Understanding Human Mobility Patterns from Digital Traces

by

Yingxiang Yang

B.E., Southeast University (2011)

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 2013

Author .....
Department of Civil and Environmental Engineering
May 10, 2013

Certified by .....
Marta C. González
Gilbert Winslow Career Development Assistant Professor
Thesis Supervisor

Accepted by .....
Heidi M. Nepf
Chairman, Department Committee on Graduate Theses

Understanding Human Mobility Patterns from Digital Traces

by

Yingxiang Yang

Submitted to the Department of Civil and Environmental Engineering
on May 10, 2013, in partial fulfillment of the
requirements for the degree of
Master of Science in Transportation

Abstract

Our current digital age is characterized by the shift from traditional industry to an economy based on the information computerization. The sweeping changes brought about by digital computing have provided new data sources for transportation modeling. In this thesis, two mainstream trends in utilizing digital traces in transportation modeling are explored.

The first approach is to incorporate mobile phone records and digital map point of interests into commuting flow prediction models such as the gravity model and the radiation model. An extension to the radiation model is proposed to adjust to the different degrees of homogeneity of opportunities when the scale of the study region changes. The density of the point of interests is a suitable proxy for commuting flow attraction rates at all the scales. Moreover, the parameter α in the extension to the radiation model is predictable given the size of the study region. When traditional data sources are not available, mobile phone records is shown to be an ideal alternative. Home and work locations can be inferred at individual level and then aggregated to show its equivalence to the census data. This method is applied to Rwanda, Dominican Republic and Portugal.

The second approach is using low-frequency bus GPS records to evaluate transit service. The analysis under such data scarcity requires careful data handling. This thesis demonstrates that how the data pre-processing procedure, namely map-matching and kernel density estimation, step by step turns the raw GPS data into information for service evaluation. Bus service quality is analyzed by measuring statistics of headway and in-vehicle travel time. The headway analysis helps to identify bottlenecks caused by the road network layout and passenger volumes while the comparison of peak vs. off-peak hour travel speed helps to identify bottlenecks caused by traffic conditions.

To sum up, the thesis explores new digital data sources and methods in transportation modeling. The purpose is to provide analysis procedures that are of lower costs, higher accuracy and are readily applicable to different countries in the world.

Thesis Supervisor: Marta C. González

Title: Gilbert Winslow Career Development Assistant Professor

Acknowledgments

First and foremost I would like to express my sincere gratitude to my advisor Marta C. González for her continuous support of my study and research. What she brings to the whole research group is not only wisdom, but also happiness, enthusiasm and motivation. I'd like to thank Prof. González especially for her insightful guidance in my research and her patient correction of my writing. Her immense knowledge and patience help me to appreciate the beauty of science.

To my parents, there is never enough words to fully express my gratitude to you. I learn from you the principle of life, the pursuit of dreams, and the meaning of responsibility. You've always been supportive to my decisions. Though we are thousands of kilometers apart, my heart will always be with you.

To all the members of the HuMNet research group: Christian, Jameson, Serdar, Carlos, Vitally, Pu, Shan and Gaston for the illuminating discussions and the help in everyday life. I also want to thank the other lab mates in room 1-151 and 1-249. I'll never forget the joy in such a multi-culture lab.

This dissertation would not have been possible without the collaboration and the help from the following individuals: David, Dietmar, and Peter. I want to thank our Austrian collaborators: Thank you for the wonderful summer in Vienna.

Last but not the least, I would like to thank the CEE department for providing such a delightful environment. All the courses I've taken and all the talks with the professors will give me a lifetime of benefits.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction and Overview

The transportation industry encompasses movements of people, goods, and information, which are fundamental components of human societies. Millions of employees in this huge industry work to ensure the daily movement of the societies. In the US, more than 10 million people are working in the transportation industry (US Department of Transportation, BTS, G-7 Countries: Transportation Highlights.), contributing 11% of the total GDP (U.S. Department of Transportation, Bureau of Transportation Statistics, Pocket Guide to Transportation). The transportation industry is essential in economic development because it provides mobility needs and insures access to markets and resources for economic activities. From the industrial revolution in the 19th century to the globalization in the late 20th century, every important transformation in the human societies is accompanied by a heap in the transportation industry.

It is because the transportation industry's unique position in the development of societies that transportation modeling has always attracted the attention of planners, engineers, and economists. Transportation modeling is used to develop information to help make decisions on the future development and management of transportation systems, especially in urban areas. They are used to estimate the number of trips that will be made on a transportation alternative at some future date. These estimates are the basis for transportation plans and are used in major investment analysis,

environmental impact statements and in setting priorities for development (14).

As mentioned before, transportation encompass the mobility of human, goods, and information, among which human mobility is the most directly sensible one. People may not be aware that how commodities are transported from factories to their local stores, but they all care about the quality of their own travel activities. That's why human mobility patterns have been broadly studied. Having a clear map of the travel demands and daily activities of the diverse urban groups can help us to design better cities. This is becoming a pressing need as traffic congestion and pollution are becoming worldwide issues. Even the total travel demand keeps increasing, we can still adopt strategies to alleviate such problems. Such strategies may include: re-balance the spatial temporal distribution of travels; spread out the peak of travel, re-allocate the land use in a city to avoid long distance separation of residential and industrial regions. Understanding daily travels and activities could also help us to organize cities as sustainable systems: using less energy, less water and producing less waste per-capita comparing to sprawling alternatives.

Data is the prerequisite of accurate modeling. If the input data is of low quality, the output will inevitably also of low quality however sophisticated a model is. Therefore in human mobility modeling a large portion of time and money are spent on data collection. The most classic modeling method is the four step model, which dates back to the 1950s and is still widely used all around the world. The four steps are: trip generation, trip distribution, mode choice, and route assignment. The four step model has significant data demands in addition to that required to define the activity and transportation systems. The primary need is for data that define travel behavior, and these are gathered via a variety of survey efforts. Household travel surveys with travel-activity diaries provide much of the data that are required to calibrate the four step model. These data and observed traffic studies (counts and speeds) provide much of the data needed for validation. Nowadays a 24 hour survey of approximately 3,000 households costs the city about $1.5 million. Because they are expensive and intrusive, surveys only describe typically a sample day, which is a very limited time period.

The alternatives to overcome the limitations of surveys are using the digital devices around us. Our current digital age is characterized by the shift from traditional industry to an economy based on the information computerization. The sweeping changes brought about by digital computing and communication technology during the latter half of the 20th century have provided new data sources for transportation modeling. Automatic Data Collection Systems (ADCS) on public transit systems provide abundant long-term, high accuracy data through vehicle GPS systems, smartcard transaction records, and automatic passenger counting systems. These systems can not only be used to monitor real time system operation, but also be used to improve system efficiency and diagnose service bottlenecks.

While the aforementioned automatic data collection systems can provide data only for public transit, there is another data source, mobile phone transaction data, which can provide information for nearly all the travelers. Mobile phones records can pinpoint a user's location, either to the nearest cell tower or within meters using GPS or WIFI sensors, which has the potential to deepen our understanding of human mobility patterns within a city. Questions about the micro-structure of a city such as where individuals live and work, which previously could only be answered with surveys, can now be explored using mobile phone data. Such a 24/7 information source has broad impacts on both urban planning and epidemiology. Now we can track how the population moves during different times of day and thus infer how land is used dynamically, which is a mission impossible for static and often antiquated zoning and regulatory data. Furthermore, knowing how people move within and between cities will provide needed insights into social contact networks used by epidemiologists to model disease spread in urban environments.

The geo-located digital information helps us to create an Internet that connects real world objects: Internet of Things. Together with the techniques of cloud computing, crowd sourcing, and wireless sensor networks, a city that has higher efficiency, sustainability, and livability is no longer a distant view.

1.2 Literature Review

Transportation modeling requires interdisciplinary cooperation. Different communities have different modeling approaches which have both similarities and differences. These communities can be mainly classified into three categories: the transportation engineering community; the statistical physics community; and the computer science community. These three communities have different characteristics. The transportation engineering community relies more on surveys and is gradually using more automatically collected data. The researches usually try to build a connection between people's travelling patterns and their socio-economic characteristics. The statistical physics community captures collective trips at large scales. The computer science community focuses on using data mining and machine learning techniques to extract patterns from large amount of data. Nowadays as the cooperation between different disciplines is getting closer, the boundary between different communities is diminishing.

1.2.1 Approaches from the Transportation Engineering Community

When it comes to transportation modeling, an inevitable model is the four step model. It has a long history and is still widely used all over the world. It is also where this section will start at. As the digital traces around us are becoming more abundant, automatically collected data is gradually taking the place of manual surveys. The most widely used automatically collected data in transportation modeling and planning is in the public transit system. The vehicle GPS data, smartcard swiping records and automatic passenger counting data can be used to infer passenger flows, track vehicle locations, adjust scheduling, and even reroute bus routes.

The Four Step Model

The four step model originated in the 1950s (86). It decomposes the whole travelling procedure into the following four steps:

1. Trip generation, which determines the frequency of origins or destinations of trips in each zone by trip purpose, as a function of land uses and household demographics, and other socio-economic factors.

2. Trip distribution, which matches origins with destinations.

3. Mode choice computes the proportion of trips between each origin and destination that uses a particular transportation mode.

4. Route assignment, which allocates trips between an origin and destination by a particular mode to a route.

It fits well into the transportation analysis framework proposed by Manheim (82) and other scholars (47). The first comprehensive application of the four step model is in the Chicago Area Transportation Study (114). Several federal legislation requirement in the 1960s and 1970s helped to institutionalize the four step model and bring environmental concerns into it. A flow chart of the four step model is shown in Figure 1-1.

For the first step, the core question is to answer: How many trips per family? To answer this question, data need to be collected from surveys such as: Number of trips as a function of number of people per household, of number of cars, type of dwelling, residential area, distribution among trip purposes, distribution between motorized and non-motorized, distribution between chained and un-chained trips, number of captive public transport users, etc. The models used in this step are usually regression analysis (46; 84; 73; 88) or cross-category analysis (96).

The second step calculates the elements in the Origin-Destination Matrix from the marginal values derived from the first step. Various aggregate models of trip distribution have been proposed (89; 128; 118; 101; 117; 116). Among all these efforts, the gravity model, which assumes that the number of movements between an origin-destination pair decays with their distance, is the most widely used one(128; 42; 33). When an empirical OD is available the iterative proportional fitting (IPF) method can be used directly (98).

After the flow of each OD pair is assigned, the third step further splits the flow into different transportation modes. This step is dominated by discrete choice models (4; 16; 83; 17). It's a method for modeling choices from discrete alternatives. It's components include: decision-makers and their socio-economic characteristics; alternatives and their attributes. The decision maker selects the alternative with the highest utility among all the alternatives in the choice set. The utility of an alternative is often a function of the characteristics of the traveller and the features of the travel mode.

The previous steps are all on zoning level. The last step assigns the flow of different modes between different OD pairs to road networks. Often (for high volume assignment) Wardrop's principle of user equilibrium is applied (equivalent to a Nash equilibrium), wherein each driver (or group) chooses the shortest (travel time) path, subject to every other driver doing the same (1; 74). The difficulty is that travel times and demand rely on each other so that it's hard to determine where the equilibrium point is. Other more computationally economic methods include all-or-nothing method which do not consider congestion and assign all travellers to the shortest path. This method is hardly used because it deviates from reality. Other more realistic methods adjust the congestion level of the road network gradually and assign the flows dynamically (37; 7; 87).

The four step model has limitations such as:

- Demands are for trip making rather than for activities.
- The purpose of making trips should not be just making trips but to perform certain kind of activities.
- Person-trips as the unit of analysis which lacks household interactions.
- Sequential nature of the four-step process makes behavior modeled in earlier steps unaffected by choices modeled in later steps.
- Limited types of policies that can be analyzed.

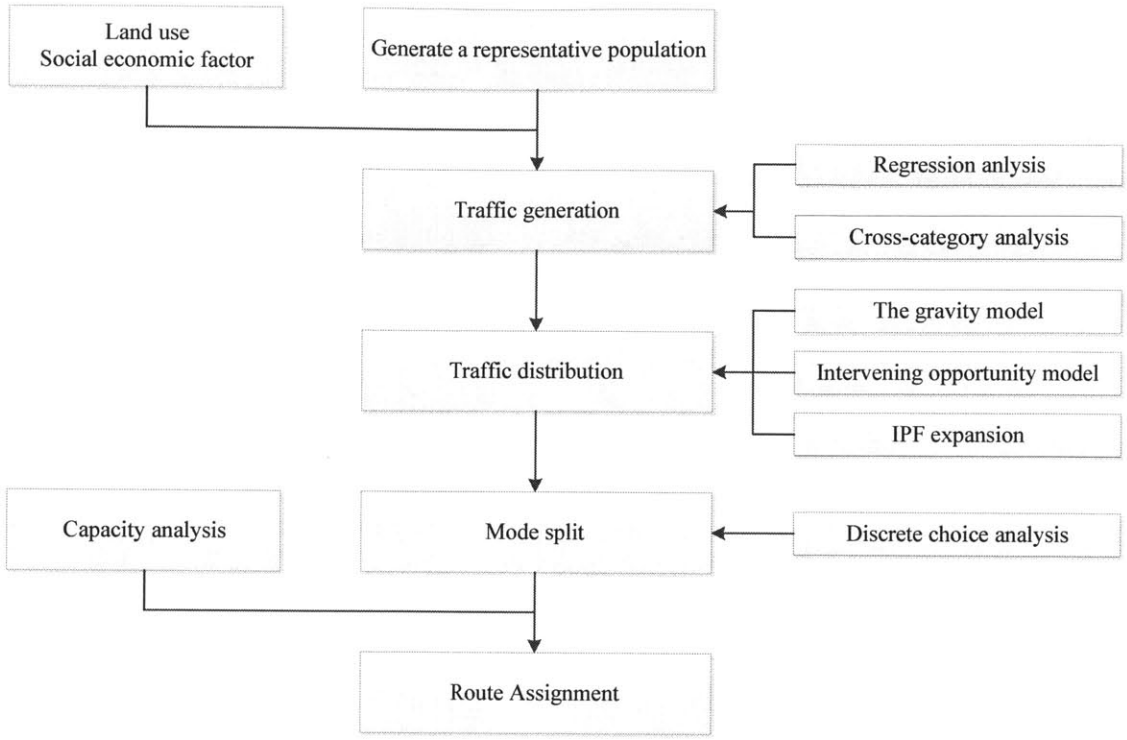


Figure 1-1: Four step flowchart and the mainstream models used in each step

All the disadvantages of the trip based four step model lead to the activity based models in which travel demand is derived from demand for activities; tours are inter-dependent; people face time and space constraints that limit their activity schedule choice; activity and travel scheduling decisions are made in the context of a broader framework (19; 6; 106; 55).

Using Digital Traces in Public Transit

Geo-located digital traces are more prevalent in transportation engineering, especially public transportation. The automatic data collection systems such as the smartcard transaction record system are originally implemented just for collecting revenues. But later transportation researchers find that the data has much more potential values. The data can be used on estimating passenger flows, optimizing scheduling, planning new routes, etc.

Automatic Data Collection Systems for public transit (ADCs) are classified as AVL/GPS (Automatic vehicle location) systems, AFC (Automatic Fare Collection) systems and APC (Automatic Passenger Count) systems (22; 27). AVL systems record vehicle location information; AFC systems are often designed for the purpose of revenue management; APC systems count the numbers of boarding and alighting passengers.

It is generally agreed that automated data collection systems present the opportunity to do statistically valid analyses on service reliability for the first time (50; 70). Furth *et al.* (52) reviews past and potential applications of automatic data collection systems in service planning, scheduling, performance evaluation and system management. The study describes a number of analysis and decision support tools that have been, or could be, developed using the output of these systems.

Examples of the application of these ADC systems include bus schedule improvement (70; 51), travel time prediction (109), identifying causes of performance issues in bus schedule adherence (81), service reliability measurements (52), real time controls(40; 60).

The abundance of ADC data differ from system to system, this poses restrictions on the type of analysis could be done on each system (27; 22). Furth *et al.* (52) define four key dimensions in archived data: level of spatial and temporal detail, complete vs. exception data, fleet penetration and sample size, and data quality control.

Data detail increases from level A to E. With merely AVL data, only level A and B could be reached (though some AVL systems may pull at higher frequency).

AFC data is used mainly for OD Matrix deduction: A set of algorithms was applied to Metro Card trips to infer the destination stop for each origin stop (11); fare collection data was used to estimate Rail OD Matrix in London (30); GIS system was integrated with AFC and AVL system to enable study in more detailed transfer trips (123; 10); Multi-day AFC transaction record improved deduction accuracy of bus OD matrix (108). The study results could be further applied to transit assignment (104) and making transportation policies (43).

AVL data is used mainly on service quality evaluation. Abkowitz *et al.* (2; 3)

Table 1.1: Hierarchical ADC data detail

Level	Description	Event-Independent Records	Event Records	Between-Stop Performance Data
A	AVL without real-time tracking	infrequent (typically 60 to 120 s)	-	-
B	AVL with real-time tracking	infrequent (typically 60 to 120 s)	each timepoint	-
C	APC or event recorder	-	each stop	-
D	event recorder with between-stop summaries	-	each stop and between-stop events	recorded events and summaries
E	event recorder / trip recorder	very frequent (every second)	all types	all events, full speed profile

proposed a series of passenger-centric reliability metrics that capture the distribution of travel time, schedule adherence and headway distribution. The distributions are characterized by the mean, coefficient of variation (mean divided by the standard deviation), and the percentage of observations beyond a certain value. The new level of data available from AVL systems permits a refinement of the metrics. Camus *et al.* (25) systematically discussed the limitations of TCQSM's method for LOS estimation and proposed a new service measure called weighted delay index. This method is validated by AVL data. Chen *et al.* (32) proposed three bus service reliability measures: punctuality index based on routes, deviation index based on stops, and evenness index based on stops, which measures respectively evaluates bus service at stop, route and network level. AVL data is an ideal source to perform such analysis. Byon *et al.* (24) used low frequency GPS data to estimate bus headway distribution and adherence to schedule.

1.2.2 Approaches from the Statistical Physics Community

Statistical physics techniques go a long way in improving models of mobility networks. The circulation of bank notes is analyzed to show human travelling statistics (21). It

is shown that the distribution of travelling distances decays as a power law, indicating that trajectories of bank notes are reminiscent of scale-free random walks known as Lévy flights. Second, the probability of remaining in a small, spatially confined region for a time is dominated by algebraically long tails that attenuate the superdiffusive spread. Human travelling behaviour is described mathematically on many spatiotemporal scales by a two-parameter continuous-time random walk model.

We still have to notice that the movement of banknote is very different from the movement of people. On the other hand, mobile phone data is a much closer approximation to human trajectories. González *et al.* analyzed human mobility patterns by studying the trajectory of 100,000 anonymized mobile phone users whose position is tracked for a six-month period (56). They found that human trajectory is actually different from Lévy flight predicted by random walk. Human trajectories show a high degree of temporal and spatial regularity, each individual being characterized by a time independent characteristic travel distance and a significant probability to return to a few highly frequented locations. It is shown that the distribution of displacements over all users is well approximated by a truncated power-law:

$$P(\Delta r) = (\Delta r + \Delta r_0)^{-\beta} exp(-\Delta r/\kappa) \tag{1.1}$$

$\beta = 1.75 \pm 0.15$, $\Delta r_0 = 1.5km$ and the cutoff values $\kappa = 400km$ and $\kappa = 80km$. The radius of gyration r_g also follows a truncated power-law:

$$P(r_g) = (r_g + r_g^0)^{-\beta_r} exp(-r_g/\kappa) \tag{1.2}$$

$r_g^0 = 5.8km$, $\beta_r = 1.65 \pm 0.15$ and $\kappa = 350km$. Moreover, after correcting for differences in travel distances and the inherent anisotropy of each trajectory, the individual travel patterns collapse into a single spatial probability distribution.

Song *et al.* (99) show that the mechanisms of exploration and preferential return help to recover many important scaling laws in human mobility. Exploration mean that with probability

$$P_{new} = \rho S^{-\gamma} \tag{1.3}$$

the individual moves to a new location. S is the number of location he/she has already visited. Preferential return means that with the complementary probability $P_{ret} = 1 - \rho S^{-\gamma}$ the individual returns to one of the S previously visited locations. In this case, the probability to visit location is chosen to be proportional to the number of visits the user previously had to that location. These mechanisms help to grasp three important properties in human mobility:

- The number of distinct locations $S(t)$ visited by a randomly moving object is expected to follow $S(t) \sim t^{\mu}$, $\mu = 0.6 \pm 0.02$.

- The visiting frequency f of the kth most visited location follows Zipf's law $f_k \sim k^{-\xi}$ where $\xi \approx 1.2 \pm 0.1$.

- The convergence of the mean square displacement (MSD) predicted by the continuous time random walk is too slow when compared with empirical data.

Other applications include modeling social networks (57; 26; 39); finding the predictability of human mobility (100; 112); modeling the spreading of epidemics (115; 91; 92). For more detailed review of applications of statistical physics, please refer to (28; 12).

1.2.3 Approaches from the Computer Science Community

Machine learning, a branch of artificial intelligence, is about the construction and study of systems that can learn from data. An important application is about recognizing patterns from data, which makes it a perfect tool to analyze human mobility traces, to detect daily routines and anomaly behaviors. The mainstream models in machine learning can be classified as clustering, classification, graphical models, and reinforcement learning. It is intuitive to use graphical model and reinforcement learning algorithms to analyze human trajectories since a human trajectory is a time series and these algorithms can deal with the causal relationships of actions between consecutive time steps.

A graphical model is a probabilistic model for which a graph denotes the conditional dependence structure between random variables. Markov chain and hidden Markov chain are the simplest form of graphical models. But they are already quite powerful in modeling human daily routines (34; 68). The input data is usually GPS human traces, the output is the conditional probability distribution of a daily routine. More complex graphical models are usually called Markov random field for undirected graphs and Bayesian network for directed graphs. The more flexible graph structure means more modeling power, under the premise that the required input data quality can be reached. Hierarchical Markov model and Bayesian networks enable features such as mode choice and activity pattern choice (38; 77; 69; 76).

Another approach is using reinforcement learning, in which an agent take actions in an environment so as to maximize some pre-defined cumulative reward. Reinforcement learning integrates the concepts of reward (utility) maximization and context-dependent choice heuristics. The application of reinforcement learning is in many fields including robotics (80), game theory (105) and dispatching system (13). Applying reinforcement learning to human mobility has several advantages. First, the imitation of human learning through trial-and-error interactions with a dynamic environment helps to explain behavioral mechanisms (67). Secondly, it doesn't need an expert-system to inform it what selection is right and what is wrong. Thirdly, it could react to unforeseen events and take both long term learning and short term dynamics into account. Among the first attempts, Charypar and Nagel built the basic model of activity time plans using q-learning and got quite realistic results (31). This model was then modified to allocate both time and location choice of activity-travel pattern (66). Because q-learning generally takes a long time to converge and "the curse of dimensionality" limits the feasible dimensionality of the problem, q-learning was combined with regression tree to form the new algorithm called q-tree (110).

With the close collaboration between the urban planning and computer science, urban computing is emerging as a concept where every sensor, device, person, vehicle, building, and street in the urban areas can be used as a component to probe city dynamics to further enable city-wide computing for serving people and their cities

(125). The approach usually combines GPS trajectories with data sources from urban planning such as the land use within a city and census data. The aim is usually trying to infer potential interesting locations for different groups of people for location based services (127; 124) or to explore the influence of social networks on peoples activity-travel patterns (126; 121).

1.3 Thesis Outline

The rest of the thesis will show the applications of different digital traces in different aspects of transportation modeling.

Chapter 2 shows the application of mobile phone data and digital map point of interests in commuting flow estimation. The models used include the gravity model and the radiation model. This chapter shows that at large scales the radiation model's performance is comparable to models relying on adjustable parameters. But when zooming in from the inter-city scale to the inner-city scale, some extensions must be applied to the model. The radiation model is extended by adding one parameter α which reflects the influence of the size of the study region and the heterogeneity of the distribution of opportunities. The extended radiation model gives close commuting flow predictions to the census data at all the scales. For regions without detailed census data but with available cell phone records, a cell phone user OD matrix expansion method is proposed to gain insights into these regions' commuting flow characteristics. This method is validated in the Bay Area and then applied to three different countries: Rwanda, Dominican Republic, and Portugal, through which some special commuting flow characteristics are observed. These characteristics are also captured by the radiation model with the α parameter extension.

Chapter 3 shows the application of low frequency bus GPS data. The potential of "low-frequency" bus localization data for the monitoring and control of bus system performance is investigated. Data with a sampling rate as low as one minute, when processed appropriately, can provide ample information. In particular, accurate estimates of stop arrival and departure times which in turn allow the analysis

of headways and travel times can be obtained. A three parameter gamma family of distributions is fitted for headways at the stops along a bus line. The evolution of the parameters demonstrates critical points on the line where bus bunching is significantly increased. Moreover, this analysis allows to differentiate problems associated with varying passenger demand from uncertainties associated with traffic conditions. Furthermore, both expected travel time and travel time variability can be calculated from low-frequency localization data. The above results can be used to calibrate a simulation model which can test bus control strategies. The methods are applied to data obtained from bus route number 1 in Boston.

The final chapter concludes the paper.

Chapter 2

A Multi-Scale Multi-City Study of Commuting Patterns Incorporating Digital Traces

2.1 Introduction

In order to describe the commuting flow patterns of people, various aggregate models of trip distribution have been proposed (89; 128; 118; 101; 117; 116). Among all these efforts, the gravity model, which assumes that the number of movements between an origin-destination pair decays with their distance, is the most widely used one (128; 42; 33). On the other hand, the intervening opportunity model argues that the trip volume is more related to the number of opportunities between the origin and the destination rather than to just their distance (101; 102). Inspired by the intervening opportunity model, the recently proposed radiation model (97) has an analytical formulation that can estimate trip volumes using only population density. This is indeed a remarkable achievement of the radiation model, since previous models require existing Origin-Destination (OD) data for parameter calibration, and such data is not generally available worldwide.

In this chapter, first the performance of the doubly constrained gravity model

with the radiation model (97) at different scales from San Francisco to the entire west coast of the US is compared. Either model is applicable to certain scales. Then based on these two models, an extension to the radiation model with one parameter α is proposed. This extension provides enough flexibility to reach good performance at all scales. Moreover,a functional relationship, $\alpha = 0.34 * log_{10}(l) + 0.22$, between the parameter α and the linear scale (square root of the area) of the study region l is established.

This functional relationship was found by randomly selecting 1000 different regions in the US with different region centers and region sizes (ranging from a few kilometers to 1000 kilometers). Then the census commuting OD statistics is used to calibrate the best α value for each region. The results show a clear functional relationship between the parameter α and the size of a region, which is rooted in the difference in the heterogeneity of the distribution of opportunities. This makes the α parameter predictable so that the model is applicable to regions without empirical OD matrices for parameter calibration.

The gravity model, the intervening opportunities model and the radiation model commonly use population density as a proxy for both the trip generation and trip attraction rates. While this approximation is reasonable at large scales, at the inner city scale it does not hold. When a city is divided into block groups, population density can only represent the trip generation but not the attraction rate. However, the availability of various kinds of urban digital traces provide us with more choices of data sources. Digital geo-located information such as the point of interests (POIs) is a good representation of trip attraction rates. In the US and some other countries, the validity of the newly proposed model can be tested by comparing its prediction to empirical census data on commuting. However, since a detailed census on commuting flows is costly and time consuming, many countries lack this information. As an alternative opportunity, cell phone providers serve nowadays almost all populated regions in the world. Cell phone records in some cities are sufficient to provide a seed commuting OD matrix, which can be expanded to recover the full commuting OD matrix for the whole population under study. Thus cell phone records, together

with population and POI densities are suitable and very economical alternatives for generating commuting flow patterns of regions that lack traditional survey data.

2.2 Data Description

2.2.1 Census Data

The LEHD Origin-Destination Employment Statistics (LODES) datasets (23) used by OnTheMap version 6 were reported using 2010 census blocks. Data files are state-based and organized into three types: Origin-Destination (OD), Residence Area Characteristics (RAC), and Workplace Area Characteristics (WAC), all at census block geographic detail. Data is available for most states for the years 2002-2010. The sources of data include:

- Unemployment Insurance (UI) Wage Records reported by employers and maintained by each state.
- The Office of Personnel Management (OPM) provides information on employees and jobs for most Federal employees.
- The Quarterly Census for Employment and Wages (QCEW) provides informaiton on firm structure and establishment location.

What is used in this study is the Origin-Destination (OD) data. The structure of the OD files is in Table 2.1.

2.2.2 Bay Area Cell Phone Data

The Bay Area cell phone data are collected by a US cell phone operator and contain about half a million customers. Each time a person uses a phone (call/text message/web browsing) the time and the cell phone tower providing the service is recorded. This altogether generates 374 million location records in the three week observational period. A Voronoi tessellation is used to estimate the service area of a

Table 2.1: Data format description for the OD files

Pos	Variable	Type	Length	Description
1	wgeocode	Char	15	Workplace Census Block Code
2	hgeocode	Char	15	Residence Census Block Code
3	S000	Num	8	Total number of jobs
4	SA01	Num	8	Jobs of workers age 29 or younger
5	SA02	Num	8	Jobs for workers age 30 to 54
6	SA03	Num	8	Jobs for workers age 55 or older
7	SE01	Num	8	Jobs earnings $1250/month or less
8	SE02	Num	8	Jobs earnings $1251/month to $3333/month
9	SE03	Num	8	Jobs earnings greater than $3333/month
10	SI01	Num	8	Jobs in SI01 sectors
11	SI02	Num	8	Jobs in SI02 sectors
12	SI03	Num	8	Jobs in SI03 sectors
13	createdate	Char	8	Date on which data was created

cell phone tower. It provides the rough region where a cell phone user can be located by his/her phone usage. Among these half a million users, The 189,621 most frequent users are selected to study the commuting flows of the Bay Area (113). For each user, the most frequently connected tower during day time (6 am to 6pm) is assigned as the tower of the working location while the most frequently connected tower during night (6 pm to 6 am) is assigned as the home tower location.

2.2.3 Rwanda, Lisbon and Santo Domingo Cell Phone Data

The Rwanda cell phone data are collected by a phone company and contain more than 1 million users. Each time a person calls the time and the cell phone tower providing the service is recorded. There are around 215 million records over a period of 40 days. The entire Rwanda is covered by 196 towers while the capital city Kigali is covered by 47. The 410,309 most frequent users are selected for this study. The cell phone data from Portugal and Dominican Republic are similar. Lisbon has 62790 frequent users while Santo Domingo has 52125 frequent users.

2.3 The Radiation and the Gravity Model

Different forms of the gravity model are used when applied to different areas of study. In epidemics study (9; 41; 111; 95) it usually takes the form:

$$T_{ij} = \gamma \frac{n_i^{\alpha} n_j^{\beta}}{C(r_{ij})} \tag{2.1}$$

where T_{ij} is the flow between zone i and j. n_i and n_j are the population of the two zones. r_{ij} is the distance between them and C is a distance decay function. α and β are parameters to be fitted from data. γ is an adjustment parameter controlling the sum of the flows. This is usually called the unconstrained gravity model because it does not guarantee the obtention of the desired marginals (the total production and attraction at each zone).

In transportation planning, gravity model usually takes the form(5; 42; 119):

$$T_{ij} = \frac{\alpha_i \beta_j O_i D_j}{C(r_{ij})} \tag{2.2}$$

where O_i and D_j are the total trip production and attraction volumes of zones i and j respectively. For a study region with N zones, there are $2N$ parameters α_i and β_j. These parameters are calculated by iteratively applying the condition:

$$\alpha_i = 1/\sum_j \beta_j D_j C(r_{ij}) \tag{2.3}$$

$$\beta_j = 1/\sum_i \alpha_i O_i C(r_{ij}) \tag{2.4}$$

This is called the doubly constrained gravity model because it ensures consistent values of the trip production $\sum_j T_{ij} = O_i$, and trip attraction $\sum_i T_{ij} = D_j$ per zone. In order to calibrate the α_i and β_j parameters, the model requires accurate input of the total trip production and attraction volumes O_i and D_j. Since the number of parameters grows with the number of zones, when the number of zones is large it is computationally hard to get the calculation to converge.

An alternative for trip distribution is the recently proposed radiation model, which is inspired by the intervening opportunity model but has a closed analytical form of the T_{ij} distribution. It takes the form:

$$T_{ij} = O_i \frac{n_i n_j}{(n_i + s_{ij})(n_i + n_j + s_{ij})} \tag{2.5}$$

where s_{ij} is the population within the circle of radius r_{ij} centered at zone i (not including the population in zones i and j) and the rest of the notations are the same as in the gravity model.

First the suitability of the doubly constrained gravity model and the radiation model on predicting commuting flows is explored at three different scales: the west coast of the US, the Bay Area, and San Francisco. The three regions are shown in Fig. 2-1(a). The west coast of the US is divided into 183 counties while the two smaller regions are divided into 100 zones. Each zone is a cluster of blocks determined by applying the k-means clustering method on the 7,348 census blocks in San Francisco and 117,219 blocks in the Bay Area. The unconstrained gravity model is not compared because it is often prevailed by the doubly constrained gravity model. Detailed comparisons between these two models are in the appendix.

A basic assumption in both models need to be tested: the population density could represent well both the commuting trip generation and attraction rates at different scales. The 2010 census LEHD Origin-Destination Employment Statistics (LODES) (23), which provides home and employment locations for the entire US population at block level, is used as a benchmark. Fig. 2-1(b-d) shows the densities (volumes per unit area) of commuting flow generation, attraction and population at the west coast of the US. All the three of them have similar distributions, so at this scale the assumption holds. Fig. 2-2(a, b) shows the commuting trip generation and attraction rates in San Francisco. Their distributions are different. The correlations between them at the three scales are shown in Table 2.2. This evidence shows that the smaller the scale, the lower is the correlation between the commuting trip attraction rate and the population density. Thus another proxy for the commuting trip attraction

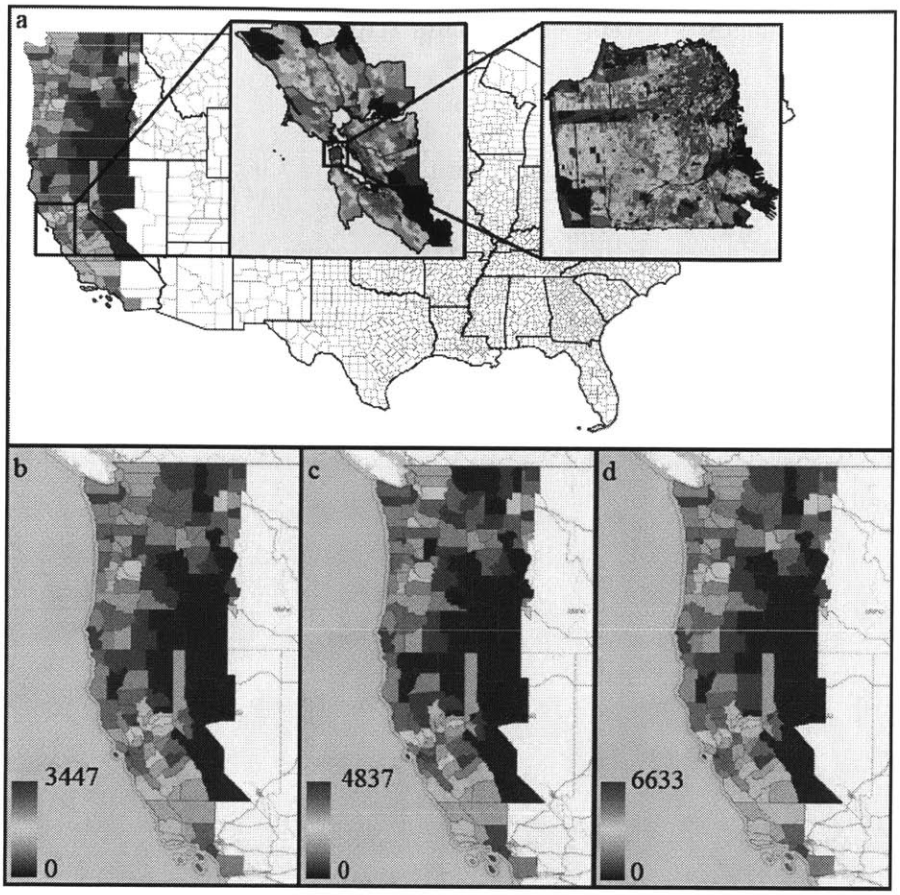


Figure 2-1: Different features of trip production and attraction at country level. (a) Three scales of study: West coast of the US, the Bay Area, and San Francisco. (b-d) Commuting trip generation rate, trip attraction rate and the population density in $(\#/km^2)$ in the west coast of US. Their distributions are similar.

rate is needed. Digital traces of facilities are available online, which provide novel sources of information for modelling human mobility (103; 58). The density of point of interests (POIs), which is defined as point locations that someone may find useful or interesting on digital maps or GPS software, is a suitable proxy for the attraction rate of commuting trips at different study region scales. For example, Google Places provides the name, the longitude, the latitude and functions of each POI. The three study regions contain 1,774,154; 319,170 and 85,230 POIs respectively. According to Table 2.2, all the scales the POI density has high correlation with the commuting trip attraction rate. Therefore in the following calculation the density of POIs is used to represent the commuting trip attraction rate. To be more specific, the POI density is used to calculate D_j in the gravity model and n_i, n_j, s_{ij} in the radiation model.

Table 2.2: Correlation between commuting trip generation, attraction, population and POI

	West Coast		Bay Area		San Francisco	
	Population	POI	Population	POI	Population	POI
Generation	0.99267	0.92607	0.97065	0.49074	0.95638	0.29172
Attraction	0.98907	0.92958	0.41712	0.85853	0.15693	0.88003

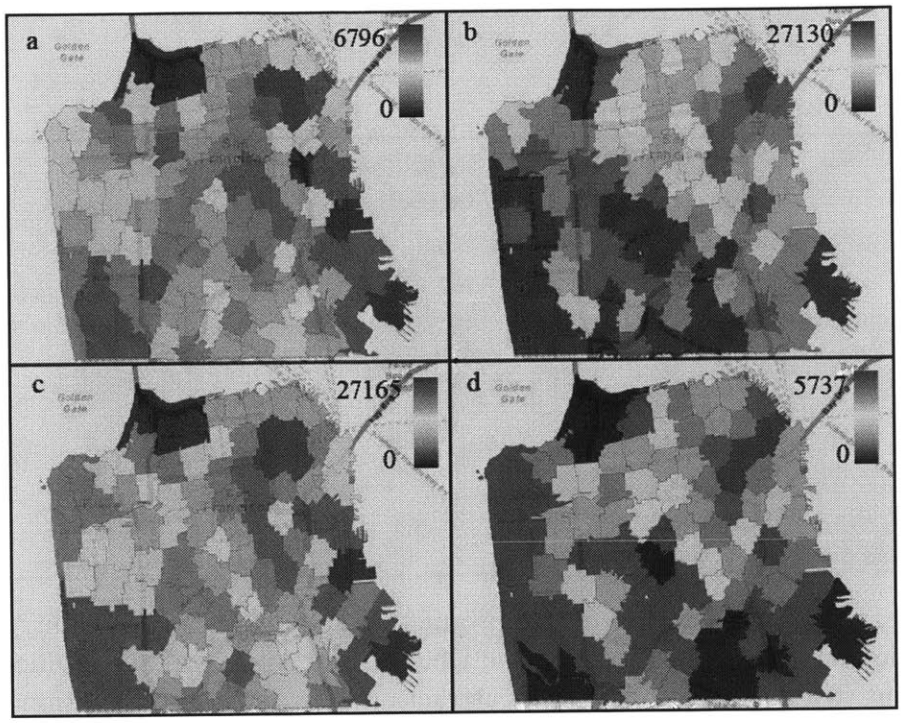


Figure 2-2: Different features of trip production and attraction at city level. (a-d) Commuting trip generation rate, trip attraction rate, the population density and the POI density in $(\#/km^2)$ in San Francisco. In San Francisco the commuting trip generation and attraction rate distributions are different. While the population density has high correlation with the commuting trip generation rate, the POI density has high correlation with the commuting trip attraction rate. See details in Table 2.2.

The doubly constrained gravity model with power distance decaying function $C(r) = r^k$ is compared with the radiation model. Fig. 2-3 shows the comparison of the census data, the gravity model, the radiation model, and the extended radiation model. The 3 rows represent San Francisco, the Bay Area, and the west coast of the US respectively. The first column shows the commuting distance $(P(r))$ distribution, the second column is the commuting destination zone rank $(P(k))$ distribution.

For each traveler, the closest zone to the origin zone is of rank 1, the second closest is rank 2, etc. The third column is the $P(s_{ij}+n_i)$ distribution. s_{ij} and n_i are defined in the formulation of the radiation model. The sum of them represents the total number of opportunities within the circle defined by the origin and the destination zone. The radiation model gives satisfying prediction only at the west coast scale. At the two smaller scales the radiation model predicts too few relatively long distance trips. The doubly constrained gravity model gives closer predictions to the census data but has much more parameters. The extended radiation model, with one parameter α, gives close predictions to all the scales at all the measurements.

At San Francisco and the Bay Area scale, the radiation model constantly underestimates relatively long distance trips. In other words, it chooses mostly the top ranked locations. This result is in agreement with the approximations of the predictions reported by (97) which states that the radiation model is equivalent to a gravity model with $C(r) = r^4$ under homogeneously distributed opportunities. This indicates a high cost for long distance trips, while empirically the gravity model cost functions have values in the range $r^{0.5}$ and r^2). The radiation model has been proved to be successful at country scale since the population distribution is generally flat but with a few high peaks corresponding to highly urbanized areas, which indicates higher heterogeneity in the distribution. In contrast, at inner city scale, both the population and the POIs are distributed more homogeneously which causes short tail of the commuting distance distribution for the radiation model.

The parameter free feature of the radiation model becomes a double-edged sword in such multi-scale studies. As long as the density of distributions of opportunities and population are the same, the radiation model would give the same prediction regardless of the size of the study region.

The doubly constrained gravity model gives very close predictions to the census data at the two smaller scales since the large number of parameters makes the model flexible. At the US west coast scale, even adjusting the α_i, β_j and the parameter k in the power distance decay function cannot fit the model well in both short distance trips and long distance trips. At this scale, the doubly constrained gravity model

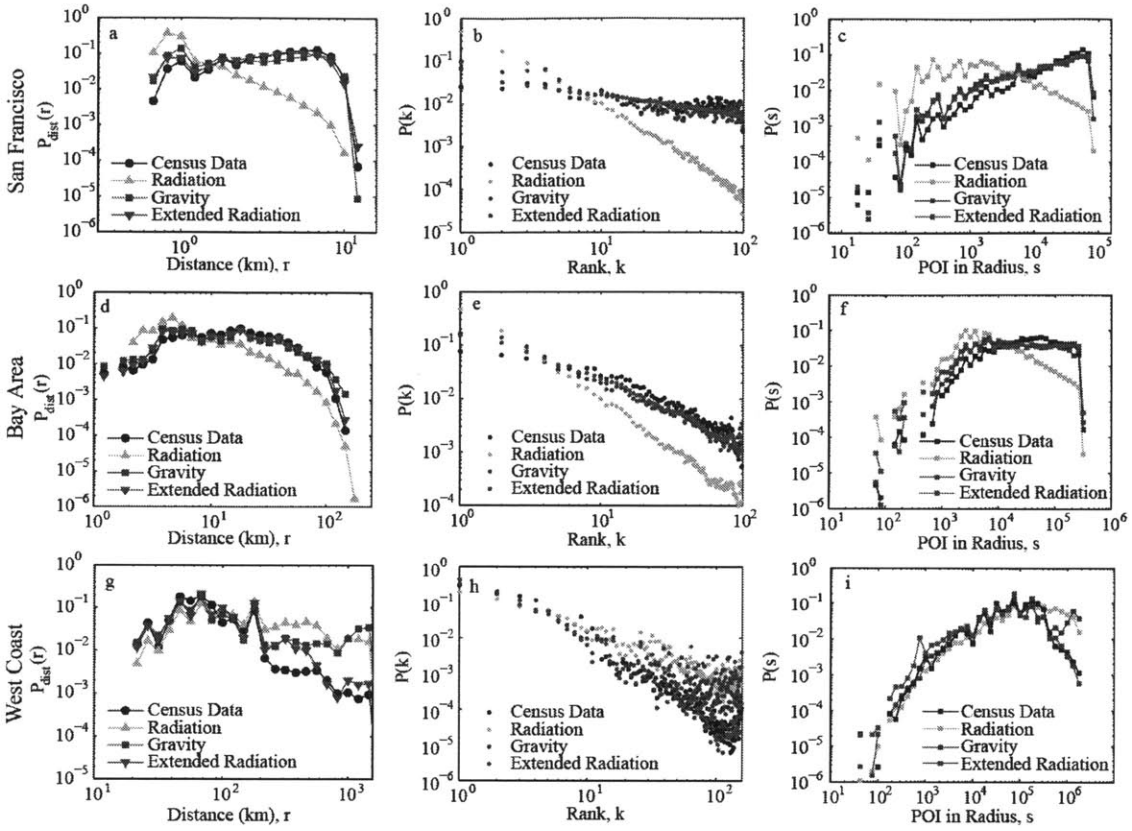


Figure 2-3: Comparison of the census data, the gravity model, the radiation model, and the extended radiation model.

gives similar prediction results to the radiation model.

2.4 Extension to the Radiation Model

It is desirable to introduce the scale dependency on the homogeneity of opportunities in order to keep the advantages of the radiation model. Although compared with the radiation model, the doubly constrained gravity model gives a better or equivalent prediction at the three scales, it still has the following disadvantages:

1. The large number of parameters indicates high flexibility of the model. While this gives very good fits to a particular dataset, it generalizes poorly, meaning that the parameter values cannot be applied to different regions of study.

2. It needs an empirical OD matrix as a seed for parameter calibration.

3. Compared with the radiation model, it needs a much longer time for the iterative fitting to converge.

4. It cannot account for the effect of the number of intervening opportunities between the origin and the destination. Gravity models make the same prediction for 10 km trips in rural Iowa than in New York City, while empirical data shows people in rural Iowa are more likely to move further in other to satisfy their needs.

The proposed extension of the radiation model is inspired by the way the intervening opportunity model handles the homogeneity of opportunities (102). The model defines the probability that one opportunity cannot satisfy the traveler's need as $e^{-\lambda}$. For a given region with S opportunities, the probability that none of them can satisfy the traveler's need is $e^{-\lambda S^{\alpha}}$. The modification factor of α is introduced to represent the equivalent number of independent opportunities. The parameter α is introduced similarly to the intervening opportunity model so that the new model takes the form:

$$T_{ij} = O_i \frac{n_i n_j^{\alpha}}{(n_i + s_{ij}^{\alpha})(n_i + n_j^{\alpha} + s_{ij}^{\alpha})} \tag{2.6}$$

α represents the correlation of the attractiveness of different opportunities in one zone. In one limiting case, when $\alpha \to 0$, $s^{\alpha} \to 1$, which means that the correlation of different opportunities in a zone is 1, so that the effective number of opportunities in a zone is always 1. In this case the number of opportunities in a zone doesn't influence the traveler's decision while only the zone's relative distance to the origin matters. This parameter is sufficient to successfully introduce the desired features because:

1. Homogeneity of Opportunities: since s_{ij} is usually one order of magnitude larger than n_i and n_j, it dominates the denominator, particularly in regions where opportunities are homogeneous. In the original radiation model, regions with more homogeneously distributed opportunities give too much cost to long distance trips, so we'd expect α to be smaller than 1 to correct this feature.

2. Multi-scale Expression: the larger the region size, the larger the α should be in order to introduce higher costs and to limit mobility.

3. Scale and Homogeneity of Opportunities: These two quantities are usually highly correlated. Large regions have only a few peaks of opportunities which represents urban centers while smaller regions have more homogeneous opportunities. Fig. 2-1 and 2-2 clearly shows this effect.

The detailed derivation is as: In the original radiation model (97) which mimics the mobility pattern by a particle emission and absorption process, the closed form probability to travel from an origin zone with m opportunities to a destination zone with n opportunities with S opportunities in between is:

$$P(1m, n, s) = \int_0^\infty P_m(Z)P_s(< Z)P_n(> Z)dz \tag{2.7}$$

z can be an arbitrary random variable. The calculation of $P(1m, n, s)$ is composed of three parts. $P_m(Z)$ is the probability to emit a particle with absorption threshold Z. $P_s(< Z)$ is the probability that none of the S opportunities in between absorbs the particle. $P_n(> Z)$ is the probability that the particle is finally absorbed by a zone with n opportunities. These three elements can be expressed as:

$$P_m(z) = \frac{dP_m(< z)}{dz} = mp(< z)^{m-1})\frac{dp(< z)}{dz} \tag{2.8}$$

$$P_s(< z) = p(< z)^s \tag{2.9}$$

$$P_n(> z) = 1 - p(< z)^n \tag{2.10}$$

The two flaws of the radiation model have been identified: the performance is not good under homogeneously distributed opportunities and cannot deal with different sizes of the study region. These are issues in the particle transmission and absorption processes. So the $P_m(z)$ is unchanged and a modification factor α is introduced to the

transmission and absorption processes. S^α and n^α are used to represent the effective number of opportunities. So that:

$$P_s(< z) = p(< z)^{s^\alpha} \tag{2.11}$$

$$P_n(> z) = 1 - p(< z)^{n^\alpha} \tag{2.12}$$

Now the probability of travelling between two zones becomes:

$$\begin{aligned} P(1m, n, s) &= \int_0^\infty P_m(z) P_s(< z) P_n(> z) dz = \\ &= \int_0^\infty m p(< z)^{m-1} \frac{dp(< z)}{dz} p(< z)^{s^\alpha} (1 - p(< z)^{n^\alpha}) dz \\ &= m_i \int_0^\infty dz \frac{dp(< z)}{dz} [p(< z)^{m+s^\alpha-1} - p(< z)^{m+n^\alpha+s^\alpha-1}] \\ &= m[\frac{1}{m+s} - \frac{1}{m+n^\alpha+s^\alpha}] = \frac{mn^\alpha}{(m+s^\alpha)(m+n^\alpha+s^\alpha))} \end{aligned} \tag{2.13}$$

Now it still has a closed form solution but is flexible enough to adapt to different opportunity homogeneities and region sizes.

Fig. 2-3 shows that with one parameter the extended radiation model is flexible to work at all the three scales. The α values are 0.62, 0.88 and 1.15 respectively.

Secondly, how the parameter α systematically changes as the size of the study region changes is studied, measured as the square root of the region area. 1,000 study regions are selected randomly. They all have population 200,000 or more in order to avoid unpopulated regions such as national parks. The linear scale of the regions range from a few kilometers to about 1,000 kilometers. The census commuting OD data is used to obtain the best α value in the extended radiation model for each case. Fig. 2-4 shows how α increases as the region size increases. The linear relation on the log-log plot indicates that α increases as a power function of the region size: $\alpha = 0.34 * log_{10}(l) + 0.22$. This means that the parameter α, as expected, reasonably depend on the scale of the study region and the model is applicable to regions without empirical OD matrices for parameter calibration.

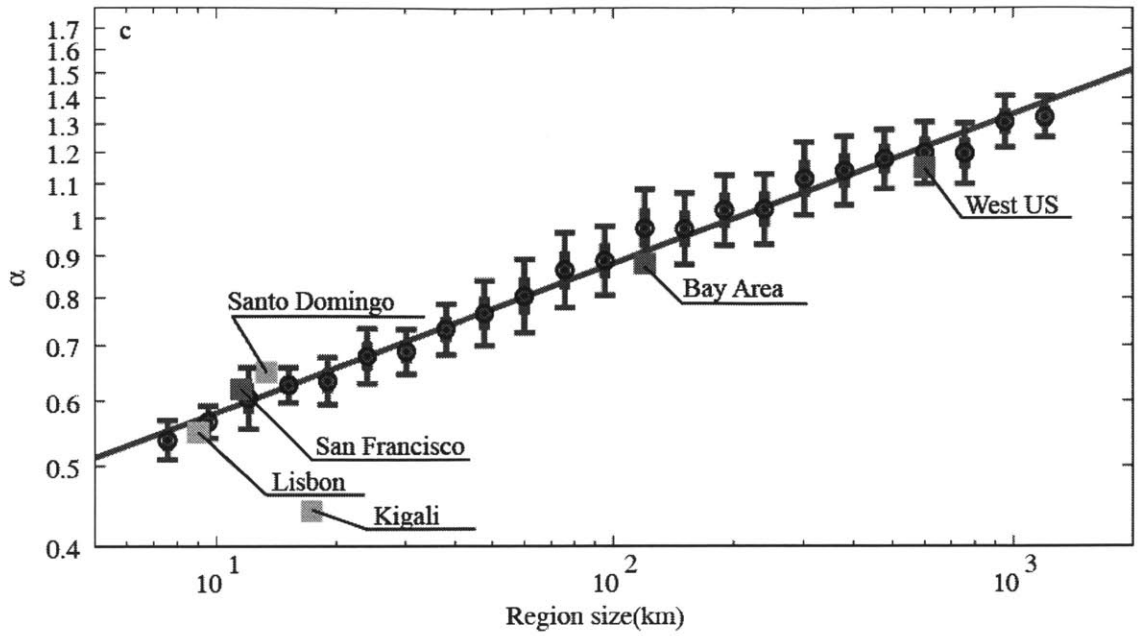


Figure 2-4: Functional relationship between the parameter α and the region size. 1000 regions with random centers and sizes are seleted. The sizes range from a few kilometers to 1000 kilometers and the population in each region is limited to be greater than 200,000. In each case the corresponding α value is calibrated by the census commuting OD statistics. The relationship between α and the size of the region is in the red line and the error bar is in blue. Three special cases in the US: San Francisco, the Bay Area, and the west coast of the US are marked in magenta. The α values for the cell phone users in Lisbon, Santo Domingo, and Kigali are in green. All of them, except Kigali, conform to the functional relationship $\alpha = 0.34 * log_{10}(l) + 0.22$. l is the linear size of the region. In Kigali the actual α value is smaller than the predicted value because a much smaller cell phone tower density and higher percentage of unpopulated area.

2.5 Multi-city Study and the Role of Mobile Phone Data

Not many countries in the world have detailed census data for commuting flow prediction and model calibration. Those countries with data scarcity are often developing countries that need this kind of modeling the most. For these countries finding an alternative data source to provide guidance for their urban growth, economic planning and epidemics controlling is a pressing need. Cell phone records are increasingly showing the potential to become such a data source of valuable information (78; 115; 8) since most populated areas have cell phone service coverage and the value of cel-

l phone data in modeling human mobility has recently been highlighted in various studies (56; 100; 99; 9). For instance, in Rwanda there is no detailed commuting census data available yet. Even if there were, the high migration rate of people would make the census outdated quickly. Luckily, the country has 215,030,420 cell phone records from one cell phone service provider in just three months. Cell phone records can be used to provide a seed cell phone user commuting OD matrix, which can be expanded to recover the full commuting OD matrix for the whole population under study.

Each cell phone record has a time stamp and a corresponding cell phone tower. For each user, the most frequently used tower between 6PM and 6AM is assigned as the home location and the most frequently used tower during day is assigned as the work location. Again using the 2010 census home and employment location data as a benchmark, the Bay Area is used as an example to validate that the cell phone data could provide accurate predictions to commuting flow patterns. The sample includes the 189,621 most active cell phone users in order to eliminate the bias of having too few records. The 892 cell phone towers in the Bay Area are mapped to the previously defined 100 block clusters to form the commuting OD matrix for the cell phone users. In order to compare this with the census commuting OD matrix of the entire population Iterative proportional fitting (IPF) method is used to expand the cell phone user OD matrix (44). The basic procedure is first getting the distribution of population and POIs to represent the marginal distributions of commuting trip generation and attraction rates for each block cluster. Then iteratively adjust the elements of the seed matrix to let them match the desired marginals. Fig. 2-5(a-c) shows the comparison results of the distribution of commuting distance, the distribution of the number of commuters between O-D pairs, and the comparison of the census commuting flow T_{ij} and the expanded cell phone user commuting flow T'_{ij}. The close fitting in all the three figures shows that the commuting patterns of the whole population can be recovered from the seed matrix provided by cell phone records. For countries that do not have population density census statistics for the IPF expansion, the Landscan (18) population density estimation is available worldwide at $1km^2$ resolution. This

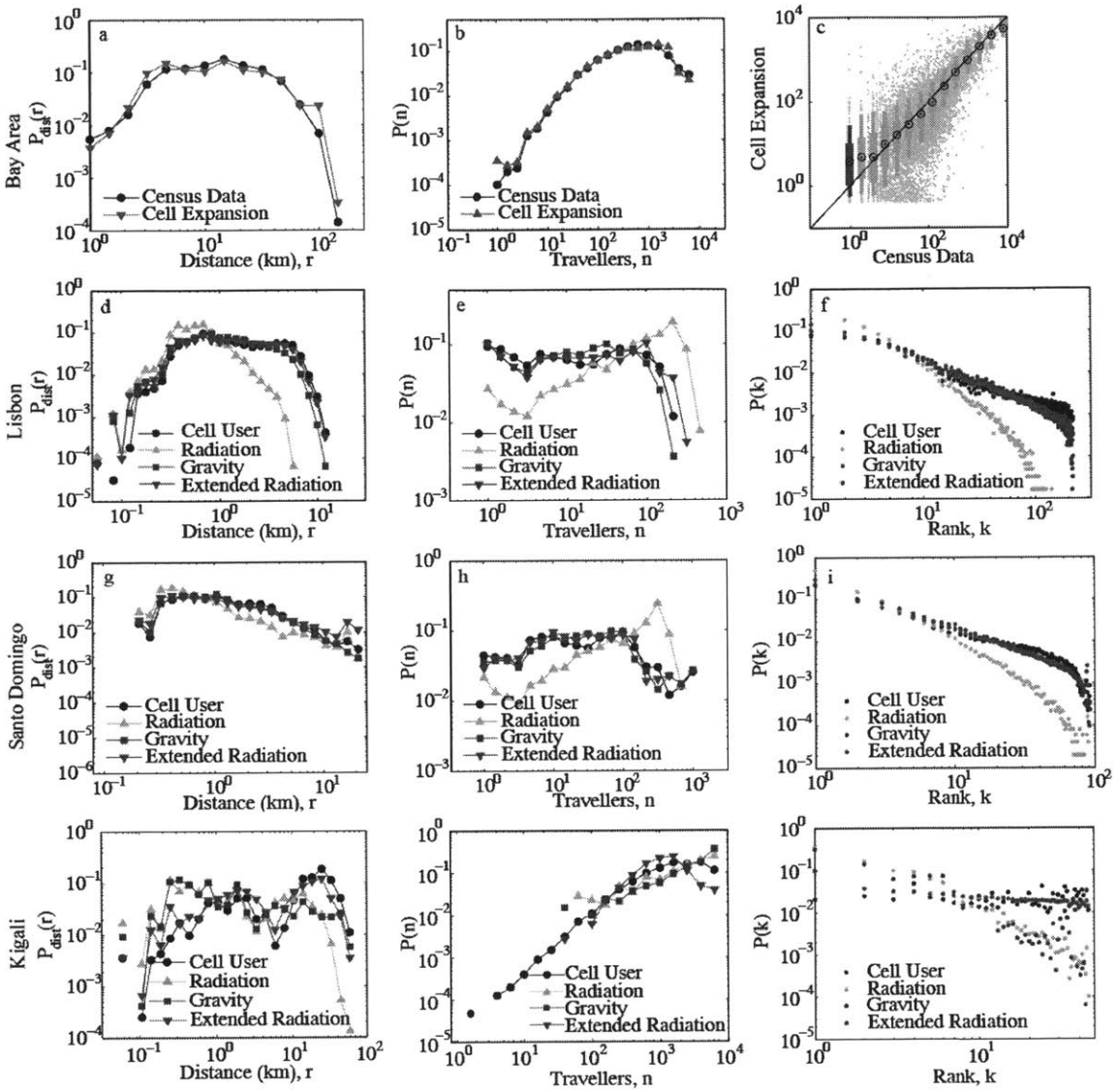


Figure 2-5: Cell phone users' commuting patterns in different cities.

method is extended to three different countries: Portugal, Dominican Republic and Rwanda. The capitals in the three countries, Lisbon, Santo Domingo and Kigali, are selected as examples. Their records contain 40,999; 49,502; and 31,532 frequent users respectively. Since for these three regions there are no empirical commuting OD matrices for benchmarking, in the next step instead of doing the IPF expansion to get the commuting OD for the whole population under study, the gravity model, the radiation model and the extended radiation model are applied to the cell phone users and test how much can they recover cell phone users' commuting patterns. The

inferred home and work locations for each user are aggregated to get the commuting trip generation O_i and attraction D_i for each tower location. The three models' results are shown in Fig. 2-5(d-i). The lineal scales of these three regions are respectively 9 km, 13 km and 16 km. The best α values for the extended radiation model are respectively 0.55, 0.65 and 0.45. Their scale vs. α relationships are marked in green on Fig. 2-4. Two of them agree well with the fitting curve observed from the US regions. The functional relationship between α and the size of the study region seems to be generalizable across different cities. Kigali is an exception. Several reasons can account for this. First, a large proportion of Rwanda's territory is unpopulated land, so the effective size of the study region should be much smaller than the city's area. Secondly, this area has considerably less number of towers and POIs per area.

Some particular commuting characteristics can be observed in different cities. Santo Domingo and Lisbon have similar city sizes and frequent user numbers, but their $P(r)$ and $P(n)$ distributions are quite different. In Santo Domingo the commuting distance distribution has a fatter tail, which indicates that people generally have to travel longer to work. The ranking plot also shows this fact since the slope in Santo Domingo is smaller than the one in Lisbon. In the case of Kigali the $P(n)$ plot shows a different pattern from the other two cities. There are more OD pairs with relatively large flows while much less OD pairs with small flows. The rank plot trend is more flat presenting two peaks. This is probably because in Rwanda the commuting flows are more agglomerated; the residential and working opportunities are highly concentrated at a few places, unlike the cases in Lisbon and Santo Domingo where the opportunities are more scattered across the region. In consequence, the commuting destination is less influenced by the distance but more influenced by where the *hubs* are. In this sense the commuting flows in Rwanda have higher predictability once the *hubs* are identified.

As is proposed in some previous studies (12; 62; 63; 65), there may exist some simple scalings for a given region of total area A and population P. The length scale of a region is represented by $\sqrt{A}$. The expected scaling of the total distance travelled

by all the population l_d should be of the form:

$$\frac{l_d}{\sqrt{A}} P^{\beta} \tag{2.14}$$

In one limiting cases, if every individual is going to the nearest neighbor (with a typical distance $1/\sqrt{\rho}$ while $\rho = P/A$ is the average density of the city), $\beta = 1/2$. In another case, if everyone goes randomly, $\beta = 1$. The empirical cases show that the β value is usually around 0.6.

The same measurement for the 1000 different regions in the US and the cell phone users in Rwanda, Santo Domingo, and Lisbon is calculated. The results are shown in Fig. 2-6. The corresponding β value is 0.75. Since only the commuting distance is counted here, the larger β value shows that people would travel longer for working than doing other activities.

2.6 Discussion

In summary, in this chapter an extension of the radiation model is proposed that could predict the network of commuting flows at different spatial scales and in different cities worldwide. In addition, the spatial density of the distribution of facilities as downloaded from Internet's digital maps, also known as POIs, together with the density of population are the two basic ingredients for modelling commuting networks.

The proposed extension uses one parameter to adjust to the different degrees of the homogeneity of opportunities when the scale of the study region changes. In contrast, while due to the large number of fitting parameters, the doubly constrained gravity model also fits the number of trips in most of the studied datasets, the obtained parameters cannot be generalized and depend on empirical flows of the particular region under study for their calibration. The parameter α in the extended radiation model, takes into account the effects of both the scale and the heterogeneity of opportunities, which are highly correlated. Thus, α to a large extent can be estimated by knowing only the size of the region under consideration.

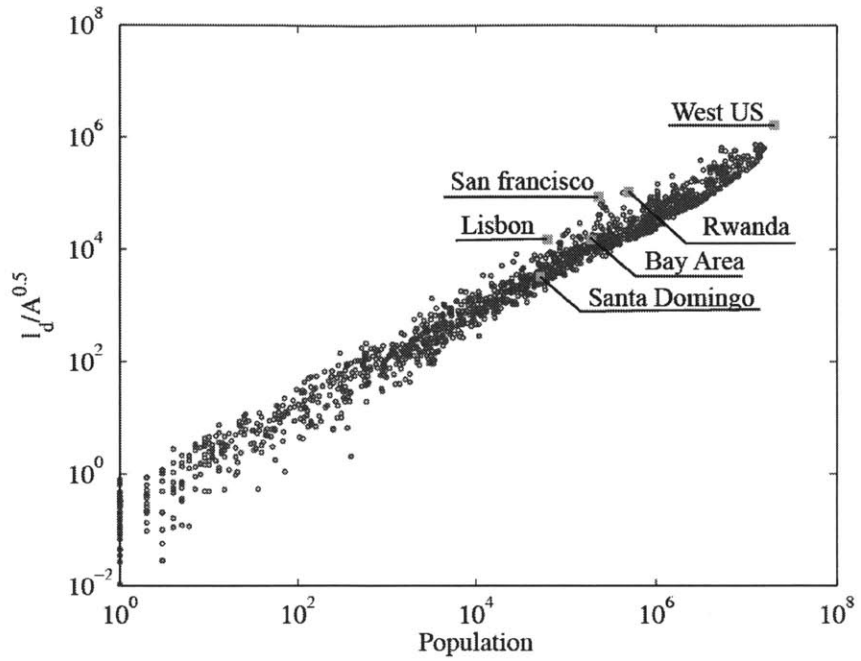


Figure 2-6: A scaling measurement of $l_d/\sqrt{A}$ vs. β. l_d is the total distance travelled by all the population in a region. A is the total area. P is the population. They should follow the scaling relationship: $\frac{l_d}{\sqrt{A}}$ P^{β}. The blue dots are 1000 randomly selected regions from the US with different sizes and population. Three special cases: the west coast of US, the Bay Area, and San Francisco are marked in green. The measured values for cell phone users in Rwanda, Santo Domingo and Lisbon are also marked. The β value is 0.75, larger than the empirical result of 0.6 which measures travels for all activities, which indicates that people are willing to travel further for commuting than for other activities.

In order to explore the validity of the new model in diverse cities that lack of census data, first the Bay Area is used as an example to show that cell phone records are a very good alternative data source to extract commuting patterns. Home and work locations are inferred at individual level from the cell phone records and then aggregated to show its equivalence to the commuting information obtained by census data. Then commuting flows are extracted from cell phone users in three different countries where census commuting data is not available. These results show not only the applicability of the proposed model to successfully model the commuting patterns for the cell phone users in cities from these three countries, but also show some unique commuting characteristics in each city.

Taken together, the proposed extension to the radiation model is readily available to be incorporated into mobility modeling at different scales and different regions worldwide under different data availability. The sample of US county level commuting flow prediction is shared on a webpage to help in this direction (64).

Chapter 3

Using Low-frequency AVL Data for the Monitoring and Control of Bus Performance

3.1 Introduction

Modern advanced traveler information systems (ATIS) are capable of providing information on expected travel times for all modes and any given origin-destination pairs in real time by incorporating many different data sources including past measurements of vehicle trajectories, passenger demands and current traffic conditions. However, collecting these data is still costly and many data sources are not universally available. In particular automatic fare collection (AFC) or boarding and alighting counts (both of which play a significant role in some ATIS) are currently not available in many systems, whereas automatic vehicle location (AVL) information is widespread.

However, many AVL systems provide location-at-time data with a low sampling frequency (on the order of a minute). In this situation travel time prediction and bus arrival time prediction can only be based on such AVL data sets. For example, Nextbus is a major provider of AVL data services to transit agencies across North America and offers those agencies web services so that agencies can release their data

to the public. The agencies control the amount, quality, and frequency of the updates, and many agencies chose to provide low-frequency data due to budget constraints.

Transit agencies need to collect data for performance evaluation. Traditional manual data collection is both time consuming and expensive. The implementation of automatic data collection systems (ADC) has provided new alternatives for data collection and system evaluation. Many transit companies have implemented both automatic vehicle location systems (AVL) and automatic fare collection systems (AFC) on transit vehicles. These AVL data are also used by transit agencies to evaluate their transit system performance. It provides the opportunity to gather large amount and long term transit operation information at relatively low costs. This kind of data has the potential to enable operators to evaluate transit system performance, diagnose service bottlenecks, and improve the system level of service (53; 71).

While low-frequency AVL data are widely available, The main limitation of AVL data are measurement errors due to the GPS devices and recording or transmission errors. To be more specific, the limitations of AVL data include: GPS devices have drifts of 1 meter to 40 meters; Errors may occur when recording and transmitting data; GPS pulling interval might be as long as 60s. These limitations pose higher requirements on public transportation researchers. How to transform raw AVL data into useful information for transit riders and operators is not trivial.

The low sampling frequency implies that stop dwell times and on-route travel times are not trivial to separate. Linear interpolation methods that are often used (see e.g. (24) and (35)) distort travel speed and headway measurements significantly. Therefore better alternatives are desirable.

In this chapter a new methodology for re-sampling of low frequency location-at-time AVL records is adopted. The methodology is based on the distribution of GPS measurements providing information on typical dwell times. It is shown using both experimental and real world data that the new method performs superior in comparison to other resampling schemes. This consists of: data prepossessing which refines data quality, service performance analysis which diagnoses service bottlenecks, and finally shows how to apply the measured statistics in the calibration of bus

movement models.

The data prepossessing step includes map-matching and re-sampling. For the map-matching problem, Ref. (94) provided a formal and up-to-date review of the multiple existing techniques. The problem here is the low-frequency of record polling, for which two researches (122; 120) both propose to first collect the subset of likely matches found by point-to-point and point-to-curve matching for each GPS coordinate and then searching for the most likely route through them. This is the map-matching method adopted here for being the most suitable to the Nextbus data.

For the re-sampling problem, data interpolation is needed because both passengers and transit agencies are more interested in when buses arrive at certain points (such as stops) while the available AVL records are only recorded every 60s. For high frequency AVL data, it is relatively easy to judge when a bus is staying at stops. But for low frequency AVL data, stop arrival and departure times are not apparent. In previous studies, most AVL data interpolation methods were performed on time-at-location data which AVL records are pulled when buses pass certain points on their routes (52). For location-at-time data, some researches (24; 35) assumed that between pulled AVL records buses are travelling at a constant speed and then performed analysis on bus speed distributions, headway regularity and adherence to schedule. This assumption made the analysis convenient. But in order to get more accurate results a more realistic interpolation method is required. Here a convenient method is presented that does not assume constant velocity but correct by pauses at stops.

Using the re-sampled data, analysis on transit service quality is performed. Among all the service quality measurements, headway distribution (75; 29), adherence to schedule and in-vehicle travel time are the most studied (3). For high frequency urban transit service, headway distribution is the measurement directly related to operations (75; 2). It determines passengers' experienced waiting time and could be effectively improved by operational strategies such as holding and stop skipping (90; 60). The variation in the distribution of headway across the entire route is chosen as a proxy of the deterioration in service quality.

In this chapter the goals are threefold:

- to provide a more accurate data preprocessing methodology which enhances low frequency AVL data (see section 3.2) into data which is applicable;
- to show that the such obtained data can be used to evaluate bus service quality (including travel time uncertainty) and diagnose service bottlenecks (see section 3.3) using the preprocessed data;
- to use the acquired statistics to calibrate a bus movement model, which subsequently can be used to evaluate different control strategies for mitigating bus bunching effects (see section 3.4).

To achieve these goals, first map-matching is performed on low-frequency location-at-time AVL data to assign transit vehicles to their route shapes. Second, re-sampling is performed incorporating both buses' actions of traveling on route segment and staying at stops. Kernel density estimator observed from the preprocessed AVL data is used to calibrate the interpolation method. Stop arrival and departure times are inferred. Based on the interpolated data, statistics of in-vehicle travel time could show how bus headway deviation propagates along the route. Bottlenecks are identified and investigated to show underlying causes. Methods to estimate route travel time and travel time variability are provided. The final section shows how the results could be used to calibrate bus movement models.

The research in this chapter is also contributed by David Gerstle, Peter Widhalm and Dietmar Bauer. David did the map-matching analysis, Peter analyzed the travel time variability and Dietmar did the re-sampling procedure. Here I want to acknowledge their contributions to this research.

3.2 Data Prepossessing

In this chapter location-at-time AVL data provided by the Nextbus service is used. NextBus, Inc., provides data for a large number of US and Canadian transit companies (including LA Metro, MBTA, NYC MTA, San Francisco Muni, and the Toronto Transit Commission). The Nextbus server is polled every 60 seconds and returns the

bus locations in the form of an XML document. Additionally, schedule and route information are given in the form of general transit feed specification (GTFS) format introduced by Google. Both data sources are combined in order to match the observed locations to the route specified in the GTFS. Locations of bus stops from the GTFS are used in order to derive arrival and departure times from the AVL records.

To illustrate the methodology, the AVL records of all the workdays between May 1st 2011 and June 15th 2011 from the route 1 of the MBTA (Massachusetts Bay Transit Authority), totalling 4624 trips during weekdays and 796 trips on weekends, are used. MBTA route 1 runs from Dudley station in Boston to Harvard university in Cambridge. The outbound and inbound stops near Harvard university are not symmetric because of the one way streets. The data used contains outbound runs' records which have 33 stops starting at Dudley station and ending at Quincy St at Harvard St. The average distance between stops is about 250m. A map of MBTA route 1 is shown in Fig. 3-1 where also significant stops are indicated: between stops 8 and 9 the route turns into a main arterial. Between stop 18 and 19 there are 3 intersections and the bus transfers with a metro line here. At stops 12 and 25 large traffic volumes during peak hours are observed.

The scheduled headway during morning peak hours equals 8-9 minutes, during the afternoon peak 7-8 minutes while at off-peak times a 12 to 13 minute interval is scheduled.

Next two sub sections describe the two steps of data prepossessing in order to obtain a dataset suitable for further analysis.

3.2.1 Map-matching

The first step in the data preprocessing is map-matching, see ref. (94) for an up-to-date review. Service performance evaluations require correct distance along shape information. The available AVL data contains different kind of errors. Thus performing map-matching on the raw AVL data to rule out as many errors as possible is necessary. In order to open the discussion for future research on this type of data, here the main types of errors found in the AVL data are summarized.

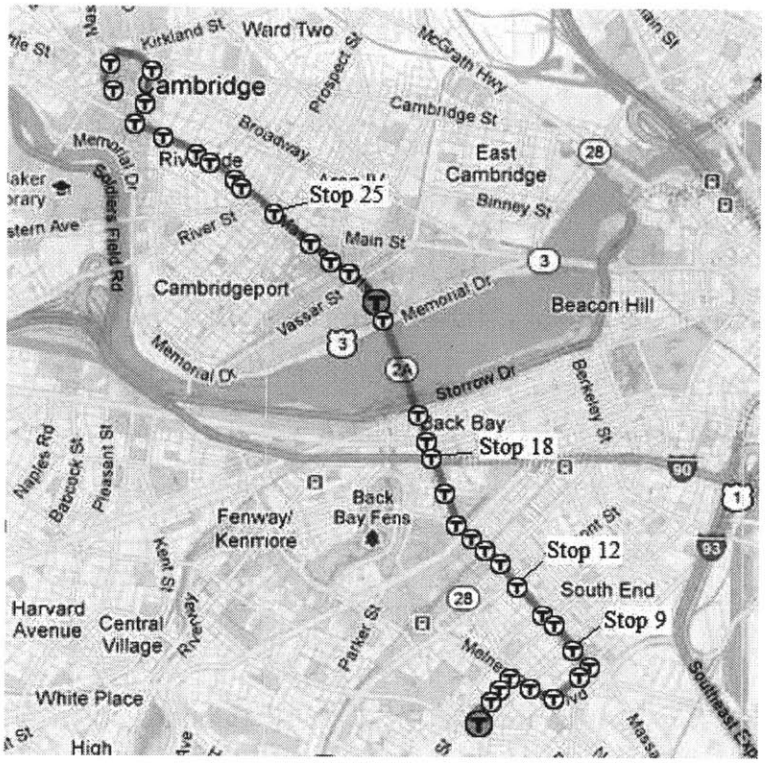


Figure 3-1: MBTA route 1: Significant stops discussed in the text are marked. Map obtained from Google maps.

In the data set at hand map-matching has to deal with the following most frequent problems:

1. Wrong temporal order: time stamp and location pairs appear to be in the wrong order. This makes buses appear to be going backwards along the shapes.

2. No matching from the shape: the locations provided in AVL records are far from the route shape. Such points occur in particular at the start or the end of trips.

3. Wrong interval: some trips only have a few data points recorded at irregular intervals. This may indicate the buses are out of service or the GPS devices are broken.

After ruling out the erroneous records, the rest records are matched to bus route shapes using point-to-point and point-to-curve matching methods. Point-to-point method matches each GPS coordinate to the closest node on the route shape while

point-to-curve method takes the route as a polyline and match each GPS coordinate to the polyline. For detailed descriptions please refer to (122) and (120). On average 97.6% of the observations could be map-matched directly. For the rest 2.4% erroneous observations, 0.6% could be fixed using obvious heuristics, which are mainly wrong temporal order observations. 1.8% of the observations are finally discarded, two thirds of which correspond to a no matching from the shape.

3.2.2 Re-sampling Procedure

The output of the map-matching procedure are sequences of $(t_{i,j}, x_{i,j})$ pairs ($t_{i,j}$: time stamp for j-th observation of trip i, $x_{i,j}$: distance along shape) sequences for all trips on a given shape for given route. Since the GPS tracks of the AVL data is only available approximately every 60 seconds, a re-sampling scheme is needed in order to obtain information on arrival and departure times at stop locations.

Two different approaches could be followed: Viewing time as a function of the distance along the shape, re-sampling may use interpolation (e.g. linear) or smoothing methods such as spline smoothing. This has the disadvantage that the result reflects the properties of the interpolation scheme which might not be desirable. E.g. in the case of three consecutive observations of a bus, the middle one occurring while the bus was in a stop s (defined as a small interval $[X(s) - G, X(s) + G]$ around the stop location $X(s)$ according to the shape of length, $2G = 30\text{m}$) while the remaining two are on the road linear interpolation implies that the bus is at the stop only for a short duration. For spline smoothing the stopping time will depend heavily on the smoothing parameter. This behavior clearly is undesirable.

As an alternative, explicit modeling of bus travel explains time $t(x)$ as a function of distance along shape x. The simplest model for bus movement is constituted by assuming constant speed v_s for the travel between stops s and $s+1$ leading to travel time $\widehat{TT}_{i,s} = (X(s+1) - X(s) - 2G)/v_s$ for trip i on route segment s and non-negative stopping times $\widehat{ST}_{i,s}$ inside the stop (seen as intervals $[X(s) - G, X(s) + G]$ around the location of stop s). For a shape with S stops (not counting the start of the trip) this leads to $2S$ parameters. For this model $\widehat{TT}_{i,s}$, $\widehat{ST}_{i,s}$ and the time $t_{i,1}$ of the start

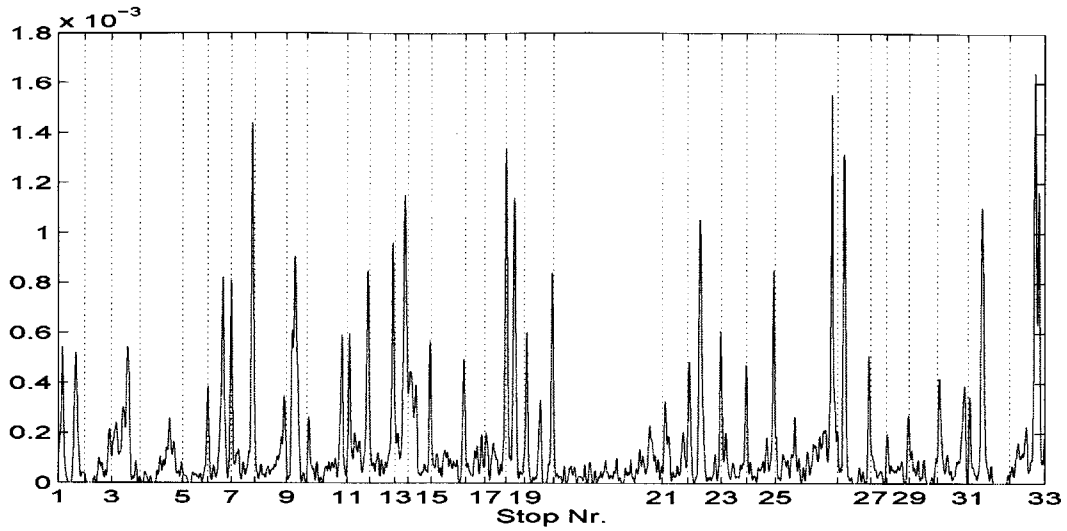


Figure 3-2: Kernel density estimator for observed location of buses of route 1 in Boston in the inbound direction. Dashed lines indicate places of stops.

of the trip fully determine the trajectory of the bus according to

$$\hat{t}(x) = t_{i,1} + \sum_{s:X(s)<x} (\widehat{TT}_{i,s} + \widehat{ST}_{i,s}) + \frac{x - X(s) - G}{X(s+1) - X(s) - 2G} \widehat{TT}_{i,s+1} \tag{3.1}$$

for each $x \notin [X(s) - G, X(s) + G]$ not in a stop. Inside the stops linear progression between entering the interval and exiting the interval can be assumed without loss of generality.

This model needs to be calibrated with real world data in order to be useful. In this respect the following observation will be used: For an observer that searches for the location of a bus at a random time instant, the chance to find the bus in an interval of 10m, say, is proportional to the share of time the bus spends in this interval during the observation period. The same holds true for more frequent sampling of bus location. Fig. 3-2 provides a snapshot of the location of observations of buses on route 1 in the inbound direction (i.e. in direction of Harvard).

It can be seen that at some stops buses are more frequently found. Other spikes occur at traffic lights. Sampling this distribution on a spatial grid of grid size 10m one obtains the probability to find a bus at a random time point in the corresponding region. This can be related to average dwell time percentages $\bar{ST}_s$ inside the stop

intervals as well as on the segments between the stops.

In order to calibrate this model the observations are split into two groups: Observations in the stops $[X(s)-G, X(s)+G]$ are termed in-stop observations and denoted as $z_{i,k}, k = 1, \ldots, K$ with corresponding time $t_{i,k}$ and stop $s_{i,k}$. The remaining ones are termed on-road observations and denoted as $y_{i,l}, l = 1, \ldots, L$ with corresponding time $t_{i,l}$ and segment $s_{i,l}$. In-stop observations impose restrictions as $\hat{t}(X(s_{i,k}) - G) \leq t_{i,k}$ and $\hat{t}(X(s_{i,k}) + G) \geq t_{i,k}$, i.e. the trip must arrive at the stop prior to being observed in the stop and depart after being observed there. The on-road observations should be replicated as good as possible using the model. From the assumption of constant speed between stops it is clear that there will be no perfect match in particular in situations where the bus needs to wait at a traffic light.

At the same time the model is desired to match the dwell time profile as closely as possible in order to incorporate information on typical dwell times. Hence the re-sampling is achieved by finding the parameters minimizing the squared distance to the scaled (with the actual total travel time TTT_i for the whole trip) dwell time profile and the weighted on-road observations subject to the restrictions on arriving and departure times implicit in the on-stop observations:

$$\begin{aligned} \min L(t_{i,1}, \widehat{TT}_{i,s}, \widehat{ST}_{i,s}, s = 1, \ldots, S) \quad := \quad & \sum_{l=1}^{L} (t_{i,l} - \hat{t}(y_{i,l}))^2 + w \sum_{s=1}^{S} (\widehat{ST}_{i,s} - TTT_i * \bar{ST}_s)^2, \\ s.t. \qquad & \hat{t}(X(s_{i,k}) - G) \leq t_{i,k}, \\ & \hat{t}(X(s_{i,k}) + G) \geq t_{i,k}, k = 1, \ldots, K, \\ & \widehat{TT}_{i,s} \geq (X(s+1) - X(s) - 2G)/V, s = 1, \ldots, S, \\ & \widehat{ST}_{i,s} \geq 2G/V, s = 1, \ldots, S. \end{aligned} \tag{3.2}$$

Here $w > 0$ is a weighting factor. Large w results in closer fit to the average dwell times, small w puts emphasis on being close to measured observations. V imposes a maximal travel speed. This leads to a linear least squares problem with linear restrictions which can be efficiently solved using general purpose optimizers.

The re-sampling procedure uses the assumption that the expected dwell times

are identical for all buses, i.e. that there are no systematic deviations from the expectations. This is not realistic for a full day, while it appears tenable for time intervals across different days. Consequently the re-sampling is calculated separately for a segmentation of the day into ten time intervals.

Below the re-sampling procedure is validated in two ways: First the results from the re-sampling are compared to the observations in a synthetic simulated data set. Second, real data using a GPS logger with a higher temporal resolution has been obtained.

Validation Using Synthetic Data

In order to validate the re-sampling procedures a synthetic data set of buses running on route 1 has been generated using a microscopic simulator implementing the optimal velocity model (OVM) as presented in (107) with a discrete time update of one second and a total duration of 10 hours. Only one direction with no overtaking is simulated. The shape contains 33 bus stops. If a bus reaches a stop, a random integer is drawn simulating uncertain boarding and alighting processes. The stop duration is distributed discretely uniform $\{0, 1, 2\}$ seconds except for stops 6, 21 and 31 where the range is 30 to 149 seconds and stops 7 to 20, where the range is 5 to 19. On the route twenty intersections with traffic signals are simulated. The signal timing is coordinated using the maximum allowed speed with red and green time split evenly at 30 seconds each. During the red light periods of signal 2, 7, 12, 17 and 18 cars enter the road segment with intensity $t/36000 * 0.75$ according to a Poisson arrival process. Other than that cars enter the road at the start of the shape at an arrival rate of 0.2. Every three minutes a bus is drawn. The stopping of a bus in a stop is not modelled in detail, but rather buses stop immediately when reaching the stop and leave after boarding and alighting is completed. In between cars pass the bus. The added complexity of deceleration into the stop is included in the random boarding and alighting, the reintegration into traffic follows the rule that cars need to stop for reentering buses.

195 bus trajectories are generated at a sampling frequency of one second. Sub-

sequently the trajectories are sub-sampled to a sampling frequency of sixty seconds using random starting time stamp. The two re-sampling strategies (simple linear as well as the procedure proposed above) are applied and stopping times as well as travel times between bus stops are calculated with the two approaches. For the distribution based re-sampling the 10 hours are partitioned into three intervals.

A sample of the output of the resampling procedure can be found in Figure 3-3 below. It can be seen that the resampling follows the true observations more closely than linear interpolation. The results show that the more complicated re-

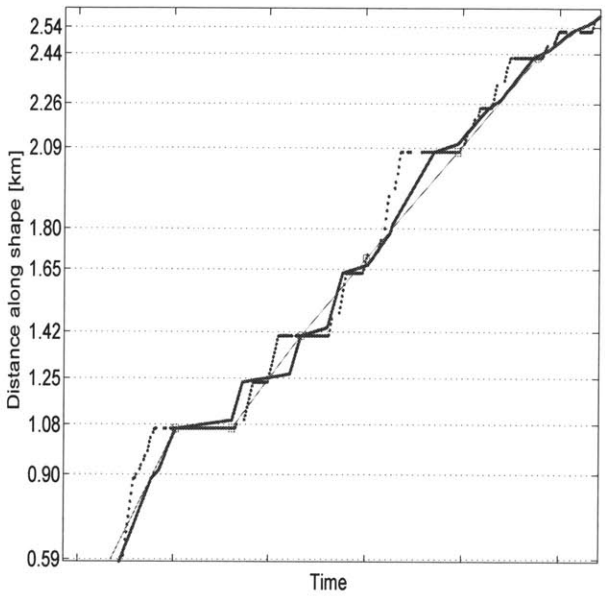


Figure 3-3: MBTA route 1: Sample output of the resampling scheme. Blue dots: high frequent observations. Magenta boxes: Observations with sixty second temporal resolution. Thin magenta line: linear resampling. Thick red line: Proposed resampling.

sampling pays off resulting in a smaller mean absolute deviation for stop duration of 8.8 seconds compared to 11.2 seconds for the simple linear interpolation based re-sampling. Also the travel times between the stops are replicated with a higher accuracy (mean absolute deviation of 10.3 compared to 12.0 second). Additionally note that the absolute performance is high in comparison to the sampling interval 60 seconds. This is mainly due to a better capturing of the long stops. For the three long stops 6, 21 and 31 the mean absolute deviation of the re-sampling equals 21.4

seconds compared to 46.3 seconds for the simple method.

Validation Using High-Frequency GPS Data

In order to validate the methodology, high-frequency GPS records with a sampling frequency of one second are collected on bus route 1 on two days, Thursday February 23rd 2012 and Saturday February 25th. A total of 15 trips provide data which is map matched and converted to sequences of (time stamp, distance along shape) pairs. The one second interval GPS data provides ground truth against which the two sampling strategies are validated. To this end the trips are separated into different regimes according to weekend or weekday as well as five intervals during the day.

The two re-sampling schemes have been applied to sub-sampled (using each sixtieth observation) copies of each of the 15 trips. In order to remove random effects due to the starting point all 60 sub-sampled versions are used.

The results are less pronounced than for the synthetic data but still an advantage of the more complex re-sampling scheme compared to the simple method can be observed. The mean absolute deviation in stopping time over all stops in direction 1 totals 4.83 seconds for the presented re-sampling method and 7.7 seconds for the simple interpolation. In direction 0 the numbers of 6.5 for the proposed method compared to 7.7 for the simple method results in a slight advantage of the proposed method.

Note in this respect that the errors are comparable but smaller than in the synthetic data set. Thus, in the following the presented re-sampling method will be used.

3.3 Analysis of Service Quality

The resampled data represents a great opportunity to make statistics of the bus service performance. The usual service quality measurements relate strongly to variations of headway, travel time and variability of travel time. Transit operators are interested to optimize indicators based on these components which include elements

not under the influence of the transit operators such as traffic conditions and demand fluctuations. In this section I'll demonstrate that the resampled data set can be used in order to extract useful information about these three components of service quality measurement.

The total travel time in a transit trip can be decomposed into walking access time (time from origin to bus stop/ train station), waiting time, in-vehicle travel time, and transfer time (walking time from the alighting stop of the first route to the boarding stop of the second route). Among these components, waiting time and in-vehicle travel time are determined by transit operations and traffic road conditions. It is those dynamic components that transit operators are interested in optimizing. In this section the variations of headway, the travel time and the variability of travel time are analyzed.

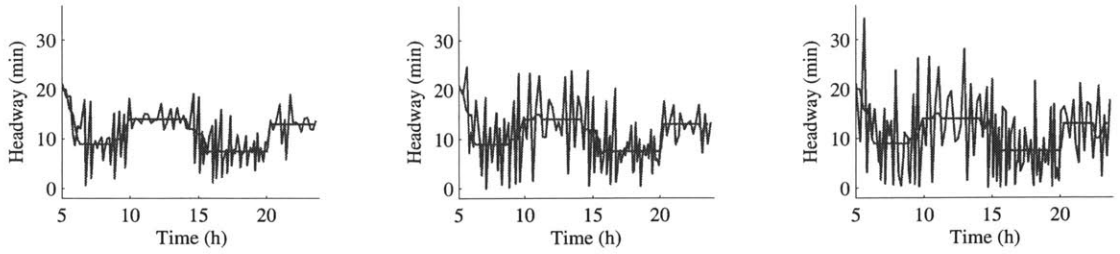
3.3.1 Headway

Headway is defined as the time interval from the tip of one vehicle to the tip of the next one behind it arriving at a certain place (usually a stop). The expectation μ and the variance σ^2 of headway influence expected waiting times at stops according to the following formula (see (61)):

$$W = \mu \times \frac{(1 + \frac{\sigma^2}{\mu^2})}{2} \tag{3.3}$$

W is the expected waiting time, μ is the average headway and S is the standard deviation of headway. When $S = \mu$, the expected waiting time is twice as when $S = 0$. This shows that the variance of headway also has a significant influence on passenger's waiting time.

For MBTA route 1, Fig. 3-4(a) to Fig. 3-4(c) show how the actual headways compare to the scheduled ones at the initial stop, the 15th stop, and the last stop on one of the workdays. The headway deviation is defined as the difference between the actual headway and the scheduled headway. Even at the initial stop the deviations are significant. The deviations at the initial stop are caused by operation issues:



(a) Actual VS. Scheduled Headway at initial stop

(b) Actual VS. Scheduled Headway at 15th stop

(c) Actual VS. Scheduled Headway at the last stop

Figure 3-4: Actual vs. scheduled headway at the initial, the middle and the last stops.

either because there are buses available at the terminal but the operators fail to dispatch them in time, or because bus slack time at the terminal is not enough so that buses are not ready at their scheduled departure time. For the first case, more timely dispatching is needed. For the second case, bus slack time at terminals should be re-optimized or simply more buses are needed. To explore how headway changes from the first to the last stop during the evening peak, the headway distribution at each stop is calculated and various distributions such as exponential, Erlang, gamma and normal distribution (54; 15) are fitted. The best statistical fit is obtained by the three parameter gamma distribution which is recommended by the traffic engineering handbook (93): The corresponding probability density function equals

$$f(x) = \frac{(x-\gamma)^{\alpha-1}}{\beta^{\alpha}\Gamma(\alpha)} exp(-(x-\gamma)/\beta), x \geq \gamma, f(x) = 0, x < \gamma. \tag{3.4}$$

Here $\alpha > 0$ is the continuous shape parameter. When α is 1 the distribution becomes an exponential distribution and when α is 4 or 5 the shape is close to a normal distribution. $\beta > 0$ is the continuous scale parameter. The larger the scale parameter, the more spread out the distribution is. γ is the continuous location parameter which determines the center of the distribution. Γ is the Gamma function. The comparison of headway histograms and fitted distributions are shown in Fig. 3-5. The three parameter gamma distribution fits quite well from the initial stop to the last stop. Fig. 3-6 shows how the three parameters change from the first to the last stop. At the

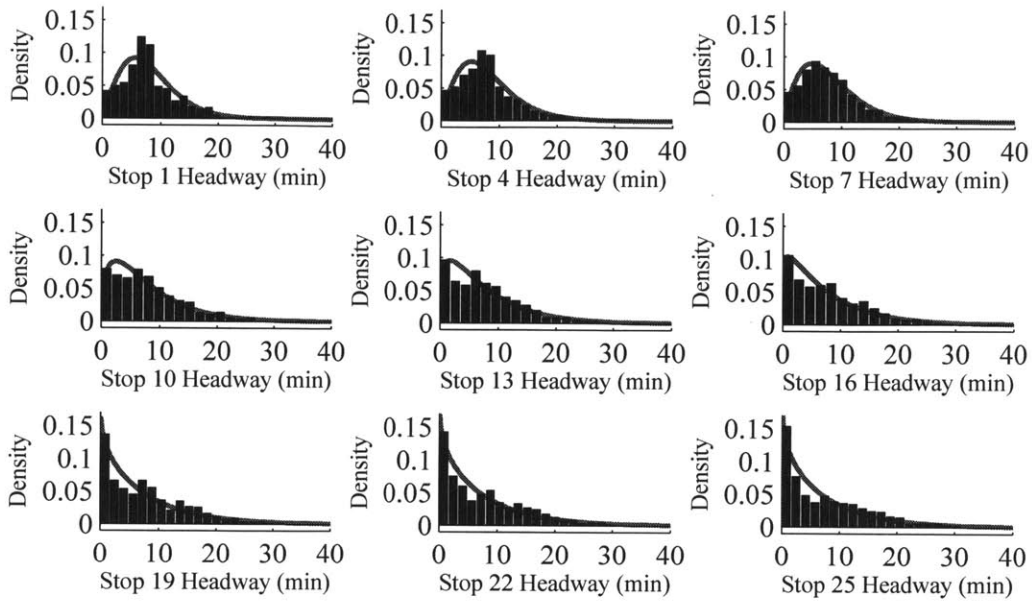


Figure 3-5: Headway fitting results at different stops using the three parameter gamma distribution.

initial stop, the shape parameter is close to 4 which shows that it is relatively close to a normal distribution. The shape parameter decreases quickly along the route. After stop 9 it stabilizes at around 1 which indicates that it is close to an exponential distribution. The location parameter also stabilizes at around 0 after stop 9. This means that after stop 9 a large proportion of headways are close to 0, which indicates that bus bunching is severe. After the shape and the location parameter stabilize, the scale parameter keeps increasing which shows that the variance of headway keeps increasing. These three parameters show that the headway changes from a rather deterministic manner (close to schedule) at initial stops to a random manner (severe bunching) at around stop 9. The analysis below will explain why stop 9 is a critical point. To further understand the headway variations, the headway coefficient of variation C_{vh}, a measurement proposed in TCQSM (72), is calculated at different stops:

$$C_{vh} = \frac{Standard\ deviation\ of\ headway\ deviations}{Mean\ scheduled\ headway} \tag{3.5}$$

Fig. 3-7 shows how the headway coefficient of variation at each stop changes along the

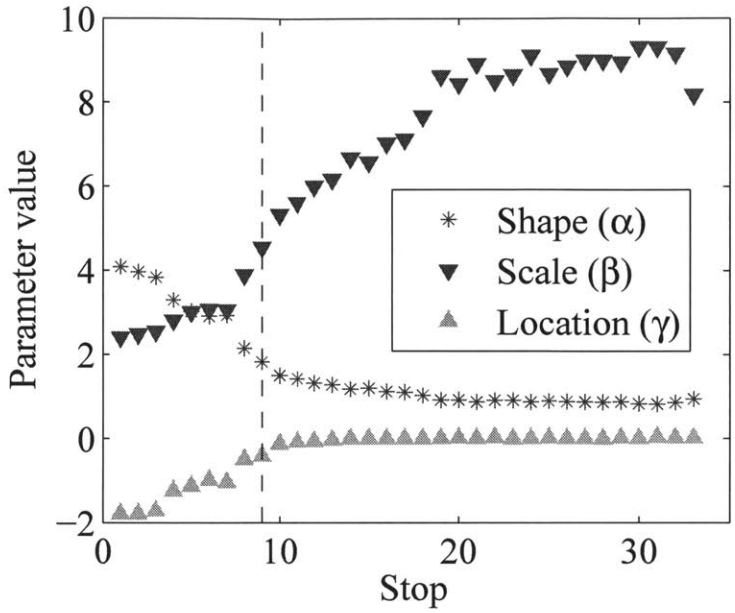


Figure 3-6: Evolution of the parameters of the gamma distribution across the bus route during the evening peak.

route. The general trend is that the headway coefficient of variation keeps increasing, but the rate of increase varies from stop to stop. Between some stops it increases more quickly, which shows that in these segments travel times (between arriving at consecutive stops) are more unstable. In order to observe the change rate, the gradient of headway coefficient of variation at each stop is shown in Fig. 3-7. At stops 8 and 19 it increases the most. Inspection from the map shows that between stop 8 and 9 the bus turns from a secondary road (Albany St.) to the main artery connecting Boston and Cambridge (Massachusetts Avenue). The traffic signal waiting time at this intersection could vary a lot which causes higher headway variance. Bus priority at this intersection hence could greatly increase headway regularity. Between stop 18 and 19 there are three closely spaced intersections. The Metro Green line also transfers with route 1 at stop 18. Hence both the waiting times at intersections and varying passenger flows make the headway unstable here. Two possible ways to improve the service quality are better bus priorities at these traffic lights and holding strategies in order to better synchronize buses and the metro line. Notice that while the headway variance increases from the first to the last stop, the average headway

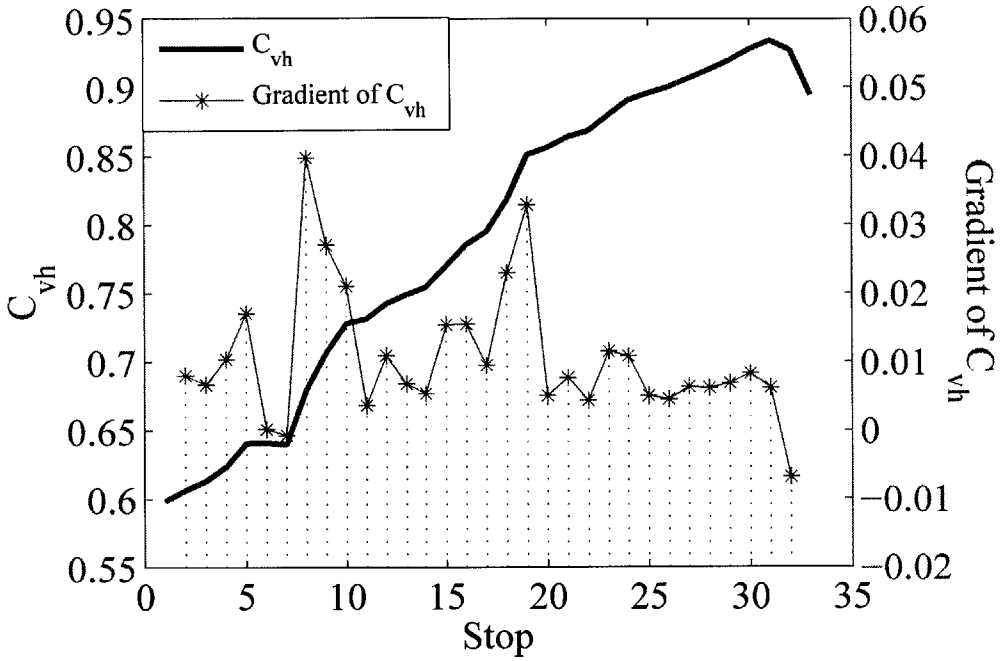


Figure 3-7: Headway coefficient of variation (left axis) and its gradient (right axis) showing bottlenecks with high headway variation increase at stop 9 and 19.

remains almost the same. The average headway at the first, 15th and the last stop are respectively 485s, 487s, 490s. Passengers at latter stops would experience much longer average waiting time not because there are not enough buses, but because the variance of headway is higher. Solutions for this situation typically include bus holding strategies and stop skipping strategies which could be further explored.

3.3.2 In-vehicle Travel Time

Another component of the total travel time is in-vehicle travel time which is composed of two parts: travel time between stops and stop dwelling time. In-vehicle travel time is largely determined by traffic conditions and the road network structure. Fig. 3-8 shows how the average total trip time, running time and stop dwelling time change during different times of day. The total trip time has clear morning and evening peaks at 8am and 5:30pm respectively. Running time and stop dwell time show identical peaks, which means that the increase in the total peak trip time is caused by both

increasing passenger volumes and slower travel speed. Running time has a larger influence on the increase of the total peak trip time.

Fig. 3-9 compares the average travel speed (stop dwell time not included) and the percentage decrease from off-peak to peak hours at each segment. Segment i is the road between stop i and $i+1$. At segment 20 the average speed is always the highest because this segment is at Harvard bridge and on the bridge there are no traffic lights or stops. The peak hour speed is generally lower than the off-peak hour speed. The highest percentage decreases are at segment 12 and 25 which are respectively at the intersection of Massachusetts Avenue and Tremont St. and at Central Square, both are crowded commercial areas in Boston and Cambridge respectively. Traffics could be guided to parallel roads during peak hours to relief the burden on these segments. It is interesting to notice that stops with the most percentage decrease in peak hour

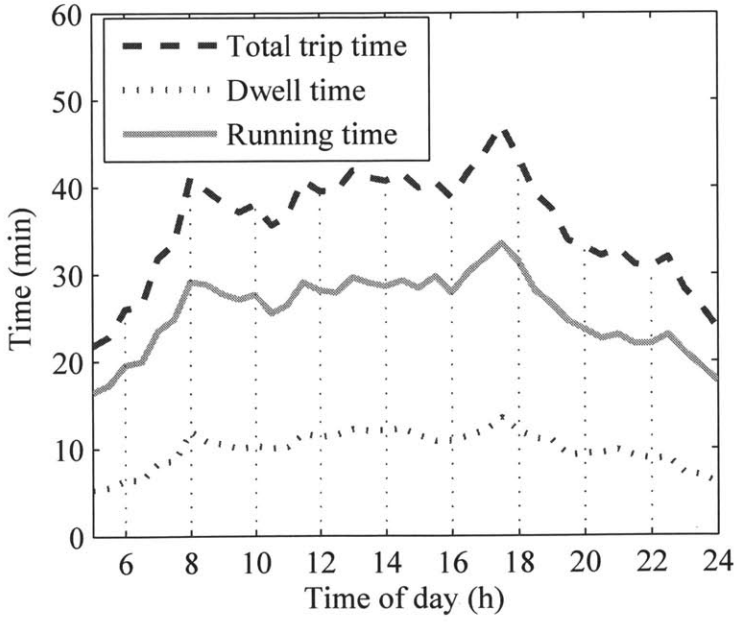


Figure 3-8: Trip time composition at different times of day. Values represent the sum over the entire route.

speed do not correspond to stops where the statistics of headway varies the most. This is because the headway coefficient of variation is a measurement of stability while the in-vehicle travel time is a measurement of the average speed performance.

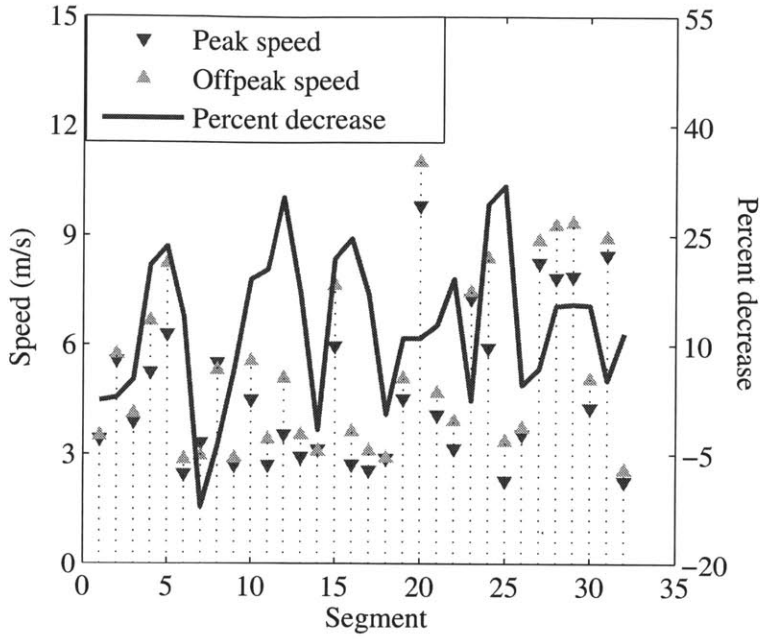


Figure 3-9: Trip time at different times of day and peak/ off-peak travel speed comparison.

3.3.3 Variability of Trip Travel Time

Another component of bus performance is constituted by travel time variability. High variability implies low predictability and hence uncertain travel times. Thus alongside the expected travel time, travel time reliability is one of the most important factors when selecting a route to a desired destination. The expected travel time can be calculated easily by summing up the mean segment travel times along the route in the transportation network. On the other hand, in order to estimate variability, it is necessary to account for correlations between the individual segments composing the trip. Figure 3-10(a) shows the correlation matrix of the segment travel times along the bus route. The probability distribution of the trip travel time can be approximated by building clusters of highly correlated segments and assuming full correlation within each cluster and independence between segments in different clusters. The quantile function, i.e. the quasi-inverse of the cumulative distribution function (CDF), of the sum of fully correlated segment travel times $\widehat{TT}_s, s \in C$ in cluster C is computed as

the sum of the quantile functions of the individual constituents, i.e.

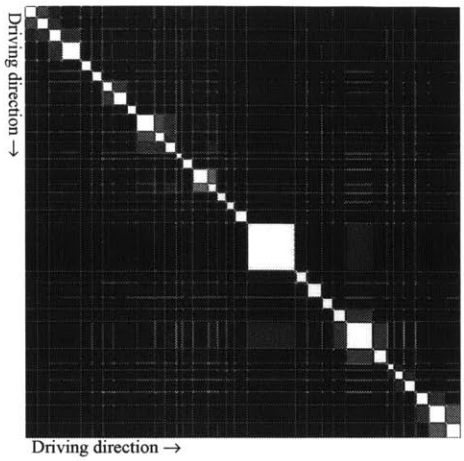
$$Q_C(p) = \sum_{s \in C} Q_{\widehat{TT}_s}(p) \tag{3.6}$$

where $Q_c(p)$ denotes the p-th ($0 \leq p \leq 100$, percent) quantile of the travel time for cluster C. The distribution of the sum of independent cluster travel times is approximated by Monte Carlo simulation, where cluster travel times are drawn repeatedly and randomly according to the previously calculated probability distributions of the respective clusters. Figure 3-10(b) compares the resulting estimations with empirical travel time distributions for trips from the first to the last stop of MBTA route 1. For comparison, also depicted are the results under the assumption that all segments are independent and assuming that every segment is fully correlated with every other. These results show that not accounting for dependencies leads to underestimation of travel time variability, whereas the proposed approximation yields good agreement with the observed distribution. Thus the low frequency localization data can be used in order to infer route travel time reliability for arbitrary routes along the line providing another way to investigate bus performance.

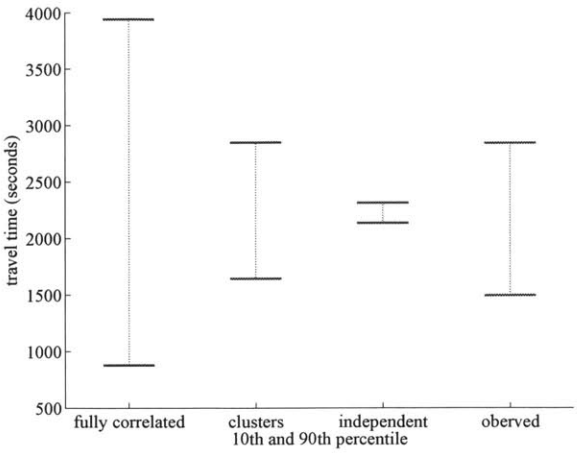
3.4 Application: Calibration of a Bus Movement Model

Beside performing service quality analysis and bottleneck diagnosis, transit agencies may be interested to evaluate the effect of different measures to improve service quality. This section shows that the headway and the travel time statistics calculated in the previous section can be used to calibrate bus movement simulation models.

The bus movement is affected by traffic conditions, traffic signals and the number of passengers boarding and alighting. The number of passengers is both related to passenger arrival rates and the arrival time of the previous bus. In ref. (36) a very convenient bus movement model is built incorporating the effect of the previous bus and the random noise caused by road conditions and traffic signals. Using the



(a) Correlation matrix of segment travel times on MBTA Route 1. White indicates high correlation and black indicates low correlation. Red lines mark the positions of stops.



(b) From the left to the right: the spread between the 10^{th} and 90^{th} percentile of the estimated travel time distribution assuming (a) full correlation between all segments, (b) full correlation within and independence between segment clusters, (c) independence between all segments and (d) the observed travel time distribution.

Figure 3-10: Correlation between segment travel times and comparison of estimated travel time under various assumptions to observed travel time.

measured statistics of service performance this model can be calibrated with more accurate statistics from headway and travel time of different segments in a route, which are in general introduced as random variations.

3.4.1 The Model

The bus movement model of (36) can be expressed as:

$$U_{n,s} = C_s + \beta_s(h_{n,s} - H) + v_{n,s+1} \tag{3.7}$$

Here $U_{n,s}$ is the nth run's segment travel time from stop s to $s+1$. The dwell time at stop s is included while the dwell time at stop $s+1$ is not. C_s is the scheduled travel time from s to $s+1$. H is the scheduled headway while $h_{n,s}$ is the actual headway for the nth run at stop s. β_s is a dimensionless parameter expressing the effect of the deviation from the scheduled headway on the dwell time. If a headway is longer than the scheduled value and the passenger arrival rate remains constant, there will

be more passengers arriving than expected which causes longer than expected stop dwell time. The inclusion of β_s makes two buses "attract" each other when their headway is shorter than H and "repel" each other when the headway is longer than H. it is mentioned (36) that β_s typically ranges from 10^{-2} to 1. The noise term $v_{n,s+1}$ incorporates effects such as road conditions and traffic signals. It is assumed to have zero mean, variance σ^2_{s+1} and to be independent of h_{ns}.

$a_{n,s}$ represents the arrival time of the nth run at stop s. The above equation could be transformed as:

$$a_{n,s+1} - a_{n,s} = C_s + \beta_s(a_{n,s} - a_{n-1,s} - H) + v_{n,s+1} \tag{3.8}$$

3.4.2 Model Calibration

As $a_{n,s+1}$, $a_{n,s}$, and $a_{n-1,s}$ can be acquired directly from the interpolated data, estimates of β_s can be obtained using regression. Since the random noise term $v_{n,s+1}$ is

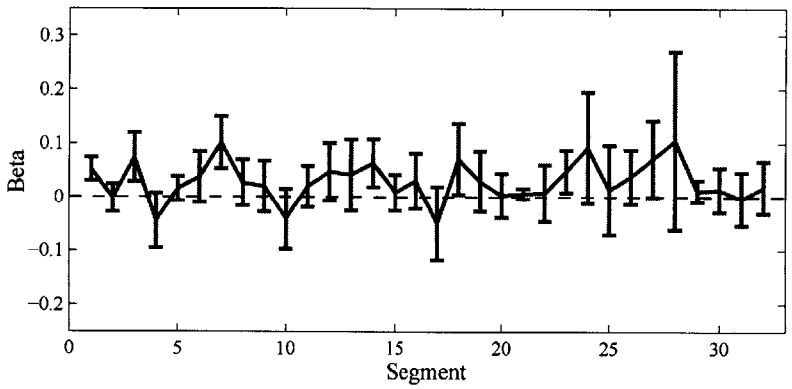


Figure 3-11: Regression result of β_s (see equation (3.7)) at each segment.

assumed to have zero mean and to be independent of h_{ns}, $a_{n,s+1} - a_{n,s} - C_s$ can be regressed on $a_{n,s} - a_{n-1,s} - H$ to obtain estimates for β_s.

The regression results with error bars, provided in Fig. 3-11, show the expected positive signs for 27 out of 32 segments. None of the negative coefficients are significant. All coefficients are of the expected magnitude. Of all the 32 segments, 27 are above zero (the upper bounds are all above zero). The rest 5 show unexpected

negative signs. They are respectively segment 2, 4, 10, 17, 31, which are all stops with light passenger flows. This may also be caused by unstable passenger arrival rates. The maximum β_s is 0.1047 at segment 28 while the average value of β_s is 0.0287. These values all agree with the typical β_s values (36). which ranges from 10^{-2} to 1. Note that the low number of significant coefficients may be an indication of insufficient sample size and hence small power of the tests.

With β_s known, the distribution of the residual random noise term could be estimated. First the assumption that v_{s+1} and h_s are independent needs to be tested. The correlation between v_{s+1} and h_s for each segment is calculated. The maximum value is $4.74 * 10^{-16}$ while the minimum value is $-4.43 * 10^{-16}$. This confirms that v_{s+1} and h_s are independent. It can be observed that when using the interpolated data the shape of v_{s+1} estimated at each segment separately fits well a lognormal distribution.

Notice that with the interpolated AVL data all the components needed for the movement simulation model are available. C_s and H can be acquired from the schedule. Headway distributions using the Gamma family of densities have been fitted above, v_{s+1} is approximated using a log-normal distribution.

Fig. 3-12 provides a comparison between the results of three different ways to calculate segment travel time U_s are compared. The first one calculates U_s directly from the interpolated AVL data. It is regarded as the true segment travel time. The second one, the red line on each plot, calculates U_s from the right hand side of equation (3.8) using the observed three parameter gamma distribution for h_s and the lognormal distribution for v_{s+1}. The third one, the green line on each plot, shows the model without calibration, that is when using the most common form, random values (e.g. a normal distribution) for both v_{s+1} and h_s. The latter two can be regarded as simulated values and the calibrated model gives clearly more accurate results to all the segments along the route. Kolmogorov-Smirnov statistics is used to test quantitatively how close the simulation results are to the true values. The legend on each plot shows Kolmogorov-Smirnov test values for each of the two simulation methods. The calibrated model has smaller Kolmogorov-Smirnov test values at all

the segments. This shows that important corrections can be obtained applying the two presented functions (the three parameter gamma and the lognormal distribution) into bus movement models.

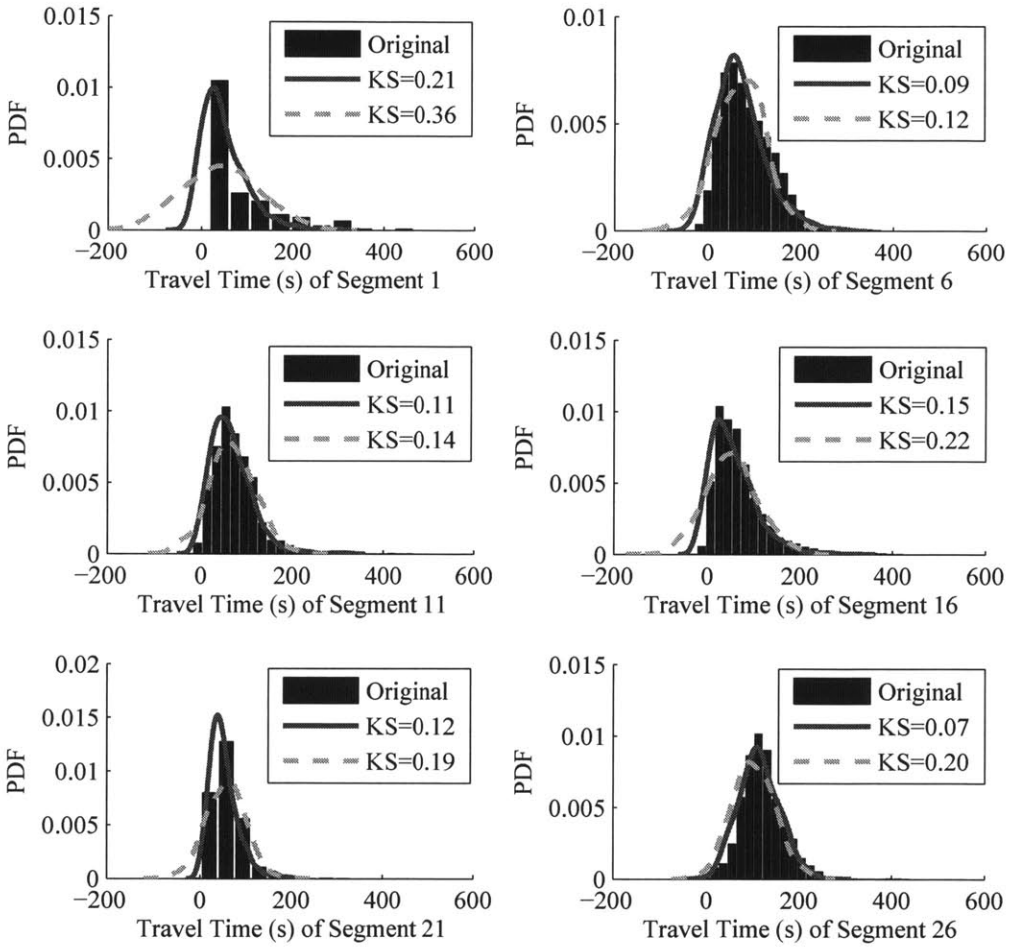


Figure 3-12: Comparison of bus travel time (U_s) at different segments of the route. The original data and two simulated bus models are shown. One (green dashed line) assuming normally distributed random noise term and the other one (red solid line) calibrated with the statistical distributions observed in this study. Kolmogorov-Smirnov test values are shown on the legends.

3.5 Conclusions and Outlook

Public transportation is playing an increasingly important role in urban transportation systems since limited road capacities largely restricts the private car growth. ADC provides a new but not yet perfect source of data for transit agencies. How to transform the data to overcome the limitations and to extract useful information becomes an interesting topic.

In this chapter a low-frequency AVL data analysis procedure is proposed which allows service performance evaluation and the calibration of a bus movement model. It is demonstrated how this procedure step by step turns the raw AVL data into information for service bottleneck diagnosis and bus movement simulation. The data used are low-frequency location-at-time AVL records as provided e.g. by Nextbus. In particular the point-to-point and point-to-curve map-matching are applied to rule out errors such as wrong temporal order, no matching from the shape, and wrong interval. Kernel density estimator observed from the preprocessed AVL data is used to estimate arrival and departure times at stops. Once the stop arrival and departure times are estimated, service quality is analyzed by measuring statistics of headway and in-vehicle travel time. Headway evolve from a normal distribution at the initial stops to an exponential distribution at the last stops. The headway analysis helps to identify bottlenecks caused by the road network layout and passenger volumes while the comparison of peak vs. off-peak hour travel speed helps to identify bottlenecks caused by traffic conditions. Finally, these observations are used to calibrate a model of bus movement to show its further application prospects.

The main contributions of this study to the state-of-the-art research can be concluded as:

1. A more robust and accurate data prepossessing methodology is provided which is demonstrated to be superior to the widely applied linear interpolation method. Different kinds of errors in AVL data are discussed and ruled out. More accurate stop arrival and departure time estimates are obtained by using a kernel density estimator of bus dwell time.

2. Using this preprocessing method, headway distribution evolution along one bus route is studied in detail. It is demonstrated that the method can be used in order to detect bottlenecks caused by both road layout and traffic conditions separately so that they can be treated differently to improve service quality.

3. Route travel time variability can be inferred from clustering of segments based on segment travel time correlations. This delivers hints on bus performance problems via increases in variability thus providing a more complete view on the performance of the bus line.

4. These results can be used to calibrate bus simulation models which in turn can be further applied to evaluate various bus control strategies.

Therefore this chapter demonstrates the potential of widely available (and hence low cost) low-frequency AVL data to improve bus service and to provide valuable information for the passengers in terms of travel time predictions including travel time reliability. In particular it is shown show that such data provides an alternative means for monitoring and controlling bus performance for transit authorities not willing to invest in more expensive solutions.

Chapter 4

Conclusion

Our current digital age is characterized by the shift from traditional industry to an economy based on the information computerization. The sweeping changes brought about by digital computing and communication technology during the latter half of the 20th century have provided new data sources for transportation modeling.

In this thesis, two mainstream trend in utilizing digital traces in transportation modeling are exploreds. The first is to use mobile phone data, combined with digital map point of interests, to calculate commuting OD matrix. The second is to use pre-processed low frequency bus GPS data to evaluate public transit service quality and diagnose service bottlenecks.

Human mobility modeling is an essential component of various areas of study ranging from epidemiology to urban and transportation planning. Commuting flows take up a large proportion of the total flows of a population. Current mainstream models for commuting flow prediction are the 'gravity model' and the 'intervening opportunity model'. These models require previous Origin-Destination (OD) matrices as input for parameter calibration. The recently proposed 'radiation model' is parameter-free, has a closed analytical form and presents the potential to become an universal model for OD generation.

In chapter 2 an extension of the radiation model is proposed that predicts the network of commuting flows at different spatial scales and in different cities worldwide. In addition, the spatial density of the distribution of facilities as downloaded from

Internet's digital maps, also known as POIs, together with the density of population are the two basic ingredients for modelling commuting networks.

The proposed extension uses one parameter to adjust to the different degrees of the homogeneity of opportunities when the scale of the study region changes. In contrast, while due to the large number of fitting parameters, the doubly constrained gravity model also fits the number of trips in most of the studied datasets, the obtained parameters cannot be generalized and depend on empirical flows of the particular region under study for their calibration. The parameter α in the extended radiation model, takes into account the effects of both the scale and the heterogeneity of opportunities, which are highly correlated. Thus, α to a large extent can be estimated by knowing only the size of the region under consideration.

In order to explore the validity of the new model in diverse cities that lack of census data, first the Bay Area is used as an example to show that cell phone records are a very good alternative data source to extract commuting patterns. Home and work locations are inferred at individual level from the cell phone records and then aggregated to show its equivalence to the commuting information obtained by census data. Then commuting flows are extracted from cell phone users in three different countries where census commuting data is not available. These results show not only the applicability of the proposed model to successfully model the commuting patterns for the cell phone users in cities from these three countries, but also show some unique commuting characteristics in each city.

In chapter 3 a low-frequency AVL data analysis procedure is proposed which allows service performance evaluation and the calibration of a bus movement model. It is demonstrated how this procedure step by step turns the raw AVL data into information for service bottleneck diagnosis and bus movement simulation. The data used are low-frequency location-at-time AVL records as provided e.g. by Nextbus. In particular the point-to-point and point-to-curve map-matching are applied to rule out errors such as wrong temporal order, no matching from the shape, and wrong interval. Kernel density estimator observed from the preprocessed AVL data is used to estimate arrival and departure times at stops. Once the stop arrival and departure

times are estimated, service quality is analyzed by measuring statistics of headway and in-vehicle travel time. Headway evolve from a normal distribution at the initial stops to an exponential distribution at the last stops. The headway analysis helps to identify bottlenecks caused by the road network layout and passenger volumes while the comparison of peak vs. off-peak hour travel speed helps to identify bottlenecks caused by traffic conditions. Finally, these observations are used to calibrate a model of bus movement to show its further application prospects.

The research conducted in this thesis opens the door to many future research topics:

The potential of both the cell phone records and the point of interests can be further exploited in future studies. From cell phone records we can observe long term regularities in daily activity patterns at individual level, which are very hard to acquire by traditional survey methods. Future mobility models should include not only commuting trips, but also trips of other purposes. With digital footprints such as the point of interests or Foursquare records (49), non-commuting trips, which have higher flexibility, can also be traced.

For the public transit analysis, further questions need to be answered: How to expand the performance analysis from one single line to the entire transit network?. How does translate in the practice the advantage of having a calibrated bus movement simulation model compare?, or How to use the model to develop strategies improve service quality?

Taken together, the thesis explores new digital data sources and methods in transportation modeling. The purpose is to provide analysis procedures that are of lower costs, higher accuracy and are readily applicable to different countries in the world.

Appendix A

Comparison of the No Constrained and Doubly Constrained Gravity Model

In this section we compare the estimation results of no constraint gravity model and doubly constrained gravity model on cell phone user commuting OD at city level. Here we use the cell phone records because the data is available at different countries so that we can perform a cross culture comparison. We choose 9 cities from the Bay area, Rwanda, Portugal and Dominican Republic: San Francisco, Oakland, San Jose, San Rafael, Lisbon, Kigali, La Romata, Santo Domingo, and Santiago. For each cell phone user we can estimate his/ her home and work location. Aggregate such results gives us the cell phone users' commuting OD matrix. Use the marginals (cell phone user commuting trip production and attraction number for each tower) as inputs for the following models.

No constraint gravity model takes the form:

$$T_{ij} = \frac{n_i^{\alpha} n_j^{\beta}}{f(r_{ij})} \tag{A.1}$$

T_ij is the flow between location i and j. Each location is a tower. n_i is the number of cell phone users whose home location is tower i, n_j is the number of cell phone

Table A.1: Regression parameters for the 9 cities

parameter	α	β	γ
San Francisco	0.17	0.09	0.46
Oakland	0.21	0.18	0.67
San Jose	0.16	0.15	0.56
San Rafael	0.21	0.23	0.73
Lisbon	0.21	0.23	0.73
Kigali	0.16	0.11	0.43
La Romata	0.53	0.33	0.91
Santo Domingo	0.25	0.20	0.68
Santiago	0.34	0.27	0.73

users whose working location is tower j. r_{ij} is the distance between them and f is the distance decay function. α and β are parameters to be fitted from data. We adopt the power distance decay function:

$$f(r_{ij}) = r_{ij}^{\gamma} \tag{A.2}$$

The model turns into:

$$T_{ij} = \frac{n_i^{\alpha} n_j^{\beta}}{r_{ij}^{\gamma}} \tag{A.3}$$

The parameters α, β, and γ could be estimated using least square linear regression (48) after a simple transformation:

$$log(T_{ij}) = \alpha log(n_i) + \beta log(n_j) - \gamma log(r_{ij}) \tag{A.4}$$

The inputs of the regression model are T_{ij}, n_i, and n_j, the outputs are estimation results of α, β, and γ. The α, β, and γ regression results for the 9 cities are:

In (111; 9) a similar regression method is applied. The difference is that trips are divided into short and long trips and the parameters are estimated separately. In [nature] the estimations of [α, β, γ] are [0.30,0.64,3.05] for short distances (r¡119km) and [0.24,0.14,0.29] for long distances. The doubly constrained gravity model takes the form:

$$T_{ij} = \frac{\alpha_i \beta_j O_i D_j}{r_{ij}^{\gamma}} \tag{A.5}$$

Table A.2: Seed sample matrix without expansion

Zone	1	2	3	4	O_i
1	0	1.5	2	3.5	150
2	1.5	0	2.5	3	200
3	2	2.5	0	2	100
4	3.5	3	2	0	50
D_j	30	70	250	150	Total=500

O_i and D_j are total trip production and attraction volumes at location i and j. For a study region with n locations, there are $2n$ parameters of α_i and β_j, and one parameter of γ. Unlike the no constrained gravity model, those it has $2n+1$ parameters, only one parameter, γ, needs to be predetermined. α_i and β_j can be estimated even without knowing T_{ij} by iterating:

$$\alpha_i = 1/\sum_j \beta_j D_j C(r_{ij}) \tag{A.6}$$

$$\beta_j = 1/\sum_i \alpha_i O_i C(r_{ij}) \tag{A.7}$$

Let's use a very simple example to illustrate the algorithm. Suppose an area with 4 zones. Their distance Matrix and O_i, D_j are as followed: Initially α_i and β_j are all set to 1 and β_j are updated as:

$$\beta_1 = \frac{1}{1*200*1.5^2 + 1*100*2^2 + 1*50*3.5^2} = 0.00848$$

$$\beta_2 = \frac{1}{11*150*1.52 + 1*100*2.52 + 1*50*32} = 0.01134$$

$$\beta_3 = \frac{1}{1*150*2^2 + 1*200*2.5^2 + 1*50*2^2} = 0.01220$$

$$\beta_4 = \frac{1}{1*150*3.5^2 + 1*200*3^2 + 1*100*2^2} = 0.01681$$

Table A.3: Converged sample OD matrix

Zone	1	2	3	4	O_i	α_i
1	0	45	86	19	150	0.71159
2	22	0	120	58	200	1.16540
3	7	20	0	73	100	1.31410
4	1	5	44	0	50	1.07820
D_j	30	70	250	150		
$\beta + j$	0.00710	0.01343	0.01291	0.01482		

Then α_i are updated:

$$\alpha_1 = \frac{1}{0.01134 * 70 * 1.5^2 + 0.01220 * 250 * 2^2 + 0.01681 * 150 * 3.5^2} = 0.00848$$

$$\alpha_2 = \frac{1}{0.00848 * 30 * 1.5^2 + 0.01220 * 250 * 2.5^2 + 0.01681 * 150 * 3^2} = 0.01134$$

$$\alpha_3 = \frac{1}{0.00848 * 30 * 2^2 + 0.01134 * 70 * 2.5^2 + 0.01681 * 150 * 2^2} = 0.01220$$

$$\alpha_4 = \frac{1}{0.00848 * 30 * 3.5^2 + 0.01134 * 70 * 3^2 + 0.01220 * 250 * 2^2} = 0.01681$$

After 4 iterations α_i and β_j values converge. The final OD matrix is:

Here we compare the results from: no constraint gravity model with parameters estimated in this study, no constraint gravity model with parameters estimated in previous study (97), doubly constrained gravity model with parameters estimated in this study. For each model we compare the model estimation results with the cell phone user commuting OD matrix and calculate the correlation between them. Fig. A-1 shows how the correlation changes from city to city and from model to model. In all cities the doubly constrained gravity model outperforms the no constraint gravity model. It has correlation more than 0.8 in all the cities except in Kigali, the capital city of Rwanda. We've mentioned that the commuting flow in Rwanda is special because it's more agglomerated: a few OD pairs have very large flows and these OD pairs are not necessarily close to each other. This makes it hard for gravity model prediction. The comparison of the doubly constraint gravity model and the no gravity model with parameters estimated in this study are in Fig. A-2 and A-3. Again the

doubly constraint gravity model prevails at each measurement. The 8 rows represent the 8 different cities and the three columns show: 1) the comparison between the actual and estimated flow volume from the doubly constraint gravity model; 2) the comparison between the actual and estimated flow volume from the no constraint gravity model; 3) the travel distance $P(r)$ distribution. Again the doubly constraint gravity model prevails at each measurement.

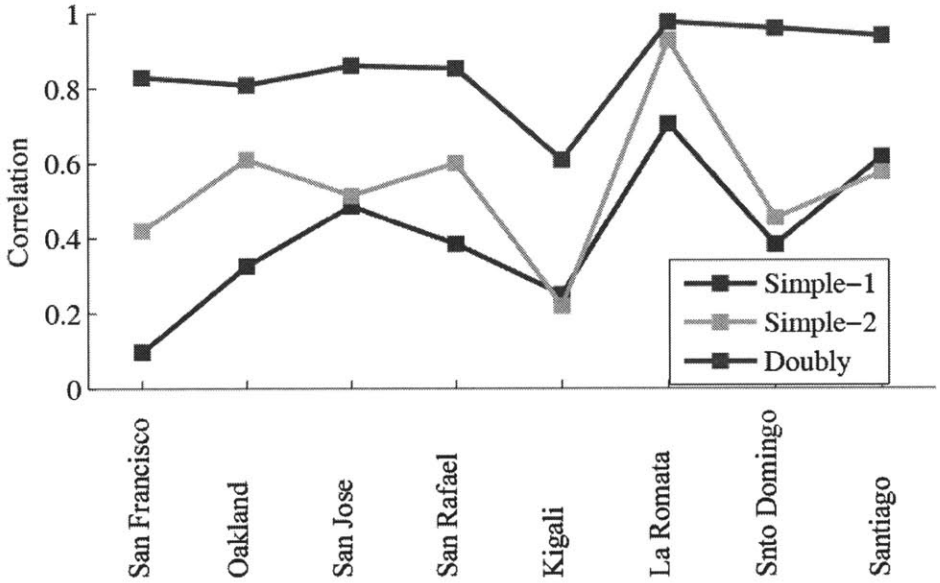


Figure A-1: The correlation between the census commuting OD pair volumes and results from different models. The doubly constraint gravity model's result is in red. The no constraint gravity model with parameters estimated from a previous study is in blue. The no constraint gravity model with parameters estimated in this study is in green. In all cities the doubly constraint gravity model outperforms the no constraint gravity model. It has correlation more than 0.8 with the actual census data in all the cities except Kigali

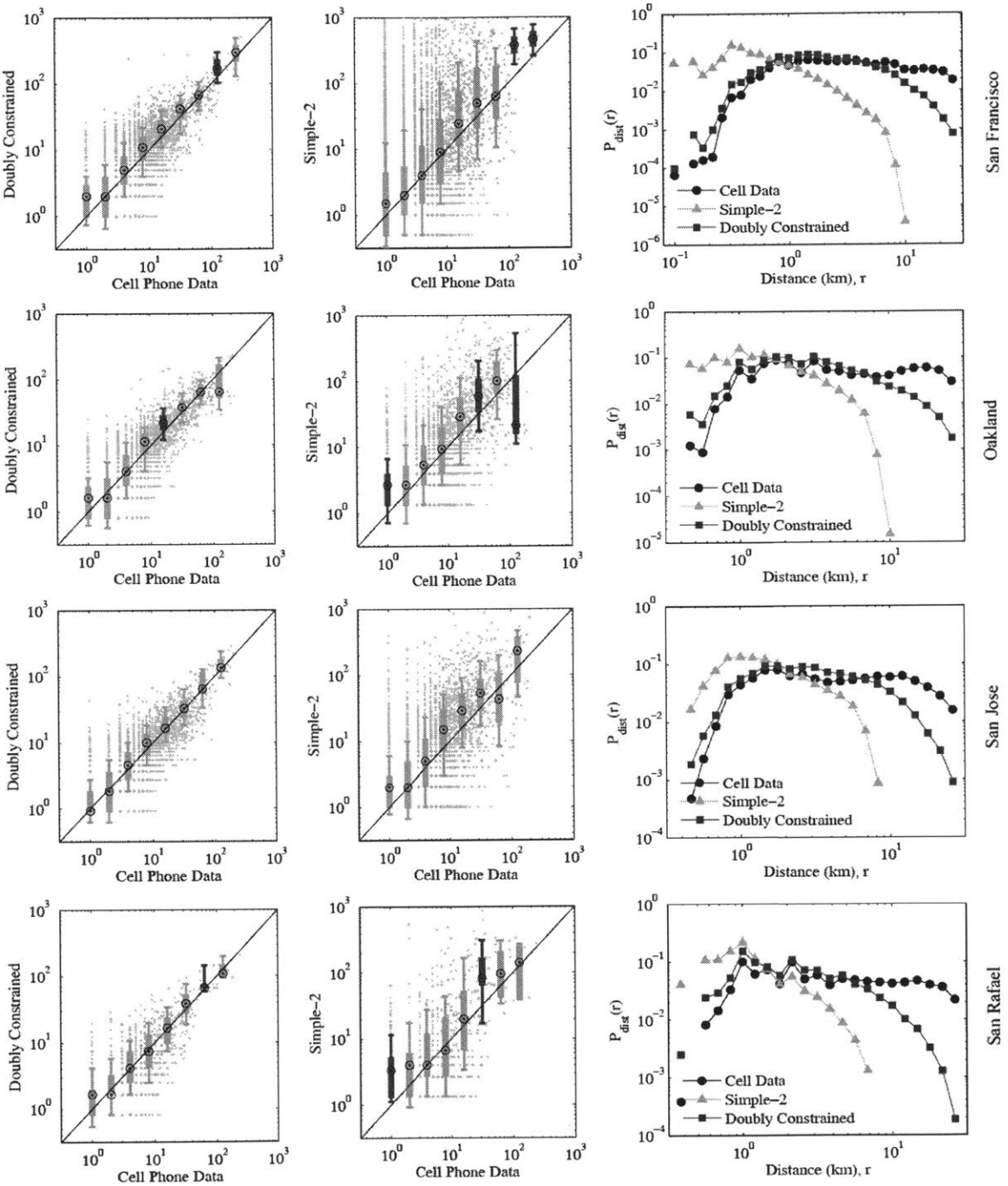


Figure A-2: Further comparison of the two gravity models in San Francisco, Oakland, San Jose and San Rafael

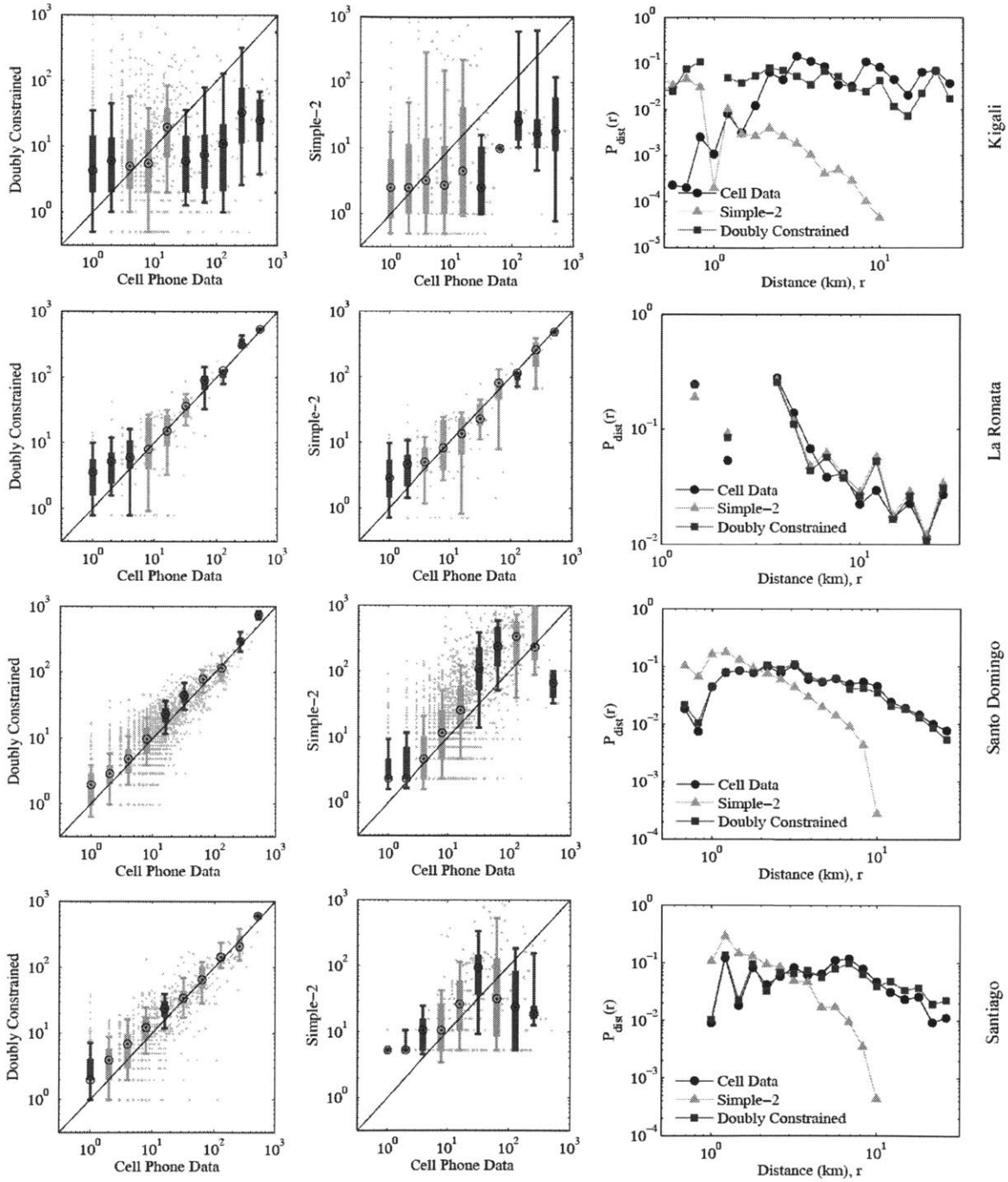


Figure A-3: Further comparison of the two gravity models in Kigali, La Romata, Santo Domingo and Santiago

Appendix B

Methods

B.1 K-means Clustering of Blocks

The 2010 Census LEHD Origin-Destination Employment Statistics (LODES) datasets contains home and work location counts at block level. San Francisco has 7348 blocks while in transportation planning a city is often divided into a much less number of regions (85; 79). To make the estimation results at different scales comparable, here we adopted k-means clustering (59; 20) to divide each study region into 100 locations. The blocks are clustered according to their geographical locations. The procedure is performed in the following way:

Randomly pick 100 (lon,lat) coordinate pairs in the study region to represent the centers of the clusters. They are denoted as $_k, k = 1, , 100$. Each block's center location is denoted as a vector $X_i, i = 1, , numberofblocks$. The goal is to find an assignment of X_i to clusters, as well as a set of vectors μ_k, such that the sum of the squares of the distances of each data point X_i to its closest vector μ_k, is a minimum. Use 1-of-K coding scheme to represent which cluster each data point X_i should belong to. For each data point X_i, we introduce a corresponding set of binary indicator variables $r_{ik} \in 0, 1, k = 1, , 100$, describing which of the 100 clusters the data point X_i is assigned to. If data point X_i is assigned to cluster k then $r_{ik} = 1$, and $r_{ij=0}$ for $j \neq k$. The objective function, J, is to minimize the sum of the squares of

the distances of each data point to its assigned vector μ_k:

$$J = \sum_{i=1}^{N} \sum_{k=1} 100 r_{ik} \left\| X_i - \mu_k \right\|^2 \tag{B.1}$$

Here the distance measure $\left\| X_i - \mu_k \right\|^2$ is the distance of the two coordinate pairs on earth. To find the values for r_{ik} and μ_k, iteratively perform:

1. Keep μ_k fixed, find the r_{ik} values to minimize J. This is simply to find the closest μ_k to each data point X_i.

2. Keep r_{ik} fixed, find the μ_k values to minimize J. J is a quadratic function of μ_k. Take the derivative and with respect to μ_k and set it to zero shows that:

$$\mu_k = \frac{\sum_i r_{ik} X_i}{\sum_i r_{ik}} \tag{B.2}$$

Iteratively perform these two steps until converge. Fig. B-1 and B-2 shows the comparison of San Francisco's blocks before and after clustering.

Figure B-1: The 7348 blocks of San Francisco.

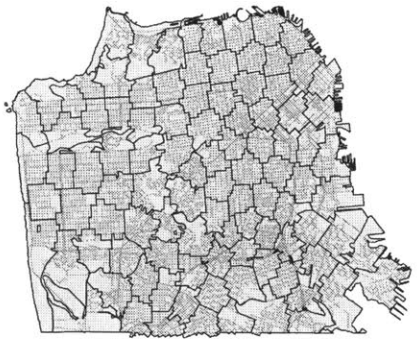


Figure B-2: 100 block clusters acquired from k-means clustering

B.2 IPF Procedure for OD Expansion

We use the Bay area as an example to show that cell phone data could provide a good commuting OD seed matrix. We have deduced home and work location for each user. Here a location is a cell phone tower. There are 892 towers in the Bay area while in previous methods we divided San Francisco into 100 locations. In order to match these two different types of divisions we mapped the 892 cell phone towers to the previously defined 100 block clusters to form the 100100 commuting OD matrix for the cell phone users. We should notice that the cell phone users we chose are like a sample from the whole population and the sampling rates in different block clusters may differ from one another. In order to get the commuting OD matrix for the whole population from the cell phone user commuting OD matrix, we need to reweight or perform seed matrix expansion on the cell phone user commuting OD matrix. The iterative proportional fitting method is adopted (45).

Iterative proportional fitting is a procedure for adjusting a table of data cells such that they add up to selected totals. Unadjusted data cells may be referred to as "seed", and the selected totals may be referred to as "marginal". In our two dimensional case, the "seed" is the cell phone user commuting OD matrix denoted as t_{ij}, i is the home location while j is the work location. We've shown that population and POI are good representations of trip generation and attraction. We use them to represent the "marginals". The column marginal D_i is the trip attraction of each location and the row marginal O_i is the trip generation of each location. D_i are normalized to have the same sum as O_i. The numerical solution is:

1. $\hat{T}_{ij}^{m} = t_{ij}, m = 0$

2. a1) For $i = 1, , N$

i. Solve for $\alpha : \sum_j \hat{T}_{ij}^{m} \alpha = O_i$

ii. $\hat{T}_{ij}^{m+1/2} = \hat{T}_{ij}^{m} \alpha$

a2) $m = m + 1/2$

b1) For $j = 1, , N$

i. Solve for $\alpha : \sum_i \hat{T}_{ij}^{m} \alpha = D_j$

ii. $\hat{T}_{ij}^{m+1/2} = \hat{T}_{ij}^{m}\alpha$

b 2) $m = m + 1/2$

3. Repeat step 2 until converge.

Some may doubt that the close fit of the expanded Bay Area cell phone user seed matrix to the actual census data is because we used quite accurate marginal (in this case the population density and the density of POIs), so that the seed matrix do not have much influence. We test this assumption by doing the following comparison: compare the travelling distance $P(r)$ distribution of: 1) the census commuting OD data; 2) the cell phone user seed OD matrix without IPF expansion; 3) the IPF expanded cell phone user seed matrix; 4) the IPF expanded random seed matrix. The result is shown in Fig. B-3. Among all others, only the IPF expended cell phone user seed matrix gives close fit to the census data. As for the IPF expanded random seed matrix, even though it has accurate marginal, it still deviates from the actual $P(r)$ distribution. In this way the value of both the IPF method and the cell phone user seed matrix are shown.

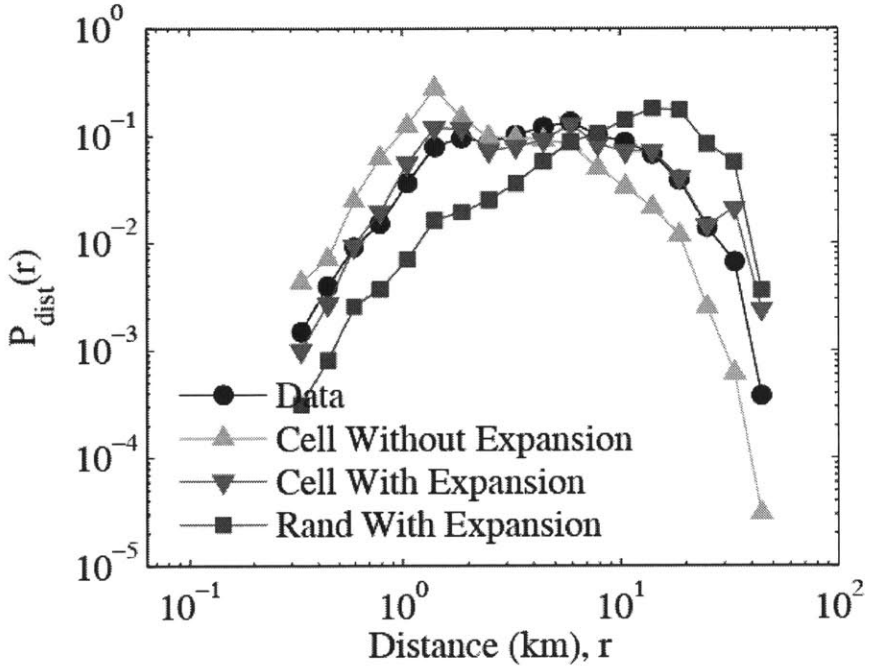


Figure B-3: Comparison of the travelling distance $P(r)$ distributions. The census commuting OD data is in black. The cell phone user seed OD matrix without IPF expansion is in green. The IPF expanded cell phone user seed matrix is in purple. The IPF expanded random seed matrix is in red. Only the IPF expended cell phone user seed matrix gives close fit to the census data. As for the IPF expanded random seed matrix, even though it has accurate marginal, it still deviates from the actual $P(r)$ distribution.

Bibliography

[1] Mustafa Abdulaal and Larry J LeBlanc. Methods for combining modal split and equilibrium assignment models. *Transportation Science*, 13(4):292–314, 1979.

[2] Mark Abkowitz. Transit service reliability. Technical report, Urban Mass Transportation Administration, 1978.

[3] Mark Abkowitz and John Tozzi. Research contributions to managing transit service reliability. *Journal of advanced transportation*, 21(1):47–65, 1987.

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

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

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

[7] Y Arezki and D Van Vliet. A full analytical implementation of the partan/frank–wolfe algorithm for equilibrium assignment. *Transportation Science*, 24(1):58–62, 1990.

[8] J.P. Bagrow, D. Wang, and A.L. Barabási. Collective response of human populations to large-scale emergencies. *PloS one*, 6(3):e17680, 2011.

[9] D. Balcan, V. Colizza, B. Gonçalves, H. Hu, J.J. Ramasco, and A. Vespignani. Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences*, 106(51):21484–21489, 2009.

[10] James J Barry, Robert Freimer, and Howard Slavin. Use of entry-only automatic fare collection data to estimate linked transit trips in new york city. *Transportation Research Record: Journal of the Transportation Research Board*, 2112(1):53–61, 2009.

[11] James J Barry, Robert Newhouser, Adam Rahbee, and Shermeen Sayeda. Origin and destination estimation in new york city with automated fare system data. *Transportation Research Record: Journal of the Transportation Research Board*, 1817(1):183–187, 2002.

[12] Marc Barthélemy. Spatial networks. *Physics Reports*, 499(1):1–101, 2011.

[13] AG Barto and RH Crites. Improving elevator performance using reinforcement learning. *Advances in neural information processing systems*, 8:1017–1023, 1996.

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

[15] G. Bellei and K. Gkoumas. Transit vehicles headway distribution and service irregularity. *Public Transport*, 2(4):269–289, 2010.

[16] Moshe Ben-Akiva and Bruno Boccara. Discrete choice models with latent choice sets. *International Journal of Research in Marketing*, 12(1):9–24, 1995.

[17] Moshe Ben-Akiva, Daniel McFadden, Kenneth Train, Joan Walker, Chandra Bhat, Michel Bierlaire, Denis Bolduc, Axel Börsch-Supan, David Brownstone, David S Bunch, et al. Hybrid choice models: Progress and challenges. *Marketing Letters*, 13(3):163–175, 2002.

[18] B. Bhaduri, E. Bright, P. Coleman, and M.L. Urban. Landscan usa: a high-resolution geospatial and temporal modeling approach for population distribution and dynamics. *GeoJournal*, 69(1):103–117, 2007.

[19] 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):1–28, 2001.

[20] Paul S Bradley and Usama M Fayyad. Refining initial points for k-means clustering. In *Proceedings of the Fifteenth International Conference on Machine Learning*, volume 66. San Francisco, CA, USA, 1998.

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

[22] A Brown, RF Casey, M Foy, B Kaplan, LN Labell, B Marks, L Moniz, D Parker, JW Royal, CL Schweiger, et al. Advanced public transportation systems: the state of the art update 2000, federal transit administration, 2000, usdot pub no. Technical report, FTA-MA-26-7007-00-1. URL http://www. itsdocs. fhwa. dot. gov/3583. pdf, 1994.

[23] US Census Bureau. https://explore.data.gov/labor-force-employment-and-earnings/lehd-origin-destination-employment-statistics-lode/zvvq-y3uj/. `https://explore.data.gov/Labor-Force-Employment-and-Earnings/LEHD-Origin-Destination-Employment-Statistics-LODE/zvvq-y3uj/`, 2012. [Online; accessed 26-Feb-2013].

[24] Young-Ji Byon, Cristián Eduardo Cortés, C Martinez, Francisco Javier, Marcela Munizaga, and Mauricio Zuniga. Transit performance monitoring and analysis with massive gps bus probes of transantiago in santiago, chile: Emphasis on

development of indices for bunching and schedule adherence. In *Transportation Research Board 90th Annual Meeting*, 2011.

[25] Roberto Camus, Giovanni Longo, and Cristina Macorini. Estimation of transit reliability level-of-service based on automatic vehicle location data. *Transportation Research Record: Journal of the Transportation Research Board*, 1927(1):277–286, 2005.

[26] 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.

[27] Robert F Casey. Advanced public transportation systems deployment in the united states: update, january 1999. Technical report, 1999.

[28] Claudio Castellano, Santo Fortunato, and Vittorio Loreto. Statistical physics of social dynamics. *Reviews of modern physics*, 81(2):591, 2009.

[29] L.C. Cham. Understanding bus service reliability: a practical framework using avl/apc data. In *MS Thesis, Prof. Nigel Wilson, Massachusetts Institute of Technology*, 2006.

[30] Joanne Chan et al. *Rail transit OD matrix estimation and journey time reliability metrics using automated fare data.* PhD thesis, Massachusetts Institute of Technology, 2007.

[31] David Charypar and Kai Nagel. Q-learning for flexible learning of daily activity plans. *Transportation Research Record: Journal of the Transportation Research Board*, 1935(1):163–169, 2005.

[32] Xumei Chen, Lei Yu, Yushi Zhang, and Jifu Guo. Analyzing urban bus service reliability at the stop, route, and network levels. *Transportation research part A: policy and practice*, 43(8):722–734, 2009.

[33] J.M. Choukroun. A general framework for the development of gravity-type trip distribution models. *Regional Science and Urban Economics*, 5(2):177–202, 1975.

[34] Pau-Choo Chung and Chin-De Liu. A daily behavior enabled hidden markov model for human behavior understanding. *Pattern Recognition*, 41(5):1572–1580, 2008.

[35] C.E. Cortés, J. Gibson, A. Gschwender, M. Munizaga, and M. Zúñiga. Commercial bus speed diagnosis based on gps-monitored data. *Transportation Research Part C: Emerging Technologies*, 2011.

[36] C.F. Daganzo. A headway-based approach to eliminate bus bunching: Systematic analysis and comparisons. *Transportation Research Part B*, 43:913–921, 2009.

[37] Robert Barkley Dial. A probabilistic multipath traffic assignment model which obviates path enumeration. *Transportation Research/UK/*, 5, 1971.

[38] Thi V Duong, Hung H Bui, Dinh Q Phung, and Svetha Venkatesh. Activity recognition and abnormality detection with the switching hidden semi-markov model. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 838–845. IEEE, 2005.

[39] Nathan Eagle, Alex Sandy Pentland, and David Lazer. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*, 106(36):15274–15278, 2009.

[40] Xu Jun Eberlein, Nigel HM Wilson, and David Bernstein. The holding problem with real–time information available. *Transportation science*, 35(1):1–18, 2001.

[41] Rosalind M Eggo, Simon Cauchemez, and Neil M Ferguson. Spatial dynamics of the 1918 influenza pandemic in england, wales and the united states. *Journal of The Royal Society Interface*, 8(55):233–243, 2011.

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

[43] Cheng-Min Feng and Cheng-Hsien Hsieh. Effect of resource allocation policies on urban transport diversity. *Computer-Aided Civil and Infrastructure Engineering*, 24(7):525–533, 2009.

[44] S.E. Fienberg. An iterative procedure for estimation in contingency tables. *The Annals of Mathematical Statistics*, pages 907–917, 1970.

[45] Stephen E Fienberg and Michael M Meyer. Iterative proportional fitting. Technical report, DTIC Document, 1981.

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

[47] Michael Florian, Marc Gaudry, and Christian Lardinois. A two-dimensional framework for the understanding of transportation planning models. *Transportation Research Part B: Methodological*, 22(6):411–419, 1988.

[48] Robin Flowerdew and Murray Aitkin. A method of fitting the gravity model based on the poission distribution. *Journal of Regional Science*, 22(2):191–202, 1982.

[49] Foursquare. http://www.foursquare.com. `http://www.foursquare.com`, 2012. [Online; accessed 26-Feb-2013].

[50] Peter G Furth. *Data analysis for bus planning and monitoring*, volume 34. Transportation Research Board, 2000.

[51] Peter Gregory Furth et al. *Using archived AVL-APC data to improve transit performance and management.* Transportation Research Board, 2006.

[52] Peter Gregory Furth, Brendon J Hemily, Theo HJ Muller, and James G Strathman. *Uses of archived AVL-APC data to improve transit performance and management: Review and potential.* Transportation Research Board Washington, DC, 2003.

[53] PG Furth, B. Hemily, THJ Muller, and JG Strathman. Tcrp report 113: Using archived avl-apc data to improve transit performance and management. *Transportation Research Board of the National Academies, Washington, DC*, 2006.

[54] G. Gentile, S. Nguyen, and S. Pallottino. Route choice on transit networks with on-line information at stops. *Transportation Science, Technical Report TR-03-14, Universitadi Pisa, Dipartimento di Informatica, Pisa, Italy*, 2003.

[55] Reginald G Golledge, Mei-Po Kwan, and Tommy Gärling. Computational process modeling of household travel decisions using a geographical information system. *Papers in regional science*, 73(2):99–117, 1994.

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

[57] Marta C González, Pedro G Lind, and Hans J Herrmann. System of mobile agents to model social networks. *Physical review letters*, 96(8):088702, 2006.

[58] Michael Gould, Max Craglia, Michael F Goodchild, Alessandro Annoni, Gilberto Camara, Werner Kuhn, David Mark, Ian Masser, David Maguire, Steve Liang, et al. Next-generation digital earth: A position paper from the vespucci initiative for the advancement of geographic information science. *International Journal of Spatial Data Infraestructures Research (17250463)*, 3, 2008.

[59] John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Applied statistics*, pages 100–108, 1979.

[60] Mark D Hickman. An analytic stochastic model for the transit vehicle holding problem. *Transportation Science*, 35(3):215–237, 2001.

[61] EM Holroyd and DA Scraggs. *Waiting times for buses in central London.* Printerhall, 1966.

[62] Yanqing Hu, Yougui Wang, Daqing Li, Shlomo Havlin, and Zengru Di. Maximizing entropy yields spatial scaling in social networks. *arXiv preprint arXiv:1002.1802*, 2010.

[63] Jean-Paul Hubert and Philippe L Toint. From average travel time budgets to daily travel time distributions: appraisal of two conjectures by koelbl and helbing and some consequences. *Transportation Research Record: Journal of the Transportation Research Board*, 1985(1):135–143, 2006.

[64] Humnetlab. http://humnetlab.mit.edu/extendedradiation/, 2013. [Online; accessed 17-Apr-2013].

[65] Mark L Huson and Arunabha Sen. Broadcast scheduling algorithms for radio networks. In *Military Communications Conference, 1995. MILCOM'95, Conference Record, IEEE*, volume 2, pages 647–651. IEEE, 1995.

[66] Davy Janssens, Yu Lan, Geert Wets, and Guoqing Chen. Allocating time and location information to activity–travel patterns through reinforcement learning. *Knowledge-Based Systems*, 20(5):466–477, 2007.

[67] Leslie Pack Kaelbling, Michael L. Littman, and Andrew W Moore. Reinforcement learning: A survey. *arXiv preprint cs/9605103*, 1996.

[68] Eunju Kim, Sumi Helal, and Diane Cook. Human activity recognition and pattern discovery. *Pervasive Computing, IEEE*, 9(1):48–53, 2010.

[69] Minkyong Kim, David Kotz, and Songkuk Kim. Extracting a mobility model from real user traces. In *Proc. IEEE Infocom*, volume 6, pages 1–13, 2006.

[70] T Kimpel, James G Strathman, and Steve Callas. Improving scheduling through performance monitoring using avl/apc data. *Submitted to University of Wisconsin-Milwaukee as a Local Innovations in Transit project report under the Great Cities University Consortium*, 2004.

[71] T.J. Kimpel, J.G. Strathman, and S. Callas. Improving scheduling through performance monitoring. *Computer-aided Systems in Public Transport*, pages 253–280, 2008.

[72] Kittelson Asssociates, Inc., KFH Group, Inc., Parsons Brinckerhoff Quade, Douglass, Inc., and K. Hunter-Zaworski. Tcrp report 100: Transit capacity and quality of service manual. Technical report, Transportation Research Board, 2003.

[73] FS Koppelman and EI 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.

[74] William HK Lam and Hai-Jun Huang. A combined trip distribution and assignment model for multiple user classes. *Transportation Research Part B: Methodological*, 26(4):275–287, 1992.

[75] D.H. Lee, L. Sun, and A. Erath. Study of bus service reliability in singapore using fare card data. In *The 12th Asia Pacific ITS Forum*, 2012.

[76] Lin Liao, Dieter Fox, and Henry Kautz. Extracting places and activities from gps traces using hierarchical conditional random fields. *The International Journal of Robotics Research*, 26(1):119–134, 2007.

[77] Lin Liao, Donald J Patterson, Dieter Fox, and Henry Kautz. Learning and inferring transportation routines. *Artificial Intelligence*, 171(5):311–331, 2007.

[78] X. Lu, L. Bengtsson, and P. Holme. Predictability of population displacement after the 2010 haiti earthquake. *Proceedings of the National Academy of Sciences*, 109(29):11576–11581, 2012.

[79] Thomas L Magnanti and Richard T Wong. Network design and transportation planning: Models and algorithms. *Transportation Science*, 18(1):1–55, 1984.

[80] Sridhar Mahadevan and Jonathan Connell. Automatic programming of behavior-based robots using reinforcement learning. *Artificial intelligence*, 55(2):311–365, 1992.

[81] Michael Mandelzys and Bruce Hellinga. Identifying causes of performance issues in bus schedule adherence with automatic vehicle location and passenger count data. *Transportation Research Record: Journal of the Transportation Research Board*, 2143(1):9–15, 2010.

[82] Marvin L Manheim. *Fundamentals of Transportation systems analysis; Volume 1: Basic concepts.* 1979.

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

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

[85] Michael D Meyer and Eric J Miller. *Urban transportation planning: A decision-oriented approach.* 2001.

[86] Robert Buchanan Mitchell and Chester Rapkin. Urban traffic–a function of land use. 1954.

[87] Sang Nguyen. An algorithm for the traffic assignment problem. *Transportation Science*, 8(3):203–216, 1974.

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

[89] J. Ortuzar and L.G. Willumsen. *Modelling transport.* Wiley, 1994.

[90] C. Pangilinan, N. Wilson, and A. Moore. Bus supervision deployment strategies and use of real-time automatic vehicle location for improved bus service reliability. *Transportation Research Record: Journal of the Transportation Research Board*, 2063(-1):28–33, 2008.

[91] Romualdo Pastor-Satorras, Alexei Vázquez, and Alessandro Vespignani. Dynamical and correlation properties of the internet. *Physical review letters*, 87(25):258701, 2001.

[92] Romualdo Pastor-Satorras and Alessandro Vespignani. *Evolution and structure of the Internet: A statistical physics approach.* Cambridge University Press, 2007.

[93] J.L. Pline. *Traffic engineering handbook.* Prentice Hall, 1992.

[94] M.A. Quddus, W.Y. Ochieng, and R.B. Noland. Current map-matching algorithms for transport applications: State-of-the art and future research directions. *Transportation Research Part C: Emerging Technologies*, 15(5):312–328, 2007.

[95] José J Ramasco and Alessandro Vespignani. Commuting and pandemic prediction. *PNAS*, 106(51):21459–21460, 2009.

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

[97] F. Simini, M.C. González, A. Maritan, and A.L. Barabási. A universal model for mobility and migration patterns. *Nature*, 484(7392):96–100, 2012.

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

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

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

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

[102] S.A. Stouffer. Intervening opportunities and competing migrants. *Journal of Regional Science*, 2(1):1–26, 1960.

[103] Daniel Sui, Sarah Elwood, and Michael Goodchild. *Crowdsourcing geographic knowledge: volunteered geographic information (VGI) in theory and practice.* Springer Publishing Company, Incorporated, 2012.

[104] WY Szeto, Muthu Solayappan, and Yu Jiang. Reliability-based transit assignment for congested stochastic transit networks. *Computer-Aided Civil and Infrastructure Engineering*, 26(4):311–326, 2011.

[105] Gerald Tesauro. Td-gammon, a self-teaching backgammon program, achieves master-level play. *Neural computation*, 6(2):215–219, 1994.

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

[107] Arne Treiber and Martin Kesting. *Verkehrsdynamik und -simulation*. Springer, 2010.

[108] Martin Trépanier, Nicolas Tranchant, and Robert Chapleau. Individual trip destination estimation in a transit smart card automated fare collection system. *Journal of Intelligent Transportation Systems*, 11(1):1–14, 2007.

[109] L Vanajakshi, SC Subramanian, and R Sivanandan. Travel time prediction under heterogeneous traffic conditions using global positioning system data from buses. *IET intelligent transport systems*, 3(1):1–9, 2009.

[110] Marlies Vanhulsel, Davy Janssens, and Geert Wets. Calibrating a new reinforcement learning mechanism for modeling dynamic activity-travel behavior and key events. 2007.

[111] Cécile Viboud, Ottar N BjØrnstad, David L Smith, Lone Simonsen, Mark A Miller, and Bryan T Grenfell. Synchrony, waves, and spatial hierarchies in the spread of influenza. *science*, 312(5772):447–451, 2006.

[112] Dashun Wang, Dino Pedreschi, Chaoming Song, Fosca Giannotti, and Albert-Laszlo Barabasi. Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1100–1108. ACM, 2011.

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

[114] Edward Weiner. Urban transportation planning in the united states: an historical overview (revised edition). Technical report, Department of Transportation, Washington, DC (USA). Office of the Assistant Secretary for Policy and International Affairs, 1986.

[115] 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.

[116] 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.

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

[118] A.G. Wilson. Land-use/transport interaction models: Past and future. *Journal of Transport Economics and Policy*, pages 3–26, 1998.

[119] A.G. Wilson and ML Senior. Some relationships between entropy maximizing models, mathematical programming models, and their duals*. *Journal of Regional Science*, 14(2):207–215, 1974.

[120] D. Wu, T. Zhu, W. Lv, and X. Gao. A heuristic map-matching algorithm by using vector-based recognition. In *Computing in the Global Information Technology, 2007. ICCGI 2007. International Multi-Conference on*, pages 18–18. IEEE, 2007.

[121] Xiangye Xiao, Yu Zheng, Qiong Luo, and Xing Xie. Inferring social ties between users with human location history. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–17, 2012.

[122] JS Yang, SP Kang, and KS Chon. The map matching algorithm of gps data with relatively long polling time intervals. *Journal of the Eastern Asia Society for Transportation Studies*, 6:2561–2573, 2005.

[123] Jinhua Zhao, Adam Rahbee, and Nigel HM Wilson. Estimating a rail passenger trip origin-destination matrix using automatic data collection systems. *Computer-Aided Civil and Infrastructure Engineering*, 22(5):376–387, 2007.

[124] Vincent W Zheng, Yu Zheng, Xing Xie, and Qiang Yang. Collaborative location and activity recommendations with gps history data. In *Proceedings of the 19th international conference on World wide web*, pages 1029–1038. ACM, 2010.

[125] Yu Zheng. *Computing with spatial trajectories.* Springer Science+ Business Media, 2011.

[126] Yu Zheng, Yukun Chen, Xing Xie, and Wei-Ying Ma. Geolife2. 0: a location-based social networking service. In *Mobile Data Management: Systems, Services and Middleware, 2009. MDM'09. Tenth International Conference on*, pages 357–358. IEEE, 2009.

[127] Yu Zheng, Like Liu, Longhao Wang, and Xing Xie. Learning transportation mode from raw gps data for geographic applications on the web. In *Proceedings of the 17th international conference on World Wide Web*, pages 247–256. ACM, 2008.

[128] G.K. Zipf. The p 1 p 2/d hypothesis: on the intercity movement of persons. *American sociological review*, 11(6):677–686, 1946.