin by outlining the system architecture in section 4.2.1. In section 4.2.3 we explain our methods of extracting, cleaning, and storing road network information from a variety of sources. We discuss recent advances in OD creation from mobile phone data in section 4.3 and implement a simple, parallel incremental traffic assignment algorithm for these trips in section 4.3.1. We present comparisons of these results to estimates from traditional survey methods in section 4.4.1. Finally, in sections 4.4.2, 4.4.3, 4.4.4 we present a variety of measurements that can be made with the proposed system as well as an online, interactive visualization for conveying these results to researchers, policy makers, and the public. To demonstrate the flexibility of the system, we perform these analyses for five metro regions spanning countries and cultures: Boston and San Francisco, USA, Lisbon and Porto, Portugal, and Rio de Janeiro, Brazil.

4.1.1 Description of Data

Large telecommunications companies, private applications, and network providers collect and store enormous quantities of data on users of their products and services, presenting computational challenges for storing and analyzing them. Billions of phone calls must be processed, data from open- and crowd- sourced repositories must be parsed, and results must be made more accessible to individuals that generated them. At the same time, it is critical that measurements from these new sources are statistically representative and corrected for biases inherent in new data. This process requires integration of new pervasive data with reliable (though less extensive) traditional data sources such as the census or travel surveys. We combine the following data sets to illustrate the capabilities of the system architecture here proposed:

1. *Call Detail Records (CDRs):* At least three weeks of call detail records from mobile phone use across each subject city. The data includes the timestamp

and the location for every phone call (and in some cases SMS) made by all users of a particular carrier. The spatial granularity of the data varies between cell tower level where calls are mapped to towers and triangulated geographical coordinate pairs where each call has a unique pair of coordinates accurate to within a few hundred meters. Market shares associated with the carriers that provide the data also vary. Personal information is anonymized through the use of hashed identification strings. For reference, 6 weeks of CDR data from the Boston area containing roughly 1 billion calls made by 1.6 million unique users consumes roughly 70 gigabytes of disk space in its raw format. In cities with longer observation periods, data size quickly becomes a performance issue.

2. *Census Data:* At the census tract (or equivalent) scale, we obtain the population and vehicle usage rate of residents in that area. For US cities, the American Community Survey provides this data on the level of census tracts (each containing roughly 5000 people). Census data is obtained for Brazil through IBGE (Instituto Brasileiro de Geografia e Estatística) and for Portugal through the Instituto de Nacional de Estatística. All cities analyzed in this work have varying spatial resolutions of the census information.

3. *Road Networks:* For many cities in the US, detailed road networks are made available by local or state transportation authorities. These GIS shapefiles generally contain road characteristics such as speed limits, road capacities, number of lanes, and classifications. Often, however, these properties are incomplete or missing entirely. Moreover, as such road inventories are expensive to compile and maintain, they simply do not exist for many cities in the world. In this case, we turn to OpenStreetMaps (OSM), an open source community dedicated to mapping the world through community contributions. For cities where a detailed road network cannot be obtained, we parse OSM files and infer required road characteristics to build realistic and routable networks. At this time, the entirety of the OSM database contains roughly 4 terabytes of geographic features related to roads, buildings, points of interest, and more.

4. *Survey and Model Comparisons:* Wherever possible, we obtain the most recent travel demand model or survey from a particular city and compare the results to those output by our methods. In Boston, we use the 2011 Massachusetts Household Travel Survey (MHTS) and upscale trips according to standard procedures, in San Francisco, the 2000 Bay Area Transportation Survey (BATS), in Rio de Janeiro, a recent transportation model output provided by the local government, and in Lisbon, the most recent estimates from the MIT-Portugal UrbanSim LUT model that uses the 1994 Lisbon transportation survey as input [36]. We found no recent travel survey or model for Porto.

Table 4.1 compiles descriptive statistics for these data sources for each city we explore in the latter sections of this paper.

Table 4.1: A comparison of the extent of the data involved in the analysis of the subject cities.

	City				
	Boston	SF Bay	Rio	Lisbon	Porto
Population (mil.)	4.5	7.15	12.6	2.8	1.7
Area ($1000km^2$)	4.6	18.1	4.5	2.9	2.0
# of Users (mil.)	1.65	0.43	2.19	0.56	0.47
# of Calls (mil.)	905	429	1,045	50	33
# of cell towers	N/A	892	1421	743	335
# of Edges (ths.)	21.8	24.3	22.7	28.1	15.1
# of Nodes (ths.)	9.6	11.3	22.1	16.1	8.6
# of Tracts	732	1139	729	295	272

4.2 System Architecture and Implementation

4.2.1 Architecture

The system architecture to integrate the data sources above must be flexible enough to handle different regions of the globe which may have different data availability and quality and efficient enough to analyze massive amounts of data in a reasonable amount of time. The proposed system must also be modular, so that components

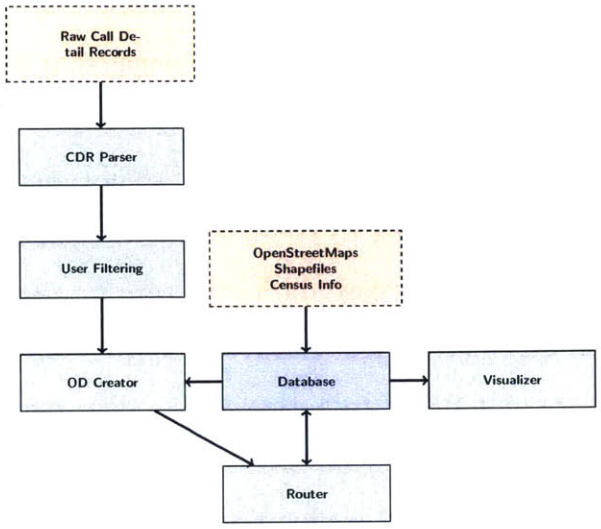

Figure 4-1: A flowchart of the system architecture.

can be updated easily as new technologies and algorithms become available. To meet these requirements, we choose an object-oriented approach with loose schema requirements. A final object is to make results accessible to a range of end users via online, interactive visualization. To satisfy these constraints, we propose the system architecture depicted in Figure 4-1.

4.2.2 Parsing, Standardizing, and Filtering User Data

One of the biggest challenges in parsing and analyzing travel survey data is the incredible variety in data schema, collection, and reporting practices. Each planning organization typically constructs its own set of data codes and definitions and provides data in unique formats. This makes it very difficult to compare surveys done in different cities. Call detail records, on the other hand, are typically available for many cities from the same provider and in the same format, and in most cases, translating between the formats of different carriers is simply a matter of shuffling columns. The first component of our system is a simple architecture to convert all CDR data to a standard format that can be expected by the rest of the components.

Given the size of these data sets and the rapidly evolving schema requirements of new models, choosing the proper data structure is critical. Google's open source

Protocol Buffer library[1] is an ideal choice as they provide fast serialization for speed and space efficient file storage as well as flexible schemas that can be changed without compromising backwards compatibility. These structures were designed to serve some of the largest databases in the world and are more than enough for our task.

We take a user centric approach to CDR data. We define a *user_data* protocol buffer message that will form the core data structure for our custom User class in an object-oriented programing model. Each User object can be assigned a number of attributes such as the number of calls they make, their home and work locations, and mobility characteristics such as the average time between calls or the average distance traveled on each trip. More sophisticated methods can compute the number and distribution of their trips and even expand them based on census information. We define similar structures and classes for OD matrices, trips, and census data. The serialization routines built into the protocol buffer library ensures that storage of raw data is efficient. To analyze a new city, the user only needs to write two simple routines, one to parse a single line of the CDR file and populate relevant user attributes and one to populate census data objects. Standardizing the CDR data format in this way makes it very easy to compare the output of our estimation models across different cities.

4.2.3 Creating and storing geographic data

A relational database is used to store road network and census information for every city in a standard format. Given the current cost of computing resources, these systems provide adequate performance for storing static GIS and census data and have convenient, mature interfaces for easy access. We also use this database to store aggregated results from our estimates so that they can be made available to interactive web APIs and visualization platforms. We use a Postgres and the open source spatial extension PostGIS to store and manipulate census and road network data.

While census tract or TAZ (Traffic Analysis Zone) polygons and demographic

[1]Google Protocol Buffers `https://developers.google.com/protocol-buffers/`

information are stored in this database, it is computationally inefficient to perform point-in-polygon calculations for each user or call record in our CDR dataset. To dramatically speed these computations, we rasterize polygons into a small pixel grid, where pixel values is a unique identifier for the census tract covering that pixel. This raster is then used as a look-up table to convert the latitude and longitude of calls into census tract IDs. The rasterization introduces some error along the borders of tracts, but these errors are minimized by making pixel sizes much smaller than the size of the raster and resolution of the location estimates of calls (between 10m and 100m).

While the platform supports road networks supplied by local municipalities in the form of shapefiles, we have implemented a parser to construct routable road networks from OpenStreetMap (OSM) data due to its global availability. Transportation networks in OSM are defined by *node* and *way* elements. Nodes represent points in space that can refer to anything from a shop to a road intersection, while ways contain a list of references to nodes that are chained together to form a line. In our context, relevant ways are those used by cars and relevant nodes are intersections within the road network. Ways and nodes may also contain a number of tags to denote attributes such as "number of lanes" or "speed limit". Many roads, however, do not include the whole set of attributes necessary for accurate routing. For example, city roads often lack speed limit information required to estimate the time cost, which in turn is used to find shortest paths based on total travel time. To infer this missing data, our system supports the creation of user-defined mappings between highway types and road properties. For example, ways tagged as "motorways" are generally major highways and have a speed-limit of 55 mph in the Boston area. They tend to have 3 lanes in each direction. "Residential" roads, on the other hand, have a speed-limit of 25mph and 1 lane in each direction. Each road segment is also given a capacity based on formulas suggested by the US Federal Highway Administration. Using these mappings, we parse the OSM xml data to create a routable, directed road graph with all properties required to estimate realistic costs driving down any given road.

We implement two additional cleaning steps to improve efficiency. The first filters

out irrelevant residential roads. These small local roads are filtered from our network, as they are not central to the congestion problem, yet tend to increase computation time significantly. Finally, in OSM data, a node object can refer to many things, for example an actual intersection or simply a vertex on a curve used to draw a turn. The latter case results in a network node with only one incoming and one outgoing edge (assuming U-turns are not allowed). These nodes are superficial and increase network size and routing algorithm run times needlessly. We simplify networks by removing these nodes from the network and only connecting true intersections, keeping the geographic coordinates of the nodes so that link costs still reflect actual geographic length of roads rather than straight line distances between start and end points. The parsed and cleaned edges are then loaded into the Postgres database, preserving attributes and geometry. Pseudo-code of the algorithm to parse and simplify OSM networks can be found in algorithm A.1 in Appendix A.

4.3 Estimating Origin-Destination Matrices

The following sections review algorithms for transforming billions of geo-tagged data points into origin destination matrices and assigning these flows to transportation infrastructure. Some of these algorithms are important for their deviation from traditional approaches and some are important for their computational efficiency, a requirement when faced with such massive data sets. We adapt (if necessary, for computational efficiency) and integrate the methods presented and validated in Chapter 2 into a full implementation of travel demand estimation for cities.

First, we adapt the spatial and temporal clustering method used to extract stay locations for Boston CDR data in Chapter 2. Given a user's trajectory of spatiotemporal points $P = \{p_1(x_1, y_y, t_1), \ldots, p_n(x_n, y_n, t_n)\}$, the goal is to discover meaningful locations at which a user repeatedly stays for a significant amount of time. The algorithm begins by considering each call in a time ordered sequence. Two consecutive (p_i, p_{i+1}) points are considered to form the start of a *candidate set* of points at the same semantic location if the distance between them is less than a thresh-

old $\Delta r_{i,i+1} < \delta$. Subsequent points are added to this candidate set if they also meet this criteria, e.g. p_{i+2} is added if $\Delta r_{i+1,i+2} < \delta$. The result is a candidate set $S = \{p_s(x_s, y_s, t_s), \ldots, p_t(x_t, y_t, t_t)\}$ containing a number of consecutive calls. A candidate set is considered to represent a single *candidate stay* if time between the first and the last observation in the subsequence S are separated by a time greater than a threshold $\Delta t_{m,n} > \tau$. The geographic location of a candidate stay is set to be at the centroid of points in S. Due to noise in locations and daily call frequencies, multiple candidate stays that are actually the same place may be estimated at a slightly different geographic coordinate on different observation days. To account for this, a final agglomerative clustering algorithm is used to consolidate candidate stays to a single semantic location regardless of the temporal sequence of individual calls. Though many agglomerative clustering algorithms exist, we implement a simple, efficient grid based approach by assigning each filtered location to a grid cell and then defining a final *stay point* as the centroid of all filtered locations in each cell. A final pass through the original calls assigns any call within a distance δ from a stay point to that stay point regardless of whether or a not a consecutive call was recorded from that location. This algorithm removes noisy or spurious outliers from the data set while preserving as much information on visits as possible. It may also be run on both triangulated and tower-based CDR data, in the latter case it removes noise associated with calls from the same location being routed through different nearby towers due to environmental factors. Pseudo-code can be found in algorithms A.2 through A.5 in Appendix A.

With de-noised trajectories of stay points, the next step is to infer contextual information about each location such as the activity or purpose. Using visit frequencies and times, we implement the same methods discussed in Chapter 2 to designate stay points as *home*, *work*, and *other* and estimate average daily trips between consecutive stay points. We next assign a random departure time for these trips based on the conditional probability that user departed during an hour between the time they were last observed at the origin and the time they were first observed at the destination. This conditional probability function for departure time can be derived from surveys

such as the National Household Travel Survey (as done for Boston in Chapter 2) or estimated empirically using observed call frequencies of all users over the course of the day. Having assigned departure times and purposes to each trip, we can construct trips made by a given user. Generally, we are interested in trips between geographic areas such as towns or census tracts so here we convert origin and destination points to IDs of the tract of zone they are in. The result is a vector of trips between locations in the city for each user in our data set.

While a trip represents an observation of movement of at least one person between two locations, we must still be careful to control for differences in market share and usage rates across a city. We first scale trips based on how often an individual uses their phone. For each user, we calculate the average number of trips made during a given time window by dividing the number of trips counted by the number of days that user was observed making a call, as outlined in Chapter 2. This step effectively measures the average number of trips a user makes between two locations on a day given that they are observed in our data set. Due to differences in daily usage of mobile phones among the population, not every user makes enough calls on a typical day to infer their movement patterns. For this reason, we must filter out users that do make enough calls. This step requires trade-offs between sample size and amount of data we have on each selected user. Because we will eventually be routing these trips through the transportation network, it is important to correctly estimate the total number of trips taken as well as the distribution of trips across the city. In practice, we find that filtering out users who we measure to make fewer than 2.5 trips per day leaves a large sample size of active users and results in valid estimates of trip tables and OD matrices as shown in subsequent sections. Those implementing these methods may find that different filtering criteria produce samples suited for different tasks.

We then expand the average trip counts of filtered users to account for market penetration rates. As with survey participants, the ratio of cell phone users to the population is not uniform within the region. Each user is assigned a home census tract and expansion factors are computed for each tract by measuring the ratio of the

number of users assigned there and the reported population. As we saw in Figure 2-2 in Chapter 2, these expansion factors tend to be less than 25 in Boston, but can be higher in places with lower market share. They are generally much lower than surveys which may only choose two or three individuals to represent hundreds or thousands in an area. Each user's typical daily trip volumes are then multiplied by the expansion factor corresponding to their home tract and the now represent the movements of some fraction of the tracts population.

Finally, we may wish to consider only trips via a certain mode, e.g. vehicle trips. Though CDR data does not provide resolution required to measure mode choice, vehicle trips can be approximated by methods proposed in Chapter 3, or if data on mode shares at the origin-destination level is not available, weighting person trips by vehicle usage rates in the home census tract of users. In this way, full OD matrices for vehicle or person trips are computed by summing the expanded trip volume computed for all users between all pairs of census tracts. We also construct partial OD matrices containing only trips of a certain purpose during a certain time window. Due to the relative consistency of CDR data around the world, we can adopt this same OD creation procedure in all cities. Pseudo-code to generate OD matrices can be found in algorithms A.6 and A.7 in Appendix A.

4.3.1 Trip Assignment

Having estimated OD flows, our next task is to efficiently assign these trips to transportation infrastructure, in this case a road network [10]. The traffic assignment module in the software platform takes tract-to-tract OD matrices and distributes trips among nodes, or intersections. A trip originating in a census tract is assigned uniformly at random to an intersection in that tract and to an intersection within its destination tract. This distributes flows such as not to create artificial congestion points and reflects general uncertainty in the exact origin of trips. Future iterations of the platform could also incorporate additional approaches, such as the centroid connector method used in Chapter 3 and many travel demand modeling frameworks. With intersection to intersection flows, the next task is to assign traffic to routes.

Our system is modular so that it may implement any number of traffic assignment algorithms. Here, however, we take a simple ITA approach (as in Chapter 3), since it is computationally efficient for many trip pairs in detailed road networks and allows us to keep track of each vehicle as it is routed through the network. We develop a set of tools to perform large scale routing and traffic assignment using parallelization for speedups. First, the parsed and optimized road network is loaded into a graph object. In our implementation, we use the Boost Graph Library for its flexibility and efficiency. We can then compute shortest paths based on a user defined cost (in this case travel time on road segments). We choose the A* algorithm among the wide range of shortest path algorithms, as it's widely used in routing on geographic networks for its flexibility and efficiency. The A* algorithm implements a *best-first-search* using a specified heuristic function to explore more promising paths first. The euclidian distance between nodes provides an intuitive heuristic that ensures optimal solutions are found. While this algorithm provides the same results as Dijkstra's algorithm, we find that it becomes more efficient to compute paths one by one for sparse OD matrices.

A simplified schematic explaining our implementation of the ITA procedure can be seen in Figure 4-2. Though increments must be routed in serial, all routes discovered within an increment are independent. To speed up the routing process, we divide all trips in an increment into batches and send these batches to different threads for parallel computation. Because the road network remains fixed in each increment, we only need to store a single graph object shared by all threads. When a shortest path is found, we walk that path and increment counts of the number of vehicles that were assigned to each road and sum the counts from all batches after the increment has finished. We also keep track of the origin and destination census tracts of the assigned vehicles in a bipartite graph for later analysis. After all trips have been routed, we compute final V/C ratios and other metrics of each segment and update these values in the database so they can be used for other applications or visualization. Pseudo code for this ITA procedure can be found in algorithm A.8 of Appendix A.

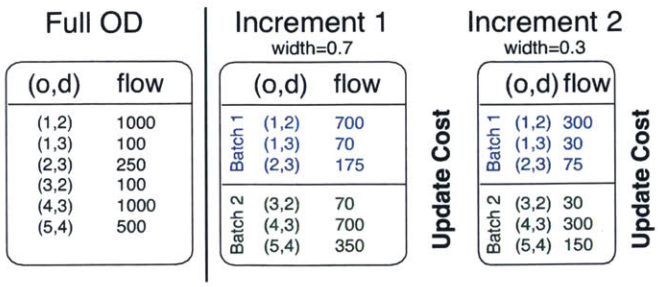

Figure 4-2: Our efficient implementation of the incremental traffic assignment (ITA) model. A sample OD matrix is divided into two increments and then split into two independent batches each.

4.4 Results

In the following sections we demonstrate the range of outputs provided by our system. We first report trip tables and compare origin-destination matrices produced by our system to available estimates made using travel surveys. We then report road network performance as well as characteristics of road usage patterns enabled by the construction of a bipartite road usage network.

4.4.1 Trip Tables and Survey Comparison

In order to understand when and where these new data will be effective and how the results differ from traditional approaches, we compare the output of our system to previous travel surveys wherever possible. In four of the cities studied, we find estimates of travel demand from surveys: the 2011 Massachusetts Household Travel Survey (MHTS) in Boston, the 2000 Bay Area Travel Survey (BATS) in San Francisco, a 2013 transportation plan in Rio de Janeiro, and estimates from a 2012 LUT model in Lisbon [36]. While these surveys do not always produce all estimates we are able to generate with our system, we make comparisons wherever possible.

Trip tables report the total number of trips of a given purpose or during a given time of day for a city and represent the total load placed on transportation infrastructure. In Table 4.2, we report trip tables for each city in this study. We find close agreement with trip tables estimated using CDR data and surveys in Boston and the San Francisco Bay Area and less agreement in Rio de Janeiro. We note, however,

Table 4.2: Trip tables estimates. Where possible, our results are compared to estimates made using travel surveys. For each city, we report the number of person trips in millions for a given purpose or time. Trip purposes include: home-based word (HBW), home-based other (HBO), and non-home-based (NHB). Trip periods include: 7am-10am (AM), 10am-4pm(MD), 4pm-7pm (PM), and the rest of the day (RD). We note that the exact boundaries of the surveys do not exactly coincide with those used in our estimation so direct comparisons are not exact. No comparisons could be found for Porto. *Note that the Lisbon Survey only contains estimates of vehicle trips in millions.

City	HBW	HBO	NHB	AM	MD	PM	RD	Total
Boston	5.76	8.99	6.72	3.71	7.68	5.75	4.33	21.47
MHTS	3.22	12.83	9.49	5.32	8.87	8.20	3.15	25.54
SF Bay	4.07	10.05	7.04	4.47	7.81	5.35	3.53	21.16
BATS	4.60	11.54	4.66	4.18	6.90	4.22	3.00	20.80
Rio	9.92	17.17	11.46	7.71	14.09	10.47	6.29	38.55
Survey	2.06	–	–	1.31	1.19	1.24	–	3.74
Lisbon	1.08	2.01	1.21	0.79	1.67	1.26	0.58	4.30
Survey*	0.61	–	–	–	–	–	–	–
Porto	0.49	0.87	0.46	0.32	0.70	0.54	0.27	1.83
Survey	–	–	–	–	–	–	–	–

that the 3.74 million person trips estimated for Rio is far too low given the population of the region and highlights the difficulty in finding reliable planning resources in many areas. Finally, we note that in Lisbon, the survey results represent vehicle trips only, while we report person trips. When adjusting for mode car ownership rates in Portugal, our numbers align more closely. We were unable to find a survey or model for comparison in Porto. Note that differences in the size of the study area for Boston and trip period definitions, and minor modifications to the algorithms to produce origin-destination trips, result in trip counts different than those presented in Table 2.2.

In addition to trip tables, it is also necessary to compare the distribution of trips from place to place around the city. In order to make this comparison, the area unit of analysis for the survey and our model must be aligned. Given the resolution of mobile phone data, our system is designed to create ODs at the census tract (or equivalent) level while many surveys aggregate to larger traffic analysis zones or super districts. For comparison, we aggregate the OD matrices from CDRs to the coarser grained resolution provided by the survey and compare results. Figure 4-3

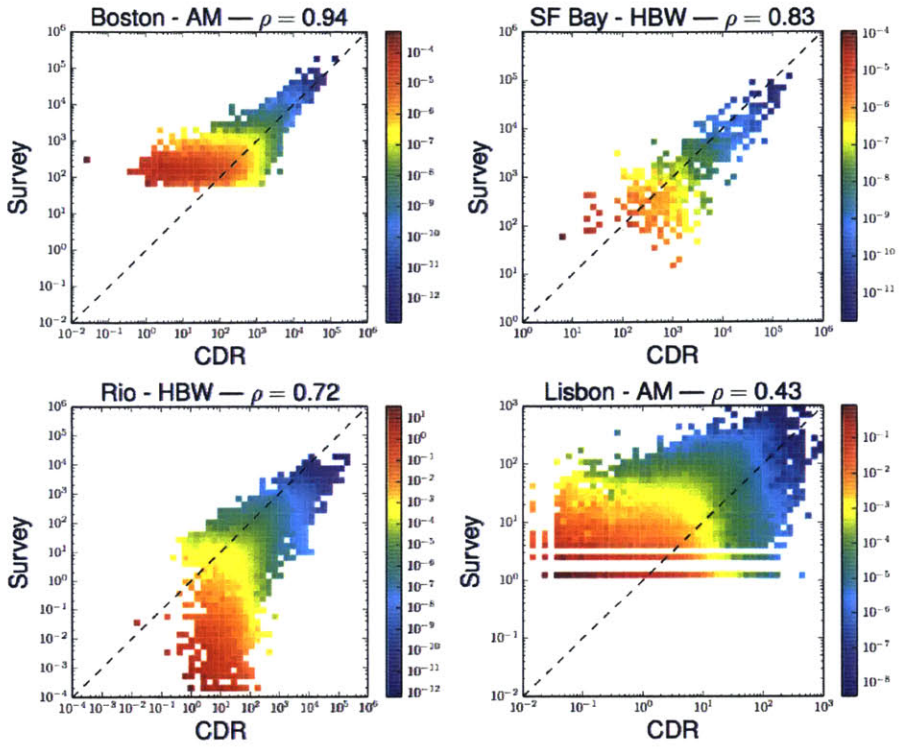

Figure 4-3: Correlations between OD matrices produced by our system and those derived from travel surveys at the largest spatial aggregation of the two models. In Boston, this is town-to-town, in San Francisco, MTC superdistrict-to-super district, in Rio, census superdistrict-to-superdistrict, and in Lisbon, freguesia-to-freguesia. The larger of these area units (e.g. towns in Boston), the better our correlations, while correlations at the smallest aggregates(e.g. freguesias in Portugal), correlations are lower.

show correlation histograms comparing OD matrices at the largest spatial aggregation available produced by our methods and those produced by traditional methods. In general we find very high correlations in Boston, San Francisco, and Rio, with lower correlations in Lisbon. Lisbon, however, has the smallest units of aggregation and these results demonstrate the limitations of these comparisons at very high spatial resolutions.

4.4.2 Road Network Analysis

The first output of this procedure is volume, congestion (volume-over-capacity), and travel times for all road segments. Using the outcomes of our analyses, we calculated the distributions of volumes on roads, along with V/Cs in Figure 4-4. Interestingly,

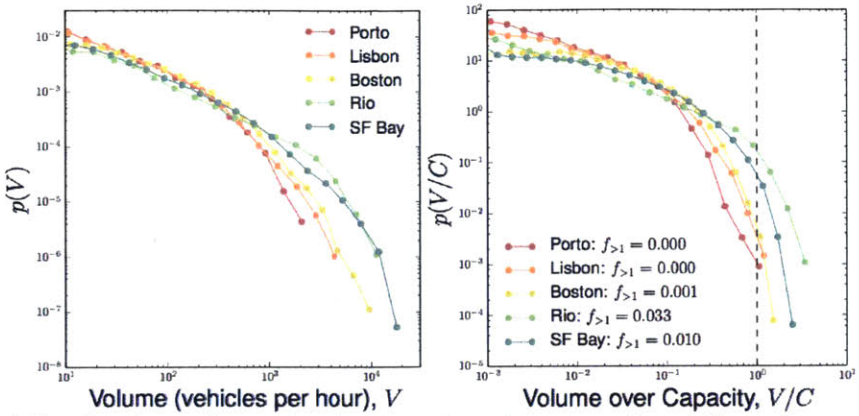

Figure 4-4: Distributions of travel volume assigned to a road and the volume-over-capacity (V/C) ratio for the five cities. The values presented in the legend refers to the fraction of road segments with $V/C > 1$.

the results suggest qualitatively similarly distributed volumes and V/Cs for our five subject cities. Moreover, our findings are consistent with general congestion studies that identify Rio de Janeiro as one of the most congested cities in the world and the San Francisco Bay Area not far behind. Smaller cities such as Boston and Porto have fewer problems with congestion.

4.4.3 Bipartite Road Usage Graph

In addition to measuring physical network properties of roads, the system architecture enables detailed analysis of individual road segments and neighborhoods within a city. To this end we create a bi-partite usage graph. Every time a route between two locations is assigned, we traverse the path and keep a record of how many trips from each driver source (census tract) used each road. This record is then used to construct a bipartite graph containing two types of nodes: road segments and driver sources, as shown in Figure 4-5. Roads are connected to driver sources that contribute traffic to that segment and census tracts are connected to roads that are used by people who live here.

This bipartite framework of analysis allows us to augment visualizations of congestion maps in two ways. The first focuses on a single road segment. For example, when we identify a segment of a highway that becomes highly congested with traffic jams each day, we can easily query the bipartite graph to obtain a list of census tracts

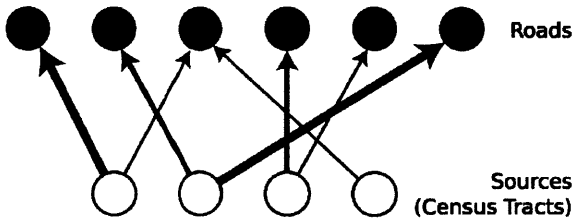

Figure 4-5: A graphical representation of the bipartite network of roads and sources (census tracts), with edge sizes mapping the number of users using the connected road in their individual routes.

where drivers sitting in that traffic jam are coming from and where they are going to. The census tract nodes can also be given attributes from containing any demographic data a user wishes. With this information, it is possible to identify leverage points where policy makers can offer alternatives to these individuals or even power applications such as car sharing, by notifying drivers that others sharing the same road may be going to and from the same places. Moreover, businesses considering products or services based on who may be driving by or near different locations may find value in these detailed breakdowns.

Rather than selecting a road segment node, we may also select a single census tract, and check its neighbors to construct a list of all roads used by individuals moving to or from that location. For example, for a given neighborhood in a city we can identify all major arteries that serve that local population. This information provides a detailed look at a central location based on how much road usage it induces. Moreover, geographic accessibility, critical to many socio-economic outcomes, can now be measured in locations that were previously understudied.

4.4.4 Visualization

To help make these results accessible to consumers and policymakers, we build an interactive web visualization to explore road usage patterns in each city. Most GIS platforms can connect directly PostGIS databases to visualize and analyze road networks with our estimated usage characteristics. While these platforms are preferred by advanced users familiar with GIS data, they are opaque to many consumers who may benefit from more detailed information on road usage. A simple API is imple-

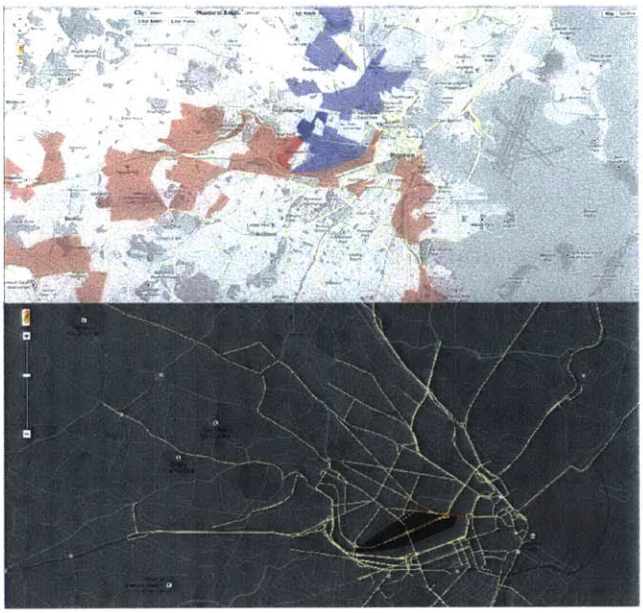

Figure 4-6: Two screen images from the visualization platform. (a) The trip producing (red) and trip attracting (blue) census tracts using Cambridge St., crossing the Charles River in Boston. (b) Roads used by trips generated at the census tract including MIT.

mented to query the database and generate standard GeoJSON objects containing geographic information on roads as well as computed metrics such as level of service. We also implement queries to answer questions such as "What are all the census tracts used by drivers on a particular road?" or "What are all roads used by a given location in the city?". These data are then parsed and displayed on interactive maps using any of the available online mapping APIs and D3js allowing users, with functionality that enables one to select individual roads and areas. Two screen images of this system is shown in Figure 4-6.

4.5 Conclusion

This chapter has presented a full implantation of a travel demand model that uses new, big data resources as input. We have presented a system that combines and improved upon many disparate advanced in recent years to produce fast, accurate, and inexpensive travel demand estimates. We began by outlining methods to extract meaningful locations from noisy call detail records and estimate origin-destination

matrices by counting trips between these places. Normalized and scaled trips counts are compared to estimates made using survey data in both trip tables and at the OD pair level. These flows are then assigned to road networks constructed from OpenStreetMap data using an incremental traffic assignment algorithm. As routes are assigned, a number of metrics on road usage are measured and stored.

Chapter 5

Assessing the impact of real-time ridesharing on urban traffic

5.1 Introduction

In 2014, ridesourcing services Uber[1], Lyft[2], and Sidecar[3] launched ridesharing programs in the US that match customers making similar trips. Ridesharing offers monetary incentives to customers who pay a reduced rate for their trip, as well as drivers who are able to carry more passengers more efficiently. Ubiquitous technologies have allowed for the emergence of these real-time ridesharing services, with GPS providing driver and customer locations and route navigation, smartphone apps affording real-time ride requests, and social networks establishing trust and accountability between customers and drivers. Further, advances in computing speed and data storage has enabled the development of platforms to run rideshare optimization algorithms in real-time.

Ridesharing has garnered support for its potential to reduce private automobile use by providing a convenient and affordable alternative to driving alone, translating to reduced roadway congestion and vehicle emissions in the short-term, and reduced

[1] www.uber.com
[2] www.lyft.com
[3] www.side.cr

automobile ownership in the longer term. A recent intercept survey of ridesharing customers in San Francisco reveals that although ridesharing often substitutes for longer transit trips, it otherwise complements transit, with many observed trip origins and destinations near transit stations [66]. However, ridesharing critics are skeptical about the likelihood that ridesharing will decrease vehicle congestion and emissions, due to the potential of ridesharing to divert trips from transit or other non-motorized modes and induce new trips altogether. Furthermore, safety and liability concerns tied to inadequate driver training and insurance, as well as direct competition with highly regulated taxi companies, has led some to call for regulations of the app-based ridesharing industry. As city leaders and policy makers are faced with decisions about regulating the growing rideshare market, it is becoming increasingly important that the overall impacts of ridesharing are understood.

Whether or not ridesharing adds to vehicle traffic depends on the balance of competing forces. On the one-hand, ridesharing may increase traffic by replacing non-driving modes such as transit, walking, or cycling or inducing new trips. On the other hand, ridesharing may decrease traffic by increasing vehicle occupancy, serving the first/last mile of potential transit trips, and reducing private car ownership and use. This research focuses on understanding the impact of two of these key drivers: diversion of non-drivers and diversion of travelers from private, single occupancy cars or taxis. The other factors are likely to occur on a longer time scale and therefore harder to quantify.

With the uncertainty surrounding the impacts of ridesharing in mind, this research aims to answer two questions unresolved in existing literature:

- What proportion of trips can be matched by a real-time ridesharing service given the temporal and spatial distribution of all urban trips and travel modes?
- What is the change in the number of vehicles and traffic congestion given relative adoption rates of ridesharing from auto and non-auto travelers?

To help answer these questions, we again turn to mobile phone data as a resource for travel demand estimation. In Section 5.2, we motivate our approach and the

use of mobile phone data for this application within the context of related work. We first compare and contrast our data and methods with recent rideshare research coming from the urban computing domain, demonstrating how we differ from and add to this body of work. Then, we relate this approach to methods typically in the transportation industry to evaluate the demand for and impact of a new travel alternative.

The rest of the chapter follows with discussions of the data, methods, and results. Using the procedures developed in Chapter 2 and implemented in Chapter 4, we first estimate average daily origin-destination (OD) trips from mobile phone records. We then use methods described in Chapter 3 to estimate the proportions of these trips made by driving and other non-driving modes. Next, we match spatially and temporally similar trips, and explore a range of adoption rates for drivers and non-drivers, in order to distill rideshare vehicle trips, and by extension, total vehicle trips. Finally, we use algorithms described in Chapter 3 and efficiently implemented in Chapter 4 to allocate these vehicle trips to a road network and evaluate the impacts of ridesharing on urban congestion.

5.2 Related Work

To-date, much of the research related to ridesharing has focused on understanding the characteristics of ridesharing trips and users. In a recent survey of app-based, on-demand rideshare users in San Francisco, researchers found that 45% of ridesharers stated they would have used a taxi or driven their own car had ridesharing not been available, while 43% would have taken transit, walked, or cycled [66]. The authors conclude there is a need for research to explore the impact of modal shift on vehicle congestion and emissions. This work is a step in that direction and leverages mobile phone data to understand underlying travel demand not captured by small-scale intercept surveys.

Santi et al. developed a framework to compute maximum matching of shareability

networks constructed from an OpenStreetMaps[4] (OSM) road network and 172 million taxi trips made in New York City in 2011 [71]. We add to this body of work by introducing a method that takes into account trip-making by all modes, rather than just taxis, which represent only a portion of potential rideshare demand. Further, authors explicitly assume that traffic conditions—which impact the travel time criteria used by their matching algorithms—will remain largely unaltered by the emergence of ridesharing. Given their finding that ridesharing could cut total taxi vehicle miles by 40%, we revisit this assumption in this work; using traffic assignment algorithms commonly used in transportation planning applications to route total vehicle demand, we assess the impact of ridesharing on network-wide congestion.

Cici et al. used mobile phone and social network data to evaluated demand for ride-sharing between strangers, friends, and friends-of-friends in four cities in Spain and the US [29]. As in this work, the authors use CDR data, however, they focus on commuting trips between home and work locations inferred from CDR, Twitter, and Foursquare data. We build on this research by (i) using trips across all purposes, and (ii) estimating auto mode shares by origin and destination rather than selecting a portion of the trips as vehicle trips based on a single city-wide mode share, and (iii) estimating the impacts of ridesharing on urban congestion and travel times rather than using an online mapping service for static routing and travel time characteristics.

Researchers in [29, 71] focused on addressing the computational challenges of trip-matching—an NP-hard optimization problem—in real-time and developed heuristics to quantify potential ride-sharing demand. These algorithms re-route trips in order to match them with similar, overlapping trips, explicitly capturing demand for ridesharing relative to passenger's willingness to experience prolonged travel time. We believe that development of such heuristics are crucial for the effective implementation of a real-time ridesharing system. In this work we focus on other aspect that affect rideshare demand and urban congestion, namely:

- Total network-wide trips by mode

[4] An open source mapping community supporting data on road networks all over the world. www.openstreetmap.org

- Rideshare adoption by mode, and the impact of this mode-shift on the number of network-wide vehicles, and

- Dynamic relationship between demand and traffic congestion.

With this framework, we introduce a novel application to the field of urban computing by evaluating the impact of a new transportation option on urban traffic using mobile phone location data. With mobile phone data available in real-time, these methods can be adapted to support real-time rideshare matching applications.

Within the transportation domain, evaluating the demand for and impact of a new travel mode or option traditionally involves acquiring, adapting, and running a travel demand model. In particular, mode choice models are often used to estimate the diversion of trips to new modes or travel alternatives [6, 14, 16]. However, such models are expensive to develop and calibrate and their availability may be limited. Instead, we propose a framework to assess overall demand and congestion impacts of a new mode based on a range of hypothetical adoption levels. This approach offers an alternative end-to-end solution to quickly and economically perform transportation scenario analyses in any city for which mobile phone data is available.

5.3 Data

5.3.1 Mobile Phone

To estimate travel demand patterns we utilize mobile phone CDR data in the Boston metropolitan area, as in Chapters 2, 3, and 4. The CDR dataset contains more than eight billion mobile phone records for roughly two million anonymized users over two months in the Spring of 2010. Each record contains an anonymous user ID, longitude, latitude, and timestamp at the instance of a phone call or other types of phone communication (such as sending SMS, etc.). The coordinates of the records are estimated by service providers based on a standard triangulation algorithm, with an accuracy of about 200 to 300 meters.

5.3.2 GIS/Survey

As in Chapters 2, 3, and 4, we rely upon a variety of spatial and survey data sources, summarized below.

- Road network: For traffic simulation, we use a GIS shapefile from the local transportation authority containing road characteristics such as speed limits, road capacities, number of lanes, and classifications [26].

- Census tracts: CDR trips are aggregated to the spatial resolution of 974 study area Census tracts, which contain roughly 5000 residents each. We use a GIS shapefile as well as population estimates from the American Community Survey to expand observed users to total population [1, 2].

- Communities: CDR trips are aggregated to 174 areas (163 study area towns and 11 Boston neighborhoods), referred to as communities in this chapter as well as Chapter 3. Note that in Chapters 2 and 4, the aggregation level that we referred to as town consisted of 164 areas–the study area towns, including Boston. We divided Boston into 11 neighborhoods for this analysis to match rideshare trips at a finer resolution within the city of Boston. We used a GIS shapefile of town boundaries developed by MassGIS to map Census tracts to these communities, but further split Boston into neighborhoods using local knowledge of neighborhood boundaries[55].

- Census commuting trips: The 2006-2010 Census Transportation Planning Products (CTPP) Part 3 provides commuting characteristics between Census tract pairs [85]. This nationally-available dataset provides tabulations across 16 different travel modes, which we use to infer mode shares.

5.4 Methods

5.4.1 Trip Estimation

As described in detail in Chapter 2, we estimate average daily trips using the CDR dataset in Boston. CDRs are first converted into clustered locations or *stay points* at which users engage in activities for an observed duration. These locations are inferred to be home, work, or other depending on observation frequency, day of week, and time of day, and represent a user's origins and destinations. Next, we construct trips between two consecutive stay points in a day. Since the arrival time and duration at these locations reflect the *observed* (based on phone usage) rather than *true* arrival time and duration, we probabilistically infer departure hour h using the NHTS survey data on trips in major US cities.

For each user u, we generate trip matrices t_{ij} by summing the number of trips from origin Census tract i to destination tract j. By dividing these trips by the number of days n on which we observed the user, we compute average daily transition matrices with the probabilities that a user makes a trip between any origin and destination pair ij on an average weekday. Lastly, user trips are multiplied by expansion factors w based on the population of a user's home Census tract. Summing across all individuals, we compute average daily trip matrices T_{ij}, as summarized in Equation 5.1.

$$T_{ij}(h) = \sum_{u=1}^{U} t_{ij}(u,h)/n(u) * w(u) \tag{5.1}$$

5.4.2 Mode Share Estimation

Capturing spatial variation in mode shares is essential to estimating reasonable distributions of vehicle trip patterns from the person trips inferred from CDR data. Further, whether or not ridesharing reduces vehicle traffic depends on the extent to which ridesharing diverts customers from different modes.

With that in mind, we want to determine the number of trips made by three travel

modes: drive-alone or taxi, carpool, and non-driving modes. The fraction of travelers for each mode is calculated using CTPP commuting data. Averaging this data across Boston, 70

We use the methods presented and validated in Chapter 3 to infer these travel mode shares. First, we use the CTPP commuting data aggregated from Census tract pairs ij to community pairs IJ in order to minimize the effects of matrix sparsity and small sampling size. We compute (i) drive-alone and taxi mode share d_{IJ}, (ii) carpool mode share c_{IJ}, and (iii) non-auto mode share o_{IJ} such that $\sum(d_{IJ} + c_{IJ} + o_{IJ}) = 1$. Community pairs with sampling issues, including few total trips or zero auto trips, are assigned average mode shares depending on their geography: Urban-Urban, Suburban-Urban, or Suburban-Suburban. Communities lying within Boston's I-95 highway ring are designated as Urban, and all other communities are designated Suburban.

5.4.3 Rideshare Vehicle Estimation

Given OD trips inferred from the CDR data and mode shares inferred from census data, we next estimate drive-alone, carpool, and non-auto trips and match candidate ridesharing trips together. Since rideshare adoption among drive-alone and taxi travelers would reduce vehicles, and adoption among other non-auto travelers would increase vehicles, we want to capture rideshare adoption from these modes seperately. We assume existing carpoolers would not adopt ridesharing since they already coordinate their trip with at least one other traveler; however, we take into account the contribution of carpool vehicles to total vehicle traffic.

We explore different adoption rates for the ride sharing service by travelers that use taxi or drive-alone a_d and travelers that use non-driving modes a_o. While total rideshare trips will increase with increasing values of a_d and a_o, vehicle traffic will reduce for vehicle traffic will reduce for $a_d >> a_o$, and increase under $a_d << a_o$.

Additionally, we introduce a parameter s, representing the maximum number of rideshare customers that can be matched within a vehicle. In this study, we assume $s = 2$ since existing rideshare services are matching two trip requests currently.

However, if and when rideshare services allow larger rideshare passenger occupancy, we can increase s to match more potential ridesharers in fewer rideshare vehicles. This parameter also gives the flexibility to assess the potential of dynamic shuttle services, such as Bridj[5], which currently serves Boston-area commuters using sprinter passenger vans with 12 seats.

We require finer temporal resolution to match potential ridesharing trips than departure hour as estimated in previous steps. We assume that trips occur uniformly throughout each hour, and compute the number of trips made within time window Δ. For $\Delta = 6$ *minutes*, for example, hourly demand is split into 10 intervals, and potential ridesharing trips are matched within an interval Δ. From the ridesharer's perspective, Δ represents the maximum allowable change in departure time the customer would be willing to incur to take ridesharing. A larger Δ will enable more trips to be matched by a ridesharing service, but will impose higher level of inconvenience to customers, which may hinder adoption.

Lastly, we define the spatial resolution for which ridesharing trips can be matched. Because Census tracts are the size of a few city blocks in downtown Boston, it is too restrictive of an assumption to only match trips beginning and ending in the same Census tracts. Accordingly, we match potential ridesharing trips based on the study area communities. Figure 5-1a illustrates the probability distribution function of community area, with a median of 15.0 miles^2 and a mean of 16.2 miles^2 (by comparison, for a circle, this implies a radius of 2.27 miles). Figure 5-1b illustrates the spatial distribution of community areas, with the majority of those in the urban core having areas less than 10 miles^2, while communities with the greatest area lie on the southern border of the study area.

It should be noted that when we refer to ridesharing we explicitly mean *end-to-end ridesharing*[6]. However, by using spatial resolution of communities we also implicitly capture *en-route ridesharing*[7] trips with the origins and destinations of

[5] `www.bridj.com`

[6] ride-sharing between users with similar origins and destinations

[7] ride-sharing between users sharing portions of their paths between dissimilar origins and/or destinations, such that additional passengers can be picked up en-route

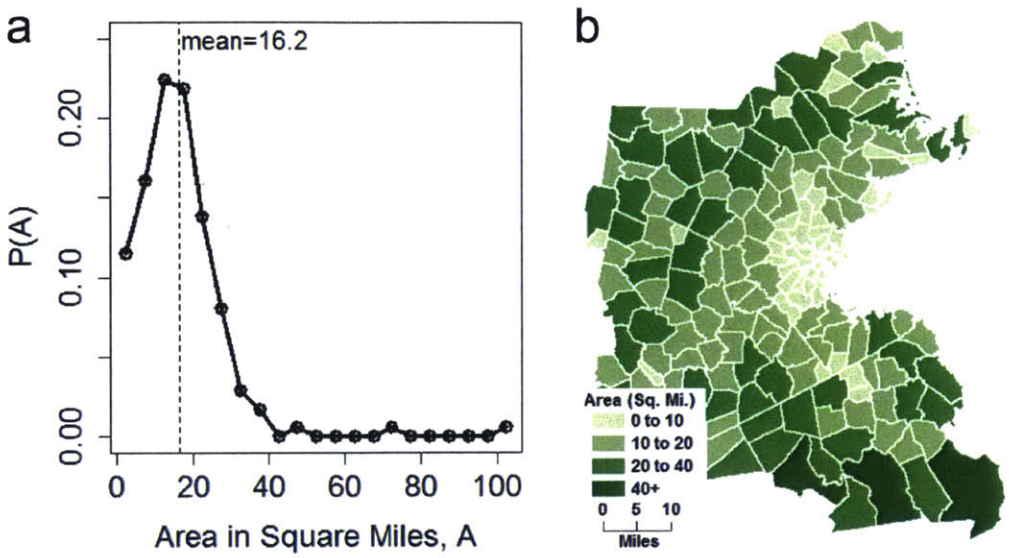

Figure 5-1: (a) Probability distribution of community areas in miles^2 and (b) Spatial distribution of community areas, with area increasing from light to dark shades of green.

matched ridesharers falling within the same communities. As reported in [5] and Section 2.4.3, another benefit of using a larger spatial resolution is that the correlation between CDR and survey trips increases by reducing the noise and/or spatial error present in the tract pair trips.

Finally, we estimate the number of vehicles $V_{IJ}(h)$ needed to satisfy trips $T_{IJ}(h)$ by first mapping tract pair trips to community pairs, $T_{ij}(h) \rightarrow T_{IJ}(h)$. Next, driver adopters $f_{D,IJ}(h)$ per Δ are calculated using the driver adoption rate a_d and share of drivers d_{IJ} as shown in Equation 5.2. Similarly, non-driver adopters $f_{O,IJ}(h)$ per Δ are calculated using the non-driver adoption rate a_o and share of non-drivers o_{IJ} as shown in Equation 5.3. Potential ridesharers $f_{IJ}(h)$ per Δ are simply the sum of driver and non-driver adopters (Equation 5.4).

$$f_{D,IJ}(h) = T_{IJ}(h) * a_d * d_{IJ} * \Delta/60 \tag{5.2}$$

$$f_{O,IJ}(h) = T_{IJ}(h) * a_o * o_{IJ} * \Delta/60 \tag{5.3}$$

$$f_{IJ}(h) = f_{D,IJ}(h) + f_{O,IJ}(h) \tag{5.4}$$

In practice, however, a ridesharing system may not be able to capture all of these willing adopters $f_{IJ}(h)$. Given the number of adopters traveling between two communties within a time step δ, we will reject adopters if: (i) there are less than s travelers, or (ii) there is a residual in the division of the number of trips by s. Accordingly, we can measure the efficiency $e_{IJ}(h)$, or percentage of potential demand that can be realized, of the rideshare system.

We refer to ridesharers that are unable to be matched in a rideshare vehicle as rejected ridesharers $r_{IJ}(h)$. Further, we assume that these rejected ridesharers will instead take their "original" mode, meaning that driver adopters who are rejected will drive just as they would if ridesharing were not available. Accordingly, single-occupancy vehicles carrying driver adopters who are rejected $v_{X,IJ}(h)$ are accounted for in the calculation of total vehicles in subsequent steps.

Under this framework, rejected ridesharers $r_{IJ}(h)$ are calculated as the remainder of the potential ridesharers $f_{IJ}(h)$ per Δ divided by group size s (Equation 5.5). By extension, matched ridesharers $m_{IJ}(h)$ are calculated as the difference between the potential and rejected ridesharers (Equation 5.6), and the efficiency of ridesharing $e_{IJ}(h)$ can be computed as the ratio of matched to potential ridesharers (Equation 5.7).

$$r_{IJ}(h) = f_{IJ}(h) \bmod s * 60/\Delta \tag{5.5}$$

$$m_{IJ}(h) = f_{IJ}(h) * 60/\Delta - r_{IJ}(h) \tag{5.6}$$

$$e_{IJ}(h) = m_{IJ}(h)/f_{IJ}(h) * \Delta/60 \tag{5.7}$$

The total number of vehicles traveling between two communities in a given hour $V_{IJ}(h)$ is the sum of all types of vehicles, namely: rideshare vehicles ($v_{W,IJ}(h)$), un-matched drivers ($v_{X,IJ}(h)$), drivers not adopting ($v_{Y,IJ}(h)$), and carpool vehicles ($v_{Z,IJ}(h)$), as shown in Equation 5.8. Since vehicles carrying matched ridesharers have a vehicle occupancy of s, the number of rideshare vehicles $v_{W,IJ}(h)$ is simply the number of matched ridesharers $m_{IJ}(h)$ divided by group size s (Equation 5.9).

Single-occupancy vehicles carry rejected driver adopters $v_{X,IJ}(h)$, equal to the ratio of driver adopters to total potential ridesharers $f_{D,IJ}/f_{IJ}(h)$ multiplied by rejected ridesharers (Equation 5.10), as well as by drivers who did not adopt $v_{Y,IJ}(h)$ (Equation 5.11). Lastly, carpool vehicles $v_{Z,IJ}(h)$ are computed using the average vehicle occupancy of carpool vehicles in the study area (in Boston, $p = 2.18$), as shown in Equation 5.12.

$$V_{IJ}(h) = v_{W,IJ}(h) + v_{X,IJ}(h) + v_{Y,IJ}(h) + v_{Z,IJ}(h) \tag{5.8}$$

$$v_{W,IJ}(h) = m_{IJ}(h)/s \tag{5.9}$$

$$v_{X,IJ}(h) = r_{IJ}(h) * f_{D,IJ}/f_{IJ}(h) \tag{5.10}$$

$$v_{Y,IJ}(h) = T_{IJ}(h) * d_{IJ} * (1 - a_d) \tag{5.11}$$

$$v_{Z,IJ}(h) = T_{IJ}(h) * c_{IJ}/p \tag{5.12}$$

5.4.4 Traffic Assignment

By simulating traffic under various rideshare adoption scenarios, we assess the impact of ridesharing on urban travel conditions. To allocate vehicle trips to the road network we perform traffic assignment, the final step of traditional four-step travel demand models used for transportation planning. Using Incremental Traffic Assignment (ITA), as described in Chapter 3 and implemented in Chapter 5, we distribute trips to the roadway network [3].

In ITA, a fraction of total OD trips are routed via shortest path (minimizing travel time), then road segment travel times are updated to reflect congestion, and the process is repeated until all OD trips are assigned. This procedure therefore enables us to capture the relationship between vehicle flow and travel time under congested conditions. For example, when a road segment's volume-over-capacity ratio V/C (i.e. the ratio between the number of cars using a road and its maximum flow capacity) is small, drivers can easily travel at free-flow speeds. As roads become congested and V/C increases, however, the speed at which drivers can travel decreases. This

relationship is often captured by the Bureau of Public Road's (BPR) volume delay function described by Equation 5.13. As in Chapter 3, we select a value of 0.85 for the α parameter (the default is 0.15) and keep the default value of 4 for β.

$$t_{current} = t_{freeflow} \cdot (1 + \alpha(V/C)^{\beta}) \tag{5.13}$$

Accordingly, we use the ITA procedure to route four batches of vehicle trips in 40%, 30%, 20%, and 10% increments, and update road segment travel times between each batch using Equation 5.13.

5.5 Results

5.5.1 Change in Vehicles

The efficiency of rideshare matching in different areas and hours of the day varies widely depending on the magnitude of trips, mode share breakdowns, and adoption rates. However, identifying aggregate trends across all hours of the day and OD pairs enables us to draw conclusions that may help us understand the impacts of ridesharing in other cities with different travel patterns and behavior. With this in mind, we estimate a model to capture aggregate daily impacts of ridesharing on the number of network-wide vehicle trips.

To help define the functional form of this model, we first analytically derive the percent change in vehicles $\Delta V_{IJ}(h)$ for a given OD pair and hour relative to the baseline scenario with no rideshare adoption. Specifically, the percentage change in vehicles is derived using Equations 5.8 through 5.12, with the numerator equal to the total vehicles under ridesharing ($V_{IJ}(h)$) minus the sum of the baseline drive-alone vehicles ($V_{Y,IJ}(h)$ for $a_d = 0$) and carpool vehicles ($V_{Z,IJ}(h)$), and the denominator equal to the sum of the baseline drive-alone vehicles ($V_{Y,IJ}(h)$ for $a_d = 0$) and carpool vehicles ($V_{Z,IJ}(h)$). This simplifies to the formulation shown in Equation 5.14.

$$\Delta V_{IJ}(h) = \beta_{IJ}(h) * [(s-1) * a_d * d_{IJ} - a_o * o_{IJ}] \tag{5.14}$$

$$\text{where, } \beta_{IJ}(h) = \frac{e_{IJ}(h)}{-s * (d_{IJ} + c_{IJ}/p)}$$

In other words, Equation 5.14 shows that the change in vehicles for a given hour and OD pair is proportional to the difference between the number of driver and non-driver adopters of the ridesharing service, and a parameter $\beta_{IJ}(h)$. $\beta_{IJ}(h)$ describes the relationship between the efficiency of the rideshare system $e_{IJ}(h)$ (as defined in Equation 5.7) and the share of vehicle trips given no rideshare adoption. Accordingly, $\beta_{IJ}(h)$ will be larger for OD pairs and time periods with higher rideshare matching efficiency and higher shares of non-driving trips, and result in a greater change in vehicles due to ridesharing.

While we can use Equation 5.14 to calculate the change in vehicles for a given OD pair and time period, we next generalize this relationship (for $s = 2$) to model aggregate results in any city using the model described by Equation 5.15. With the change in vehicles calculated in Boston for all hours and OD pairs, we empirically estimate $\beta = -0.5922$, minimizing the mean squared error between the data and model predictions. Equation 5.16 describes this model, such that the total change in vehicles ΔV can be calculated for adoptation rates a_o and a_d, given $\beta = -0.5922$, and aggregate Boston mode shares of $d = 0.7003$ and $o = 0.0846$.

$$\Delta V = \beta * (a_d * d - a_o * o) \tag{5.15}$$

$$\Delta V \approx -0.5922 * (a_d * 0.7003 - a_o * 0.0846) + 0.0175 \tag{5.16}$$

Note that the model in Equation 5.16 also includes an intercept parameter, which increases the estimated percentage change in vehicles by 1.75%, providing a better fit of the Boston data. At the extremes, the model suggests a 43% decrease in vehicles for 100% driver adoption and 0% non-driver adoption, and a 14% increase in vehicles for 0% driver adoption and 100% non-driver adoption in Boston (as compared with -40% and +13% change in vehicles from the data, respectively).

Moreover, $\beta = -0.5922$ captures the inefficiency of the ridesharing system in

Boston. Generalized from $\beta_{IJ}(h)$ in Equation 5.14, the β parameter in Equation 5.15 is described by Equation 5.17. Assuming perfectly efficient ($e = 1$) rideshare matching in Boston, it follows that $\beta_e \approx -0.6765$, as shown in Equation 5.18. As expected, $\beta = -0.5922$ is smaller in magnitude than $\beta_e = -0.6765$, equating to an average efficiency e of approximately 88%.

$$\beta = \frac{e}{-2 * (d + c/p)} \tag{5.17}$$

$$\beta_e \approx \frac{1}{-2 * (0.7003 + 0.0846/2.18)} \approx -0.6765 \tag{5.18}$$

Figure 5-2 shows the total percent change in vehicles δV (relative to $a_d, a_o = 0$) across all OD pairs and hours of the day under these adoption rate scenarios on the y-axis versus the ratio of driver to non-driver adoption rates (a_d/a_o) on the x-axis. As illustrated by the dashed lines, the model approximates the data points very well. As shown in the right side of the figure, when the number of ridesharing adopters from drivers is greater than from non-drivers ($a_d * d/a_o * o > 1$), there is a reduction in vehicles ($\delta V \leq 0$). Given the average mode shares of drivers ($d = 70.0\%$) and non-drivers ($o = 21.5\%$) in Boston, this relationship results in an overall reduction in vehicles for $a_o \lesssim 3.26 * a_d$, as illustrated by the point on the x-axis at which the data crosses the $\delta V = 0$ in the left side of the figure.

The relationship between a_d, a_o, and δV is further illustrated by Figure 5-3, with δV as estimated by the model and calculated from the data shown in Figure 5-3a and Figure 5-3b, respectively. White cells have no change in vehicles ($\delta V = 0$), and the black line with a slope equal to 3.26 is a contour line approximately representing scenarios with no change in vehicles from the model. Figure 5-3c illustrates the percentage of ridesharers that diverted from non-driving modes for each combination of driver and non-driver adoption rates. Here, white cells illustrate scenarios where ridesharers diverted from driving and non-driving modes equally.

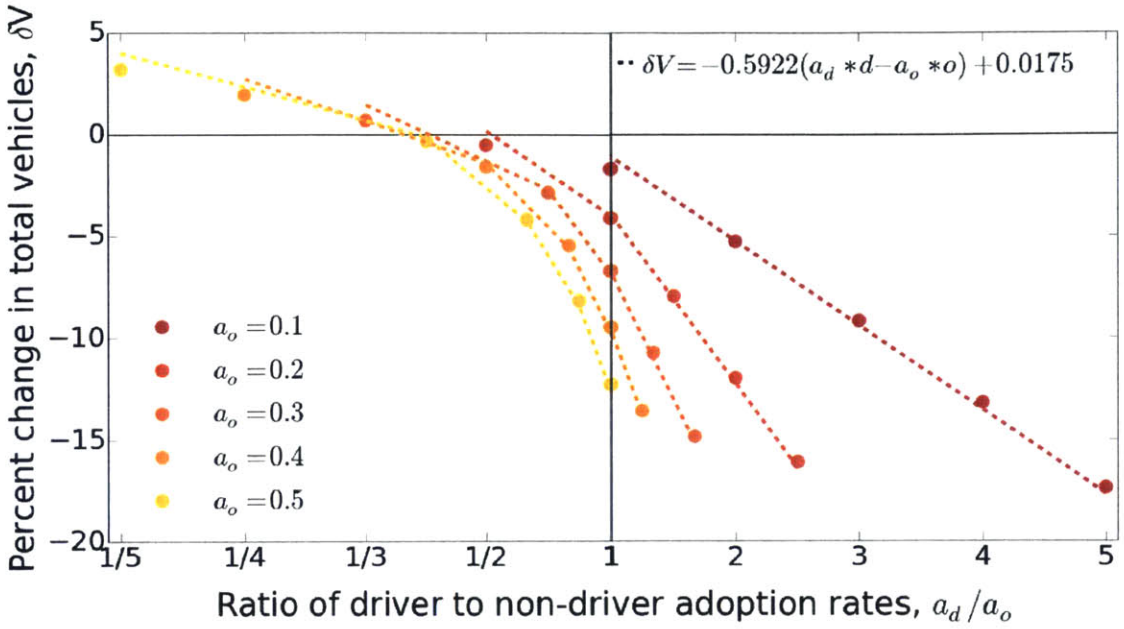

Figure 5-2: Percent change in total vehicles δV relative to the ratio of driver and non-driver adoption rates a_d/a_o. δV is proportional to the difference between driver and non-driver rideshare trip shares $(a_d * d - a_o * o)$, as described by the model: $\delta V = -0.5922 * (a_d * d - a_o * o) + 0.0175$. In other words, there is a reduction in vehicles ($\delta V \leq 0$) when the number of ridesharers diverted from drivers is greater than those diverted from non-drivers $(a_d * d \geq a_o * o)$. Given the average mode shares of drivers ($d = 70.0\%$) and non-drivers ($f_o = 21.5\%$) in Boston, this relationship results in an overall reduction in vehicles for $a_o \lesssim 3.26 * a_d$, as illustrated by the data and model.

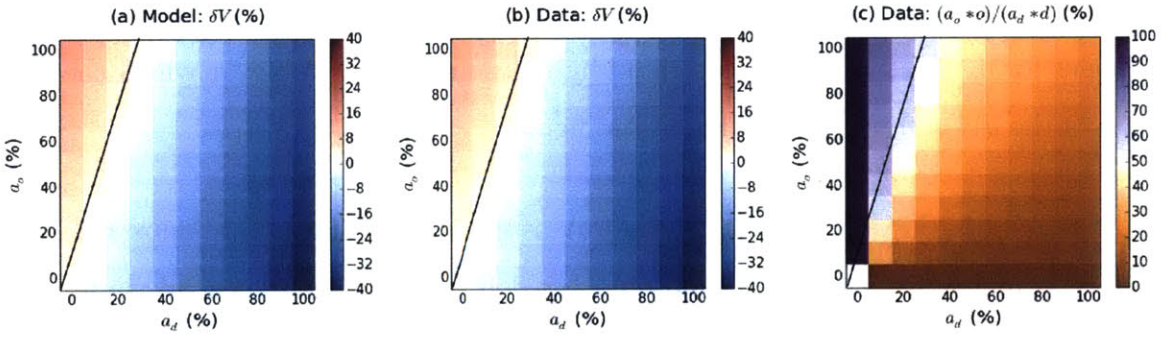

Figure 5-3: (a) Percent change in total vehicles δV relative to the ratio of driver a_d and non-driver adoption rates a_o as estimated by the model $\delta V = -0.5922 * (d * a_d - o * a_o) + 0.0175$. (b) Percent change in total vehicles δV relative to the ratio of driver a_d and non-driver a_o adoption rates from the data. (c) Percentage of ridesharers that diverted from non-driving modes $(o * a_o)/(d * a_d + o * a_o)$ relative to the ratio of driver a_d and non-driver a_o adoption rates from the data. The black line on each plot is described by $a_o = 3.26 * a_d$, approximately representing $\delta V = 0$.

a_d (%)	a_o (%)	Vehicles (%)	VMT (%)	VHT (%)	Congested TT (%)
0	50	5.99	1.83	3.02	7.16
10	10	-1.83	-0.85	-1.43	-2.98
50	0	-19.17	-11.57	-17.55	-37.30

Table 5.1: Percent change in vehicles, vehicle miles traveled (VMT), vehicle hours traveled (VHT), and congested travel time (TT) relative to drive-alone/taxi and other non-auto adoption rates $a_d, a_o = 0$. Results are for peak hourly evening (3-7pm) trips, $s = 2$, and $\Delta = 6$.

5.5.2 Change in Traffic

Next, we simulate traffic patterns to assess the network-wide impacts of ridesharing in the peak weekday evening hour for a few adoption scenarios. Table 5.1 summarizes the resulting percent change in vehicles, vehicle miles traveled (VMT), vehicle hours traveled (VHT), and congested travel time (minutes spent in non-free flow driving conditions). Again, the actual percent change in vehicles for the three adoption scenarios shown in Table 5.1 (5.99%, -1.83%, and -19.17%) are similar to those we can estimate using the model in Equation 5.16 (4.26%, -1.90%, and -18.99%).

We see a smaller change in total VMT than total vehicles, suggesting that ridesharing is more efficient in shorter distance, urban markets. Meanwhile, the percent changes in VHT are more significant than for VMT, suggesting that the increase in rideshare efficiency in these markets is somewhat counteracted by the fact that they experience more congestion than longer distance markets.

Lastly, the percent changes in congested travel times reflect the relationship between road segment volume and travel time as captured by the BPR function; changes in vehicle demand have an exponential impact on travel times under congested conditions. This trend is demonstrated in the third adoption scenario, with a decrease in congested travel time (37%) nearly double the decrease in vehicles (19%). The change in congested travel time for this third adoption scenario is equivalent to reducing the percentage of travel time spent in congestion from 17% to 13% for an average vehicle.

In general, these trends suggest that under moderate to high levels of rideshare adoption, ridesharing services would have a noticeable impact on urban traffic conditions.

5.6 Conclusions

This research explores the extent to which ridesharing services impact network-wide congestion using mobile phone records. To-date, other research efforts have used partial travel demand (i.e. of taxi or commuting trips) to estimate the proportion of trips that can be pooled for ridesharing under explicitly-defined spatio-temporal constraints. In contrast, we estimate aggregate, total daily travel patterns using Call Detail Records (CDRs) and explore different scenarios of adoption rates to estimate ridesharing demand.

Further, we assess the impact of relative levels of rideshare adoption from auto and non-auto travelers on vehicle usage and traffic congestion. When the number of ridesharing adopters from drivers is greater than from non-drivers, there will be a reduction in total vehicles, and vice versa. In Boston, given the aggregate mode shares of drivers and non-drivers, this translates into a reduction in vehicles when the non-driver adoption rate is less than about three times the driver adoption rate.

However, the magnitude of the change in vehicles varies spatially and temporally, depending on the distribution of trips and mode shares. This stems from the fact that a ridesharing service will not be able to match all potential trips with one another, resulting in a system that cannot operate at perfect efficiency. In this work, we assume that any customer who cannot be matched will be turned away, representing uncaptured rideshare demand. Using data for Boston, we estimate a parameter to capture the average efficiency of the rideshare service across all OD pairs and hours, enabling us to define a model to estimate the total change in vehicles given auto and non-auto rideshare adoption rates and aggregate mode shares. Future research should explore this relationship in other cities with different demand and travel mode distributions.

Lastly, by simulating traffic for several rideshare adoption scenarios, we evaluate the impact of ridesharing on cumulative vehicle travel time and distance. We find that under moderate to high adoption rate scenarios, ridesharing would likely have noticeable impacts on congested travel times, indicating the importance of incorpo-

rating traffic simulation into ridesharing studies. Future work could explore variable adoption rates dependent on trip attributes, such as distance or time of day, as well as socioeconomic characteristics of travelers or trip origins.

Chapter 6

Conclusion

Today, vast amounts of data are generated and collected unobtrusively from mobile phones around the world. Moreover, advances in computing speed and storage capabilities have dramatically decreased the cost of storing and analyzing data. This environment has opened the door for transformative shifts in the way we model travel behavior.

For transportation applications, travel demand modeling has historically relied on travel surveys for development, calibration, and validation. Although sophisticated, these models have been limited by the scope, scale, and quality of infrequent, expensive household travel surveys. Mobile phone data, on the other hand, doesn't have detailed information about respondents and their travel decisions, but captures orders of magnitude larger sample sizes and can be collected in quickly and cheaply. In conjunction, these two data sources could fundamentally improve our understanding of human mobility.

But to leverage mobile phone data for travel demand estimation, an interdisciplinary approach that brings data mining, computing, and statistical methods together with the transportation community's expertise in transportation system dynamics is necessary. Incorporating big data resources into transportation modeling frameworks, and developing new methods and tools altogether, was precisely the goal of this research.

Although researchers from the urban computing community have developed meth-

ods to extract mobility patterns from mobile phone data, we did so here with transportation planning applications and existing limitations specifically in mind. Further, we presented a comprehensive set of algorithms to transform raw mobile phone data into origin-destination trips and finally road usage, with these methods validated against survey data. We then presented a portable, efficient, and flexible software platform that implements these methods such that they can be quickly and easily applied to any city. Lastly, we demonstrated an application of mobile phone data utilizing and adapting these methods to evaluate ridesharing.

We showed that using mobile phone data, we could extract origin-destination trips by purpose and time of day–key market segments analyzed by transportation modelers–comparable to information reported in local surveys. Furthermore, our methods of estimating vehicle trips and assigning these vehicles to road segments produced origin-destination travel times, as well as road segment volumes and congestion levels similar to surveys. While these results show great progress in making big data useful for transportation engineering, there are still limitations inherent in this data and our models. Specifically, we highlight three areas that are ripe for further study.

1. We have shown the level of aggregation applied to OD matrices can affect the correlation observed between model outputs. This is a standard manifestation of the modifiable area unit problem and a more detailed exploration may indicate which levels of analyses were better suited for different data sources. Moreover, a more detailed analysis of uncertainty in model estimates may make it easier to assess their correlation and validity.

2. Our traffic assignment algorithm is efficient, but simple. In the future, a stochastic dynamic user equilibrium assignment methods should be explored and compared. Moreover, route choice modeling may be significantly improved by the availability of high resolution GPS trajectories of drivers. We believe our system's modular design makes it easy to incorporate these new models.

3. Our mode choice model remains simple and will likely require more sophisti-

cation for modeling trips not taken in private vehicles. This, combined with improvements in route choice, may make it possible to estimate multi-modal trip demand, as public transportation, bike lanes, and even water transportation networks are included in OpenStreetMap data.

We hope future work will address these limitations and improve on the methods presented here. Furthermore, as more data becomes available in the form of calls, GPS traces, or real-time traffic monitoring systems, there is room for our methods to be improved by incorporating these new data sources. In particular, combining these other data sources with mobile phone data will capture a more complete picture of human mobility. Lastly, this data enables real-time or near real-time transportation demand management applications, but much research is needed to adapt available methods to handle real time data feeds and processing requirements.

Appendix A

Algorithms

ALGORITHM A.1: Parsing OpenStreetMap Networks

```
{OSM files are XML based and contain way and node objects}
ways = set of ways in an OSM file
nodes = set of nodes in an OSM file
graph = an empty graph
---
{Add each pair of consecutive nodes to the edge list}
for way in ways do
    for i = 0 to i = way.nodes.size() − 2 do
        graph.addNode(way.nodes[i])
        graph.addNode(way.nodes[i + 1])
        graph.addEdge(way.nodes[i], way.nodes[i + 1])
---
{Simplify the network by merging road segments }
for way in ways do
    startNode = way.nodes[0]
    for node in way.nodes do
        if all edges into and out of node are segments of the same way then
            graph.removeNode(node)
            remove all edges to or from node
        else
            endNode = node
            graph.addEdge(startNode, endNode)
            FillEdgeAttributes()
            startNode = endNode
---
{Notes}
*FillEdgeAttributes() fills in missing data such as speed limits or number of lanes based on way attributes
*graph.addNode(node) and graph.addEdge(node1, node2) only add objects if they do not already exist
*graph.removeNode(node) also removes all edges containing that node
*when simplifying the network, proper geographic lengths are kept even when nodes are deleted
```

ALGORITHM A.2: Stay Point Algorithm - Step 1 - Initialize

1: {Each user object has a number of attributes}
2: $call$ = a call object with an associated latitude, longitude, stay index
3: $calls$ = vector of a user's calls ordered by timestamp
4: $candidateSet$ = empty set of consecutive calls that meet criteria for a stay
5: $candidateStays$ = a vector of centroids from candidate sets
6: δ = distance threshold between consecutive calls (in meters)
7: τ = time threshold between entry into and exit from the stay (in seconds)
8: ds = a grid size for the agglomerative clustering algorithm (in meters)
9: $stayCalls$ = an empty vector of calls from stay points
10: {Notes}
11: *$Centroid(callSet)$ returns an object whose latitude and longitude are the centroid of all points in the input
12: *$DistanceBetweenCalls(call1, call2)$ returns the geographic distance between calls in meters
13: *$TimeBetweenCalls(call1, call2)$ returns the time between call in seconds

ALGORITHM A.3: Stay Point Algorithm - Step 2 - Candidate Stays

1: {For each user, loop through all calls and find candidate stays}
2: $candidateIndex = 0$
3: $candidateSet = \{\}$
4: **for** $i = 0$ **to** $i = calls.size() - 2$ **do**
5: **if** $DistanceBetweenCalls(calls[i], calls[i+1]) < \delta$ **then**
6: $candidateSet.append(calls[i+1])$
7: **else**
8: **if** $TimeBetweenCalls(candidateSet[0], candidateSet[end]) > \tau$ **then**
9: **for** $call$ **in** $candidateSet$ **do**
10: $call.stayIndex = candidateIndex$
11: $candidateStay = Centroid(candidateSet)$
12: $candidateStays.append(candidateStay)$
13: $candidateSet = \{calls[i]\}$
14: $candidateIndex = candidateIndex + 1$

ALGORITHM A.4: Stay Point Algorithm - Step 3 - Agglomerative Clustering

```
1: grid = construct a uniform grid that covers all of a user's calls with cell dimensions
   ds × ds
2: stayIndex = 0
3: for grid cells containing a candidateStay do
4:     candidateStays = {listofcandidateStayincell}
5:     stay = Centroid(candidateStays)
6:     for call made from a candidateStay in this cell do
7:         call.longitude = stay.longitude
8:         call.latitude = stay.latitude
9:         call.stayIndex = stayIndex
10:        stayCalls.append(call)
11:    stayIndex = stayIndex + 1
```

ALGORITHM A.5: Stay Point Algorithm - Step 4 - Final Pass

```
1: {Final pass to add any remaining calls to the stay}
2: for i = 0 to i = calls.size() do
3:     if call not part of a stay and DistanceBetweenCalls(call, stay) < δ for any
   stay then
4:         call.longitude = stay.longitude
5:         call.latitude = stay.latitude
6:         call.stayIndex = stayIndex
7:         stayCalls.append(call)
8: Sort stayCalls by timestamp
9:
```

ALGORITHM A.6: OD Creation Algorithm - Step 1 - Home / Work Expansion

{Data objects}
tracts = census tract data objects containing demographic variables
$OD(o, d, p, t) = 0$ for origin o, destination d, purpose p, and period t

{Detect home and work for all users and compute expansion factors}
for *user* **in** *users* **do**
    user.stays = vector of calls at stay points sorted by time
    user.home = index of stay point visited the most between 8pm and 7am on weekdays
    user.work = index of non-home stay point visited the most between 7am and 8pm on weekdays
    if user visits work less than once per week **then**
        user.work = null
    for *stay* **in** *user.stays* **do**
        stay.label assigned as *home*, *work*, or *other*
    user.weekdays = number of weekdays a user records a stay
    user.workdays = number of weekdays a user records a stay at work
    $tract[user.home].numUsers = tract[user.home].numUsers + 1$
for *tract* **in** *tracts* **do**
    $tract.expansionFactor = tract.population/tract.numUsers$

ALGORITHM A.7: OD Creation Algorithm - Step 2 - Trip Counting

{Count and expand trips}
for *user* **in** *users* **do**
    trips = empty vector to store trips taken by a user
    for $i = 1$ **to** $i = user.stays.size()$ **do**
        $s0 = user.stays[i - 1]$
        $s1 = user.stays[i]$
        if $s0 == S1$ **then**
            continue
        if $s0$ and $s1$ are on the same effective day **then**
            trip = new trip from $s0$ to $s1$
            $trip.purpose = PurposeFromLabels(s0, s1)$
            trip.workday = true if workday for user, false otherwise
            $trip.departure = GetConditionalDepartureTime(s0, s1)$
            $trips.append(trip)$
        else$s0$ and $s1$ are not on the same effective day
            morning = create trip from home to first recorded stay
            night = create trip from last recorded stay to home
            $trips.append(morning)$
            $trips.append(night)$

```
    for trip in trips do
        o = trip.origin
        d = trip.destination
        p = trip.purpose
        t = trip.departure
        if trip.workday == true then
            flow = tract[user.home].expansionFactor/user.workdays
        else
            flow = tract[user.home].expansionFactor/user.weekdays
        OD(o, d, p, t) = OD(o, d, p, t) + flow
```

{Notes}

*$PurposeFromLabels(s0, s1)$ returns a trip purpose (HBW, NHB, HBO) based on the label of origin and destination stays

*$GetConditionalDepartureTime(s0, s1)$ returns a departure time based on the observation times at origin and destination

*an effective day is defined as a period between 3am today until 3am on the next consecutive morning

ALGORITHM A.8: Incremental Traffic Assignment

```
graph = road network
OD(p, t) = origin-destination matrix for purpose p and time window t
B = a bipartite network containing roads and census tracts
incrSize = vector of increment sizes, e.g. [0.4, 0.3, 0.2, 0.1]
nBatches = number of threads to use
for i = 0 to i < incrSize.size() do
    for b = 0 to b < nBatches do
        create new thread
        batch = GetBatch(OD, b)
        for all o, d pairs in batch do
            flow = OD[o, d].flow · incrSize[i]
            route = A*(o, d, graph)
            for all segment s in route do
                s.flow = s.flow + flow
                B_{e→o} = B_{s→o} + flow
    wait for all threads to finish
    for segment s in graph do
        s.cost ← s.freeFlowTime · (1 + α((s.volume / s.capacity)^β))
```

* GetBatch(OD, B) returns only the subset of OD pairs pertaining to a batch

* $A^*(o, d, graph)$ returns the shortest path between o and d if a path exists

Bibliography

[1] *2014 TIGER/Line Shapefiles.* `www.census.gov/cgi-bin/geo/shapefiles2014/main`.

[2] American Community Survey, 2006-2010 5-year estimates. `http://www.census.gov/acs/www/data_documentation/2010_release/`.

[3] Transportation Research Board National Cooperative Highway Research Program. Report 716: Travel Demand Forecasting: Parameters and Techniques, 2012.

[4] Rahmi Akcelik. Travel time functions for transport planning purposes: Davidson's function, its time dependent form and alternative travel time function. *Australian Road Research*, 21(3), 1991.

[5] Lauren Alexander, Shan Jiang, Mikel Murga, and Marta C. Gonzalez. Origin-destination trips by purpose and time of day inferred from mobile phone data. *Transportation Research Part C: Emerging Technologies*, 2015.

[6] Alex Anas. Discrete choice theory, information theory and the multinomial logit and gravity models. *Transportation Research Part B: Methodological*, 17(1):13–23, 1983.

[7] Theo Arentze and Harry Timmermans. *Albatross: a learning based transportation oriented simulation system.* Eirass, 2000.

[8] Yasuo Asakura and Eiji Hato. Tracking survey for individual travel behaviour using mobile communication instruments. *Transportation Research Part C: Emerging Technologies*, 12(3):273–291, 2004.

[9] Hillel Bar-Gera. Evaluation of a cellular phone-based system for measurements of traffic speeds and travel times: A case study from israel. *Transportation Research Part C: Emerging Technologies*, 15(6):380–391, 2007.

[10] Holger Bast, Stefan Funke, Peter Sanders, and Dominik Schultes. Fast routing in road networks with transit nodes. *Science*, 316(5824):566–566, 2007.

[11] Edward Beimborn and Rob Kennedy. Inside the blackbox: Making transportation models work for livable communities. 1996.

[12] Vitaly Belik, Theo Geisel, and Dirk Brockmann. Natural human mobility patterns and spatial spread of infectious diseases. *Physical Review X*, 1(1):011001, 2011.

[13] Michael G. H. Bell. The estimation of origin-destination matrices by constrained generalised least square. *Transportation Research Part B: Methodological*, 25:13–22, 1991.

[14] Moshe Ben-Akiva and Michel Bierlaire. Discrete choice methods and their applications to short term travel decisions. In *Handbook of transportation science*, pages 5–33. Springer, 1999.

[15] Moshe Ben-Akiva and Bruno Boccara. Discrete choice models with latent choice sets. *International Journal of Research in Marketing*, 12:9–24, 1995.

[16] Moshe E Ben-Akiva and Steven R Lerman. *Discrete choice analysis: theory and application to travel demand*, volume 9. MIT press, 1985.

[17] Boston Metropolitan Planning Organization. 1991 Boston Household Travel Survey. `http://www.surveyarchive.org/Boston/Boston_91.zip`, 1991.

[18] John L Bowman and Moshe E Ben-Akiva. Activity-based disaggregate travel demand model system with activity schedules. *Transportation Research Part A: Policy and Practice*, 35:1–28, 2001.

[19] David Branston. Link capacity functions: A review. *Transportation Research*, 10(4):223–236, 1976.

[20] Dirk Brockmann, Lars Hufnagel, and Theo Geisel. The scaling laws of human travel. *Nature*, 439(7075):462–465, 2006.

[21] M. Batty C. Roth, S. Kang and M. Barthelemy. Structure of urban movements: polycentric activity and entangled hierarchical flows. *PLoS One*, 6, 2011.

[22] N Caceres, JP Wideberg, and FG Benitez. Deriving origin destination data from a mobile phone network. *Intelligent Transport Systems, IET*, 1(1):15–26, 2007.

[23] Francesco Calabrese, Mi Diao, Giusy Di Lorenzo, Joseph Ferreira Jr, and Carlo Ratti. Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. *Transportation Research Part C: Emerging Technologies*, 26:301–313, 2013.

[24] Julián Candia, Marta C González, Pu Wang, Timothy Schoenharl, Greg Madey, and Albert-László Barabási. Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical*, 41(22):224015, 2008.

[25] Ennio Cascetta. Estimation of trip matrices from traffic counts and survey data: a generalized least squares estimator. *Transportation Research Part B: Methodological*, 18(4):289–299, 1984.

[26] Central Transportation Planning Staff (CTPS). Model-based highway data. `http://www.ctps.org/Drupal/data_resources`, 2010.

[27] Daniel McFadden Charles F Manski et al. *Structural analysis of discrete data with econometric applications.* MIT Press Cambridge, MA, 1981.

[28] J.M. Choukroun. A general framework for the development of gravity-type trip distribution models. *Regional Science and Urban Economics*, 5:177–202, 1975.

[29] Blerim Cici, Athina Markopoulou, Enrique Frias-Martinez, and Nikolaos Laoutaris. Assessing the potential of ride-sharing using mobile and social data: a tale of four cities. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 201–211. ACM Press, 2014.

[30] L. Adamic S. Aral A.-L. Barabasi D. Brewer N. Christakis N. Contractor-J. Fowler M. Gutmann T. Jebara G. King D. Lazer, A. Pentland, D. Roy M. Macy, and M. Van Alstyne. Computational social science. *Science*, 323:721–723, 2009.

[31] Carlos F Daganzo. Optimal sampling strategies for statistical models with discrete dependent variables. *Transportation Science*, 14(4):324–345, 1980.

[32] Yves-Alexandre de Montjoye, César A Hidalgo, Michel Verleysen, and Vincent D Blondel. Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, 3, 2013.

[33] S. Erlander and N.F. Stewart. The gravity model in transportation analysis: theory and extensions. *Topics in transportation*, 3, 1990.

[34] A. Pentland P. Lukowicz D. Kossmann J. Crowley F. Giannotti, D. Pedreschi and D. Helbing. A planetary nervous system for social mining and collective awareness. *The European Physical Journal Special Topics*, 214:49–75, 2012.

[35] D. Pedreschi F. Pinelli C. Renso S. Rinzivillo F. Giannotti, M. Nanni and R. Trasarti. Unveiling the complexity of human mobility by querying and mining massive trajectory data. *The VLDB Journal*, 20:695–719, 2011.

[36] Joseph Ferreira, Mi Diao, Yi Zhu, Weifeng Li, and Shan Jiang. Information infrastructure for research collaboration in land use, transportation, and environmental planning. *Transportation Research Record: Journal of the Transportation Research Board*, 2183(1):85–93, 2010.

[37] Christopher R. Fleet and Sydney R. Robertson. Trip generation in the transportation planning process. *Highway Research Record*, 1968.

[38] Marta C González, Cesar A Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[39] R. Hariharan and K. Toyama. Project lachesis: parsing and modeling location histories. *Geographic Information Science*, pages 106–124, 2004.

[40] S. Hasan, C. Schneider, S. V. Ukkusuri, and M. C. González. Spatiotemporal patterns of urban human mobility. *Journal of Statistical Physics*, 151(1-2), 2013.

[41] Martin L Hazelton. Estimation of origin–destination matrices from link flows on uncongested networks. *Transportation Research Part B: Methodological*, 34(7):549–566, 2000.

[42] Martin L Hazelton. Inference for origin–destination matrices: estimation, prediction and reconstruction. *Transportation Research Part B: Methodological*, 35(7):667–676, 2001.

[43] Martin L Hazelton. Some comments on origin–destination matrix estimation. *Transportation Research Part A: Policy and Practice*, 37(10):811–822, 2003.

[44] P. S. Hu and T. R. Reuscher. Summary of travel trends: 2001 national household travel survey. Technical report, U.S. Department of Transportation Federal Highway Administration, 2004.

[45] L. F. Huntsinger and R. Donnelly. Reconciliation of regional travel model and passive device tracking data. In *Proceedings of the 93rd Annual Meeting of the Transportation Research Board*, 2014.

[46] Md Shahadat Iqbal, Charisma F Choudhury, Pu Wang, and Marta C González. Development of origin–destination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies*, 40:63–74, 2014.

[47] Shan Jiang, Gaston A Fiore, Yingxiang Yang, Joseph Ferreira Jr, Emilio Frazzoli, and Marta C González. A review of urban computing for mobile phone traces: current methods, challenges and opportunities. In *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*, page 2. ACM, 2013.

[48] F.S. Koppelman and E.I. Pas. Estimation of disaggregate regression models of person trip generation with multiday data. *In Papers presented during the Ninth International Symposium on Transportation and Traffic Theory held in Delft the Netherlands, 11-13 July 1984*, 1984.

[49] N. D. Lane, E. Miluzzo, H. Lu, D. Peebles, T. Choudhury, and A. T. Campbell. A survey of mobile phone sensing. *Communications Magazine, IEEE*, 48(9):140–150, 2010.

[50] D. M. Levinson and A. Kumar. The rational locator: why travel times have remained stable. *Journal of the american planning association*, 60(3):319–332, 1994.

[51] HP Lo, N Zhang, and William HK Lam. Estimation of an origin-destination matrix with random link choice proportions: a statistical approach. *Transportation Research Part B: Methodological*, 30(4):309–324, 1996.

[52] Chung-Cheng Lu, Xuesong Zhou, and Kuilin Zhang. Dynamic origin–destination demand flow estimation under congested traffic conditions. *Transportation Research Part C: Emerging Technologies*, 34:16–37, 2013.

[53] Xin Lu, Linus Bengtsson, and Petter Holme. Predictability of population displacement after the 2010 haiti earthquake. *Proceedings of the National Academy of Sciences*, 109(29):11576–11581, 2012.

[54] MJ Maher. Inferences on trip matrices from observations on link volumes: a bayesian statistical approach. *Transportation Research Part B: Methodological*, 17(6):435–447, 1983.

[55] MassGIS. Community Boundaries. `http://www.mass.gov/anf/research-and-tech/it-serv-and-support/application-serv/office-of-geographic-information-massgis/datalayers/towns.html`, 2014.

[56] Gerald M. McCarthy. Multiple-regression analysis of household trip generation-a critique. *Highway Research Record*, 1969.

[57] Michael G. Mcnally. The four step model. *Handbook of Transport Modelling*, 2000.

[58] Deepak K Merchant and George L Nemhauser. A model and an algorithm for the dynamic traffic assignment problems. *Transportation Science*, 12(3):183–199, 1978.

[59] Kenneth Train Joan Walker Chandra Bhat Michel Bierlaire Denis Bolduc Axel Borsch-Supan David Brownstone David S Bunch Moshe Ben-Akiva, Daniel McFadden et al. Hybrid choice models: Progress and challenges. *Marketing Letters*, 13:163–175, 2002.

[60] Yu Nie, HM Zhang, and WW Recker. Inferring origin–destination trip matrices with a decoupled gls path flow estimator. *Transportation Research Part B: Methodological*, 39(6):497–518, 2005.

[61] NUSTATS. *Massachusetts Department of Transportation: 2010/2011 Massachusetts Travel Survey*, 2012.

[62] Walter Y. Oi and Paul W. Shuldiner. An analysis of urban travel demands. 1962.

[63] J. D. Ortúzar and Luis G Willumsen. *Modelling transport*. John Wiley & Sons, Chichester, England, 1994.

[64] Santi Phithakkitnukoon, Teerayut Horanont, Giusy Di Lorenzo, Ryosuke Shibasaki, and Carlo Ratti. Activity-aware map: Identifying human daily activity pattern using mobile phone data. *Human Behavior Understanding*, pages 14–25, 2010.

[65] A. R. Pinjari and C. R. Bhat. Activity-based travel demand analysis. *A Handbook of Transport Economics*, 1:1–36, 2011.

[66] Lisa Rayle, Susan Shaheen, Nelson Chan, Danielle Dai, and Robert Cervero. App-Based, On-Demand Ride Services: Comparing Taxi and Ridesourcing Trips and User Characteristics in San Francisco. University of California Transportation Center Working Paper, August 2014. `www.uctc.net/research/papers/UCTC-FR-2014-08.pdf`.

[67] Jonathan Reades, Francesco Calabrese, and Carlo Ratti. Eigenplaces: analysing cities using the space-time structure of the mobile phone network. *Environment and Planning B: Planning and Design*, 36(5):824–836, 2009.

[68] Lothlorien S. Redmond and Patricia L. Mokhtarian. Modeling objective mobility: The impact of travel-related attitudes, personality and lifestyle on distance traveled. 2001.

[69] Mei-Po Kwan Reginald G Golledge and Tommy Gdrling. Computational process modeling of household travel decisions using a geographical information system. *Papers in regional science*, 73:99–117, 1994.

[70] Anthony J Richardson, Elizabeth S Ampt, and Arnim H Meyburg. *Survey methods for transport planning.* Eucalyptus Press Melbourne, 1995.

[71] Paolo Santi, Giovanni Resta, Michael Szell, Stanislav Sobolevsky, Steven H. Strogatz, and Carlo Ratti. Quantifying the benefits of vehicle pooling with shareability networks. *Proceedings of the National Academy of Sciences of the United States of America*, 111(37):13290–13294, 2014.

[72] A. Schafer. Regularities in travel demand: an international perspectiv. *Journal of transportation and statistics*, 3(3):1–31, 2000.

[73] Christian M Schneider, Vitaly Belik, Thomas Couronné, Zbigniew Smoreda, and Marta C González. Unravelling daily human mobility motifs. *Journal of The Royal Society Interface*, 10(84):20130246, 2013.

[74] Andres Sevtsuk and Carlo Ratti. Does urban mobility have a daily routine? learning from the aggregate data of mobile networks. *Journal of Urban Technology*, 17(1):41–60, 2010.

[75] Paul B. Slater. Hierarchical internal migration regions of france. *Systems, Man and Cybernetics, IEEE Transactions*, 4:321–324, 1976.

[76] Michael E Smith. Design of small-sample home-interview travel surveys. *Transportation Research Record*, 701:29–35, 1979.

[77] Chaoming Song, Tal Koren, Pu Wang, and Albert-László Barabási. Modelling the scaling properties of human mobility. *Nature Physics*, 6(10):818–823, 2010.

[78] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

[79] Heinz Spiess. A maximum likelihood model for estimating origin-destination matrices. *Transportation Research Part B: Methodological*, 21(5):395–412, 1987.

[80] Heinz Spiess. Technical note–conical volume-delay functions. *Transportation Science*, 24(2):153–158, 1990.

[81] Peter R Stopher and Stephen P Greaves. Household travel surveys: Where are we going? *Transportation Research Part A: Policy and Practice*, 41(5):367–381, 2007.

[82] S.A. Stouffer. Intervening opportunities: a theory relating mobility and distance. *American sociological review*, pages 845–867, 1940.

[83] HJP Timmermans and Dick Ettema. Activity-based approaches to travel analysis. *Pergamon*, 1997.

[84] U.S. Department of Transportation Federal Highway Administration. 2009 National Household Travel Survey. `http://nhts.ornl.gov/download.shtml`, 2011.

[85] U.S. Department of Transportation Federal Highway Administration. CTPP 2006-2010 Census Tract Flows. `http://www.fhwa.dot.gov/planning/census_issues/ctpp/data_products/2006-2010_tract_flows/index.cfm`, 2013.

[86] Henk J Van Zuylen and Luis G Willumsen. The most likely trip matrix estimated from traffic counts. *Transportation Research Part B: Methodological*, 14(3):281–293, 1980.

[87] Pu Wang, Timothy Hunter, Alexandre M Bayen, Katja Schechtner, and Marta C González. Understanding road usage patterns in urban areas. *Scientific reports*, 2(1001), 2012.

[88] John Glen Wardrop. Road paper. some theoretical aspects of road traffic research. In *ICE Proceedings: Engineering Divisions*, volume 1, pages 325–362. Thomas Telford, 1952.

[89] Amy Wesolowski, Nathan Eagle, Andrew J Tatem, David L Smith, Abdisalan M Noor, Robert W Snow, and Caroline O Buckee. Quantifying the impact of human mobility on malaria. *Science*, 338(6104):267–270, 2012.

[90] A.G. Wilson. The use of entropy maximising models, in the theory of trip distribution, mode split and route split. *Journal of Transport Economics and Policy*, pages 108–126, 1969.

[91] A.G. Wilson. *Entropy in urban and regional modelling.* Pion Ltd, 1970.

[92] A.G. Wilson. Land-use/transport interaction models: Past and future. *Journal of Transport Economics and Policy*, pages 3–26, 1999.

[93] H. Yang, T. Sasaki, Y. Iida, and Y. Asakura. Estimation of origin-destination matrices from link traffic counts on congested networks. *Transportation Research Part B: Methodological*, 26:417–434, 1992.

[94] Xianyuan Zhan, Samiul Hasan, Satish V Ukkusuri, and Camille Kamga. Urban link travel time estimation using large-scale taxi data with partial information. *Transportation Research Part C: Emerging Technologies*, 33:37–49, 2013.

[95] G.K. Zipf. The p1 p2/d hypothesis: on the intercity movement of persons. *American sociological review*, 11:677–686, 1946.