University Transportation Research Center - Region 2

Final Report

The Use of Large-scale Datasets for Understanding Network State

Performing Organization: The City College of New York/CUNY

September 2013

Sponsor:
University Transportation Research Center - Region 2

University Transportation Research Center - Region 2

The Region 2 University Transportation Research Center (UTRC) is one of ten original University Transportation Centers established in 1987 by the U.S. Congress. These Centers were established with the recognition that transportation plays a key role in the nation's economy and the quality of life of its citizens. University faculty members provide a critical link in resolving our national and regional transportation problems while training the professionals who address our transportation systems and their customers on a daily basis.

The UTRC was established in order to support research, education and the transfer of technology in the field of transportation. The theme of the Center is "Planning and Managing Regional Transportation Systems in a Changing World." Presently, under the direction of Dr. Camille Kamga, the UTRC represents USDOT Region II, including New York, New Jersey, Puerto Rico and the U.S. Virgin Islands. Functioning as a consortium of twelve major Universities throughout the region, UTRC is located at the CUNY Institute for Transportation Systems at The City College of New York, the lead institution of the consortium. The Center, through its consortium, an Agency-Industry Council and its Director and Staff, supports research, education, and technology transfer under its theme. UTRC's three main goals are:

Research

The research program objectives are (1) to develop a theme based transportation research program that is responsive to the needs of regional transportation organizations and stakeholders, and (2) to conduct that program in cooperation with the partners. The program includes both studies that are identified with research partners of projects targeted to the theme, and targeted, short-term projects. The program develops competitive proposals, which are evaluated to insure the mostresponsive UTRC team conducts the work. The research program is responsive to the UTRC theme: "Planning and Managing Regional Transportation Systems in a Changing World." The complex transportation system of transit and infrastructure, and the rapidly changing environment impacts the nation's largest city and metropolitan area. The New York/New Jersey Metropolitan has over 19 million people, 600,000 businesses and 9 million workers. The Region's intermodal and multimodal systems must serve all customers and stakeholders within the region and globally.Under the current grant, the new research projects and the ongoing research projects concentrate the program efforts on the categories of Transportation Systems Performance and Information Infrastructure to provide needed services to the New Jersey Department of Transportation, New York City Department of Transportation, New York Metropolitan Transportation Council , New York State Department of Transportation, and the New York State Energy and Research Development Authorityand others, all while enhancing the center's theme.

Education and Workforce Development

The modern professional must combine the technical skills of engineering and planning with knowledge of economics, environmental science, management, finance, and law as well as negotiation skills, psychology and sociology. And, she/he must be computer literate, wired to the web, and knowledgeable about advances in information technology. UTRC's education and training efforts provide a multidisciplinary program of course work and experiential learning to train students and provide advanced training or retraining of practitioners to plan and manage regional transportation systems. UTRC must meet the need to educate the undergraduate and graduate student with a foundation of transportation fundamentals that allows for solving complex problems in a world much more dynamic than even a decade ago. Simultaneously, the demand for continuing education is growing – either because of professional license requirements or because the workplace demands it – and provides the opportunity to combine State of Practice education with tailored ways of delivering content.

Technology Transfer

UTRC's Technology Transfer Program goes beyond what might be considered "traditional" technology transfer activities. Its main objectives are (1) to increase the awareness and level of information concerning transportation issues facing Region 2; (2) to improve the knowledge base and approach to problem solving of the region's transportation workforce, from those operating the systems to those at the most senior level of managing the system; and by doing so, to improve the overall professional capability of the transportation workforce; (3) to stimulate discussion and debate concerning the integration of new technologies into our culture, our work and our transportation systems; (4) to provide the more traditional but extremely important job of disseminating research and project reports, studies, analysis and use of tools to the education, research and practicing community both nationally and internationally; and (5) to provide unbiased information and testimony to decision-makers concerning regional transportation issues consistent with the UTRC theme.

Project No: 49111-21-22

Project Date: September 2013

Project Title: The Use of Large Scale Datasets for Understanding Traffic Network Stateto Prevent Speeding-Related Accidents

Principal Investigator:

Dr. Camille Kamga
Assistant Professor, Department of Civil Engineering
The City College of New York/CUNY
Email: ckamga@utrc2.org

Dr. Satish V. Ukkusuri
Associate Professor, Purdue University
Email: sukkusur@purdue.edu

Performing Organization: The City University of New York (CUNY)

Sponsor:

University Transportation Research Center - Region 2, A Regional University Transportation Center sponsored by the U.S. Department of Transportation's Research and Innovative Technology Administration

To request a hard copy of our final reports, please send us an email at utrc@utrc2.org

Mailing Address:

University Transportation Reserch Center
The City College of New York
Marshak Hall, Suite 910
160 Convent Avenue
New York, NY 10031
Tel: 212-650-8051
Fax: 212-650-8374
Web: www.utrc2.org

Board of Directors

The UTRC Board of Directors consists of one or two members from each Consortium school (each school receives two votes regardless of the number of representatives on the board). The Center Director is an ex-officio member of the Board and The Center management team serves as staff to the Board.

City University of New York
Dr. Hongmian Gong - Geography
Dr. Neville A. Parker - Civil Engineering

Clarkson University
Dr. Kerop D. Janoyan - Civil Engineering

Columbia University
Dr. Raimondo Betti - Civil Engineering
Dr. Elliott Sclar - Urban and Regional Planning

Cornell University
Dr. Huaizhu (Oliver) Gao - Civil Engineering
Dr. Mark A. Turnquist - Civil Engineering

Hofstra University
Dr. Jean-Paul Rodrigue - Global Studies and Geography

Manhattan College
Dr. Anirban De - Civil & Environmental Engineering
Dominic Esposito - Research Administration

New Jersey Institute of Technology
Dr. Steven Chien - Civil Engineering
Dr. Joyoung Lee - Civil & Environmental Engineering

New York Institute of Technology
Dr. Nada Marie Anid - Engineering & Computing Sciences
Dr. Marta Panero - Engineering & Computing Sciences

New York University
Dr. Mitchell L. Moss - Urban Policy and Planning
Dr. Rae Zimmerman - Planning and Public Administration

Polytechnic Institute of NYU
Dr. John C. Falcocchio - Civil Engineering
Dr. Kaan Ozbay - Civil Engineering

Rensselaer Polytechnic Institute
Dr. José Holguín-Veras - Civil Engineering
Dr. William "Al" Wallace - Systems Engineering

Rochester Institute of Technology
Dr. J. Scott Hawker - Software Engineering
Dr. James Winebrake -Science, Technology, & Society/Public Policy

Rowan University
Dr. Yusuf Mehta - Civil Engineering
Dr. Beena Sukumaran - Civil Engineering

Rutgers University
Dr. Robert Noland - Planning and Public Policy

State University of New York
Michael M. Fancher - Nanoscience
Dr. Catherine T. Lawson - City & Regional Planning
Dr. Adel W. Sadek - Transportation Systems Engineering
Dr. Shmuel Yahalom - Economics

Stevens Institute of Technology
Dr. Sophia Hassiotis - Civil Engineering
Dr. Thomas H. Wakeman III - Civil Engineering

Syracuse University
Dr. Riyad S. Aboutaha - Civil Engineering
Dr. O. Sam Salem - Construction Engineering and Management

The College of New Jersey
Dr. Thomas M. Brennan Jr. - Civil Engineering

University of Puerto Rico - Mayagüez
Dr. Ismael Pagán-Trinidad - Civil Engineering
Dr. Didier M. Valdés-Díaz - Civil Engineering

UTRC Consortium Universities

The following universities/colleges are members of the UTRC consortium.

City University of New York (CUNY)
Clarkson University (Clarkson)
Columbia University (Columbia)
Cornell University (Cornell)
Hofstra University (Hofstra)
Manhattan College
New Jersey Institute of Technology (NJIT)
New York Institute of Technology (NYIT)
New York University (NYU)
Polytechnic Institute of NYU (Poly)
Rensselaer Polytechnic Institute (RPI)
Rochester Institute of Technology (RIT)
Rowan University (Rowan)
Rutgers University (Rutgers)*
State University of New York (SUNY)
Stevens Institute of Technology (Stevens)
Syracuse University (SU)
The College of New Jersey (TCNJ)
University of Puerto Rico - Mayagüez (UPRM)

** Member under SAFETEA-LU Legislation*

UTRC Key Staff

Dr. Camille Kamga: *Director, UTRC*
Assistant Professor of Civil Engineering, CCNY

Dr. Robert E. Paaswell: *Director Emeritus of UTRC and Distinguished* Professor of Civil Engineering, The City College of New York

Herbert Levinson: *UTRC Icon Mentor, Transportation Consultant and Professor Emeritus of Transportation*

Dr. Ellen Thorson: *Senior Research Fellow, University Transportation Research Center*

Penny Eickemeyer: *Associate Director for Research, UTRC*

Dr. Alison Conway: *Associate Director for New Initiatives and Assistant Professor of Civil Engineering*

Nadia Aslam: *Assistant Director for Technology Transfer*

Dr. Anil Yazici: *Post-doc/ Senior Researcher*

Nathalie Martinez: *Research Associate/Budget Analyst*

Membership as of January 2014

Project Report:

The Use of Large-scale Datasets for Understanding Network State

PI: Satish V. Ukkusuri, Camille Kamga
Student: Xianyuan Zhan, Xinwu Qian

September 16, 2013

ACKNOWLEDGMENTS

This research presented in this paper was supported by RITA/USDOT project “The Use of Large Scale Datasets for Understanding Network State” for which the authors are grateful.

TECHNICAL REPORT STANDARD TITLE PAGE

1. Report No.	2.Government Accession No. --	3. Recipient's Catalog No. --

4. Title and Subtitle The Use of Large Scale Datasets for Understanding Traffic Network State	5. Report Date September, 2013 6. Performing Organization Code --
7. Author(s) Dr. Camille Kamga, the City College of New York/CUNY, Dr. Satish V. Ukkusuri, Purdue	8. Performing Organization Report No. --
9. Performing Organization Name and Address The City College of New York, CUNY 160 Covent Ave New York, NY 10031	10. Work Unit No. -- 11. Contract or Grant No. 49111-21-22
12. Sponsoring Agency Name and Address University Transportation Research Center CCNY, 910 Marshak 160 Convent Avenue New ork,NY 10031	13. Type of Report and Period Covered Final Report 14. Sponsoring Agency Code ----

15. Supplementary Notes

16. Abstract

The goal of this proposal is to develop novel modeling techniques to infer individual activity patterns from the large scale cell phone datasets and taxi data from NYC. As such this research offers a paradigm shift from traditional transportation modeling by using large scale, disaggregate data and provides an unique perspective to understand the complex interactions among human behavior, urban environments and traffic patterns.

Urban development shapes the transportation systems, it determines what kind of transportation system a city has, and what does it look like. As an important dynamic component in urban systems, activities of transportation systems in turn captures the dynamics of the entire urban systems and enhance of our knowledge about the complex urban systems. This will ultimately contribute to the improvement of level of service and policy making on transportation systems. Taxi as a transportation tool has its unique characteristics. It is capable of capturing urban movement patterns both spatially and temporally since they serve as real-time probes in the network. Moreover, we are able to examine the pulse of the city, the gap between supply and demand, real time road congestions and even more.
On the other hand, accurate estimation and prediction of urban link travel times are important for improving urban traffic operations and identifying key bottlenecks in the traffic network. They can also benefit users by providing accurate travel time information, thereby allowing better route choice in the network and minimizing overall trip travel time. However, to accurately assess link travel times, it is important to have good real-time information from either in-road sensors such as loop detectors, microwave sensors, or roadside cameras, or mobile sensors (e.g. floating cars) or Global Positioning System (GPS) devices (e.g. cell phones). In most of these cases, only limited information is available related to speed or location, hence, one has to develop appropriate methodologies to accurately estimate the performance metric of interest at the link, path or network level. Taxis equipped with GPS units provide a significant amount of data over days and months thereby providing a rich source of data for estimating network wide performance metrics. However, currently there are limited methodologies making use of this new source of data to estimate link or path travel times in the urban network. Within this context, this study proposes a new method for estimating hourly urban link travel times using large-scale taxicab data with partial information.

The taxicab data used in this research provides limited trip information, which only contains the origin and destination location coordinates, travel time and distance of a trip. However, the extensive amount of data records compensates for the incompleteness of the data and makes the link travel time estimation possible. A novel algorithm for estimating the link travel times is also proposed and tested in this research.

17. Key Words Bridge management, bridge inspection, deterioration, service life, NDT, cost effective maintenance, preventive maintenance	18. Distribution Statement ---

19. Security Classif (of this report) Unclassified	20. Security Classif. (of this page) Unclassified	21. No of Pages 34	22. Price

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1. INTRODUCTION

1.1 Background and motivation

In public transportation system of large cities, taxicab plays a vital role in transporting travelers across urban area. As an unique but important component in urban transportation system, taxis provide accessibility and flexibility with door-to-door service. In 2007, there were around 18,000 taxis in Hong Kong serving 10 percent of total passenger transportation volume (Yang et al., 2010). As for New York City, it has the largest market for taxis in North America with a total number of 55,000 (2012) of taxis (including “for-hire” and “for-hail” taxis) transport 1.5 million passengers every day (NYCTLC 2012 annual report). The taxi service transports 25% of all fare-paying bus, subway, taxi and for-hire vehicle passengers that are traveling within Manhattan (Schaller Consulting, 2006; King et al., 2012).

In big cities like New York City, GPS devices are installed in each medallion taxicab. The taxi locations coordinates together with trip information, such as trip distance, trip fare, number of passengers, etc. are collected and archived by taxi administration agencies like the New York City Taxi and Limousine Commission (NYTLC, which is responsible for all taxi related issues in New York City). As a result, taxicab data is emerged to become a new source of information that archives the “pulse of the city”. This data is particularly interesting for several reasons: 1) taxi services is an important transportation mode in urban transportation that contribute to a considerable proportion of total travel; 2) the taxis serve as mobile probe sensors that capture the mobility characteristics of all the taxi passengers; 3) the amount of data is huge (300,000 to 550,000 trips per day); 4) rich information available (contains trip starting and ending time and coordinates, trip time and distance, trip fare, number of passengers, payment method, etc.). Thus the taxi trip data becomes an ideal source of data to analyzing and characterizing the urban dynamics and estimating urban traffic conditions. In this study, we use the New York taxi trip data from NYTLC to explore and answer two important questions: 1) What are the urban dynamics pattern revealed by the large-scale taxi trip data? 2) Is it possible to utilize this novel dataset to estimate urban traffic network condition?

Urban development shapes the transportation systems, it determines what kind of transportation system a city has, and what does it look like. As an important dynamic component in urban systems, activities of transportation systems in turn captures the dynamics of the entire urban systems and enhance of our knowledge about the complex urban systems. This will ultimately contribute to the improvement of level of service and policy making on transportation systems. Taxi as a transportation tool has its unique characteristics. It is capable of capturing urban movement patterns both spatially and temporally since they serve as real-time probes in the network. Moreover, we are able to examine the pulse of the city, the gap between supply and demand, real time road congestions and even more.

On the other hand, accurate estimation and prediction of urban link travel times are important for improving urban traffic operations and identifying key bottlenecks in the traffic network. They can also benefit users by providing accurate travel time information, thereby allowing better route choice in the network and minimizing overall trip travel time. However, to accurately assess link travel times, it is important to have good real-time information from either in-road sensors such as loop detectors, microwave sensors, or roadside cameras, or mobile sensors (e.g. floating cars) or Global Positioning System (GPS) devices (e.g. cell phones). In most of these cases, only limited information is available related to speed or location, hence, one has to develop appropriate methodologies to accurately estimate the performance metric of interest at the link, path or network level. Taxis equipped with GPS units provide a significant amount of data over days and months thereby providing a rich source of data for estimating network wide performance metrics. However, currently there are limited methodologies making use of this new source of data to estimate link or path travel times in the urban network. Within this context, this study proposes a new method for estimating hourly urban link travel times using large-scale taxicab data with partial information. The taxicab data used in this research provides limited trip information, which only contains the origin and destination location coordinates, travel time and distance of a trip. However, the extensive amount of data records compensates for the incompleteness of the data and makes the link travel time estimation possible. A novel algorithm for estimating the link travel times is also proposed and tested in this research.

1.2 Study objectives

As mentioned in previous section, this research contains two major directions, the first is characterizing the urban dynamics using large-scale taxi data; the other is to explore novel methodology of utilizing the large amount of trip data to estimate urban link travel times.

For the characterization of urban dynamics, besides the general spatiotemporal pattern of taxi movements, we explore several unrecognized but important taxi trip related patterns. Based on the data collected from NYC taxicabs, the objective is to answer the following three questions: (1) what are the spatiotemporal patterns of urban taxi trips; (2) are there any similarities among different taxi trips based on their origin, destination, time of day etc. and (3) are there any universal patterns of urban mobility as related to taxi trips and their comparison with other mobility studies using other data sources.

For the urban link travel time estimation model, the goal is to show the potential of using taxicab data as a complimentary data source in urban transportation operation and management. Currently, large-scale of taxi data has been generated and recorded every day by taxi administration agencies, however, there is no model in literature that can utilize the large-scale partial trip information from this new source of data. The novel model presented in this research will provide an important and meaninful solution for urban link travel time estimation. Efficient algorithms are proposed to solve for the hourly average link travel times, which can be implemented in real time as a measure of urban network conditions. The estimations of urban link travel times can be further fused with the information from other existing data sources such as fixed sensors in the future to provide even better information for travelers.

1.3 Organization of the research

The remainder of the research is organized as follows. Chapter 2 provides a comprehensive review of the existing works related to urban dynamics study and link travel time estimation models. Chapter 3 describes the large-scale taxi data that is used in this research and corresponding data processing efforts. Chapter 4 characterizes the urban dynamic using the large-scale taxi data. Chapter 5 presents a novel urban link travel time estimation model using partial information. Finally, Chapter 6 summarizes findings of the research and discusses about some future research directions.

CHAPTER 2. RELATED WORK

This chapter describes recent works that study urban dynamics and urban link travel time estimation using large-scale GPS data. At the end, specific contributions of this research are summarized.

2.1 Literature review

2.1.1 **Urban dynamics**

The development of transportation system is affected by the spatial structure of a city (Muller, 2004). As city growth, transportation system shapes people's travel behaviors and land use types by osmosis (Muller, 2004). Hence, understanding a city from transportation system perspective as a bottom-up process is important to understand the functioning and evolution of urban areas. In the past few decades, efforts have been made in modeling and simulating urban dynamics using data from transportation systems (Harris, 1985; Batty, et al., 1994; Guiliano, 2004).

Several pioneering studies mainly focused on mobile phone data to reveal basic urban activity and individual mobility patterns (Ratti et al., 2006; Reades et al., 2007; González et al., 2008). Based on location-based service, individual locations are collected and mapped onto actual maps to reveal urban dynamics. A case study in Milan successfully discovered the urban spatial and temporal variations of activity intensity (Ratti et al., 2006). The intensity of activity locations is further used to locate hot spots and identify city structure by analyzing spatiotemporal signatures of Erlang data, which is a measure of network bandwidth usage usually collected at the antenna level (Reades et al., 2007). From individual perspective, González et al. revealed a highly regulated human mobility pattern (González et al., 2008) from 100,000 mobile phone users' trajectories, and Calabrese et al. established a multivariate regression model to predict daily human mobility (Calabrese et al., 2013). All these studies show a promising direction in studying urban dynamics using large-scale pervasive sensing data.

In public transportation system of large cities, taxicab plays a vital role to move people across urban area. Taxis provide accessibility and flexibility with door-to-door service. In 2007, there were around 18,000 taxis in Hong Kong serving 10 percent of total passenger transportation volume (Yang et al., 2010). As for New York City (NYC), by 2012, a total number of 55,000 of taxis (including “for-hire” and “for-hail” taxis) transport 1.5 million passengers every day (NYCTLC 2012 annual report). In the last few years, most of the taxis in NYC are equipped with GPS devices, therefore the taxis become an ideal source of large urban sensing data. Taxi location data has already been widely used in studying urban dynamics, such as hot spot analysis (Chang et al., 2008), land use inference (Pan et al., 2013), and urban human mobility recognition (Veloso et al., 2011, Li et al., 2012). The related works are illustrative, but not necessarily distinctive. As urban development shapes transportation system, retrieving urban dynamics from spatiotemporal patterns of transportation system will contribute to the improvement of level of service and policy making process at planning level. Taxi as a transportation tool has its unique characteristics. It is capable of capturing urban movement patterns both spatially and temporally since they are real-time probes in the network. Moreover, we are able to examine the pulse of the city, the gap between supply and demand, real time road congestions and even more.

2.1.2 **Urban link travel time estimation**

Urban link travel time estimation is another interesting yet meaningful problem in urban transportation operation and management. Previous research on urban link travel time estimation and prediction has largely relied on various data sources, including: loop detectors (Coifman, 2002; Zhang and Rice, 2003; Oh et al., 2003, Wu et al., 2004), automated vehicle identification (AVI) (Park and Rilett,1998; Li and Rose, 2011, Sherali et al., 2006), video camera, Remote Traffic Microwave Sensors (RTMS) (Yeon et al., 2008), and automated number plate recognition (Hasan et al., 2011). All of these data collection methods require installing corresponding sensors to retrieve data. Therefore a large number of sensors are required to achieve a reasonable accuracy level based on these data sources. The cost of installing and maintaining such a large number of sensors is prohibitive. Hence predicting link travel times with reasonable accuracy and network coverage based on sensor data could be expensive.

On the other hand, there is a significant potential to use emerging large-scale data sources to estimate dynamic demand and dynamic network conditions in urban areas. For instance, GPS devices in dedicated fleets of vehicles or in users' mobile phones can be viable sources of data for monitoring traffic in large cities (Herrera, et al. 2010). Industry models, such as Inrix1, have also gained popularity in recent years where private entities install, collect, utilize and sell "large-scale" historical traffic data from GPS-equipped vehicles or mobile phones. With an increasing amount of GPS data available from taxi, transit, and mobile phones, a new option of using such large-scale decentralized data for link travel time estimation becomes realistic. Herring et al. (2010) used GPS traces data from a fleet of 500 taxis in San Francisco, CA. to estimate and predict traffic conditions. However, in this work, instead of link travel times, discrete traffic states were predicted. Zheng and Zuylen (2012) also proposed an ANN model to estimate urban link travel times based on sparse probe vehicle data (e.g., GPS traces from GPS-equipped vehicles or smartphones). Hunter et al. (2009) proposed a statistical approach for path and travel time inference using GPS probe vehicle trajectory data. The GPS data used in their study has been recorded each minute, where the inferred path consists of at most five link segments. This method is not applicable if the GPS data has a longer recording interval or only has the starting and ending coordinates. Estimating link travel times from GPS data provides a much cheaper and a larger coverage area in the urban network compared with approaches using fixed sensor data. However, all of the above mentioned approaches are only applicable for GPS trace data, in which the trajectories of vehicles are available. To the best of our knowledge, there is no study found in literature that used OD level GPS data for urban link travel time estimation, even though extensive amount of such less detailed data (e.g. taxicab data) is generated and recorded every day. It is very important and meaningful to develop new model in urban link travel time estimation utilizing this novel source of large-scale data.

[1] Inrix, Inc. http://www.inrix.com

2.2 Research contribution

This research uses large-scale taxi trip data to understand urban dynamics and estimated urban link travel times. To summarize, the contribution of this research are as follows:

1. Characterizing urban dynamics
    - Taxi data has been utilized as a novel tool to understand urban dynamics
    - Unbalanced trips in Manhattan area are observed and explored
    - Airport trips is identified as a special part of taxi trips and differ from regular taxi trip patterns
    - Land use has significant impact on taxi trip types, and different types of taxi trips are able to uncover the structure of a city
    - Moreover, the mobility of taxi trips are restricted by the urban geographical boundaries.
2. Urban link travel time estimation
    - First model in literature to estimate urban link travel time using partial trip information.
    - Efficient solution algorithm is proposed to solve the optimization problem in a short time.
    - Hourly average link travel time can be obtained from the given model.
    - Showed potential of using large-data in traffic analytics.

CHAPTER 3. CHARACTERIZING URBAN DYNAMICS USING LARGE-SCALE TAXI DATA

This chapter explores the urban dynamics using the large-scale taxi trip data collected by New York City Taxi & Limousine Commission (NYCTLC). The data is combined with census tract and land use data to provide more information in the urban dynamics analysis. Extensive spatial and temporal patterns, taxi trip classification, and taxi mobility pattern have been analyzed.

3.1 *Data*

The data used in this chapter is compiled from the large-scale taxi trip dataset collected by NYCTLC, the census tract data from TRANSCAD map data and land use data from New York City Department of City Planning. Extensive data processing involved in this process. Following is the detailed description of the data.

3.1.1 **Taxi trip record data**

The taxi trip data used in this research is collected by NYCTLC from December, 2008 to January, 2010. About 300,000 to 500,000 trips are recorded every day during the observation period. The data does not contain taxi trajectories and the only geographical information is the longitude and latitude of trip origin and destination. Other trip information available includes starting and ending time, number of passengers, trip fare (with and without tax) and travel distance.

For the urban dynamics characterization, we select one-week period's data from September 7th, 2009 to September 13th, 2009. No major social events are reported during the time. The overall statistics of this part of taxi data is given in Table 1.

Table 1 Taxi Data Set Statistics

Date	Number of Trip Recorded	Number of Trips Filtered
9.7.2009	307,528	302,888
9.8.2009	421,549	415,513
9.9.2009	480,084	473,043
9.10.2009	521,209	513,536
9.11.2009	540,529	533,817
9.12.2009	510,875	504,337
9.13.2009	450,234	443,740

3.1.2 Census tract and land use data

In addition to taxi data, census tract information and land use type are also combined in the analysis. The census tract information is extracted from the census tract area file provided in TRANSCAD[1]. On the basis of spatial distribution of taxi trips, 2211 census tracts are selected to be the study area, which cover Manhattan, Bronx, Queens, Brooklyn, Long Island, and a small portion of New Jersey. The land use map comes from New York City Department of City Planning, which divides the city into four basic zoning district: park, residential, commercial, and manufacturing. The last three types are further categorized by density from low to high.

3.1.3 Data processing

To remove errors and inconsistency, the taxi data is processed before further analysis. Firstly, invalid values in taxi data are removed, such as 0 travel distance or trip fare under the initial price. After that, all data points are mapped onto census tract map and land use map using TRANSCAD. Points that are outside the boundary are eliminated, and each piece of trip record is tagged with the corresponding census tract id and land use type.

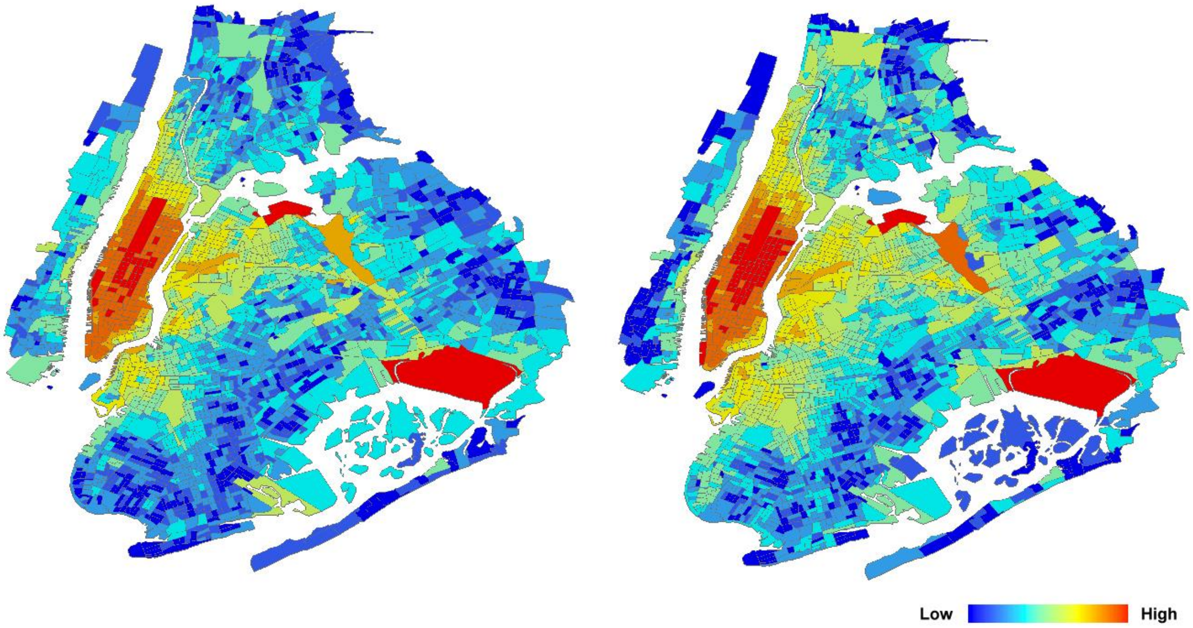
[1] TRANSCAD, a Transportation Planning GIS software by Caliper Corporation

3.2 Demand

In this section, we explore the taxi demand pattern from a geographical scope across NYC. Note that taxi trips usually fall into two categories: (1) taxi driver roams for potential passengers when empty and (2) taxi drivers have a “loaded” trip in which they are taking the passengers to the desired destination. It is reasonable to assume that taxi drivers are acquainted with the city and are profit maximizing. Therefore, they are more likely to get to places with more potential passengers when there is no passenger aboard. As for the second status, it is more representative of passenger’s choices as they will designate the destination. As a result, we view a taxi trip from two parts: trip origin and destination. Here trip origin is the place where taxi driver picks up passengers, and trip destination is for places where passengers are dropped off.

3.2.1 **Overall demand**

Around 2 million taxi trips transported 3 million passengers during the week. All taxi trips are mapped onto the census tract map by longitude and latitude of origins and destinations. Figure 1 presents an overall density plot on spatial distribution of taxi origins and destinations. It indicates that most origins and destinations are located inside Manhattan. Moreover, two very popular places lie outside Manhattan: LaGuardia Airport (LGA) and John F. Kennedy Airport (JFK). The overall demand pattern discovered is consistent with land use of NYC. Most business buildings and tourist sites are inside Manhattan, and transit locations like airport and train station always serve as big trip generators and attractors. Within Manhattan area, a majority of the trips are congregated at midtown, while lower part (main business and government center) is less preferred. Very few trips head towards upper Manhattan.



(a) Trip Origins (b) Trip Destinations

Figure 1 Aggregated Weekly Density Plot

Figure 2 provides information on overall trip structures. It is observed that 94.45% of trips origins and 92.57% of trip destinations are from Manhattan area and the two airports. With 91.98% trips origins inside Manhattan, the lower Manhattan contributes 21.22% and the midtown Manhattan dominates with 75.39%. For trip destinations, 88.52% are in Manhattan area, of which 18.23% reach lower Manhattan. Both are slightly fewer compared with trip origins. This turns out to be the fact that taxi drivers are more in favor of trips inside Manhattan, especially at midtown.

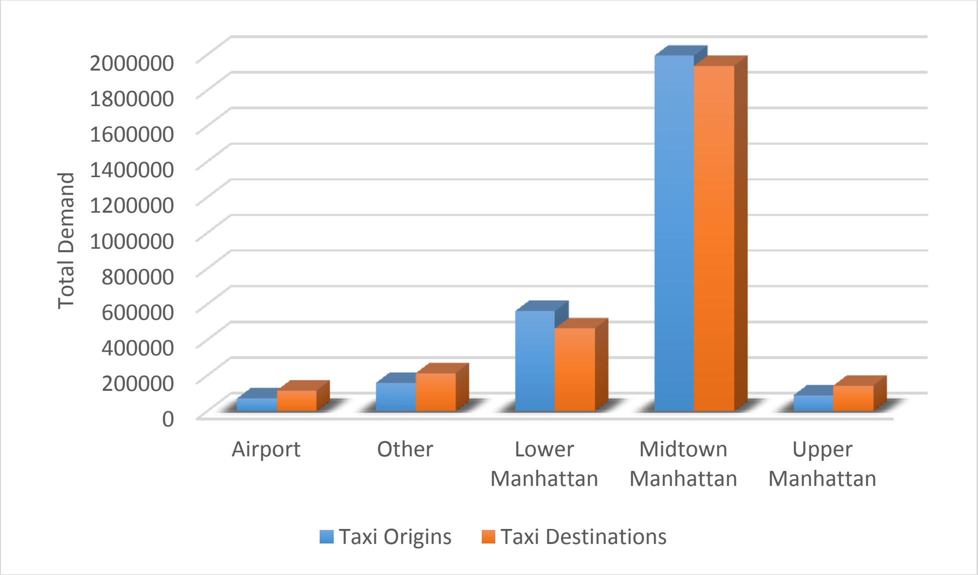


Figure 2 Weekly Manhattan Taxi Pattern

3.2.2 Hot spots analysis

Given the definition of hot spots as places with high activity intensity, here we focus on places most frequently visited across NYC. By aggregating trips of the entire week, five most popular places are identified: LGA, JFK international Airport, Penn Station, Central Park and the Fifth Avenue (between 49th street and 56th street). These places cover a wide range of land use functionality including major transit places (with different purposes), tourist site and commercial area. Figure 3 presents the temporal demand pattern of the 5 hot spots from Monday to Sunday.

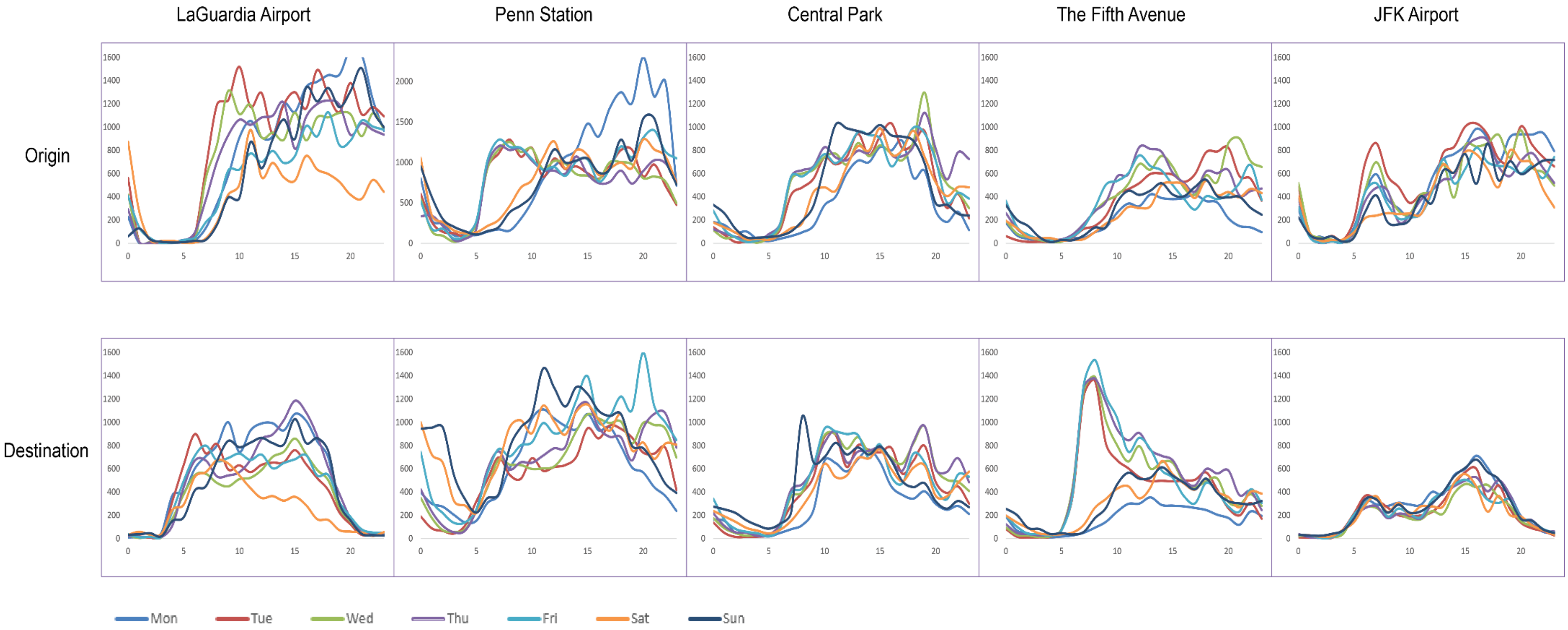


Figure 3 Weekly Hot Spots Trip Distribution (Monday to Sunday) for Taxi Origins and Destinations

Of all the 5 hot spots, Penn Station serves as the biggest taxi trip generator and attractor. Apparent origin morning peak is observed during workdays, suggesting that taxi might be quite popular as a transfer tool. It is hardly surprising that a large amount of people leave Manhattan on Friday night. However, the destination peak on Monday night indicates an unusual pattern that most people are returning to Manhattan at that time.

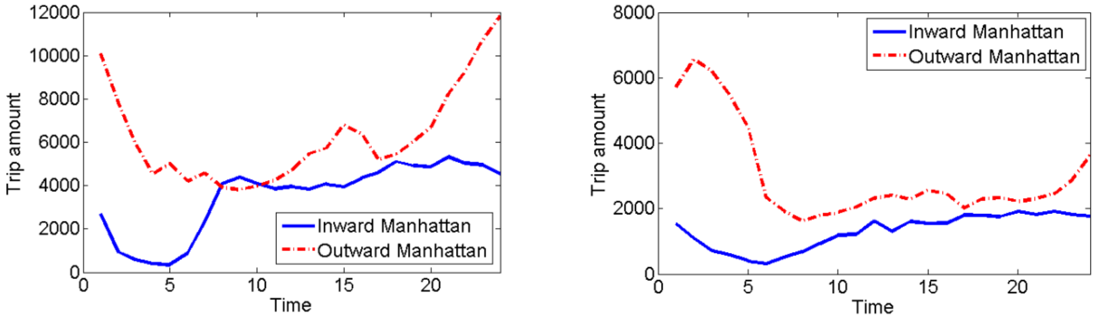
Several patterns at the two airports are also very interesting. First, fluctuation is observed for origin curves at both airports, which is closely related to the arriving pattern of flights. The amount of trip origins remain at a high level till midnight and early in the morning. One explanation could be that people have less accessibility to public transit at the time thus giving rise to the taxi demand. Moreover, the amount of taxi trips to LGA is much larger than that of JFK. The major reason is that JFK is further from downtown area and the fixed trip fare is comparatively higher. This pattern can be further studied as a potential indication of trip elasticity concerning cost.

The Central Park is a recreation and tourist site, and the demand pattern is observed to be stable. It serves as a hot spots not only for its population, but also because of the census tract covers a large area. With only one morning peak taking place on Sunday morning, most trips happens during daytime and evening peaks happen every day around 7 pm. The Fifth Avenue is a remarkable business street at midtown Manhattan. During weekdays, an apparent morning peak happens around 9 am.and most departures are concentrated at noon and after 8 pm.

3.2.3 **Unbalanced taxi trips**

For taxi trips, both passengers and drivers have their preferences of destinations, and such preferences are varying from time. This will eventually lead to an unbalanced spatial distribution of taxi resources. There are mainly two types of unbalanced trip. One is of geographical discrimination, as some destinations may be against the willingness of taxi drivers, as they may have very few potential customers at the destination. The second type is of resources shortage, which usually comes with prominent temporal characteristics. Assuming a fixed supply of taxis, it is usually hard to hail a taxi during peak hours. Both types of unbalanced taxi trips are observed in NYC taxi data.

From the overall trip pattern we found that most trip origins and destinations are located inside Manhattan, and we start by looking into patterns between trips inward and outward Manhattan area. Figure 4 gives the plot of the differences between trip origins and destinations. During daytime, the overall pattern turns out to be stable with outbound traffic slightly higher than those inbound. People may stay at Manhattan very late for entertainment and relaxation, while most public transits are out of service at that time. As a result, the huge gap observed at midnight is not surprising. The hidden fact behind the enormous gap is that taxi drivers are less likely to find potential customers outside Manhattan and return empty. Considering the large demand of taxis at that time, the first type of unbalanced trip will probably take place.



(a) Weekday (b) Weekend

Figure 4 Inward/Outward-Manhattan Unbalanced Trips

In order to reveal the unbalanced condition of inbound Manhattan trips, we extract only weekday trips and plot the distribution of trip origins and destinations in Figure 5. As discussed before, the most significant taxi shortage should occur during rush hours. Hence, the data from the morning peak and evening peak is selected. Off-peak hours are also plotted for comparison purpose. Trips are found to be unbalanced with notable geographic characteristics. Firstly, an unbalanced trip pattern is observed outside Manhattan and at upper part of Manhattan which is consistent with overall pattern. Moreover, compared with the balanced status during off-peak hours, both morning peak and evening peak display an eminent difference between trip origins and destinations. During morning rush hours, most taxi trips are merging into the center of midtown area, and is reversed during evening peak. Northeastern part of midtown Manhattan experiences the greatest shortage of taxi supply in the morning, since there is a large area of residential places. And the midtown Manhattan is undersupply in the evening.

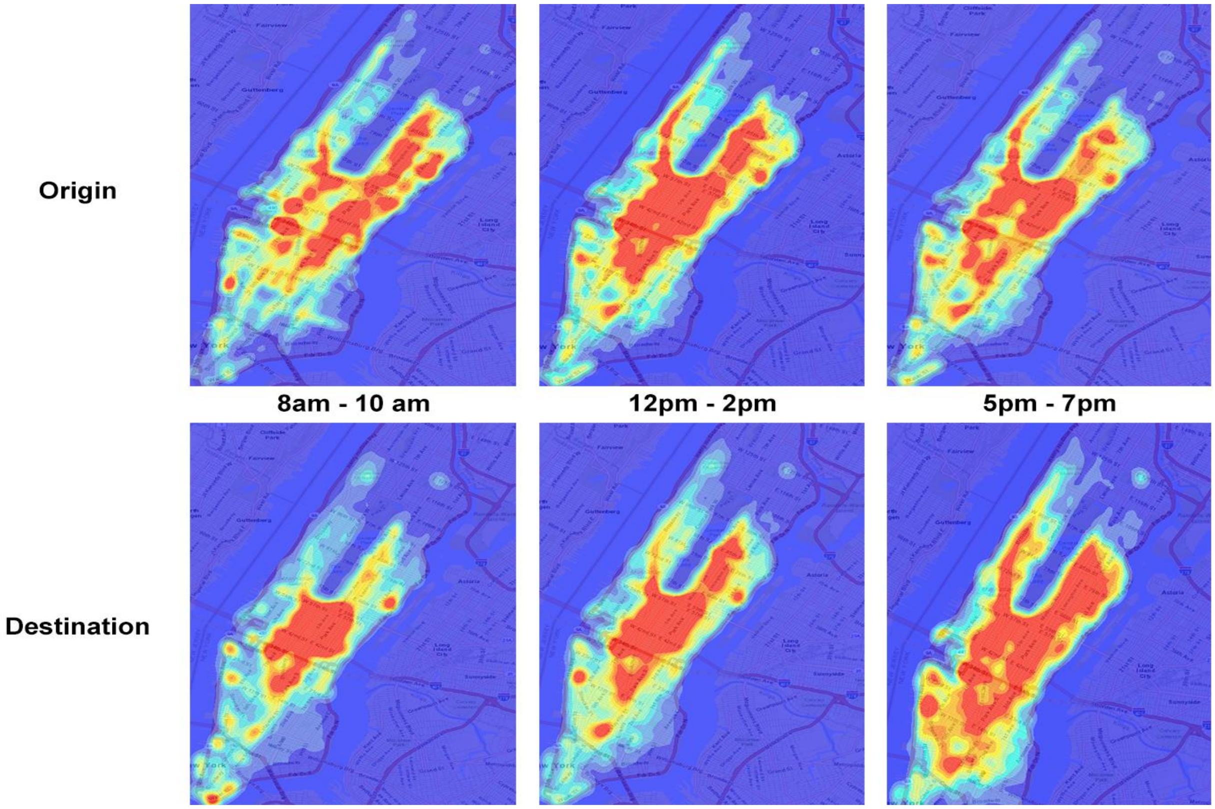


Figure 5 Trip Density Plot inbound Manhattan

(the density increases as color going from blue to red)

Based on the analysis of the unbalanced trips, several ideas might be useful to ease the problem. For nighttime unbalance of inward and outward Manhattan trips, pricing strategy should be helpful. An additional fee can be charged or a subsidy can be assigned for trips outward Manhattan only after midnight as taxi drives are less likely to leave Manhattan at that time. Moreover, since morning and evening trips have unambiguous origins and destinations, a shuttle service is believed to be effective. It can narrow the demand-supply gap of taxi service and reduce congestion at the same time.

3.3 Trip classification

It is recognized that temporal repeatability exists for taxi origins and destination at a given place. Besides, taxi trips in different parts of the city also have inherent similarities. The similarities are related to the places of taxi origins and destinations, travel distance as well as the time of the day. Clustering algorithm is implemented to classify the taxi trips.

3.3.1 **Clustering algorithm**

Clustering algorithms are widely used to disclose underlying pattern in large databases. Considering both spatial and temporal patterns of taxi trips, we use an eight-dimensional data record which covers geographic location, time of the day, travel distance and land use type as clustering input. A summary of the data inputs is given in Table 2.

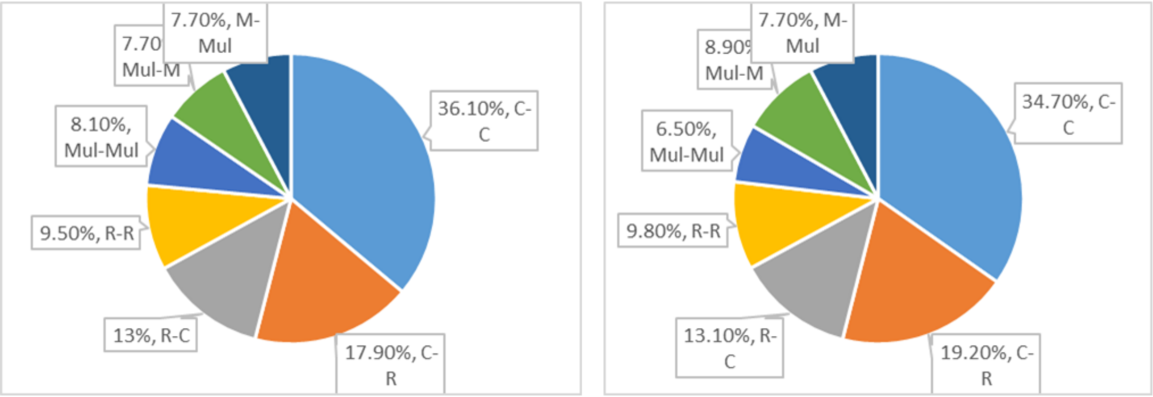
Table 2 Input Variables of Data Record

Name	Type	Example	Remark
Origin Longitude	continuous	-74.004	
Origin Latitude	continuous	40.722	
Destination Longitude	continuous	-73.981	
Destination Latitude	continuous	40.761	
Start time	continuous	19.26	In hour-scale
Trip distance	continuous	2.7	
Origin Land use type	categorical	2	1-Park, 2-Commerical 3-Residential, 4-Manufacturing
Destination land use type	categorical	4	

Conventional clustering algorithms including k-means, DBSCAN and agglomerative hierarchy clustering are sufficient dealing with continuous variables. But in our study, both continuous and categorical (land use type) variables are introduced. Therefore, a two-step clustering algorithm (Chiu et al., 2001) is implemented to process different variables in two steps, with all continuous variables assumed to be normally distributed and categorical variables to be multinomial. The first step is a pre-cluster approach which uses a sequential clustering method. The second step uses the agglomerative hierarchical approach which processes the sub-cluster from the first step recursively. Details for each step of the algorithm are referred to SPSS manual (SPSS, 2001).

3.3.2 Clustering Results

We use SPSS to run the two-step clustering algorithm. The clustering result is presented in Figure 6. For both weekday and weekend, taxi trips are classified into 7 groups. We name each cluster by its land use feature, including C-C, R-C, C-R, R-R, Multi-Multi, M-Multi, and Mul-M trips (C: Commercial, R: Residential, M: Manufacturing, Mul: the combination of the three). The spatial distribution of trip origin and destination in each cluster on weekdays is given in Figure 6.



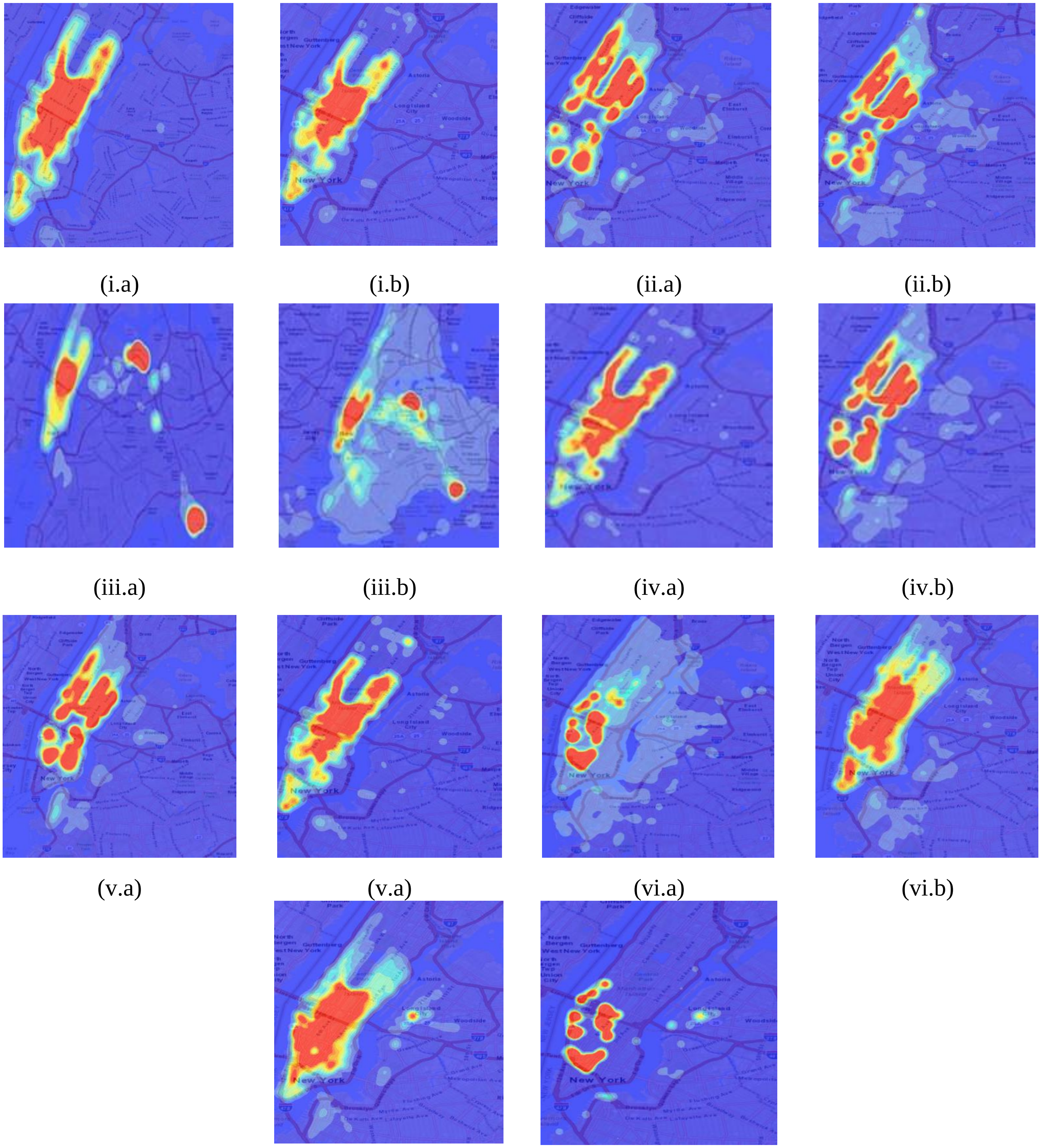
(a) Weekday (b) Weekend

Figure 6 Clustering Result

In general, C-C trips take the largest proportion with 36.1% of weekday trips and 34.7% for weekend. This turns out to be a rational pattern as commercial area attracts and produces large amount of trips. The typical urban sprawl pattern is observed from the clustering result. The commercial area are located at central and lower parts of Manhattan. Residential places are surrounding the commercial area and sprawled outside Manhattan with a lower value of the land. Taxi trips for residential places are dense by two sides of the Central Park. All these places are observed to have a higher average personal income.

The distribution of travel distance and trip starting time for each cluster is given in Figure 8. It is revealed that Mul-Mul cluster is a unique type of taxi trips. Most trip origin and destination of the cluster are positioned at midtown Manhattan, LaGuardia airport and JFK airport. The distribution of trip distance differs from other types of trip significantly. A majority of the trips have a long travel distance, and two distinct peaks are observed to be consistent with the travel distance from Manhattan to LGA and JFK. Hence, the cluster mainly account for airport trips to-and-fro Manhattan. Morevoer, these trips should be dealt with separately while analysis since short distance trips are the main component of other clusters.

For trip starting time, it is not surprising that the cluster patterns have weekday-weekend disparity. During weekdays, commercial related commuting including C-C trips, R-C trips and Mul-M trips have apparent morning and evening peaks. The evening peak usually lasts more than 4 hours starting at 5 pm. For weekend, no morning peak is observed and the amount of daytime trips increase after 10 am. The highest demand take place around midnight, which implies a shift from commercial related activities to entertainment related activities from weekday to weekend.



(i.a) (i.b) (ii.a) (ii.b)

(iii.a) (iii.b) (iv.a) (iv.b)

(v.a) (v.a) (vi.a) (vi.b)

(vii.a) (vii.b)

Figure 7 Spatial Density Plot of Cluster Origins and Destinations.

(i: C-C trip; ii: R-R trip; iii: Mul-Mul trip; iv: C-R trip; v: R-C trip;
vi: M-Mul trip; vii: Mul-M trip; a for origin and b for destination; density increases from blue to red)

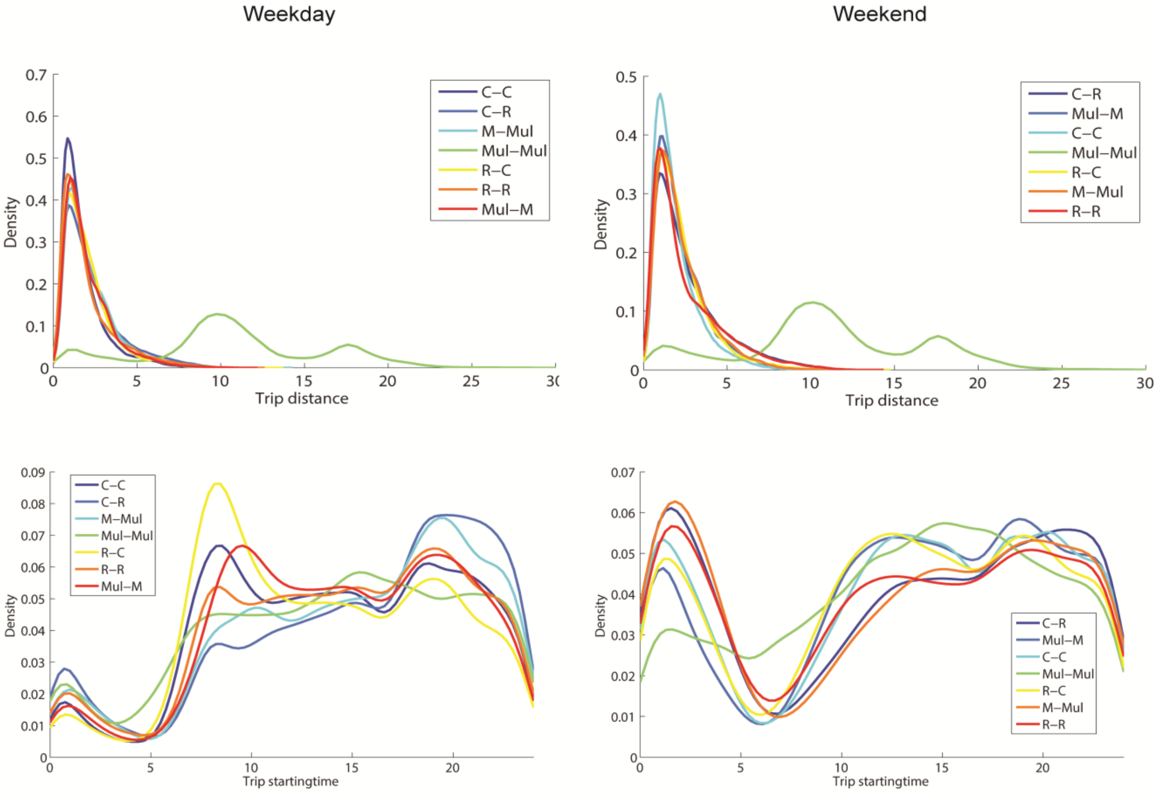
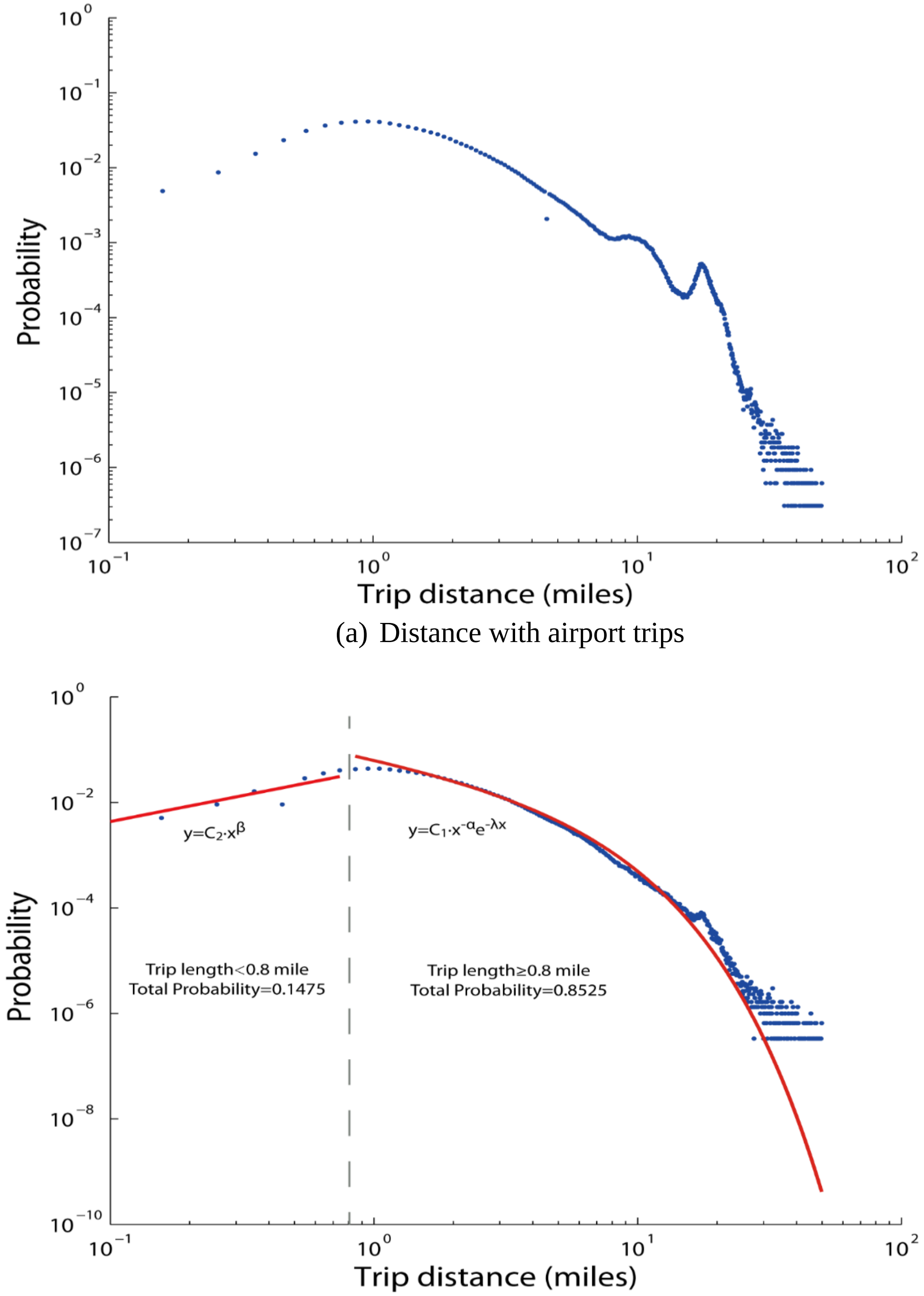


Figure 8 Travel Distance and Trip Starting Time Distribution for 7 Clusters

3.4 Taxi mobility pattern

Individual mobility pattern have been realized barely random. Several studies using data from the movement of an online game (Szell et al., 2012), the dispersal of bank notes (Brockmann et al., 2006) as well as trajectories from cellular data (Gonz ález et al., 2008) have found highly regulated pattern in human movement. The human movement is observed to follow a heavy-tailed plot under logarithmic scale and can be well approximated by scaling law. With human beings as the main participants, the taxi trips are results of human movement in an urban context as well. Hence, an effort is made to reveal the taxi mobility and examine the relationship with individual mobility.

To uncover the taxi mobility, first, the distribution of travel distance under logarithmic scale is plotted as shown in Figure 9(a). From the observation, the distribution of travel distance can be divided into parts: an ascending ranges from 0 to 0.8 mile, and then gradually descending as trip distance increases.



(a) Distance with airport trips

(b) Distance without airport trips

Figure 9 Taxi Trip Distance Distribution

Two minor peaks around 10 miles and 20 miles in the distribution are mainly caused by trips to LGA and JFK airport. The interference of airport trips has been discussed in previous section. We remove the trips to and from the two airports as they have specific purposes and unique characteristics. A refined distribution is generated in Figure 9(b).

Trips with distance less than 0.8 mile take 14.75% of total trips. As very short trips within walking radius, these trips differ from the general pattern of taxi mobility on a decision making process of whether to take taxis. The first part of the trips can be approximated with distribution:

$$P(d) \propto d^{-\beta} \tag{1}$$

where exponent β =0.9793.

The distribution resembles a power-law like distribution (straight line under logarithmic scale), however, the exponent takes a positive value. As mentioned earlier this phenomenon captures model choice process in whether take a taxi. It is intuitive that with the increase in distance, the probability of taking a taxi also increases until attaining its maximum around 0.8 miles.

The refined second part is used to capture urban mobility features of taxi trips. The trips greater than 0.8 mile contribute 85.25% of total trips. It is found that the distribution of taxi trip distance is well approximated by a power-law with exponential cut-off (also known as truncated power-law):

$$P(d) \propto d^{-\beta} e^{-\lambda d} \tag{2}$$

With exponent α =0.8652 and λ =0.3161. The distribution is found to be heavy-tailed. Unlike the power-law distribution of human movement reported (Szell et al., 2001; González et al., 2008; Brockmann et al., 2006), the taxi trip distance distribution has a faster probability decay in the tail part (the effect of the exponential cut-off term). This indicates the unique effects of urban environment on the distribution of taxi trip distance. Since the underlying size of urban area limits the distance of taxi trip, very long trip (e.g.>30 miles) are less likely to happen, and the scale-free property of a typical power-law distribution fails. It is notable that as taxi trips are important component of urban human movement, the trip distance distribution reflects a unique perspective of human mobility. That is, the taxi mobility pattern reveals the hidden role of urban geographical boundaries in limiting urban human movement.

CHAPTER 4. URBAN LINK TRAVEL TIME ESTIMATION

Taxicabs equipped with Global Positioning System (GPS) devices can serve as ubiquitous sensors monitoring traffic states in urban areas. This chapter presents a new descriptive model for estimating hourly average of urban link travel times using taxicab origin-destination (OD) trip data. The focus of this study is to develop a methodology to estimate link travel times from OD trip data and demonstrate the feasibility of estimating network condition using large-scale data with partial information. The data, collected from the taxicabs in New York City, provides the locations of origins and destinations, travel times, fares and other information of taxi trips. The new model infers the possible paths for each trip and then estimates the link travel times by minimizing the error between the expected path travel times and the observed path travel times. The model is evaluated using a testing network from Midtown Manhattan. Results indicate that the proposed method can efficiently estimate hourly average link travel times. Currently, there is a limited research on estimating urban link travel times using large-scale OD travel time data. This research provides a new possibility of fully utilizing the partial information provided in the urban taxicab data for network condition estimation purpose, which is cheap and also has a much better coverage than the usual centralized approaches.

4.1 Introduction

In the last few years, there has been a growing trend of installing GPS devises in taxicabs in urban areas. While GPS-equipped taxicabs have many advantages, including the ability to locate taxis and track lost packages, they also serve as useful real-time probes in the traffic network. Taxis equipped with GPS units provide a significant amount of data over days and months thereby providing a rich source of data for estimating network wide performance metrics. However, currently there are limited methodologies making use of this new source of data to estimate link or path travel times in the urban network. Within this context, this study proposes a new method for estimating hourly urban link travel times using large-scale taxicab data with partial information. The taxicab data used in this research provides limited trip information, which only contains the origin and destination location coordinates, travel time and distance of a trip. However, the extensive amount of data records compensates for the incompleteness of the data and makes the link travel time estimation possible. A novel algorithm for estimating the link travel times will be presented and tested in this study using a test network in New York City.

Data collected from New York City taxicabs is used to estimate the link travel times. The dataset provides an extensive amount of taxi trip data, which records the trip starting and ending geo-location, along with information about trip distance, time and fare. Unlike the detailed GPS trajectory data used in previous studies, the dataset only provides the trip origin and destination information (i.e. starting, ending location and time) without the exact trajectory of the taxicab; only path travel time and distance are known. However, the advantage of the massive amount of data (the number of observations recorded within a day range between 450,000 to 550,000) makes it possible to infer the possible routes that the taxicab is taking and further, to estimate the link travel times in the New York City network. There is potential bias associated with measuring network link travel times from taxis, as taxi drivers are just one particular group of all drivers in the network. However, given the high penetration rate of taxicabs, it is reasonable to assume that taxis are good probe vehicles and therefore taxi travel times are a good representation of the actual network condition.

In this research we propose a methodology to estimate urban link travel times based on taxi GPS data that includes only the information about the origin and destination of the trip and total travel time to reach the destination. The goal of this study is to show the potential of using taxicab data as a complimentary data source in urban transportation operation and management. The link travel times estimated from taxicabs provide an hourly aggregate measure of the urban network condition, which can be fused with the information from other existing data sources such as fixed sensors in the future.

4.2 Methodology

This section presents the proposed link travel time estimation model. We treat the path taken by a taxi as latent and derive the expected path travel time as a summation of each of the probable path travel time multiplied by the probability of taking that particular path. Link travel time estimation problem then becomes estimating the link travel times that minimize the least square error between the observed and expected path travel times. An MNL model is embedded to compute the probability that a taxi driver chooses a given path in the constructed reasonable path set, and the expected path travel time is computed for each trip record. The data are first processed to run the model, which include two steps: data mapping and constructing reasonable path set. The taxi trip origin and destination points are first mapped to the nearest links in the network. Instead of using all possible paths between each origin and destination points, we use k-shortest path algorithm to construct 20 shortest paths for each OD nodal pair of a trip, referred as the reasonable path set. The generated reasonable path sets serve as the basis for the link travel time estimation process.

4.2.1 **Link travel time estimation**

Link travel times in the network are estimated by minimizing the least squared difference between expected path travel times and the observed path travel times. We consider the actual path choice of the taxi as a latent variable and the link travel times as the model parameters to be estimated, the expected path travel time for observation i, $E(Y_i|R_i)$ can be written as:

$$E(Y_i|R_i) = \sum_{m \in R_i} g_m(\vec{t}) P_m(\vec{t}, d, \theta) \tag{3}$$

where

- Y_i is the variable of path travel time for observation i.
- R_i is the set of possible paths of a OD trip observation i.
- $\vec{t}$ is the vector of link travel times.
- d is the path distance set for R_i.

- $g_m(\vec{t})$ is the path travel time for path m.
- $P_m(\cdot)$ is the probability of selecting path m.
- θ is a positive scale parameter[1].

For a given path, the path distance is fixed, the variables to be estimated are the vector of link travel times $\vec{\mathrm{t}}$ and the scale parameter θ. Then, $E(Y_i|R_i)$ can be represented by a function of R_i, $\vec{t}$ and θ,

$$E(Y_i|R_i) = f\left(R_i, \vec{t}, \theta\right) \tag{4}$$

The error between observed path travel time y_i and expected path travel time $E(Y_i|R_i)$ is defined as the residual for observation i, which is:

$$r_i = y_i - f\left(R_i, \vec{t}, \theta\right) \tag{5}$$

Link travel times are estimated by minimizing the square difference between the expected path travel times and the actual path travel times observed in the data set D, defined as $S\left(\vec{t}, \theta\right)$,

$$S\left(\vec{t}, \theta\right) = \sum_{i \in D} r_i^2 = \sum_{i \in D} \left(y_i - f\left(R_i, \vec{t}, \theta\right)\right)^2 \tag{6}$$

$$\vec{t} = \arg\min_{\vec{t}} S\left(\vec{t}, \theta\right) \tag{7}$$

4.2.2 **Route choice model**

Due to the absence of any information on the path taken by the taxicab drivers, the actual path needs to be inferred. Thus a route choice model is developed to find the path choice of the taxicab drivers. Due to the lack of social or behavioral characteristics of taxi drivers in the dataset, traditional econometric models cannot be estimated. Hence, we build the route choice model using the limited cost variables from the dataset. We implement an MNL model to serve as the route choice model and consider the trip cost C_m in terms of both trip time and distance. The route choice model is defined as

[1] This will be further discussed in the route choice model.

$$P_m(\vec{t}, d, \theta) = \frac{e^{-\theta C_m(\vec{t}, d_m)}}{\sum_{j \in R_i} e^{-\theta C_j(\vec{t}, d_j)}} \qquad (8)$$

The parameter θ scales the perceived path cost. A large θ indicates a small perception error, and drivers will tend to select the path with minimum cost; while a small θ suggests a large perception variance, larger cost path gets more probability of being selected. In this model, θ is estimated together with the link travel times, which captures the variation in drivers' perceived path cost in different time period and network conditions.

The path cost C_m can be assumed as a function of trip fare. This is based on the assumption that each driver minimizes both trip time and distance, so that the driver can make more trips and thus make more revenue. We introduce a threshold ratio when constructing the reasonable path sets to exclude the trips that violate the aforementioned route choice behavior assumption. That is, if the taxi driver takes a much longer route to make more revenue in a single trip, then none of the paths in the reasonable path set will fall within the threshold given the observed path distance. These records are removed from the model estimation to ensure the input data matches with the route choice behavior assumption.

According to the taxicab fare rates provided by New York Taxi and Limousine Commission, the taxicab fare calculation involves both trip time and distance[1]. For standard city rate (taxi trips within Manhattan all follow this rate), fare (exclude surcharge and tax) include $2.50 upon entry, and $0.5 for each additional unit. The unit fare is:

- One-fifth of a mile, when the taxicab is traveling at 6 miles an hour or more; or
- 60 seconds when not in motion or traveling at less than 6 miles per hour.
- The taximeter shall combine fractional measures of distance and time in accruing a unit of fare.

[1] Taxicab rates from New York Taxi & Limousine Commission:
http://www.nyc.gov/html/tlc/html/passenger/taxicab_rate.shtml

The taxi rate of fare suggests a linear relationship with trip time and distance. The actual fare-time-distance relationship from the data is illustrated in Figure 10. Considering the complicated traffic condition and fare calculating method in actual situations, a linear model for the trip fare-time-distance relationship estimated from the data is used rather than the rate of fare provided by NYTLC:

$$fare = \beta_0 + \beta_1 \cdot time + \beta_2 \cdot distance \tag{9}$$

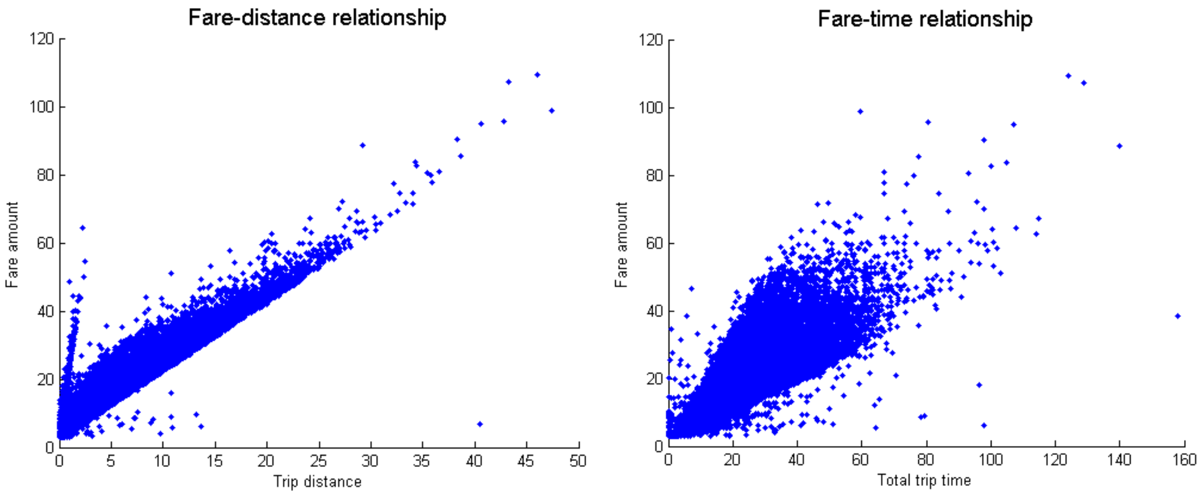


Figure 10 Fare-time-distance relationship

Table 3 Linear model for fare-time-distance relationship.

	Coefficient	Standard Deviation	P Value
β_0 (intercept)	2.143	0.00161	0.000
β_1 (coefficient for time)	0.275	0.00021	0.000
β_2 (coefficient for distance)	1.563	0.00058	0.000
Number of observations		415561	
R-squared		0.99	
Adjusted R-squared		0.99	

The estimated coefficients of β_0, β_1 and β_2 are listed in Table 3.The units for time and distance are minute and mile respectively; the fare used in the calculation does not include surcharge and tax. The estimation result shows that time and distance are highly significant in determining the trip fare. The model has a R^2 value of 0.99, suggesting that the data is well fitted using this simple linear model. The path cost used in the route choice model is therefore modeled as:

$$C_m(\vec{t}, d_m) = \beta_1 \cdot g_m(\vec{t}) + \beta_2 \cdot d_m \tag{10}$$

Where d_m is the distance for path m, and the path travel time of path m, $g_m(\vec{t})$ is defined as

$$g_m(\vec{t}) = \alpha_1 t_O + \alpha_2 t_D + \sum_{l \in L} \delta_{ml} t_l \tag{11}$$

where

- t_O is the travel time of the link where the trip starting point lies.
- t_D is the travel time of the link where the trip ending point lies.
- L is the set of the links.
- t_l is the tavel time of the link l.
- δ_{ml} is the link-path incident relationship, 1 if link l is in path m, 0 otherwise.
- α_1, α_2 are the distance proportions.

The simple linear form of the path cost function is used for two reasons: 1) the linear fare-time-distance relationship is supported by data, and distance and time are identified as significant factors that impact the trip fare; 2) a simple form of path cost function ensures the model is computationally tractable for large-scale input data and the short term link travel time estimation purpose. The constant term is not included since this common component cancels out in the MNL model. Further, as the starting and ending points lie within the starting and ending links, a taxi only experiences a part of the total link travel times to traverse those links. In this study, the proportion of this part of link travel time to the total link travel time is assumed to be the distance proportions α_1 and α_2 defined in the data mapping section.

4.2.3 **Data mapping**

It is common in urban environments such as New York City that taxicabs often travel in the GPS shadow of tall buildings causing errors in the GPS data. Thus a data mapping process is introduced to pre-process the raw GPS data. There are two purposes in this step: first, to map the data to nearest links in the road network to reduce GPS errors; second, to match the starting and ending points to the actual road network and transform the raw data into usable data for network level analysis.

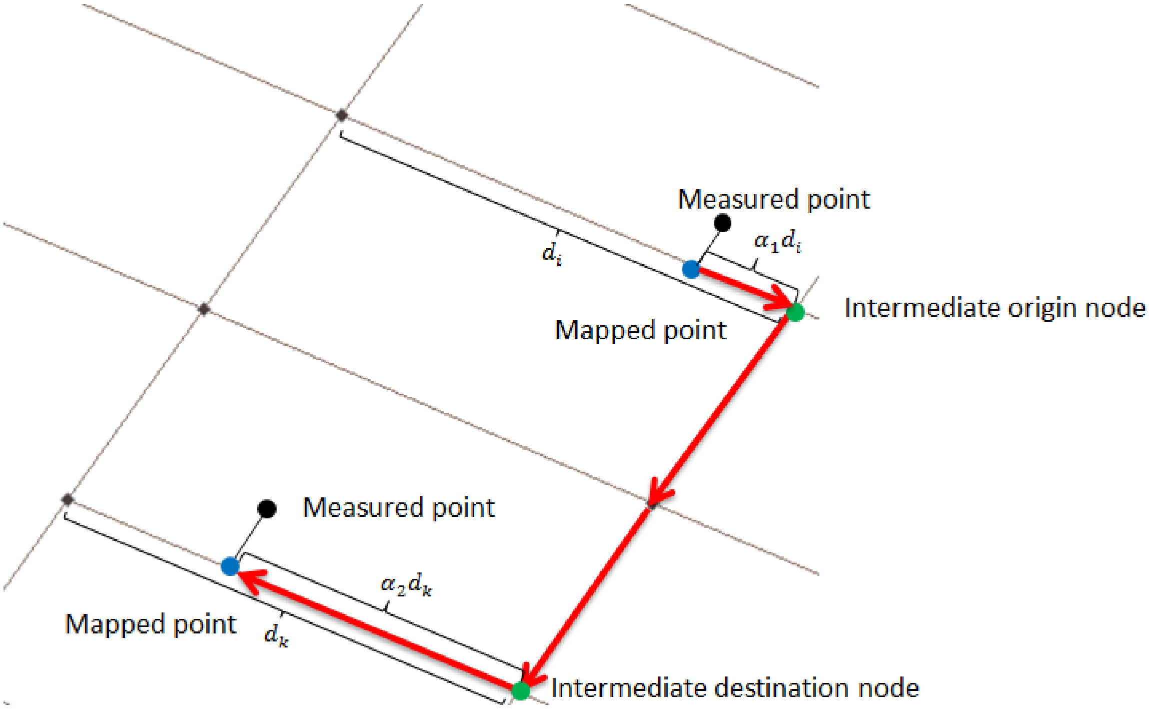


Figure 11 Illustration of data mapping

Figure 11 illustrates the data mapping procedure. The raw origin and destination points (black points in Figure 11) are mapped to the perpendicular foot of the nearest link (blue points in Figure 11), and the new points are then used in the later analysis. The locations (represented by distance ratio between two endpoints of a link) of the new points on the link are also computed to calculate accurate k-shortest path distance in later step.

The new origin and destination points correspond to four endpoints of two links. In big cities like New York, a great proportion of the links in the urban grid network are one-way streets. For origin and destination points that lie on one-way streets, the actual two intermediate nodes in abovementioned four endpoints are easily identified given the directional information of the link. For any point lies on the two-way street, both the two endpoints of this link are used as intermediate origin/destination nodes for this record. All the combination of the intermediate origin and destination nodes and the corresponding shortest path sets are then used to generate the reasonable path sets for this record. These identified intermediate points serve as intermediate origin and destination nodes. The distance proportion to the total length of the link from the new origin point to the intermediate origin node is defined as α_1, and the distance proportion from the new destination point to the intermediate destination node is defined as α_2. For points lying on the two-way streets, different combination of α_1 and α_2 are allowed for the same record, depending on the combination of intermediate origin/destination node.

4.2.4 Constructing reasonable path sets

Given the origin and destination of a taxi trip, the number of paths in urban network between the origin and destination are potentially large, especially for downtown grid networks of big cities. Since the actual path taken by a taxi driver is unknown, an important sub-question of the analysis is to infer the possible path set of a given taxi trip. Considering the large number of observations available in a large network, the overall search space for the possible path sets are huge. It is necessary to reduce the size of the possible path sets. In this study, Yen's k-Shortest Path algorithm (Yen, 1971) ($k = 20$) is used to generate the initial path sets, and the trip distance recorded in the data is then used to eliminate unreasonable paths. Only the paths that do not have excessively high or low lengths compared to the observed taxi trip distance will be used.

Because the trip distance recorded in the data is not very accurate (only accurate to 0.1 mile), a threshold ratio of 15%~25% for weekday, and 20~25% for weekend (both upper and lower) is used, depending on the amount of data available during one hour. The threshold ratio is used to filter out the unreasonable paths whose measured lengths deviate significantly from the recorded trip distance.

4.2.5 Solution approach

To solve this non-linear least square problem, the Levenberg-Marquardt (LM) method (Nocedal and Wright, 2006; Fletcher, 1971) is used. The Levenberg-Marquardt method is a widely used optimization algorithm in solving least square curve fitting and nonlinear programming problems. It outperforms the simple gradient decent method and the well-known Gauss-Newton (GN) methods in a wide variety of problems. The traditional Gauss-Newton method uses a line-search method, which is computationally expensive for solving this problem, since the objective function is huge. The updating method in Gauss-Newton method is similar to Newton's method, which has numerical issues when the approximated Hessian is near singular and easily fails to converge to the optima if improper initial value is used. Levenberg-Marquardt method on the other hand, uses a trust-region strategy instead of the line search method, which determines the step size before the updating step. The different Hessian approximation method used in LM also helps to ensure the positive definiteness of the approximated Hessian in each iteration. This results in a more robust performance, which means that in many cases Levenberg-Marquardt method finds a solution even if it starts very far off the final minimum. It is showed in Nocedal and Wright (2006) that Levenberg-Marquardt enjoys rapid local convergence near optima, and under ideal cases, the convergence is actually quadratic.

For simplicity, define

$$p_m(\vec{t},\theta) = e^{\theta(-\beta_1 \cdot g_m(\vec{t}) - \beta_2 \cdot d_m)} \tag{12}$$

Thus we can write the denominator of Eq. (12) as:

$$S_{R_i}(\vec{t},\theta) = \sum_{j \in R_i} e^{\theta(-\beta_1 \cdot g_j(\vec{t}) - \beta_2 \cdot d_j)} = \sum_{j \in R_i} p_j(\vec{t},\theta) \tag{13}$$

Then, the expected path travel time can be written as,

$$E(Y_i|R_i) = f(R_i,\vec{t},\theta) = \sum_{m \in R_i} g_m(\vec{t}) \frac{p_m(\vec{t},\theta)}{S_{R_i}(\vec{t},\theta)} \tag{14}$$

Define

$$J_{ik} = \frac{\partial f\left(R_i, \vec{t}, \theta\right)}{\partial t_k}, \qquad k = 1,2,..,N$$
$$J_{iN+1} = \frac{\partial f\left(R_i, \vec{t}, \theta\right)}{\partial \theta} \tag{15}$$

Thus J forms a $N_D \times (N+1)$ matrix, where $\mathrm{N_D}$ is the number of observations in data set D, N is the number of links in the network. The vector of link travel times and the scale parameter θ are updated iteratively using

$$\vec{t} \approx \vec{t}^{k+1} = \vec{t}^{k} + \vec{p}_t^{(k)}$$
$$\theta \approx \theta^{k+1} = \theta^{k} + p_\theta^{(k)} \tag{16}$$

$\vec{p}^{(k)} = \left(\vec{p}_t^{(k)^T}, p_\theta^{(k)}\right)^T$ is the update direction in *k*th iteration, which is obtained by solving the following linear system,

$$\left(\boldsymbol{J}^{(k)\ T}\boldsymbol{J}^{(k)} + \lambda \boldsymbol{I}\right)\vec{p}^{(k)} = \boldsymbol{J}^{(k)^T} r_i \tag{17}$$

where, $J^{(k)\ T}J^{(k)}$ is the first order approximation of the Hessian matrix of the problem, and λ is referred to as damping factor, which adjusted at each iteration under a trust-region strategy. A modified Levenberg-Marquardt method replaces the identity matrix *I* with the diagonal matrix with the diagonal element of $J^{(k)\ T}J^{(k)}$, which shows as follows

$$\left(\boldsymbol{J}^{(k)\ T}\boldsymbol{J}^{(k)} + \lambda diag\left(\boldsymbol{J}^{(k)\ T}\boldsymbol{J}^{(k)}\right)\right)\vec{p}^{(k)} = \boldsymbol{J}^{(k)^T} r_i \tag{18}$$

This study uses the modified version of Levenberg-Marquardt method, as it avoids slow convergence in the direction of small gradient. Detailed description of the updating scheme of damping factor λ and implementation is discussed by Fletcher (1971).

In the above equations, J_{ik} is computed as

$$J_{ik} = \sum_{m \in R_r} \left\{ \frac{p_m\left(\vec{t},\theta\right) S_{R_r}\left(\vec{t},\theta\right) \frac{\partial g_m\left(\vec{t}\right)}{\partial t_k}\left[1 - \beta_1 \theta g_m\left(\vec{t}\right)\right] - g_m\left(\vec{t}\right) p_m\left(\vec{t},\theta\right) \frac{\partial S_{R_r}\left(\vec{t},\theta\right)}{\partial t_k}}{\left[S_{R_r}\left(\vec{t},\theta\right)\right]^2} \right\} \tag{19}$$

in which $k = 1,2,\ldots,N$, and for $k = N+1$, the J_{iN+1} is defined as

$$J_{iN+1} = -\sum_{m\in R_r} \frac{p_m(\vec{t},\theta) S_{R_r}(\vec{t},\theta)}{S_{R_r}(\vec{t},\theta)} \left[\beta_1 g_m(\vec{t}) + \beta_2 d_m + \frac{\frac{\partial S_{R_r}(\vec{t},\theta)}{\partial\theta}}{S_{R_r}(\vec{t},\theta)} \right] \tag{20}$$

$\frac{\partial g_m(\vec{t})}{\partial t_k}$, and $\frac{\partial S_{R_r}(\vec{t},\theta)}{\partial t_k}$, are defined as follows,

$$\frac{\partial g_m(\vec{t})}{\partial t_k} = \begin{cases} \alpha_1 & if\ t_k \in L_O \\ \alpha_2 & if\ t_k \in L_{\mathrm{D}} \\ \delta_{mk} = \begin{cases} 1 & if\ link\ k\ on\ path\ m \\ 0 & if\ link\ k\ not\ on\ path\ m \end{cases} \end{cases} \tag{21}$$

$$\frac{\partial S_{R_r}(\vec{t},\theta)}{\partial t_k} = -\beta_1 \sum_{j\in R_r} \left[p_j(\vec{t},\theta) \frac{\partial g_j(\vec{t})}{\partial t_k} \right] \tag{22}$$

$$\frac{\partial S_{R_r}(\vec{t},\theta)}{\partial\theta} = -\sum_{j\in R_r} [\beta_1 g_j(\vec{t}) + \beta_2 d_j] p_j(\vec{t},\theta) \tag{23}$$

One can observe that the problem is not convex and hence may have multiple local optima. A proper initial point is needed to ensure the convergence to the most probable solution. A preprocessing step is used to search for the network wide optimal mean speed. In this step, all the links in the network are assumed to have the same mean speed $\mathrm{v_m}$, thus a 1-dimensional search algorithm can be implemented to find the $\mathrm{v_m}$ that minimizes the objective function. The obtained mean speed is then used to calculate the initial values of the link travel times. In Table 4, data from 3/15/2010 (Monday) 21:00-22:00 are used to test the choice of different initial link speeds for link travel time estimation. The result shows that using the network wide optimal mean speed as an initial point yields the lowest objective value and RMSE, which suggests that the preprocessing step is an effective approach of finding desirable link travel time estimates.

Table 4 Test results for different choices of model initial values.

Initial Speed (mph)	Objective Function value	Iteration Used	RMSE	MAPE
10[*]	779.830	20	1.372	21.87%
8	1215.410	17	1.713	29.30%
12	783.143	16	1.375	21.52%
8-12 uniformly distributed	801.487	20	1.391	22.49%
8-12 uniformly distributed	805.075	16	1.394	22.68%
6-14 uniformly distributed	805.146	27	1.395	22.55%
6-14 uniformly distributed	807.044	23	1.396	22.35%

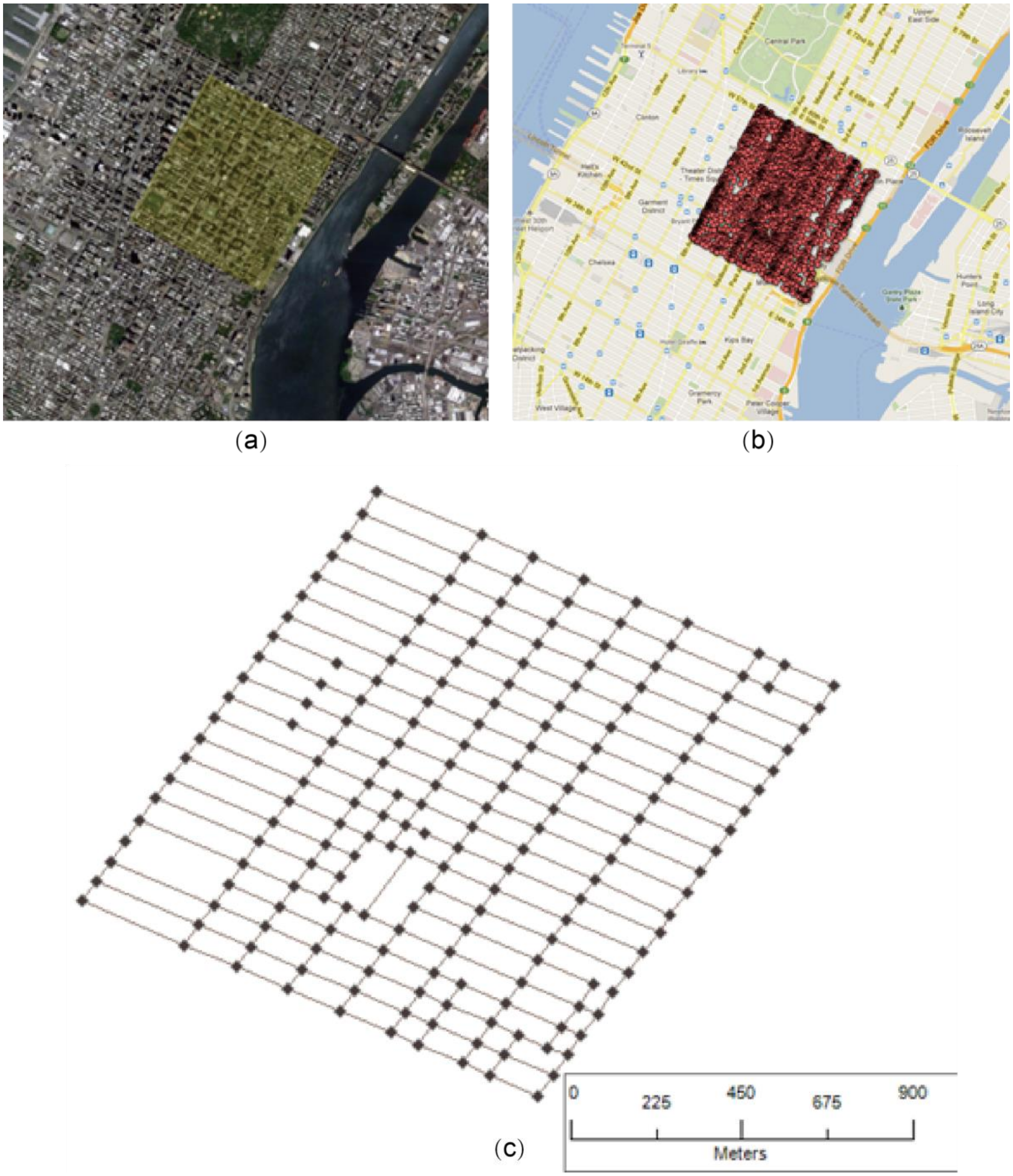
[*] Network wide optimal mean speed.

4.3 Testing data and network

The data used in this research was collected by New York City Taxi and Limousine Commission on a trip by trip basis. The data records each trip origin and destination GPS coordinate, trip distance and duration, fare, payment method, and other related information. The data set contains data from February 2008 to November 2010. In this study, a week's data (from 3/15/2010 to 3/21/2010) is selected to test the proposed method.

A small region in the southeast of Central Park of Midtown Manhattan is selected to serve as the study region, which is a $1370\text{m} \times 1600\text{m}$ rectangle area. The corresponding network is also extracted (Figure 12), which contains 193 nodes and 381 directed links. The network has 331 road segments and only 50 of them are two-way streets. From the original data set, all the records that fall within the region are extracted. Figure 13 presents the number of observations inside the study region in a typical weekday (3/15/2010, Monday) and weekend (3/20/2010, Saturday) respectively. We obtain as many as 1000 observations in one hour on a typical weekday (Monday) and about 500 observations in one hour in a weekend (Saturday) inside the study region.

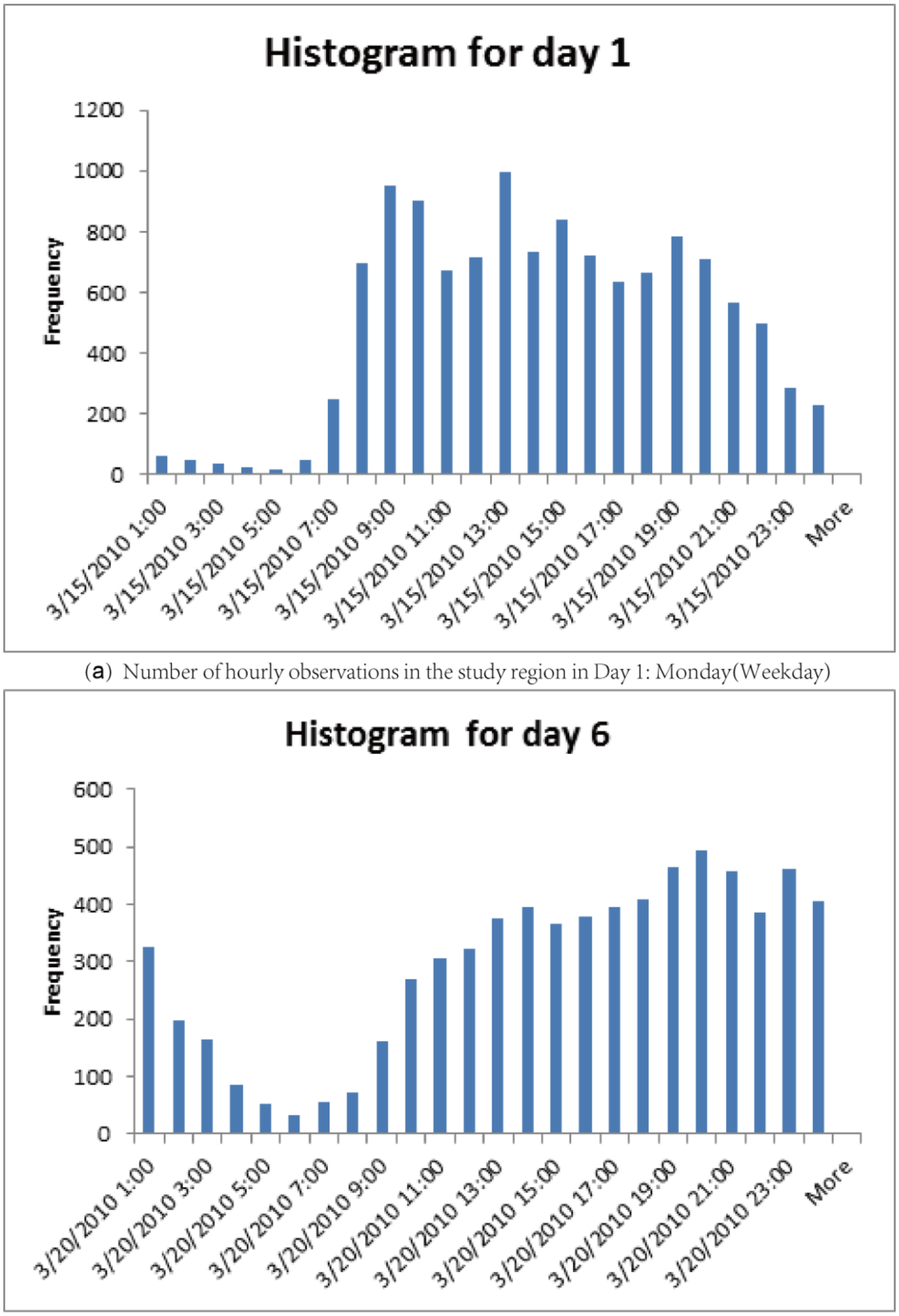
In this study, the data is split into hourly intervals, and link travel times are estimated using the data from the corresponding hour. Although traffic conditions can change rapidly during one hour, a shorter time period will not guarantee a good statistical significance due to the insufficient amount of observations given the limited information in the data.



(a) (b)

(c)

Figure 12 Testing network in the study region



(a) Number of hourly observations in the study region in Day 1: Monday(Weekday)

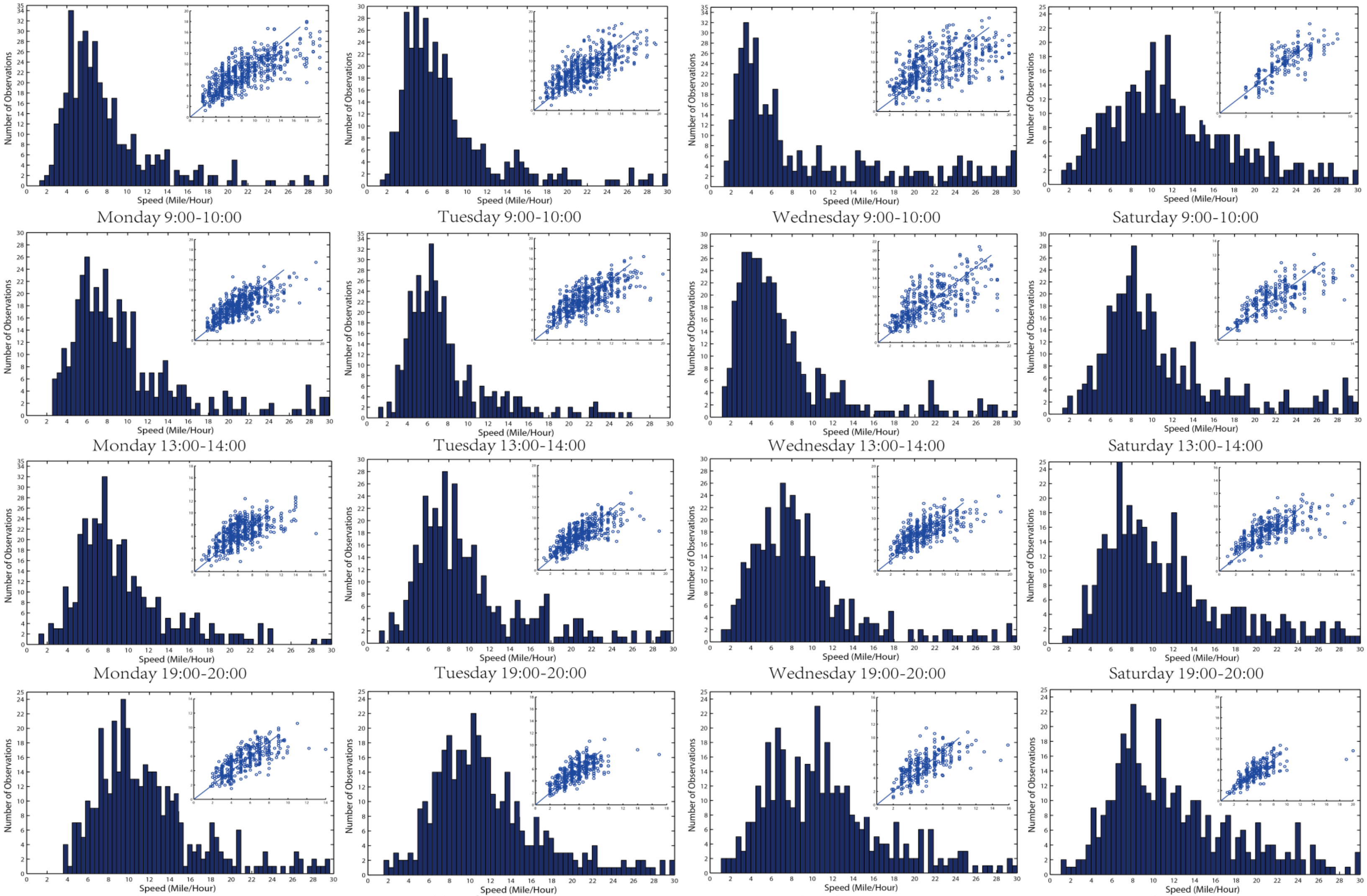
(b) Number of hourly observations in the study region in Day 6: Saturday(Weekend)

Figure 13 Histogram for number of hourly observations in the study region

4.4 Model Results

To implement the model discussed in the previous section, a Matlab code is written using Parallel Computing Toolbox. A k-shortest path set is required to be computed for each nodal pair in the network and this process takes a considerable amount of time. But once the process is complete, the path sets are stored and needs no further computation. The steps of data mapping and constructing reasonable path sets take little time to complete, as they make use of the information from already computed k-shortest path sets of the network. The Levenberg-Marquardt method provides good convergence properties, and the entire optimization process can be efficiently solved within 15 minutes using an Intel i7 CPU laptop. The computation time can be further reduced by using Matlab C/MEX code or a more powerful computer.

Link travel times for four time periods (9:00-10:00, 13:00-14:00, 19:00-20:00, and 21:00-22:00) in a day are estimated based on a week's Taxi GPS data (from 3/15/2010 to 3/21/2010). The time period from 9:00 to 10:00 represents the morning peak period, as the highest number of taxi trips are observed in this period on weekdays; while 21:00-22:00 is tested for the off-peak hour situation. A lower bound of speed (one mile per hour) is used to ensure that we do not obtain unreasonably large travel times; an upper bound of speed (30 miles per hour) is used as the free flow speed to set a lower bound for the estimated link travel times. We use the link speeds instead of the link travel times to give a more intuitive representation of the link travel time estimation results. Figure 14 presents the estimated link speeds and correlation plots of observed and estimated path travel times for Monday, Tuesday, Wednesday and Saturday, which are more representative, and the results for Thursday, Friday and Sunday are presented in Figure 15.



Monday 9:00-10:00

Tuesday 9:00-10:00

Wednesday 9:00-10:00

Saturday 9:00-10:00

Monday 13:00-14:00

Tuesday 13:00-14:00

Wednesday 13:00-14:00

Saturday 13:00-14:00

Monday 19:00-20:00

Tuesday 19:00-20:00

Wednesday 19:00-20:00

Saturday 19:00-20:00

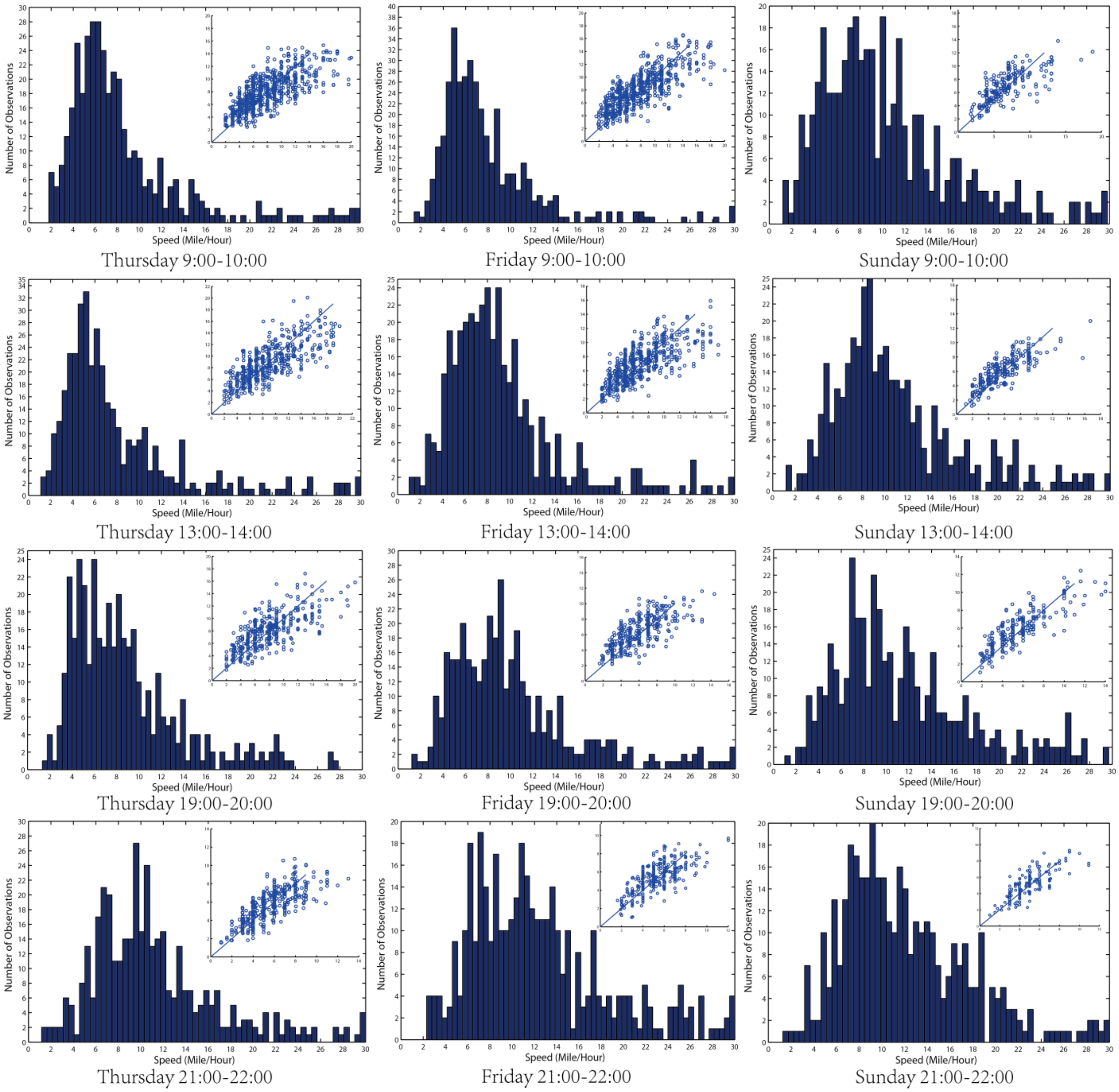
Monday 21:00-22:00

Tuesday 21:00-22:00

Wednesday 21:00-22:00

Saturday 21:00-22:00

Figure 14 Histogram of estimated link speed and correlation plot of observed and estimated path travel time for Monday, Tuesday, Wednesday and Saturday (Inside plot, X-axis: observed path travel time (min), Y-axis: estimated path travel time (min))



Thursday 9:00-10:00

Friday 9:00-10:00

Sunday 9:00-10:00

Thursday 13:00-14:00

Friday 13:00-14:00

Sunday 13:00-14:00

Thursday 19:00-20:00

Friday 19:00-20:00

Sunday 19:00-20:00

Thursday 21:00-22:00

Friday 21:00-22:00

Sunday 21:00-22:00

Figure 15 Histogram of estimated link speed and correlation plot of observed and estimated path travel time for Thursday, Friday and Sunday

(Inside plot, X-axis: observed path travel time (min), Y-axis: estimated path travel time (min))

Table 5 Model estimation error and estimated value for the scale parameter θ.

Day	Error	Time Period			
		9:00-10:00	13:00-14:00	19:00-20:00	21:00-22:00
Monday	RMSE (min)	2.614	1.981	1.937	1.372
	MAPE	29.51%	24.22%	26.27%	21.87%
	θ	0.165	0.063	0.435	0.068
Tuesday	RMSE (min)	2.461	2.302	1.827	1.437
	MAPE	29.63%	25.59%	23.33%	22.20%
	θ	1.082	0.049	0.329	0.003
Wednesday	RMSE (min)	3.827*	3.216*	2.180	1.691
	MAPE	41.32%*	34.97%*	28.73%	24.40%
	θ	1.030	0.867	1.153	0.539
Thursday	RMSE (min)	2.468	2.699	2.490	1.382
	MAPE	27.28%	27.92%	28.54%	21.05%
	θ	0.469	0.037	0.264	0.499
Friday	RMSE (min)	2.260	2.179	1.692	1.334
	MAPE	27.76%	27.04%	25.17%	22.26%
	θ	0.075	0.010	0.717	0.245
Saturday	RMSE (min)	1.034	1.690	1.839	1.584
	MAPE	16.84%	24.58%	27.14%	21.61%
	θ	0.469	0.287	0.081	0.087
Sunday	RMSE (min)	2.041	1.518	1.395	1.160
	MAPE	25.44%	23.70%	22.72%	19.87%
	θ	0.166	0.239	0.190	0.615

* Traffic disturbance caused by Patrick's Day Parade.

Based on model estimation results, for weekdays, it is found that most of the links have speeds between 4 to 8 miles/hour in the 9:00-10:00 morning peak hour. During the 13:00-14:00 period, the distribution of speed is slightly improved and peaks around 7 miles/hour. In the 19:00-20:00 period, the mean speed is observed between 6 to 8 miles/hour. However, in the 21:00-22:00 off-peak period, a great number of links are observed to have speeds around 10 miles/hour. In contrast, during weekends, a relatively higher average speed (8-10 miles/hour) is observed during 9:00-10:00 in the morning, and relatively lower average speed (about 8 miles/hour) is observed during 19:00--20:00 pm period. These values are consistent with a previous study on New York City traffic speeds where it is reported that on weekdays in the daytime, in east Midtown, average traffic speed is 6.3 mph whereas on Saturdays, the average speed is about 8.5 mph (Grynbaum, 2010).

The root mean square Error (RMSE) and mean absolute percentage error (MAPE) are used to evaluate the estimation results:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left({T_i}^{Pr} - {T_i}^{Ob}\right)^2} \qquad (24)$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{{T_i}^{Pr} - {T_i}^{Ob}}{{T_i}^{Ob}}\right| \times 100\% \qquad (25)$$

where

- ${T_i}^{Pr}$ is the model estimated trip path travel time.
- ${T_i}^{Ob}$ is the observed trip path travel time.
- n is the number of observations.

The model estimation errors and the estimated values of the scale parameter θ are presented in Table 5 As showed in the result, except for 2 time intervals (Wednesday 9:00-10:00 and 13:00-14:00), all the link travel time estimation results have MAPE below 30%, and for some off-peak hours, e.g. 21:00-22:00 pm, the MAPE is only around 22%. It is noticeable that Wednesday 9:00-10:00 and 13:00-14:00 have much larger errors and lower link travel speeds compared with other days. It is found that this Wednesday (3/17/2010) happened to have St Patrick's Day Parade. The parade was from 8:00 to 15:00 and marched down the 5th street (contained in the test network). Few roads were temporarily closed and huge crowds were drawn along the parade routes. This caused huge disruption in traffic network and explained the high estimation errors on Wednesday. It is also observed that after the parade ended, the estimation results for 19:00-20:00 and 21:00-22:00 restore to normal condition, which have RMSE under 2.5 min and MAPE under 30%. It is found that in congested time period (e.g. 9:00-10:00 on weekdays), the results have relatively higher estimation errors. This could be the effect of the rapid changes in the network condition, since the model estimates hourly average link travel times. The link speed estimates also confirm that in high estimation error time period, a greater proportion of links have relatively lower traveling speed.

The estimation results for the scale parameter θ show relatively large variance in drivers' route choice behaviors. All the estimated values for θ are smaller than 1.2, and very small θ (0.003) is observed. The change in θ reflects the relatively large variance in taxi drivers' route choice behaviors in different time periods, days and network conditions. The wide range for the estimated values of θ (0.003~1.153) could be the result of several reasons. One plausible explanation could be that as traffic conditions change in different time periods of a day, taxi drivers may have different levels of perception error, which are reflected in their route choice behaviors. However θ only considers the overall variance in taxi drivers' perceived path costs and treats all taxi drivers as homogeneous individuals, which does not capture the behavioral heterogeneity among the taxi drivers.

Three consecutive Mondays in 2009 (2009/9/14, 2009/9/21, 2009/9/28) are also investigated to see if repeatability exist across weeks (due to space limitation, these results are not included). However, no significant pattern is found in terms of link speed profile and travel time variation during a day. The findings of the three Mondays agrees with the general pattern found on weekdays discussed above, but variation in terms of the distribution of link travel speeds is also observed, and no conclusive inference can be made across weeks.

In this model, intersection delay is not modeled due to the lack of detailed vehicle trajectory information in the data. In the testing network, most of the links have lengths ranging from 80 to 300 meters; assuming the vehicle traveling at a speed of 8 miles per hour, a great number of links will have a travel time less than 1 minute. However, the intersection delay at a traffic signal sometimes can be greater than the link travel time itself. In a 10 minutes trip, it is very likely to have at least 2 minutes of intersection delay on average, which partly explains the RMSE of around 2 minutes in the model. This is a potential source of errors of the model. The intersection delay causes inconsistency in the link travel time estimation and leads to overestimation of actual link travel times. However, given only origin and destination information provided in the data, modeling intersection delay separately will introduce excessive complexity in travel time estimation, which makes the short term estimation intractable. Also, there is no guarantee on the quality of the estimated intersection delay, since too little information is available to separate the intersection delay from the total link travel time. Thus given the incompleteness of the data, we combine the intersection delay into the link travel times and focus on estimation the hourly average link travel times.

Furthermore, because the link travel times are estimated as hourly average values, variations in link travel times within one hour can introduce errors in the model estimation. The heterogeneity among the drivers' behaviors (e.g. some drivers prefer to drive fast and choose the shortest path, some drivers prefer to drive at a moderate speed and take a relatively long path, etc.) may also contribute to the estimation errors. Certain trips are observed to take as much as 20 minutes in the testing network, which involve a lot of uncertainty in path choices, leading to some errors in estimation results as well.

CHAPTER 5. CONCLUDING REMARKS AND FUTURE RESEARCH

This research exploits a large-scale taxi trip data from NYCTLC. The two major focus of this study is to understand underlying patterns of urban dynamics from the taxi trip data, and estimate urban link travel times using partial trip information.

Taxi data has been proved to be an efficient tool to understand urban dynamics and several interesting insights are raised in our study. Unbalanced trips are common in taxi industry and should be carefully investigated to improve the level of service. Airport trips is a special part of taxi trips and differ from regular taxi trip patterns. Land use has significant impact on taxi trip types, and different types of taxi trips are helpful to understand the structure of a city. Moreover, we discover that the mobility of taxi trips are restricted by the urban geographical boundaries.

A new model is also proposed to use the limited information provided in the taxi GPS data to estimate urban link travel times. The taxicab data used in this study lacks the information of actual paths taken by the taxi drivers. The proposed model treats the path taken as latent, constructs a reasonable path set, formulates an MNL model to compute the probability of a path being taken by the driver, and estimates the link travel times by optimizing a nonlinear least square problem. Model estimation results indicate that the proposed method can efficiently estimate hourly average link travel times.

It is recommended to split the whole urban region into smaller zones (e.g. 1.5km×1.5km) to implement this model, because of the following reasons: (1) Larger zones contain longer trips, which involve more uncertainties in path choices, thus long trips are less reliable in the link travel time estimation given this type of data. (2) Preparing the k-shortest path set for all the nodal pairs in a large network is computationally expensive. The number of nodal pairs grows as n^2 as the number of nodes in the network, and a greater k value is also needed to ensure a good representation of reasonable paths. By reducing the zone size, we can ensure the computational tractability for short term link travel time estimation. (3) The data provides a large number of records in an hour even in a 1.37km×1.6km size zone, thus the amount of data is enough for the model.

This model can be further verified using the actual trajectory information of the taxi trips. Although this information has been collected by NYLTC, it is currently unavailable to the researchers. The model is also applicable to use trajectory data (treating two intermediate trajectory points as origin and destination point). The accuracy of the model can be improved with more detailed data and greater number of observations.

There are still some scopes to further improve this study. From the urban dynamics perspective, the current researches are primarily focused on exploring patterns. Future study can be focused on building a model from the patterns discovered to account for human movement within urban context. Moreover, more information such as social economics can be combined into the data analysis to provide more in depth information. Furthermore, it would be interesting to develop a methodology to infer urban land use types from taxi patterns. Also, attentions can be paid on extracting travel information from taxi dynamics and provide feedbacks to users.

From the urban link travel time estimation perspective, only the data in the current time period are used in the current estimation model, and historical data are not used. Further research can be done to investigate a hybrid approach of using historical data as well as optimizing current estimation error. Another research direction in the future is to improve the route choice model to account for more realistic route choice behaviors of the taxi drivers. The current route choice model only considers drivers who minimize trip time and distance in each trip, and records that do not comply with this assumption are filtered out. A more comprehensive route choice model would utilize more data records and provide less estimation bias. Furthermore, intersection delays are important causes of irregularity of link travel times, which may lead to bias in the estimated travel times. Future research can be done to incorporate the effects of intersection delays in the link travel time estimation, and thus improve the estimation accuracy. All these efforts would provide a more accurate and reliable way to estimate urban network conditions using the partial information provided by the taxicab data.

REFERENCES

Batty, M., and Xie, Y.(1985). "From cells to cities." Environment and planning B 21, s31-s31.

Batty, M., Xie, Y. and Sun, Z. (1999). "Modeling urban dynamics through GIS-based cellular automata." Computers, environment and urban systems 23, no. 3, pp. 205-233.

Brockmann, D., Hufnagel, L. and Geisel, T. (2006). "The scaling laws of human travel." Nature 439, no. 7075, pp. 462-465.

Calabrese, F., Diao, M., Lorenzo, G. D., Ferreira, J. and Ratti, C. (2013). “Understanding Individual Mobility Patterns from Urban Sensing Data: A Mobile Phone Trace Example.” Transportation Research Part C: Emerging Technologies 26 (January), pp. 301–313.

Chang, H, Tai, Y., Chen, H. W., Hsu, J. Y. and Kuo, C. P. (2008). "iTaxi: Context-aware taxi demand hotspots prediction using ontology and data mining approaches." In Proceedings of the 13th Conference on Artificial Intelligence and Applications (TAAI).

Chiu, T., Fang, D., Chen, J., Wang, Y. and Jeris, C. (2001). “A Robust and Scalable Clustering Algorithm for Mixed Type Attributes in Large Database Environment.” Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD ’01, pp. 263–268.

Coifman, B. (2002). Vehicle reidentification and travel time measurement on congested freeways. Transportation Research Part A: Policy and Practice, 36(10), 899–917.

Fletcher, R. (1971). A Modified Marquardt Subroutine for Nonlinear Least Squares. Rpt.

AERE-R 6799, Harwell.

Fosgerau, M., Fukuda, D. (2012). Valuing travel time variability: Characteristics of the travel time distribution on an urban road. Transportation Research Part C: Emerging Technologies, 24, 83-101.

Gonz ález, M. C, Hidalgo, C. and Barab ási, A. (2008). “Understanding Individual Human Mobility Patterns.” Nature 453 (7196) (June 5), pp. 779–82.

Grynbaum, M. M. (2010). Gridlock May Not Be Constant, but Slow Going Is Here to Stay. New York Times. Retrieved July 31, 2012, from http://www.nytimes.com/2010/03/24/nyregion/24traffic.html?ref=nyregion

Guiliano, G. (2004). "Land Use Impacts of Transportation Investments-Highway and Transit."

Harris, B. (1985). "Urban simulation models in regional science." Journal of Regional Science 25, no. 4, pp. 545-567.

Hasan, S., Choudhury, C. F., Ben-Akiva, M. E., Emmonds, A. (2011). Modeling of Travel Time Variations on Urban Links in London. Transportation Research Record: Journal of the Transportation Research Board, 2260(-1), 1-7.

Herrera, J. C, Work, D., Ban, X., Herring R., Jacobson, Q and Bayen A. M. (2010). Evaluation of Traffic Data Obtained via GPS-Enabled Mobile Phones: The Mobile Century Field Experiment, Transportation Research C: Emerging Technologies 18, 568-583.

Herring, R., Hofleitner, A., Abbeel, P. (2010). Estimating arterial traffic conditions using sparse probe data. Proc. ITS, (September), 19-22.

Hunter, T., Herring, R., Abbeel, P. (2009). Path and travel time inference from GPS probe vehicle data. Neural Information Processing Systems foundation (NIPS), Vancouver, Canada, (December).

Inrix, Inc. http://www.inrix.com

King, David, Peters, J. (2012). Taxicabs for Improved Urban Mobility: Are We Missing an Opportunity? Transportation Research Board 91st Annual Meeting, 19 pages.

Li, R., Rose, G. (2011). Incorporating uncertainty into short-term travel time predictions. Transportation Research Part C: Emerging Technologies, 19(6), 1006-1018.

Li, X., Pan, G., Wu, Z., Qi, G., Li, S., Zhang, D., Zhang, W. and Wang, Z. (2012). "Prediction of urban human mobility using large-scale taxi traces and its applications." Frontiers of Computer Science 6, no. 1, pp. 111-121.

Matlab Parallel Computing Toolbox. Mathworks, Inc.

Muller, P. O (2004). Transportation and Urban Form-Stages in the Spatial Evolution of the American Metropolis. Metropolis.

New York City Taxi and Limousine Commission 2012 Annual Report. 2012.

Nocedal, J., Wright, S. J. (2006). Numerical Optimization. 2nd Edition. Springer, Pages: 258-264.

Oh, J.S., Jayakrishnan, R., Recker, W. (2003). Section travel time estimation from point detection data. In: 82nd Annual Meeting of Transportation Research Board, Washington, DC, USA.

Pan, G., Qi, G., Wu, Z., Zhang, D., and Li, S. (2013). “Land-Use Classification Using Taxi GPS Traces.” IEEE Transactions on Intelligent Transportation Systems 14 (1) (March), pp. 113–123.

Park, D., RILETT, L. R. (1998). Forecasting multiple-period freeway link travel times using modular neural networks. Journal of the Transportation Research Board, (98), 163-

170.

Ratti, C., Pulselli, R. M., Williams, S., and Frenchman, D. (2006). “Mobile Landscapes: Using Location Data from Cell Phones for Urban Analysis.” Environment and Planning B: Planning and Design 33 (5), pp. 727–748.

Reades, J., Calabrese F., Sevtsuk, A. and Ratti, C. (2007). "Cellular census: Explorations in urban data collection." Pervasive Computing, IEEE 6, no. 3, pp. 30-38.

Schaller Consulting (2006). The New York City Taxicab Fact Book, (March). www.schallerconsult.com

Sherali, H. D., Desai, J., Rakha, H. (2006). A discrete optimization approach for locating Automatic Vehicle Identification readers for the provision of roadway travel times. Transportation Research Part B: Methodological, 40(10), 857-871.

SPSS, INC (2001). "The SPSS TwoStep cluster component: A scalable component to segment your customers more effectively."

Szell, M., Sinatra, R., Petri, G., Thurner, S. and Latora, V. (2012). "Understanding mobility in a social petri dish." Scientific reports 2.

Taxicab Rates from New York Taxi & Limousine Commission: http://www.nyc.gov/html/tlc/html/passenger/taxicab_rate.shtml

Veloso, M., Phithakkitnukoon, S. and Bento, C.. "Urban mobility study using taxi traces (2011)." In Proceedings of the 2011 international workshop on Trajectory data mining and analysis. ACM, pp. 23-30.

Wu, C.-H., Ho, J.-M., Lee, D. T. (2004). Travel-Time Prediction with Support Vector Regression. IEEE Transactions on Intelligent Transportation Systems, 5(4), 276-281.

Yang, H, Fung, C.S., Wong, K.I., and Wong, S.C. (2010). “Nonlinear Pricing of Taxi Services.” Transportation Research Part A: Policy and Practice 44 (5) (June), pp.: 337–

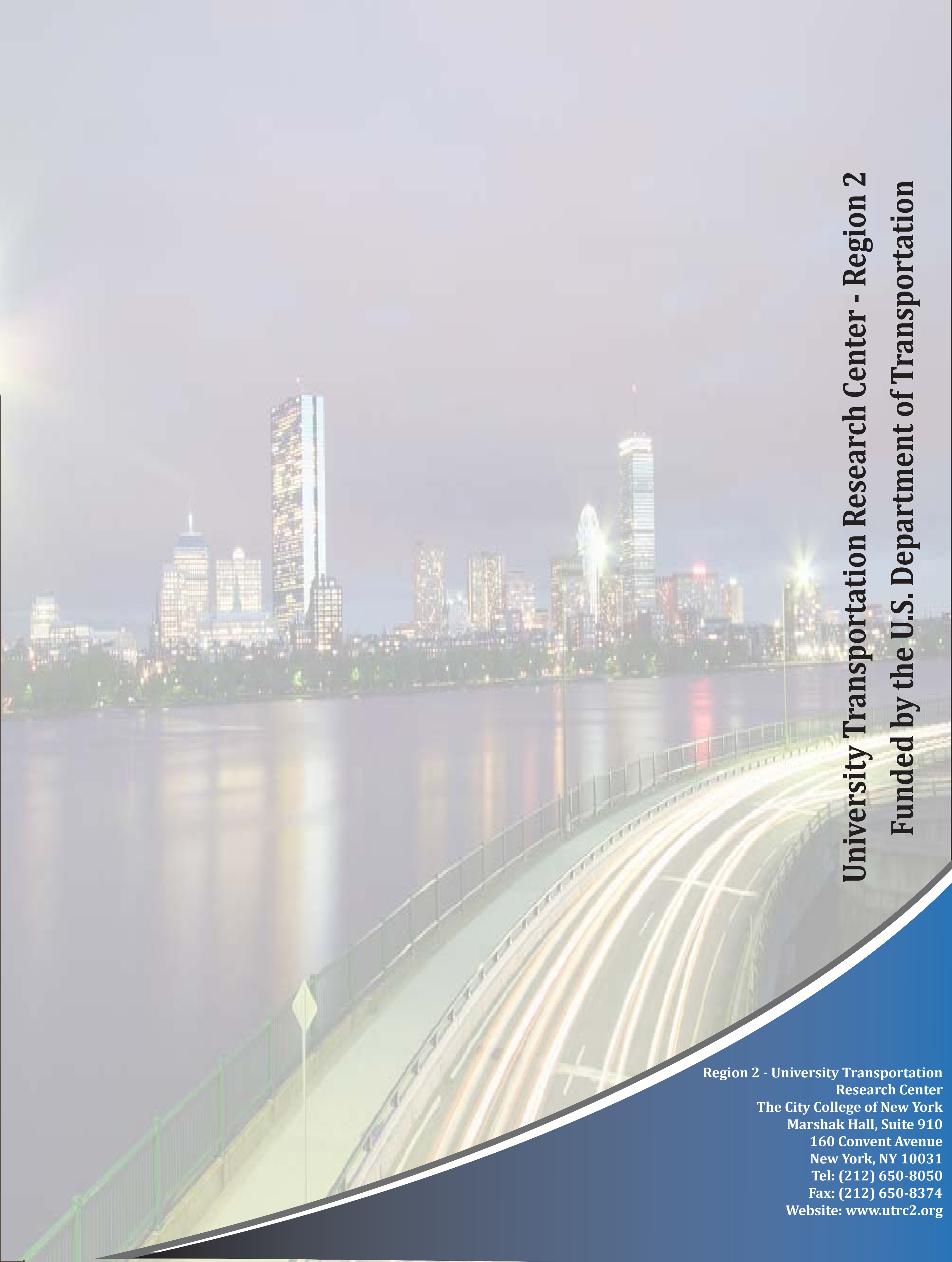
348.

Yen, J. Y. (1971). Finding the K shortest loopless paths in a network. Management Science 17:712–716.

Yeon, J., Elefteriadou, L., Lawphongpanich, S. (2008). Travel time estimation on a freeway using Discrete Time Markov Chains. Transportation Research Part B: Methodological, 42(4), 325-338.

Zhang, X., Rice, J. (2003). Short-term travel time prediction. Transportation Research Part C: Emerging Technologies, 11(3-4), 187-210.

Zheng, F., & Van Zuylen, H. (2012). Urban link travel time estimation based on sparse probe vehicle data. Accepted in Transportation Research Part C: Emerging Technologies, 13 pages. Elsevier Ltd. doi:10.1016/j.trc.2012.04.00



University Transportation Research Center - Region 2

Funded by the U.S. Department of Transportation

Region 2 - University Transportation
Research Center
The City College of New York
Marshak Hall, Suite 910
160 Convent Avenue
New York, NY 10031
Tel: (212) 650-8050
Fax: (212) 650-8374
Website: www.utrc2.org