DATA-DRIVEN MODELING OF TAXI TRIP DEMAND AND SUPPLY IN NEW YORK CITY

by

CI YANG

A Dissertation submitted to the

Graduate School-New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Civil and Environmental Engineering

written under the direction of

Eric J. Gonzales

and approved by

____

____

____

____

New Brunswick, New Jersey

October, 2015

ABSTRACT OF THE DISSERTATION

Data-Driven Modeling of Taxi Trip Demand and Supply in New York City

by CI YANG

Dissertation Director:

Eric J. Gonzales

The taxicab is an important transportation service in New York City (NYC). This dissertation addresses three primary questions related to taxi demand and supply in NYC using innovative data including: a large amount of taxi Global Positioning System (GPS) data that the NYC Taxi & Limousine Commission (TLC) has collected over a two-year period, detailed transit schedule information, and neighborhood characteristics. The three questions are: Is there variability and imbalance between taxi demand and supply across different times of day and different locations in NYC? What factors determine the distribution of demand for taxicab services in NYC by time and by location? How can the imbalance between taxicab demand and supply be identified and quantified in order to guide policies to improve transportation access in NYC?

The hypothesis is that the variability and imbalance between taxi demand and supply at different times of day and locations exists. Neighborhood characteristics are factors that are expected to be related to the distribution of taxi trips.

In the supply analysis, the taxi supply is quantified, an analysis of taxi use every day identifies factors related to the variability of taxi supply over time, and a customer search model was developed to identify areas where drivers of vacant taxis tend to go and to quantify how efficiently drivers of vacant taxis search for their next customer. The model indicates that the NYC taxicabs could be redistributed more efficiently as a system if the taxi drivers are assigned to specific passenger pickups.

In trip generation modeling, six important explanatory variables are identified that influence taxi trips: population, education, supply, income, Transit Access Time (TAT), and employment. More complex models involving count regression and accounting for spatial autocorrelation are then developed to understand the imbalance between taxi demand and supply by controlling for the above-mentioned factors. The errors of the trip generation model provide insights of when and where there is insufficient taxi supply or surplus taxi supply relative to taxi demand.

A case study is introduced to compare the total cost for two modes of transportation (transit and taxi) between NYC Penn Station and three New York area airports (JFK, LGA and EWR). Transit is found to be more cost-effective than taxi for most times of the day.

EXECUTIVE SUMMARY

The taxicab is an important transportation service in New York City (NYC). The number of yellow taxi licenses is 13,362 as of 2014, and yellow taxicabs have provided a market of 172 million trips in 2005, and 175 million trips each year in recent years. This dissertation addresses three primary questions related to taxi demand and supply in NYC using a large amount of taxi Global Positioning System (GPS) data that the NYC Taxi & Limousine Commission (TLC) has collected over a two-year period. The three questions are: Whether there is a variability and imbalance between taxi demand and supply across different times of day and different locations in New York City? What factors determine the distribution of demand for taxicab services in New York City by time and by location? How can the imbalance between taxicab demand and supply be identified and quantified in order to guide policies to improve transportation access in New York City?

The hypothesis posed in this dissertation is that there is variability of taxi demand and supply at different times of day and locations because NYC is a large urban area with various characteristics at different locations. Imbalance between taxi demand and supply is expected to exist because there are imbalances in the places and times that customers tend to request to start trips, where passengers enter taxicabs, and their destinations, where taxis again become available for hire. In theory, taxi drivers have an incentive to maximize revenue, so it is expected that vacant cabs do not reallocate themselves to match demand as efficiently as if there were centrally controlled. Neighborhood characteristics such as population density, income level, employment, land development, and transit accessibility are factors that are expected to be related to the distribution of taxi trips. By controlling those explanatory factors in the trip generation model, the imbalance of taxicab demand and supply is identified and quantified. It is expected that areas in NYC’s Outer Boroughs have insufficient supply of available taxicabs because empty taxi drivers are not likely to go to Outer Boroughs for their next pickup unless they are dropping off a passenger or seeking a customer from one of the airports.

There are three main bodies of data used in this study: 1) A complete collection of GPS taxi data for every taxi trip made in NYC within a 2-year period from December 2008 to November 2009, which is categorized by counts of pickups and drop-offs at each census tract at each hour. 2) Detailed transit schedule information for the same geographic region that is acquired using Google Transit Feed Specification (GTFS) and directional services from the Google Directions API. 3) Supplemental data such as demographic, employment, and land use data, which are expected to include key characteristics of the locations that are associated with the taxi use.

The taxi supply is quantified to prepare a taxi trip generation model. As part of the supply analysis, we identify factors related to the variability of five time series, such as daily taxi trips, shifts, and hours of operation each day. A customer search model was developed to identify areas where drivers of vacant taxis tend to go and to quantify how efficiently drivers of vacant taxis search for their next customer. It was found that areas in Manhattan, certain neighborhoods in the Outer Boroughs, and the airports are the preferred destinations for drivers of vacant taxis seeking their next customer. The model indicates that the NYC taxicabs could be redistributed more efficiently as a system if the taxi drivers are well-informed about where customers demand service and where other vacant taxis are circulating.

Identifying the factors that influence taxi demand is very important for understanding where and when people use taxis. A 10-month subset of taxi data from February 1, 2010, through November 28, 2010 is aggregated and used along with a measure for transit accessibility, Transit Access Time (TAT), as well as demographic, socioeconomic, and employment data to identify the factors that drive taxi demand using linear regression. Six important explanatory variables are identified that influence taxi trips from negative binomial regression: population, education, supply, income, TAT, and employment. More complex models involving spatial autocorrelation are then developed to understand the imbalance between taxi demand and supply by controlling for the above-mentioned factors. A novel method for identifying neighborhoods that are overserved and underserved by taxis is proposed by classifying the relationship of imbalance and residuals from the model into four categories based on their signs to help understand the demand and supply at different neighborhoods. It is discovered that areas in Harlem, Lower East Side of Manhattan, Williamsburg, and Astoria are underserved by taxis while most Outer Boroughs have balanced low demand, and Midtown Manhattan has balanced high demand.

To further illustrate the advantages and limitations of the taxi data, a case study is introduced to compare the total cost for two modes of transportation (transit and taxi) using taxi GPS data and high-resolution transit schedule information. Trips between NYC Penn Station and three New York area airports (JFK, LGA and EWR) at different times of day are used to illustrate the methods. Transit is found to be more cost-effective than taxi for most times of the day if passengers are traveling alone and the value of time is less than $40/hour, except some midnight periods when transit service has long headways that contributes a significant amount of time to waiting or transfers.

The major contributions of this study are the identification of factors that are related to taxi demand and supply and quantification of the imbalance between taxi demand and supply using taxi GPS data in a refined temporal and spatial scale. The errors of the trip generation model provide insights on when and where observed taxi demand rates are greater or less than expected demands based on socio-demographic characteristics and controlling for the supply of empty taxicabs. A novel method is developed to classify the neighborhoods using both the model error and imbalance to interpret the model results. This provides some indication of when and where there is insufficient taxi supply or surplus taxi supply relative to taxi demand. This information provides evidence on a refined geographical scale for policy-makers and regulators such as NYC Taxi and Limousine Commission (TLC) to make decisions. It also provides insights about where and when for-hire taxi providers (e.g., Uber and Lyft) are likely to be competitive and serve a transportation need not met by the conventional street-hail taxi service. Additional analyses using the customer search model and mode cost model provide insights for improving the efficiency of vacant taxicab distribution and comparing travel mode costs.

DEDICATION

To my family, who have so graciously supported me. I Love you all.

ACKNOWLEDGEMENTS

I would like to start by thanking my advisor, Dr. Eric Gonzales, for his continuous support, encouragement, and guidance throughout my academic career from my second year as a Ph.D. student to now, even after he left Rutgers University. Dr. Gonzales has led me to the transportation engineering field with his knowledge, passion, vision and experience. He has also opened more windows for my career path by introducing me to many conferences, awards, and scholarship opportunities. This work would not have been possible without his guidance and tireless commitment to his students, the university, and to the profession.

It has been an honor and a privilege to have Dr. Robert Noland, Dr. Hao Wang and Dr. Peter J. Jin serve on my committee. Thanks to Dr. Noland for his great contribution to Urban Planning education through teaching, publishing, and advising. Thanks to Dr. Wang for always making time to offer guidance and advice. Thanks to Dr. Jin for his endorsement and unique insight and passion for research.

Thanks to Dr. Shawn Taylor for his assistance in proofreading this dissertation. My appreciation is also extended to faculties in Civil Engineering department: Dr. Husam Najm, Dr. Perumalsamy Balaguru, Dr. Qizhong (George) Guo, Dr. Jie Gong, Dr. Monica Mazurek, Dr. Trefor Williams, and Dr. Ali Maher, and the departmental staff: Gina

Cullari, Linda Szary, Azam Kalantari, and Pat Szary for making the studying and workings of this institution seem simple and flawless.

I also extend my appreciation to the big family at Transportation Safety Resources Center (TSRC), Mitra Neshatfar, Andrew Kaplan, Jennifer Buison, Aimee Jefferson, Joseph Weiss, and Sally Karasov for providing a great working environment for my last year of the study, and I treasure the moments we have spent in learning, creating exciting moments and developing friendship.

Thanks to Dr. Kaan Ozbay, and Ender Faruk Morgul for their contribution and incredible effort to the New York City Taxi Project and for their friendship. Thanks also to my friends and colleagues: Dr. Dan Su, Rooholamin (Amin) Shabihkhani, Dr. Hong Yang, Farbod Farzan, Kaveh Gharyeh, Mengyang Guo, Yi Yu, Peng Lou, Kun Xie, Mahour Rahimi, and Mahyar Amirgholy for their assistance with graduate student life, research work, and friendship throughout the years.

Deep appreciation and gratitude is expressed to my family for their support and encouragement during my academic pursuits.

Finally, I would like to thank my husband Zhou Dai, my brother Gang Yang, and my daughter Angela Dai for their love, support, and patience without which none of this would be possible.

PREFACE

The work conducted in this dissertation has been presented and published in several conferences and journals. Below is the list of publication derived from this dissertation with corresponding chapter numbers:

Chapter 5

- Yang, C., & Gonzales, E. J., (2014). Modeling Taxi Trip Demand by Time of Day in New York City. In *Transportation Research Record*, No. 2429, pp. 110-120.
- Gonzales, E., Yang, C., Morgul, E., & Ozbay, K. (2014). *Modeling Taxi Demand with GPS Data from Taxis and Transit*. Mineta National Transit Research Consortium (MNTRC) Report 12-16 (Released on July 21, 2014).
- Yang, C., & Gonzales, E. J., (2014). *Modeling Taxi Demand and Supply in New York City Using Large-Scale Taxi GPS Data*. In Workshops on Big Data and Urban Informatics (BDUIC), Chicago, IL.

Chapter 6

- Yang, C., Morgul, E., Gonzales, E. J., and Ozbay, K. (2014). Comparison of Mode Cost by Time of Day for Non-driving Airport Trips to and from New

York City's Pennsylvania Station. In *Transportation Research Record*, No. 2449, pp. 34-44.

- Gonzales, E., Yang, C., Morgul, E., & Ozbay, K. (2014). *Modeling Taxi Demand with GPS Data from Taxis and Transit.* Mineta National Transit Research Consortium (MNTRC) Report 12-16 (Released on July 21, 2014).

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Taxi Market

Taxis are an important mode of transportation in urban areas, providing service between the locations of each passenger's choice. There are many cities that have taxi services, from large cities (e.g., New York City, Washington D.C., and Boston) to small towns (e.g., New Brunswick and New Jersey). New York City (NYC) has more than 8 million inhabitants (U.S. Census 2010), and it has the greatest number of taxi medallions of any city in the United States. In NYC, taxis serve an especially important role in the transportation system, because the 172 million trips completed by taxis in 2005 made up 11% of trips in the city (Schaller, 2006). According to the 2014 Taxi Fact Book (Bloomberg et al., 2014), the annual number of taxi trips has increased to 175 million, and the system has transported 236 million passengers each year from 2008 through 2013. This demand consists of trips by residents, people who work in the city, tourists, and people with disabilities.

Unlike some smaller cities that have an unregulated taxi industry (e.g., Portland, Oregon), the taxi industry in NYC has controlled entry. There are a fixed number of medallions for taxicabs, also called yellow cabs, that are issued and managed by the NYC

Taxi & Limousine Commission (TLC), which means that the number of taxicabs operating in NYC (within the 5 boroughs) is limited. As of 2014, there were 13,437 currently licensed medallions in NYC (Bloomberg et al., 2014). The TLC also regulates and licenses for-hire vehicles, known as "livery cabs" which are not allowed to pick up street hails in NYC. Effective in 2012, a new class of licenses was created to provide legal and yellow-caliber taxi service to the Outer Boroughs. They are known as green cabs or Boro taxis, and are allowed to pick up street hail customers north of East 96th Street and West 110th Street in Manhattan and in all neighborhoods outside Manhattan except at the airports. They also respond to dispatch services in those areas including the airports. About 6,000 licenses for green cabs were sold out in 2013, and the number is expected to grow to 18,000 by 2017 (Bloomberg et al., 2014).

NYC yellow taxis have been facing new challenges since app-based for-hire taxi services (commonly known as "ride-share") such as Uber became active four years ago. They operate more like unlicensed taxis or limo services than as a way to share rides with people having common destinations. There have been arguments that crowdsourcing or app-based for-hire services[1] should be more appropriate to describe Uber and similar services (Lawler, 2013). The number of Uber cars in New York City is more than 14,000, competing with only about 13,000 yellow cabs. However, some feel that Uber is violating existing taxi regulations. NYC cabs claim that Uber is picking up passengers in exclusive taxi spots. According to the news from March 31, 2015, NYC yellow cabs have been filing law suits against Uber (Harshbarger, 2015). The impacts of the emerging app-based for-hire vehicles on NYC taxi industry are motivation for my research in studying the demand and supply of yellow taxis. Understanding the demand

[1] In this dissertation, crowdsourcing or app-based for-hire services is used to describe these services.

and supply of yellow taxis may provide insights to guide future planning and regulations for all types of taxis in NYC.

Although the number of medallions fixes the supply of yellow and green cabs, the demand for taxi trips varies at different times and locations and also from day to day. The supply of vacant taxis also varies by time, location, and day. The demand and supply of taxi trips affect each other, because the demand attracts vacant taxicabs, and the supply of vacant taxis affects how easily a customer can find a ride. To understand and then forecast the demand with respect to both taxi drivers and passengers, it is necessary to model taxi demand while accounting for the availability of vacant taxis to pick up customers.

Demand and supply are fundamental concepts in economics. Demand refers to the quantity of a product or service that is desired by buyers while supply refers to the quantity of a product or service that is provided by the sellers. In a free market, the law of demand indicates that if all factors remain the same, a decrease of the price of the product or service will result in more quantity demanded. Supply is the marginal cost of supplying one additional unit at equilibrium. The market price is a reflection of demand and supply. There is a point when the economy reaches equilibrium if demand equals supply. At this point, the allocation of goods or services is at their most efficient because the amount of goods or services supplied is exactly the same as the amount of goods or services being demanded (Mankiw, 1998).

The taxi market in New York City is not a free market but a market with government policy intervention. Both the supply of taxicabs (as controlled by the number of medallions) and the fare structure and fare amount are determined by the TLC.

Although taxi drivers can choose which hours to drive and in which neighborhoods to search for passengers, the constraints on the supply limit the ability of the system to reach an efficient equilibrium with passenger demand. The taxi market is not a perfect market. Policy changes or interventions affect the taxi market to adjust the supply or demand and ideally correct market inefficiencies. However, it is a challenge to determine whether regulations are effective. Therefore the purpose of this study is to find out whether the supply meets the demand in specific locations and times of day in NYC and to provide relevant policy insights for decision makers.

The general hypothesis in this study is that there are locations and times where there is a mismatch between taxi demand and supply in New York City, and this can be systematically analyzed based on the characteristics of the neighborhoods and the patterns of taxi pickups and drop-offs. The first question is whether the taxi supply meets the demand. Our first hypothesis is that there is variability of taxi demand and supply at different times of day and locations because NYC is a large urban area with various neighborhood and transportation system characteristics. Some places may have insufficient supply and other places may have excessive supply relative to the demand. Taxi drivers have an incentive to maximize their own revenue, so vacant cabs are not expected to reallocate themselves to demand as efficiently as if they had more information, like those provided by centrally coordinated dispatch. The number of trips entering a zone can be very different from the number of trips exiting a zone, so vacant taxicabs need to travel to other parts of the city in search of their next customer. Neighborhood characteristics such as population density, income level, employment, land development, and transit accessibility are potential factors that are related to the distribution of taxi trips. By controlling for those explanatory factors, the imbalance of taxicab demand and supply can be identified and quantified. We expect to see that areas in the Outer Boroughs may have insufficient supply because empty taxi drivers are not likely to circulate in Outer Boroughs for their next customer except at airports.

There are not many studies in the literature on the topic of travel demand modeling for taxis, and there are none that focus on New York City. NYC is a large city with an extensive road network and complex traffic conditions. The scope of taxi demand analysis for this study is on the trip generation and selective modal split in modeling taxi demand. The demand and supply of taxis in New York City is of more interest, because the relationship between them is very important to understand the availability of taxis and how productive and efficient the yellow taxi system is in NYC. Taxi demand and supply modeling can help identify factors related to the distribution of taxicab services, identify locations with insufficient or excessive taxi supply, and quantify the imbalance between taxi demand and supply by time and space.

1.2 Problem Statement

The number of taxi pickups in an area represents the realized demand in that area. The number of taxi drop-offs in a region is the immediate supply of vacant taxis in that region. However, the taxi pickups and drop-offs vary at different locations of NYC, at different times of day, and at different times of the year. The problem with a mismatch in demand and supply justifies a need for modeling taxi demand and supply. We now introduce three problems that are of interest for this study: locational problem, temporal problem, and directional problem.

1.2.1 Locational Problem

There are five boroughs in New York City: Bronx, Brooklyn, Manhattan, Queens and Staten Island. Manhattan has the densest population among the five boroughs. It is the center where most of the business and tourism activities occur. The Outer Boroughs are more residential, but they also include neighborhoods with substantial employment and commercial activity. The variation in levels of activity across different locations is reflected in variations in demand for taxi service.

As shown in Figure 1.1, there are many more taxi pickups and drop-offs in Manhattan than in the other four boroughs except at the airports (John F. Kennedy International Airport and LaGuardia Airport). This is predictable, because Manhattan contains a denser concentration of activities than other parts of the city. However, the numbers of pickups and drop-offs are not equal at many locations, indicating a locational imbalance between the demand and supply. Figure 1.2 indicates that there are more pickups than drop-offs in Midtown Manhattan, and more drop-offs than pickups in residential areas in the Outer-Boroughs and outlying neighborhoods of Manhattan based on the total number of pickups and drop-offs in a 10-month period. However, activity patterns vary dramatically by time-of-day, so it is necessary to look at temporal variation as well. Understanding the imbalance among different locations will provide insights in identifying areas with excessive supply relative to demand and areas with insufficient supply, thus helping make policy decisions to improve service in those identified areas.

Problem: Where is there an imbalance between demand and supply?

Example problems include identifying the areas with insufficient supply such as northern Manhattan above 96th Street, where green taxicabs were introduced in 2013 to meet unsatisfied taxi demand. Additional questions related to this problem are:

1) What factors are related to the generation of taxi demand and supply?
2) Can a taxi demand model provide additional insights about where demand or supply appears to be over- or under-represented?
3) How does taxi demand compare with transit accessibility at different locations?

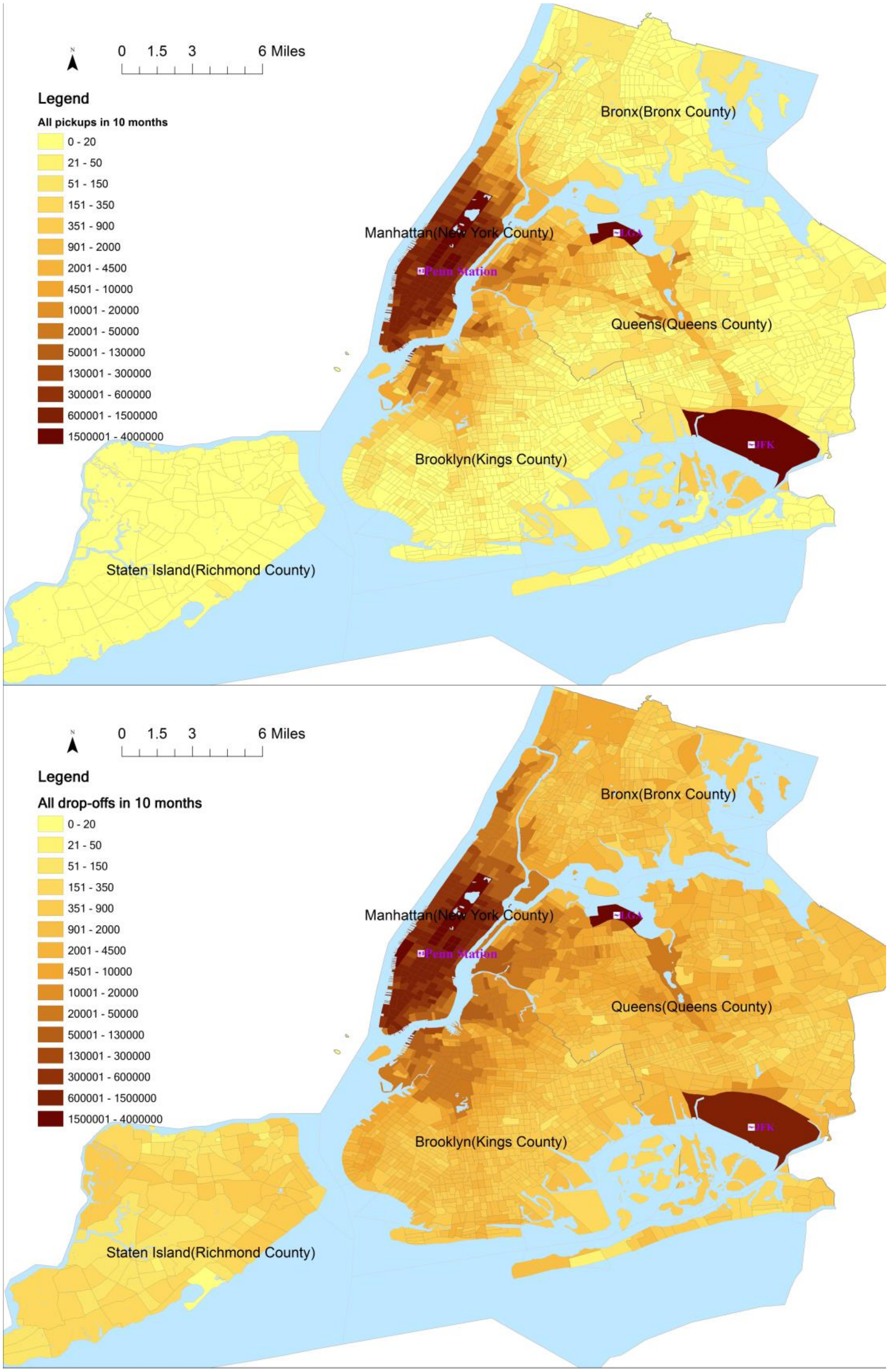


Figure 1.1 All Taxi Pickups and Drop-Offs During 10-months Period in 2010.

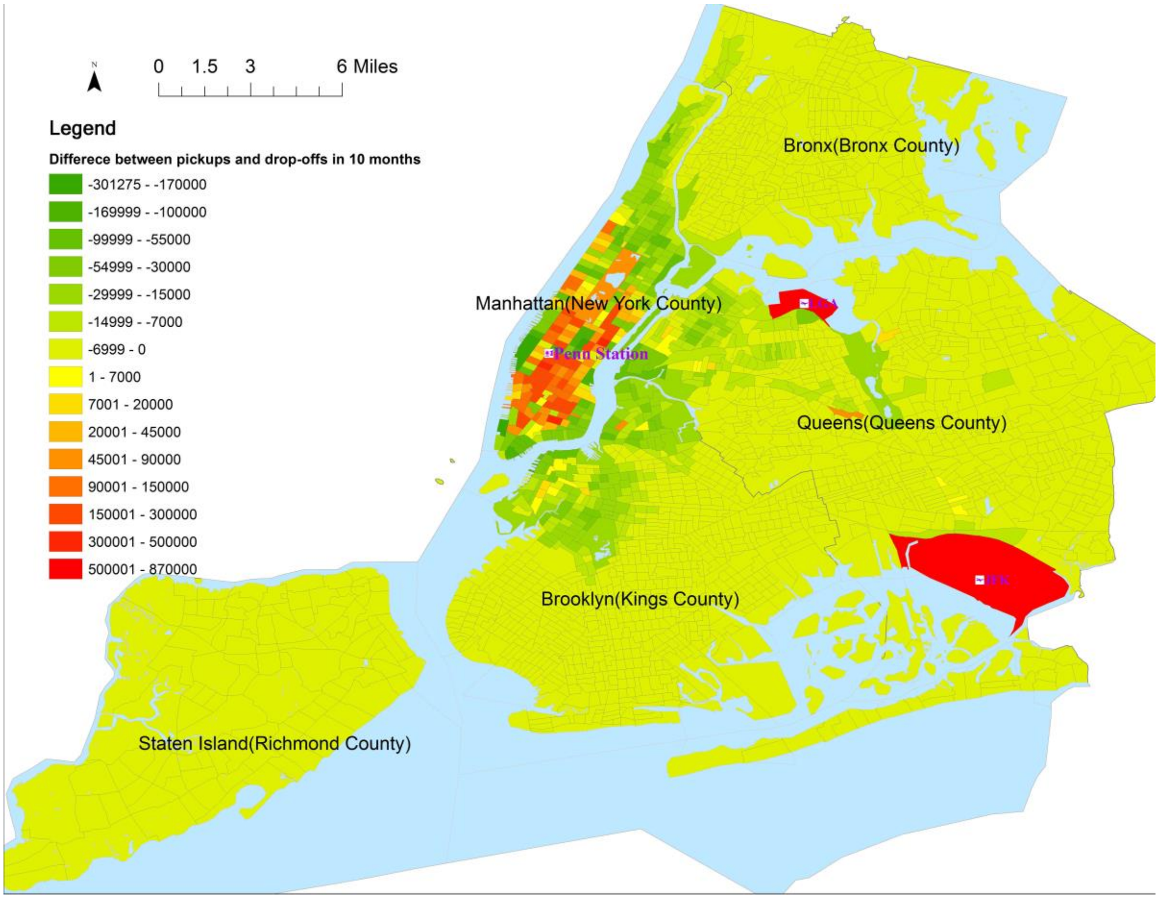


Figure 1.2 Difference (Difference = Pickups – Drop-offs) Between Pickups and Drop-Offs During 10-month Period in 2010.

1.2.2 Temporal Problem

In addition to variation in demand and supply across locations, there is also variation across time. The taxi demand (total number of pickups) and supply (total number of drop-offs) varies by time of day. Figure 1.3 shows the hourly variation of city-wide taxi demand. The figure indicates that on average, there are more taxi pickups during waking hours, from 8 a.m. to 12 p.m., compared to taxi pickups from mid-night to early morning period (1 a.m. to 7 a.m.).

Figure 1.4 shows the pickups for midnight, morning peak, noon, and afternoon peak in most of the census tracts in New York City. This gives an overview of how the

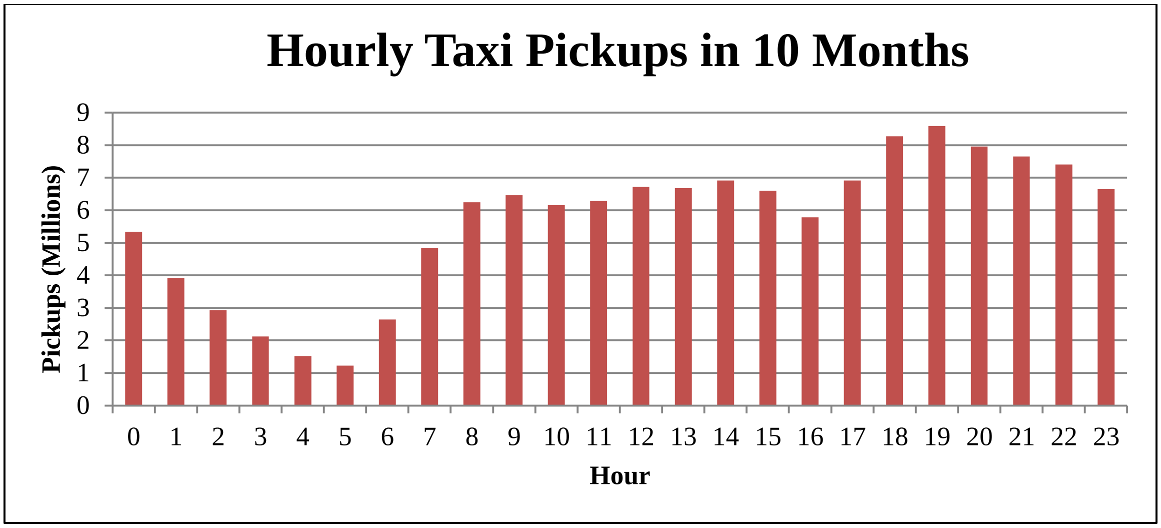


Figure 1.3 Hourly Taxi Pickups During 10-months Period in 2010.

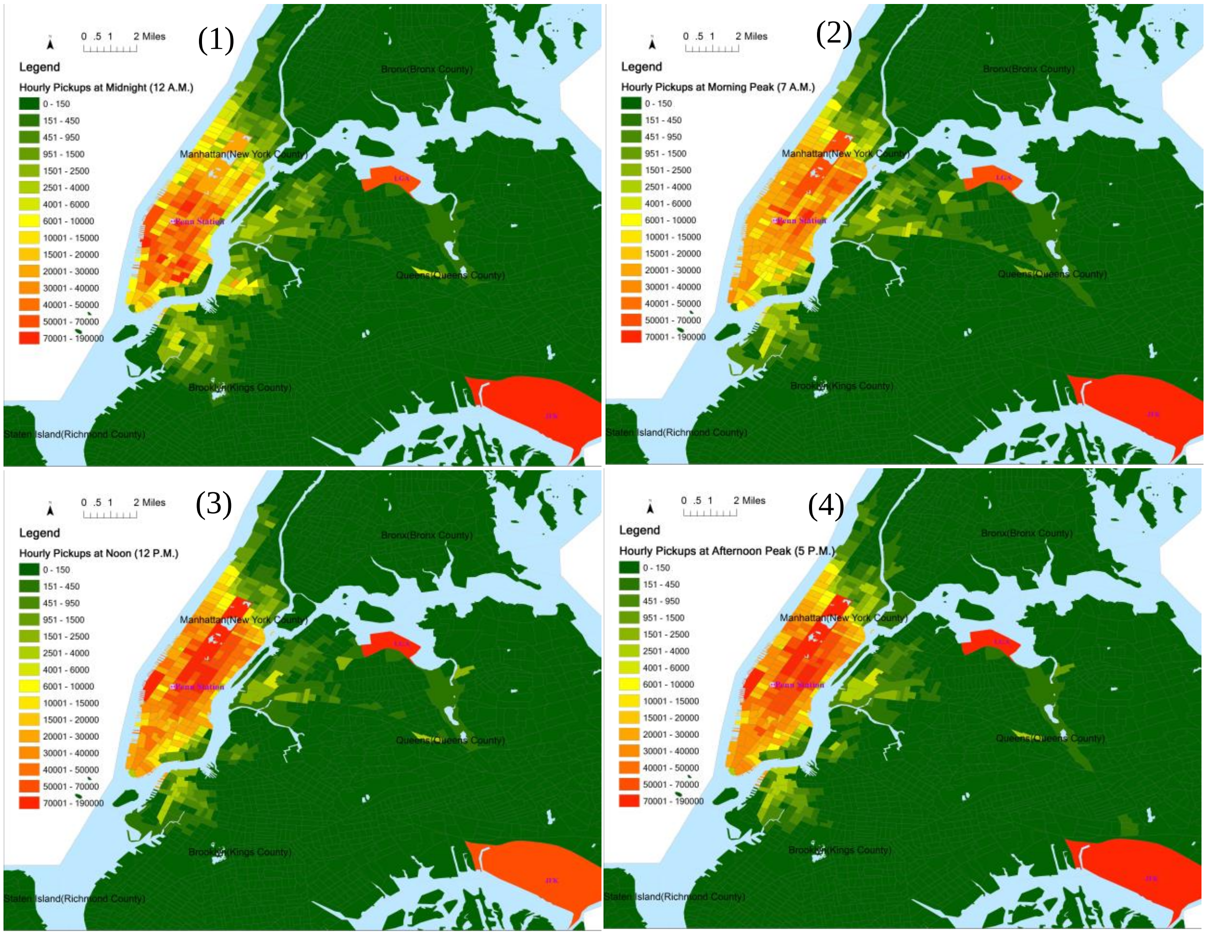


Figure 1.4 Hourly Pickups at (1) Midnight 12 A.M.; (2) Morning Peak 7 A.M.; (3) Noon 12 P.M.; (4) Afternoon Peak 5 P.M.

number of taxi pickups changes both spatially and temporally. The variation suggests that most of the census tracts in Manhattan, its surrounding neighborhood, and the airports have higher temporal variation compared to residential areas in the Outer Boroughs. During the day time, the higher demand areas are concentrated in Central and North Manhattan, while high demand areas move to Lower Manhattan.

Problem: How does the taxi demand and supply vary by time of day?

An example of a policy intervention to address the temporal variation of demand and supply would be to set a higher fare during the mid-night hours when demand exceeds supply to ensure that taxi drivers have an incentive to provide service even when fewer people use taxi services during that time period. However, demand and supply, as well as the imbalance between demand and supply, vary by both times of the day and locations, as well as the imbalance between demand and supply. It is important to understand how the spatial mismatch between demand and supply evolves over time, because there are different factors that drive taxi demand at different times of day. More questions related to this problem include:

1. What factors influence the spatial difference between demand and supply?
2. How should a model be built to consider both locational imbalance and temporal variation?
3. How are the observed taxi demand and supply different from demand and supply estimated by controlling for those factors with spatial correlations?

1.2.3 Directional Problem

The third problem is to consider the directional decisions that drivers of vacant taxis make in search of their next customer. For example, some areas have more pickups than drop-offs, and it would be useful to know where a vacant taxi came from preceding a pickup and where a vacant taxi goes following a drop-off. Do drivers of vacant taxis circulate in search of customers in or near the same area after they make a drop-off or do they travel to a more distant location in search of customers?

The relevant research task is to figure out how taxi drivers redistribute themselves to find customers where the demand is located. Coordinated centralized planning could identify the shortest redistribution distance to do this, but taxi drivers act more or less independently, seeking to maximize their own revenue. An analysis of existing and optimized redistribution behavior can reveal the effect that individual behavior has on the efficiency of the overall taxi system. Drivers of vacant taxis tend to make decisions on where to go in order to search for their customers based on their experiences and knowledge of the area. The value of an information system to facilitate matching customer demands with vacant drivers would be an improvement in matching vacant cabs to customers. The benefit can be quantified by comparing the performance of the current system to one in which coordinated centralized planning is used to identify the shortest redistribution distance.

Problem: How efficiently do empty taxis reallocate themselves to locations where customers demand service?

An example of this reallocation problem is a taxi that picks up a customer in Midtown Manhattan and drops the passenger off in a residential area that has low demand for taxi service. The taxi driver may drive to the nearest neighborhood with known demand or go back to Manhattan in search of the next customer rather than circling streets in an area where there is a low probability of finding a street-hail customer. This creates an imbalance between the two directions: one from low demand areas to high demand neighborhoods and the other from high demand neighborhoods to low demand areas. More questions related to this problem include:

- How likely is a vacant taxi to find a customer in the same area where the previous passenger was dropped off?
- Where will a taxi driver tend to go to search for their next customer if they make a drop-off outside Manhattan?
- From a systematic perspective, do drivers make efficient choices for their next pickups.

1.3 Modeling Approach

Due to the magnitude of taxi trip information from the taxi GPS dataset, this source is an example of *big data*. Big Data is usually generated automatically from electronic devices at high frequencies, providing more complete information than conventional sampled data. For example, the comprehensive set of taxi trip data provide complete spatial and temporal coverage of NYC, whereas a subsample of the tax trips may not include information about travel to neighborhoods with lower demand. Big data also helps improve data accuracy and quality, because it provides plentiful observations even though a number of faulty data points are eliminated. One of the challenges in dealing with big data is that they are too large to process using conventional tools. Using big data to develop useful models for taxi demand requires developing procedures to clean and process the information so that it can be organized in a useful way. Therefore the data sources and data treatment are considered explicitly before we perform additional analyses.

1.3.1 Supply: Efficiency of Vacant Taxis

The second analysis is conducted to define and quantify taxi supply, relate taxi supply to influential factors, and then investigate when and where drivers of vacant taxis go in reality compared with optimized central coordination. It is interesting to observe the popular places that a taxi driver tends to go after dropping off a passenger in an area with low demand. Vacant taxis tend to go to popular areas where the likelihood of finding a customer is high. It is particularly interesting to understand how efficiently taxi drivers relocate themselves after dropping off a passenger. A customer search model is proposed in this study to quantify how efficiently a taxi driver locates a passenger while driving empty.

As part of the supply analysis, taxi supply is quantified using different variables and factors related to taxi supply are identified in the analysis of variation in taxi use over time using variables chosen to quantify taxi supply, for example, the taxi trips per day. The goal of the customer search model is to identify locations that a vacant taxi driver would go after dropping off a customer in an outer-borough, and most importantly to understand how taxi drivers relocate themselves if they are well-informed about where taxi demand and other vacant taxis are located. A linear programming algorithm is proposed to perform the customer search modeling. The distance and time elapse before their next pickups are also analyzed to help rank census tracts by popularity both in the original OD-table and the modeled OD-table. The analysis of the efficiency of vacant taxis is intended to identify the optimal pattern of relocation for taxi drivers after they drop off a passenger outside Manhattan based on least vehicle-miles-traveled (VMT) and largest revenue earned.

1.3.2 Demand: Trip Generation

In order to answer the problems stated in previous section, the primary objective of this study is to identify the factors that drive demand for taxis, accounting for taxi supply, the effect of transit service availability, and the socioeconomic characteristics of the neighborhoods in which people travel. The approach that is taken is to develop demand models that acknowledge the distribution of taxi demand in space and time so that it is possible to identify how taxi demand varies from neighborhood to neighborhood and how the demand evolves over the course of a day.

A trip generation model is developed to identify the factors that determine the number of taxi pickups and drop-offs that are generated in a neighborhood. The model follows a hybrid approach, classifying taxi records by hour of the day and then using regression to model the number of pickups and drop-offs within each census tract and hour of the day. The result is a set of models that provides predictive capability and makes distinctions between location and time of day.

The goal of the trip generation model is to identify the factors that affect the demand for taxi trips at the level of a census tract, so extensive data from the U.S. Census Bureau, including characteristics related to population, age, education, income, and employment by industry sector are all considered as possible determinants of taxi demand. An additional goal is to identify what effect, if any, accessibility of transit has on demand for taxis in a neighborhood, so transit schedules are also considered for this part of the analysis. Further analysis is provided to analyze when and where there is excessive supply or insufficient supply of taxis by comparing model residuals and differences between demand and supply. This addresses both the locational and spatial problems.

1.3.3 Mode Cost Model

A third analysis is conducted to investigate the competition between taxi and transit for specific origin-destination pairs. This analysis also reveals some of the limitations of the available data. The goal is to determine how mode cost is likely to change over the course of a day. Traffic congestion affects the speed and price of taxis, and changes in transit service headways affect the amount of time that travelers can expect to wait for service. For this part of the analysis, trips between Penn Station and each of the three major airports in the NYC area are considered; John F. Kennedy International Airport (JFK), LaGuardia Airport (LGA), and Newark Liberty International Airport (EWR). Trips to and from airports constitute an important market for taxis and provide a case study for demonstrating a general modeling methodology that could be extended across other origin-destination pairs. The results provide insights about how the operating characteristics of taxis and transit as well as traveler preferences, such as value of travel time and the number of passengers, affect the tendency of people to choose one mode or the other. Since the dataset does not include the precise routing of each taxicab, we can only use origins and destinations in estimating the travel time and travel distance. For less popular origins and destinations, we may not have enough data points to make a complete comparison between the uses of the modes.

1.4 Dissertation Organization

The dissertation is organized in chapters that present the literature, data sources, and modeling efforts undertaken to address the research questions introduced in the preceding sections. A review of the existing literature on taxi markets and demand is presented in Chapter 2. Data sources and a description of how the data is processed are provided in Chapter 3. Taxi supply and the modeling of taxi supply and the efficiency of matching vacant taxis to customers is introduced in Chapter 4. Chapter 5 then develops trip generation models to estimate taxi demand at different times and locations using the NYC taxi GPS data with linear regression and count regression techniques. Chapter 6 discusses the advantages and limitations of the work that can be done with this big data from taxi GPS using a mode cost analysis. Finally, Chapter 7 concludes the dissertation and describes directions for future work.

CHAPTER 2

LITERATURE REVIEW

Taxis are an important transport mode in urban areas, because first they provide passengers with convenient, comfortable and prompt trips (Lin, 2012) and second, taxis can be used as a complement or a substitute to mass transit systems (Austin, 2011), especially in regions with less transit access. It was stated in U.S. government surveys (Bureau of Transportation Statistics, 2003) that 12% of Americans used taxi services in the previous month, and their trips generated $3.7 billion revenue a year from taxicab fares in 2003 (Li, 2006). Three types of taxi markets are generally discussed in literature: (1) the dispatch (telephone order) market, (2) the cab stand market, also known as rank place market, and (3) the hail market, also known as cruising cabs (Dempsey, 1996; Frankena & Pautlet, 1984; Schaller, 2007; Salanova, 2011). A similar on-demand service is paratransit, which includes Dial-A-Ride services that provides demand responsive services that may serve single or multiple customers in each vehicle (Lin, 2012).

In New York City, there were two types of taxicabs before 2014: yellow cabs and For-Hire-Vehicles (FHV), which include black cars, community car services and luxury limousines (NYC TLC, 2015). Effective in March 2014, a new type of taxicab called Street Hail Livery (SHL), also known as ‘Green Taxis’ or ‘Boro Taxis’, became available to pick up passengers in northern Manhattan and the Outer Boroughs (excluding the airports). NYC yellow taxis (or cabs) can be hailed on streets or stands at locations such as hotels, restaurants, and airports. The FHV cabs complement yellow cabs by providing the dispatch taxi market in NYC before the Green Taxis entered the market. According to Schaller's New York City Taxi Fact Book (2006), the hailing/standing yellow taxi market in NYC has a restricted entry by limiting the number of taxi licenses (medallions) to 12,779 in 2005, and fixed fare level and structure. Yellow taxi ridership was 241 million and generated total revenue of $1.83 billion in 2005. The annual number of taxi trips has increased to 175 million by 2013, and the yellow cab system has transported 236 million passengers each year from 2008 through 2013 (Bloomberg et al., 2014). Yellow cabs transport 11% of fare-paying passengers traveling by taxi, bus, subway, car services or black cabs in NYC, and 25% of those trips occur in Manhattan. New York City's fare is the 11th highest among the 14 U.S. cities with 1,300 or more metered taxicabs (Schaller, 2006). As a result of the important role taxis have played in NYC, it is necessary to explore the literature to understand what studies have been done on taxi demand and supply.

Modeling taxi demand is very important to understand how taxi trips are generated and distributed by time (e.g., time of day or time of year) and location, and how people choose taxis as their transportation mode. The insights we gain from taxi demand modeling are useful for planners and policy makers who want to manage taxi services effectively to meet demands. There is a large body of literature on the economic theories, models, and arguments related to taxi demand and taxi regulation. Due to the lack of data there is seldom enough empirical transportation work to support those arguments or provide evidence on a refined geographical scale to policy-makers and regulators for making decisions (Coffmann, 1977). With the development of computer technology, Global Positioning System (GPS) devices have emerged as a way of tracking vehicles including taxis, making it easy to monitor traffic, manage the taxi industry, and gather valuable geospatial data. Graphic technologies have also enhanced engineers' ability to process data in a refined geographic scale using tools such as Geographic Information Systems (GIS). Those data and tools, if available, make it possible to provide empirical analysis of taxi demand in a large urban network.

In transportation engineering, there are four steps that have been widely used in modeling large scale travel demand: trip generation, trip distribution, modal split (modal share), and trip assignment (Garber & Hoel, 2014). This dissertation focuses on the trip generation and modal split of taxi demand. Additional analysis includes an investigation of taxi supply distribution. The contribution of this dissertation is a detailed trip generation model, and some additional analyses are performed that provide insights for mode cost comparison between taxis and transit. The taxi supply distribution model allocates the empty taxis to locations where the next customer is picked up and develops a trip matrix using a combination of origins (taxi drop-offs) and destinations (next taxi pickups) to display the number of empty trips going from one origin to another destination.

The literature related to taxis is extensive. Most of the literature contributes to policies and regulations on taxi markets and taxi airport ground access. Some of these studies are governmental project reports that serve the purpose of providing insights for airport managers, regulators, and policy makers. An overview of the taxi literature helps to show where there are gaps in the existing literature that this dissertation addresses. The following sections include discussions of studies related to modeling taxi demand and supply (productivity). Those studies include travel demand modeling (trip generation models), taxi customer search modeling, and mode choice modeling. Studies on modeling travel demand on transport modes other than taxis are also helpful as a foundation of methodologies for this dissertation.

2.1 An Overview of Taxi Markets

2.1.1 Taxi Regulation/Deregulation

Historical literature related to taxis mostly covers taxi policies and regulations that can influence taxi markets. Transportation is a big part of the economy on both global and local levels; it encompasses 10% of the U.S. economy (> $1 trillion)[2] . Understanding the economic features of the taxi industry is important to understand the behaviors of taxi drivers and how that impacts the demand of taxi trips. It is also helpful to understand when and where taxi trips are distributed, and how and why people choose taxis instead of other transportation modes.

Edwin Chadwick (1859) first proposed economic regulation of taxi markets and recommended a monopoly of the market in London. As early as 1961, Turvey (1961) studied the economic features of the London Cab Trade by introducing the history of the London taxi industry. There are three groups of people that are influenced by the regulations of the cab trade: the proprietor (similar to taxi companies with fleets in the U.S.), the owner-driver, and the journeymen. The first two groups own taxis and some of

[2] The data are available from Bureau of Transportation Statistics: http://www.rita.dot.gov/bts/programs/freight_transportation/html/transportation.html

them drive taxis themselves. The journeymen are the customers who hire taxis from the proprietor and owner-driver. Journeymen are responsible for the running cost of the taxi and they earn about 39% of time and distance fares plus all extras including tips and commission paid by some hotels and night-clubs in the 1960s (Turvey, 1961).

Turvey (1961) also raises three major issues that are related to the regulation and deregulation of trade by authorities, which have remained major controversial topics for economists, planners, and taxi industry policy-makers over the last century. These three issues are: "Conditions of Fitness," regulation of fares, and limitation on taxi licenses. "Condition of Fitness" refers to setting requirements for vehicles that may be used as licensed taxi cabs in London. Turvey (1961) argued that the "Conditions of Fitness" should be relaxed so that a regular saloon car or a smaller car can be used as cabs to reduce fuel consumption and thus reduce the cost of cabs. However, this raises a question of whether to allow the cheaper and smaller car to charge lower fares. If one regulates taxi fares, the value of the fixed price is the key to maintaining the taxi market. "Thus the choice lies between low fares and low availability, on the one hand, and high fares and high availability, on the other" (Turvey, 1961), which indicates that there is a compromise between the cost and convenience of taxis as a form of transport. The limitation of taxi licenses plays a key role in the market value of a taxi medallion, and restriction on the number of licenses increases the value of a medallion but also the inconvenience for the traveling public.

Following his lead, there is intense focus on whether taxi markets should be regulated through monopoly controls, entry conditions, and fare establishment (both regulated fare level and fare structure). There are a number of case studies outside of

London that are related to theories and discussions on taxi regulation or deregulation addressing New York City (Orr, 1969; Shreiber, 1975; Shreiber 1977; Coffman, 1977; Schaller, 1993), Washington D. C. and Baltimore (Eckert, 1979), Los Angeles, Houston, Chicago, St. Louis, Boston, and Minneapolis (Eckert, 1979; Dempsey 1996) in the U.S. Other studies have addressed markets internationally in Japan (Otsuka, 1989), New Zealand (Gaunt, 1995), Sweden (Marell, 2002), and Ireland (Barrett, 2003).

People who support regulation of the taxi industry state that regulation ensures the price of taxis, thus preventing too many private vehicles from operating on the road, thereby reducing congestion and air pollution (Turvey, 1961), protecting public transit systems and medallion prices (Shreiber, 1975; Shreiber, 1977; Frankena & Pautler, 1984), providing public safety and consumer protection, and improving economic performance of the taxi market (Coffman, 1977; Schroeter, 1983; ITS 1992). Orr (1969) first raises the question of whether there should be a fixed number of licenses in NYC. The taxicab industry in NYC was unregulated until the enactment of the Haas Law in 1937, which regulates both the taxi fare and the number of medallions (i.e., the licenses to operate a cab). Shreiber (1975) compared NYC’s free taxi market before 1937 and the regulated market after 1937 and concluded that there are external costs of taxicabs associated with traffic congestion, air pollution, and social cost. Shreiber (1975) also concluded that in order to prevent the surplus of cabs created by the free market, the usage of taxicabs should be restricted. Teal (1987) argues that the deregulation of the taxicab industry is less favorable to both producers (by reducing the medallion value) and consumers (by increasing trip refusal by cab drivers). NYC has been used as a successful example in regulating the taxi industry (Shreiber, 1975).

People who favor deregulation argue that regulation also has costs and that deregulation benefits consumers (Beesley, 1973; Gaunt, 1996). They also challenge the notion that price regulation improves resource allocation (Coffmann, 1977) and that an unregulated taxicab industry will be inefficient (Williams, 1980). The New Zealand case studies have shown that a deregulated taxi market is competitive (Gaunt, 1995), and Barrett (2003) argues in favor of deregulation in a case study in Ireland. In Brisbane, it is estimated that deregulation benefited consumers by $1.47 per taxi ride but created an increase in the number of taxicabs by 15% (Gaunt, 1996). Recent work on Sydney, Australia, shows that unrestricted entry benefits the city and there is a high cost of using regulation (Abelson, 2010).

Dempsey's (1996) case study on the size and structure of the taxi industry in U.S. cities claims that the entry-free taxi market cannot perform like a regular free market because it is imperfect. Those imperfections include: 1) the absence of a competitive market for both cabstand and cruising cabs; 2) imperfect information and transaction costs; and 3) externalities. Regulations on the price and quality of taxi services are required to help diminish the effects of free-entry to the market because of those imperfections. Also, unregulated taxis without regular safety examinations or restrictions are also threats to public safety. Recent articles thus recommended using re-regulation instead of de-regulation to allow free entry to the market but retain some restrictions on taxi vehicles (Dempsey, 1996; Seymour, 2009).

In the last two decades, a collection of empirical studies concerning taxi market regulation has begun to emerge because of increased availability of taxi data, mostly collected by governmental agencies (Schaller, 1993; Dempsey, 1996; Schaller, 1999;

Maa, 2005; Schaller, 2007; Toner, 2010; Bacache-Beauvallet, 2012), while there has also been an expansion of theoretical models of taxi markets (Fernandez, 2006; Flath, 2006; Seibert, 2006; Verschoor, 2007; Kim, 2008) and review papers (Koehler, 2004; Bergantino, 2007; Salanova, 2011). Bruce Schaller (1993, 1999, 2007) has been compiling taxicab data in New York City for many years as a consultant for taxi agencies, and he has authored the New York City Taxicab Fact Book (Schaller, 2006).

There has been more taxi data available in recent years. These data come from governmental agencies as well as crowdsourcing services like Uber, Lyft, and Sidecar. For example, NYC TLC has been collecting GPS taxi data for yellow cabs as early as 2008. Since crowdsourcing services are operated on mobile phones with the use of apps, data is collected in each city where they operate. It is important for taxi agencies and commissions as well as app-based for-hire companies to share the data with the public in order to help develop some empirical models for taxi demand. The insights from taxi demand models are helpful in determining when and where taxi demand is underserved or overserved. Knowledge of these needs can potentially be used to promoting policies and regulations. For example, crowdsourcing services may serve areas underserved by yellow taxis while restrictions can be placed on crowdsourcing services in overserved areas. The details of theoretical models will be discussed in the second section of the literature review.

2.1.2 Airport Ground Service

Another big portion of recent literature related to taxis is studies on airport ground access. Approximately 65% of airport trips are made by private vehicles in the U.S. and Europe, and the remaining 35% of the journeys depend on alternative airport access modes such as taxis, buses, airport shuttles, and private vehicle services, such as rental cars, prearranged cars, limousines, company buses, and courtesy vehicles (Humphreys, 2005; Conway et al. 2011). Among those alternative airport transportations, taxis are the most important in NYC. Taxis are the primary ground carrier between Manhattan and LaGuardia Airport and provide a substantial number of trips to and from John F. Kennedy International Airport (Schaller, 2006). Taxis have the largest mode share for travel from Manhattan to New York airports, carrying about 39% of those trips (Schaller, 2006).

Because taxis are the most important airport ground access mode compared to other commercial modes (Conway et al., 2012), most of the literature has focused on airport ground access mode choice. Harvey (1986) investigated the mode choice for airport access using 1980 San Francisco Bay Area survey data and logit models. He suggests that the average air traveller is very sensitive to access travel time and travel expenses. He also demonstrated that the value of time of air travellers should be higher than for other travellers and that separate models for business and non-business travellers are necessary.

In recent decades, a number of airport mode choice studies emerged that include taxis as one of the transportation choices (Psaraki, 2002; Pels, 2003; Hess 2005, 2006, 2007; Humphreys, 2005; Gupta, 2008; Zhang, 2009; Ishii, 2009; Alhussein, 2011; Gelhausen, 2011; Akar, 2013; Chang, 2013). Most of these studies address the behavior of air passengers while only a few address the travel behavior of airport employees (Humphreys, 2005; Tsamboulas, 2012). It is necessary to consider the variability of travel time for different travel modes in utilizing logit models. There are also some discussions about the reliability of airport access time (Gosling, 2006; Tam, 2011; Koster, 2011) and about value of time or willingness to pay (Tsamboulas, 2008). In some cases, air passengers are willing to pay a premium to reduce the travel time, in particular when the trip is obligatory instead of leisurely (Tsamboulas, 2008; Gupta, 2008).

To effectively maintain airport ground accessibility, well-managed taxi operation at high-volume airports is very important (Shiner, 1999; Conway, 2012). Conway (2012) uses JFK as a case study to analyze the centralized taxi dispatching system using curbside data and survey data. Shiner (1999) states that 81% of airport respondents recognize the importance of controlling the number of taxicabs in airports. Other studies also investigate the airport taxi idling problem as an external environmental cost in the operations of airports (Lu, 2011).

2.2 Taxi Models

2.2.1 Modeling the Taxi Market

Most of the early theoretical models developed for taxi demand are economic models for the taxi market. The classic economy theory states that demand and supply will reach equilibrium in a free market. However, most taxi markets are not free markets, so the price generated by “competitive equilibrium” may be inefficient to cover the social welfare (Douglas, 1972). Economic models have related aggregated taxi demand to the price of the trip, waiting time (Orr, 1969), and expected delay (Douglas, 1972). Most literature states that taxi demand is inelastic with respect to prices (Turvey, 1961; Douglas, 1972; Shreiberi, 1975; Coffman, 1977; Beesley, 1983), which means that the percent change of taxi demand is less than one with a one percent change of taxi price.

Using the case of London, Beesley (1973) argues that the factors contributing to the number of taxis per head are:

1. the use of regulations;
2. the proportion of tourists compared to residents;
3. the percentage of people with above-average income, especially in the center of London;
4. the presence of a highly developed radially oriented railway system; and
5. the level of car ownership.

Beesley (1973) also states that cost per ride during peak hours is affected by slower travelling conditions. Beesley (1983) further expands this theory to the relationship between taxi trips and factors such as the number of cabs and price and finds that the percent increase in trips equals the sum of the percent increase in the number of cabs and the percent decrease in price. Shreiber (1975) explains the relationships between price of rides and the availability of cabs and between the number of rides per cab and the availability of cabs.

2.2.2 Taxi Customer Search Model

This taxi customer search model is a supply distribution model that shares the same fundamental structure as a demand distribution model. It is a model of the number of empty taxi trips that occur between each origin Transportation Analysis Zone (TAZ) and each destination TAZ, similar to an Origin-Destinaion (OD) model for passenger demand, such as the gravity model (Garber and Hoel, 2014). It allocates the predicted number of drop-offs in each TAZ where taxis become empty and the estimated number of pickups in each TAZ where empty taxis become occupied using an OD matrix. For a region with N TAZs, there are N^2 OD pairs. Each cell in the matrix indicates the number of empty trips that go between each pair of zones (Lavine, 2010).

Gravity models are usually used for trip distribution using the gravity function, which is an application of Newton's fundamental law of attraction, illustrating the macroscopic relationship between places (Garber and Hoel, 2014). If two TAZs are further from each other, it is less likely that there is interaction between them when all other conditions are equal. Therefore, distance acts as an approximation to impedance (cost factor) in traditional gravity models (Lavine, 2010). Gravity models have been previously applied in estimating vacant taxi distribution on the road network to maximize the profit for each taxi firm (Zhu, 2013). Similarly, we can apply the same concept by minimizing distance or maximizing profit to optimize the behavior of vacant taxi reallocation in the customer search model.

Taxi network models are also commonly used in modeling the searching behavior of vacant taxis (Yang & Wong, 1998; Wong et al, 2001; Kim et al. 2005). A network model is developed to describe how vacant and occupied taxis will cruise in a road network to search for customers and provide transportation services. As early as 1998, Yang and Wong developed the first road network taxi model with an assumption that each driver of a vacant taxi tries to minimize the searching time in finding the next customer. Later, the model was improved by Wong et al. (2001, 2002, 2005, 2008), Yang et al. (2001, 2005), and Kim et al. (2005) to include traffic congestion, taxi information system, multiple taxi modes, and multiple users.

In reality, maximizing revenue is probably the most important objective for taxi drivers. A number of studies have made assumptions that drivers of vacant taxis seek to minimize the time spent searching for their next customer by looking nearby, and local search behavior can be modeled by formulating a cell-based network and using an algorithm that assumes taxi drivers will maximize the probability of successfully finding their next customer (Wong et al., 2002, 2014). Wong et al. (2014) developed a network model of the expected profit that taxi drivers could earn in a particular zone in Hong Kong. Hu et al. (2012) designed a probabilistic vacant taxi routing model to make the routing decisions of vacant taxi drivers at intersections and claims that the probability of vacant taxi drivers finding a customer decreases if there are many vacant taxis nearby. Wong et al.'s (2008, 2014) models are built using detailed taxi GPS trace data that reveals the precise route that each taxicab drives.

2.2.3 Trip Generation Model

Trip generation is the first analytic process undertaken to forecast travel demand followed by three other processes, trip distribution, mode choice, and route assignment (O'Neill and Brown, 2001; Ben-Edigbe, 2010). Trip generation modeling is used to predict the total number of trips or the rate that trips are made originating in or destined in a certain area, which is known as a transportation analysis zone (TAZ) (O'Neill and Brown, 2011). Typically, trip generation modeling is focused on information of residences because the generated trips are often closely related to the characteristics of people who live in the area. Residential trip generation generally associates the trips with demographics and socioeconomics of the households in the TAZ (ITE manual). Recently, with geographic databases (demographics, socioeconomics) refined to the level of block group available from the U.S Census Bureau, TAZs may be defined as small as a census block group.

Trip generation models have long been used to relate total number of trips or trip rates produced from each TAZ to a number of factors such as (O’Neill and Brown, 2011; Racca and Tatledge, 2004; Kumar and Levinson, 1992; Schaller, 2005):

1. the level of service (LOS) of the travel mode;
2. accessibility of the travel mode;
3. demographics of the TAZ (such as population, race);
4. socioeconomics of the TAZ (such as income, education);
5. other characteristics of the TAZ (such as land area); and
6. land use within the TAZ.

Three methods are commonly used for modeling trip generation such as the rate method, cross classification method (Kumar and Levinson, 1992; O’Neill and Brown, 2001), and regression method (Ben-Edigbe, 2010; Mousavi, 2012). The rate method is used for traffic impact analysis on non-residential trip generation, which does not consider characteristics such as household size, income, and auto ownership. A cross classification model cross-tabulates average trip making rates with 2 or more variables, allowing us to get a clear understanding of important factors without assuming that the relationship between demands and explanatory factors follows a specific functional form or that there is independence between these factors. The regression method produces a maximum likelihood estimate (MLE) for the coefficient of each explanatory variable, typically with a model that implies a linear relationship between the explanatory variables and the dependent variable.

Regression has become a widely used statistical method to explore the relationship between response variables (i.e., dependent variables) and explanatory variables with various methods and software packages to validate the models. If enough information is available, trip generation based on regression models can be very useful to forecast and examine the travel demands in urban transportation systems at each TAZ (O’Neill and Brown, 2001; Ben-Edigbe, 2010). For example, if we know the information in a TAZ that relates to people’s travel choice for transit, we may be able to forecast the demand for transit. There are a lot of challenges to adequately covering large geographic areas in sufficient detail to produce a good analysis using regression modeling for trip generation models. Thus, a large and sufficiently detailed dataset is a necessity in obtaining good estimates of the response variables.

There is not much literature based on large taxi GPS datasets to model taxi trip generation. Schaller (2005) has performed an analysis on the total number of taxicabs in 118 U.S. cities using multiple linear regression modeling. The factors influencing the size of a city’s taxi fleet include population, employment, use of complements to taxi cabs (e.g., public transit), cost of taxi, and taxi service quality. However, the model predicts the number of taxicabs in the fleet as the dependent variable instead of the number of taxi trips generated. The influential factors that are identified include the number of workers commuting by subway, the number of households with no vehicles available, and the number of airport taxi trips. Mousavi et al. (2012) have stated that household structure, age, gender, marital status, income, employment, car ownership, population density, and distance to transit are the most influential variables on trip generation for all modes.

Taxi demand may be closely related to transit accessibility in NYC, because taxis and transit are both transportation modes that are available to the public to participate in activities. The factors that influence transit use may also influence taxi trip generation. This would depend on the accessibility of other forms of transportation such as mass transit. For this reason, factors related to transit and vehicle modal split could also be considered in the models for taxi demand. Racca and Ratledge (2004) have compiled a comprehensive list of possible factors that are used for mode choice modeling, including transit level of service, accessibility, land use, demographic characteristics, and characteristics of the trips. Their analysis on mode split versus mean age and time of day indicate that these variables affect the modes that people choose, and this means that they may also relate to taxi trip generation. Corpuz (2007) has found that socio-demographic characteristics and time of day have influenced people's choice between private vehicles and public transportation. Workers and households with higher incomes are more likely to use cars over public transit. The train and the bus are more likely to be picked during morning and late afternoon peaks when people want to avoid the time and the cost of driving in congestion (Corpuz, 2007).

Characteristics of the trip (e.g., travel purpose) and characteristics of the traveler (e.g., age, income, education) have been identified as influential factors affecting the trips generated for different travel modes (O'Neill and Brown, 2001; Kumar and Levinson, 1992; Corpuz, 2007; Chang, 2013). Trips to residential areas and non-residential areas (Chang, 2013) and trips for business and non-business purposes (O'Neill and Brown, 2001) are analyzed separately in some studies. A number of studies have been conducted about trips generated to airports (Chang, 2013) and travel to schools (Ben-Edigbe and Rahman, 2010; Ewing, 2004). Researchers have also studied trips generated by elderly people as their needs and behavior could be distinct from the general population (Chang, 2013; Schmöcker, 2005).

Without detailed information about the taxi trip purpose or the characteristics of the specific person making each trip, we will use the characteristics of the places where taxi trips start and end to gain insights about the types of people and activities that are most associated with taxi trip making. This dissertation focuses on the characteristics of the people who live and work in these places in order to develop models for taxi trip generation.

2.2.4 Mode Cost Model

Mode choice modeling is the third step in a conventional travel demand forecasting process. Mode cost modeling is a part of mode choice modeling. Both modeling techniques compare the experience that users have among various transportation options, for example public transportation and taxi (Gebeyehu, 2007), bicycling, driving using private cars, and walking (Gebeyehu, 2008). This dissertation examine the relative cost of taxis versus other mode choice options, because of the lack of ridership data and travel preference survey data prevent a complete mode choice analysis.

The most commonly used method for mode choice modeling is to estimate the cost and the utility and then use a logit model to estimate the probability of choosing one transportation mode over the others. There are many different types of logit models, for example, simple binary logit model, nested logit model (Baik, 2008), ordered logit model (Gebeyehu, 2008) and box-cox logit model (Mandel, 1997).

Studies that explicitly consider taxis as a mode choice focus mainly on modeling mode choice/cost for passengers traveling to and from airports. Harvey (1986) was one of the first studies to identify the factors influencing airport travel mode choice of departing airline passengers based on a travel survey in the San Francisco Bay Area. The analysis using a multinomial logit model shows that travel time and travel cost are two strong explanatory variables. Business travelers are more sensitive to airport access travel time than leisure travelers, and values of time for most of the individuals are estimated to be at least as high as the average wage. Extra luggage, which is defined as more than one piece per person, is mentioned as an unattractive factor for the public transportation choice. Psaraki and Abacoumkin (2002) analyzed mode split for Athens International Airport in Greece to predict future mode shares and found that international passengers are more likely to use taxicabs or to be dropped off by private cars. Pels et al. (2003) studied mode choice in the San Francisco Bay Area and reported that business travelers have higher value of time and higher access time elasticity compared to leisure travelers. The authors also reported that access time has a larger influence on mode and airport choice compared to access cost. However, the analysis is based on public transit travel time estimations drawn from train and bus schedules and for taxi cabs travel times based on distances from center of the zip code to the airport. Therefore the model does not take walking and transit transfer times into account for public transit option or the effect of traffic congestion for taxicabs and private vehicles.

Gupta et al. (2008) developed a ground access mode choice model for the New York City Metropolitan area using an air passenger survey and a nested logit model. In their analysis they prepared transit Level of Service (LOS) data exogenously using online schedules and waiting times are also taken into account. Taxi costs are approximated by a per-mile fare. Similar to the previous studies, demographic characteristics, trip cost, travel time, and trip purpose are found to be the most significant variables in passengers' mode and airport choice. They estimated the value of time for business trips as $63/hour and leisure trips as $42/hour, which are considerably higher than general purpose value of travel time estimations for the same area.

Tam et al. (2011) investigated how travel time reliability affects the mode choice of passengers using a combined dataset from revealed and stated preference surveys. In their survey, the authors define a buffer time beyond their preferred arrival time that passengers' take into account when they are making their mode choice decision. The authors stated that offering discounts for passengers travelling as groups can improve the share of public transportation and increasing reliability can attract more passengers to use bus service.

Luken and Garrow (2011) studied the airport choice problem in New York City. Their analysis based on online ticketing data showed that the accessibility of the airport significantly affects the airport choice. The authors mentioned that they did not have the flight time information therefore the existence of correlation between access time and access distance was inevitable in their case. Therefore analyzing peak and off-peak driving times can be an improvement for the model they presented.

Akar (2013) used binary logit models to evaluate the airport ground access mode choice behavior of airline passengers in Port Columbus International Airport, Ohio. The analysis results based on a passenger survey showed that the reliability of the travel mode is the dominating factor in mode choice other than automobile. An interesting finding is that the effect of cost on mode choice is not usually significant, which is not consistent with the previous studies. This fact is explained by the location of the airport which requires shorter access times from the city and comparably cheaper parking prices.

This study aims to provide insight into understanding the accurate cost information, which other works have suggested will affect the mode choice behavior of passengers by utilizing realistic data sources for taxi and transit options. De Neufville (2006) argues that planning for low-cost airports includes a significant focus on public transportation options, so more attention should be drawn on non-driving passengers.

2.3 Summary of Literature Review

Most of the historical literature addresses taxi policy and regulations. There are major issues in the arguments regarding the regulation and deregulation on the taxi markets: "Condition of fitness" (entry conditions), regulation of fares, and limitation on taxi licenses. Proponents of regulation argue that regulation ensures the price of taxi, and prevents a surplus of cabs created by the free market, consequently reducing the external costs of taxicabs. Those in favor of deregulation point out that the costs of regulation and limited competition among taxicabs hurts consumers.

A number of taxi network models have been built using traceable taxi GPS data to understand the searching behavior of vacant taxis. Detailed taxi traces include the specific route that a tax travels in a network as opposed to just the pickup and drop-off locations of passengers. Those models are based on algorithms that minimize travel time and travel distance, and maximize the revenue of empty taxicabs.

The taxi related literature identifies some important contributing factors about taxi demand, such as the price of the trip, waiting time, the use of regulation, proportion of tourists compared to residents, the percent of people with above-average income, the railway system, and the level of car ownership. In the trip generation literature, the number of trips produced using one travel mode in a transportation analysis zone (TAZ) is likely to be related to the level of service and the availability of alternative travel modes, and the demographics and socioeconomics information of those living in the TAZ.

A large portion of the recent taxi literature is related to airport ground access. It was found that taxis have the largest mode share for travel from Manhattan to New York airports. The value of time of air travelers is estimated to be higher than that of other travelers. The factors influencing airport travel mode choice include: travel time, travel time reliability, travel cost, and other benefits (e.g., luggage service). Also, business air travelers are more sensitive to airport access travel time than leisure travelers, and they tend to have higher values of time as well.

2.4 Research Gaps

The previous literature provides economic theories, models, and arguments about regulation and policies for taxi markets, and builds important theoretical foundations for taxi demand modeling. Methodologies including customer search modeling, trip generation, and mode choice models are available in other literature as well. However, there is a lack of knowledge in the application of novel data to taxi demand modeling and in factors related to taxi supply and the spatial and temporal distribution of taxi demand.

In earlier economic models, given the limitation of taxi GPS data, aggregated taxi demand data has been commonly used as the measure of interest, such as the number of taxis per head, so we lack an understanding of taxi demand at a refined temporal and spatial scale. Because of the lack of disaggregated data, most evidence is based on simple summary statistics: the total number of cabs by year (Shreiber, 1975); taxicab ownership by different fleet size (Shreiber, 1975); the medallion price by year in NYC (Shreiber, 1975); the occurrence of taxicab monopolies in 31 U.S. cities (Eckert, 1979; FTC, 1984); and the number of taxi cabs in different cities (Gaunt, 1995). Schroeter (1983) first uses taxi demand in a 1.5 hour period during 3 days to calibrate a non-linear economic model that is used to calculate the probability of certain regulatory reform benefits for both taxi drivers and customers.

Factors related to the aggregated taxi demand are also limited to summary statistics, such as the price of the trip, waiting time, expected delay, the proportion of tourists compared to residents, and the presence of a highly developed radially oriented railway system (Orr, 1969; Douglas, 1972; Turvey, 1961; Beesley, 1983).

There are gaps in the knowledge of factors that are associated with taxi demand, especially taxi trip generation and distribution in refined spatial and temporal levels of aggregation. For example, we know that taxi demand is related to income, the presence of transit, and population, but we don’t know how taxi demand is related to the accessibility of transit, age, education attainment, and other demographic characteristics. Schaller (1999) first developed an empirical time series regression model in New York City to understand the relationship between revenue per mile and economic activity (measured by employment at eating and drinking establishments), taxi supply, taxi fare, and bus fare, but the model is not spatially specific. Other works on this topic consider factors such as city size, availability and cost of private owned autos, cost of taxi usage, population, competing modes and so on to determine the number of taxis in different U.S.

cities (Schaller, 2005; Maa, 2005), but their analyses do recognize the temporal variations in travel demand.

There have been many technology developments that are beneficial to modeling taxi demand, such as in-vehicle Global Positioning Systems (GPS) that provide massive amounts of individual tracks that can be observed and recorded (Liang, 2013), and geographic analytical tools like GIS (Girardin, 2010; Balan 2011; Bai, 2013). With a two-year set of taxi GPS from the New York City Taxi and Limousine Commission (TLC), it is possible to build empirical models to understand how taxi trips are generated and distributed in both spatial and temporal scales and how they compete with other transportation modes by using large-scale taxi GPS data.

Austin and Zegras (2011) have used similar taxi GPS and demographic data to develop a taxi trip generation model in Boston. They studied how taxi pickup trips related to transit accessibility using a small sample of taxi GPS data in Boston; however, no such study is found for other cities of the U.S. We still don't know how transit access time and taxi supply are related to taxi demand, and how taxi demand and supply are distributed in a refined spatial and temporal scale using a large-scale taxi GPS data. The study also considers whether taxi services are a substitute or a complement to public transportation, but a detailed methodology to support the conclusions is not presented. The models developed in this dissertation differ from Austin and Zegras (2011) in four ways:

1. data from New York City is used instead of Boston;
2. the data sample spans 10 continuous months rather than 4 days;

3. a novel methodology is employed to estimate both the waiting and walking time to the nearest specific subway station, instead of using a function of distance from a block group to rail/bus line/route to define the transit accessibility; and
4. a novel factor, the number of taxi drop-offs aggregated by census tracts and by hours is used to represent the immediately available taxi supply at different times and locations.

For the customer search model, most of the literature uses taxi GPS data that includes traces of vehicle routes and taxi network models to estimate the probability of successfully locating the next customer for each region. These studies usually make assumptions that vacant taxi drivers seek to minimize their expected time spent searching for the next customer nearby or to maximize the probability of successfully meeting the next taxi passenger in order to formulate cell-based network and local search behavior (Wong et al., 2003, 2014; Hu et al., 2012). No literature has provided information on how to build a network model using only taxi trip origin and destination data. Therefore, there is a gap in available methodologies to build such customer-search models using limited origin-destination taxi information. Also, no literature has focused on how a taxi driver behaves after dropping off a customer in an undesirable location.

Mode choice models have a long history in the transportation demand modeling field. Much of the literature has used different types of methodologies to consider taxi use in airport ground access studies (Harvey, 1986; Psaraki and Abacoumkin, 2002). The variability of travel time (Tam et al., 2011), accessibility to airports (Luken and Garrow, 2011), and the reliability of the travel mode (Akar 2013) affect the mode choice. Gupta et al. (2008) employ transit LOS data exogenously using online schedules and waiting times to develop a ground access mode choice model for New York City. However, no literature has used massive amounts of taxi GPS data combined with transit scheduling and waiting time to predict the travel time of both taxis and other transit in mode choice modeling.

2.5 Research Contributions

Based on the gaps in the literature presented above, the research contributions of this dissertation are unique in six ways described in the following subsections.

2.5.1 Detailed & Complete Dataset

This dissertation makes use of a detailed and complete taxi GPS dataset, which is composed of about 150 million taxi trips per year in New York City from 2009-2011. The dataset includes detailed information on trip characteristics and the geographic location of pickups and drop-offs. This dataset makes it possible to analyze the number of trips by time of day and by TAZ (census tract). This fills a gap by providing empirical evidence on taxi demand modeling. A couple of studies have used the same dataset to investigate urban dynamics (Qian, 2014), travel time variability (Morgul, 2013; Yazici 2012), weather impact on travel time (Yazici, 2013a), taxi ridership (Kamga 2013), taxi-drivers' airport pickup decisions (Yazici, 2013b), and travel time estimation for different OD pairs (Zhan, 2013), but none was related to taxi demand modeling.

2.5.2 Supply Time Series

The total number of trips per day and the number of shifts worked per day are used as dependent variables in the regression time series analysis, while the temperature of each day, the precipitation (rain and snow) of each day, and whether it is a holiday are used as independent variables to understand how those factors affect the number of trips each day in New York City.

2.5.3 Customer Search Model

Another contribution of this study is a novel customer search model based on taxi GPS data consisting of trip origin and destination points. The results provide two outcomes. One provides a spatial distribution of favorable neighborhoods where taxi drivers search for customers after dropping off a customer outside of Manhattan. The other provides a measure of the efficiency of system-wide behavior of taxi drivers searching for customers.

2.5.4 Taxi Trip Generation at High Spatial and Temporal Resolution

With this GPS taxi dataset, this study uses aggregated taxi trip information by time of day and by census tract as well as aggregated demographic and socioeconomic information to perform trip generation modeling on both pickups and drop-offs. This is a hybrid model, because it uses both cross-validation and regression techniques. The temporal classification of time of day is the cross-validation model. Within each hour of the day, regression models are developed, including simple linear regression, Poison regression, negative binomial regression, and spatial regression.

2.5.5 Mode Cost Analysis for Taxis

The first contribution of this study's mode cost analysis is the time-of-day analysis which has not been considered in detail in the existing literature. Real data for peak and off-peak time periods provides useful information about the travel costs of different alternatives by time of day. Second, rather than using travel surveys for model analysis, which is the conventional method, this study uses two novel datasets to model ground access mode choice. One is a two-year taxi GPS dataset used to obtain an accurate distribution of taxi travel times. The second is Google Maps Directions API based transit travel times, which gives estimated transit travel times between specific locations by time of day based on transit schedules. The analyses are made from the perspective of a passenger who considers his or her options before taking an airport trip in dollar and time terms, which can be measured more easily than utilities.

2.5.6 Utilization of Various Techniques

Two types of Google information are used in this study. The first is the Google Transit Feed Specification (GTFS) data which provides detailed and comprehensive information on transit schedules. The second one is the Java-script Google Maps Directions API to acquire the trip route information between each origin-destination pair.

The study also employs a number of statistical tools and methods to ensure that the empirical model for taxi demand adequately accounts for correlations between explanatory variables and spatial effects. For example, various regression techniques, model selection criteria, linear programming algorithms, and statistical software (including R, SAS, and CrimeStat) are utilized. The spatial analytical tool, GIS, is also used in this study to process geographic information and to visualize model results and spatial correlations.

CHAPTER 3

DATA

In order to conduct the proposed analysis of taxi supply and taxi demand, the relevant data must be collected and organized. There are three main bodies of data used for this study. The first is a complete collection of GPS taxi data for every taxi trip made in NYC within a two-year period from December 2008 through November 2010. Second, detailed transit schedules for the same geographic region are acquired using Google Transit Feed Specification (GTFS) and optimal passenger routing using transit from one point to another is acquired using Google Maps Direction API. Finally, these transportation data are supplemented with demographic, employment, and land use data, which are expected to include key characteristics of the locations that are associated with the highest rates of taxi use. The following subsections describe these data sources and their limitations in more detail.

3.1 Taxi GPS Data

3.1.1 Data Description

The two-year database of taxi trips are separated into two datasets: a 14-month dataset composed of 195 million taxi trips made between December 1, 2008 and January

24, 2010, and a 10-month dataset containing 147 million taxi trips made between February 1, 2010 and November 28, 2010. Between 5.5 and 5.8 million taxi trips are made each day in New York City. Most analyses in this study are based on the second 10-month dataset, because it is more recent. Each record includes information about when and where a trip was made, the distance traveled, and the fare paid. Specifically, the dataset includes the following fields for each record:

1. Taxi Medallion Number, Shift Number, Trip Number, and Driver Name;
2. Pickup Location (latitude and longitude), Date, and Time;
3. Drop-off Location (latitude and longitude), Date, and Time;
4. Distance Travelled from Pickup to Drop-Off;
5. Number of Passengers;
6. Total Fare Paid, including breakdown by Fare, Tolls, and Tips; and
7. Method of Payment (e.g., cash, credit card).

These data are collected by the Taxi & Limousine Commission (TLC) using the GPS and meter devices that are installed in every licensed (yellow medallion) taxi.

3.1.2 Data Preparation

In order to prepare taxi data for the three different modeling analyses, the data has been cleaned. There are some similarities in the data filtering process for the three modeling process, such as elimination of false records (e.g., records with zero travel distance) and records without geo-coordinate information. However, the customer search data preparation is different from the trip generation process because it requires more refined information about the sequence of pickups and drop-offs in each shift. The mode cost modeling then requires more strict accuracy on the travel time and distance than the previous two types of modeling. It requires taxi travel times to be estimated as realistically as possible so that they can be compared to transit travel times.

Preparing Taxi Data for Customer Search Model

The objective of a customer search model is to identify the movements of vacant taxis and evaluate the efficiency of the customer search process. A one-week subset of the 10-month dataset between June 28, 2010 and July 4, 2010 is chosen in this part of the study because the process of constructing the sequence of trips that make up each shift is time-intensive. The one-week subset consists of around 3 million taxi trips, which has sufficient data points to identify patterns of vacant taxi movements and limits the data processing requirements for the analysis. Although the week includes Independence Day on July 4, 2010, this particular day is a Sunday, so it does not affect the normal business hours during the week considered.

To prepare the taxi data for aggregation and analysis, a filtering process similar to the preparation for trip generation modeling was employed. An additional filtering process was performed by introducing the taxi travel time and distance traveled when the taxi was vacant between a drop-off and pickup. This required new data fields to be populated in the database before the additional filtration: previous drop-off time, previous drop-off geo-coordinate, vacant taxi travel time, vacant taxi travel distance. In order to compare travel time and travel distance for both occupied and empty taxi, trip travel time, and trip travel distance are also calculated the same way as vacant taxi travel time and distance.

The overall data preparation procedure followed these steps:

1) The dataset was filtered to remove false records, which is based on the same procedures as preparing for trip generation modeling in the next subsection.
2) The dataset was ordered by the taxi medallion number, taxi shift number and trip number by creating a column "trip_order_number" in each shift so that the sequential order of trips served by each taxi can be identified and the previous drop-off information can be associated with each pickup.
3) Vacant taxi travel time is calculated based on the previous drop-off time and current pickup time. The first trip of each shift will not have vacant taxi information because it does not have previous drop off location.
4) Vacant taxi travel distance is calculated based on the previous drop-off location and current pickup location. In order to create a direct comparison between the distance traveled by vacant and occupied taxis, the both distance calculations are based on straight line distance instead of actual travel distance. This is necessary, because only drop-off and pickup locations are known, so it is not possible to know how far a vacant taxi drives on the street network.
5) As soon as the vacant taxi travel time and distance are prepared, additional filtering processes are used (Table 3.1).
6) An origin-destination table aggregated to the census tract level was constructed to show the movements of vacant taxis from locations of drop-offs to locations of pickups, however, this table is a matrix with size 2167^2, therefore the full table is not provided in this dissertation.

7) Using ArcGIS to extract only trips that had previous drop off out-side Manhattan.

Table 3.1 The Filtering Procedure in Preparing the Data for Customer Search Modeling.

	Records Deleted	**Records Remained**	**Filtering Criteria**
Raw Data		3065940	
step1	6644	3059296	The shift number is zero
step2	10551	3048745	Negative vacant trip time and associated records with same medallion and shift number
step3	42882	3005863	Duplicates on the trip number
step4	35486	2970377	Negative empty time and associated records with the same medallion and shift number
step5	920	2969457	Empty taxi travel time larger than 12 hours or trip time is larger than 12 hours
step6	2372	2967085	Either empty taxi travel distance larger than 100 miles or trip distance is larger than 100 miles.
step7	1359	2965726	Payment type is 'Dispute'
step8	6708	2959018	Payment type is 'No Charge'
step9	0	2959018	Either trip pickup or drop-off date and time is null
step10	0	2959018	Both trip time and distance are zero
step11	265	2958753	Fare amount is zero and trip distance is zero; or fare amount and trip distance are zero
step12	59305	2899448	Any coordinate is zero
step13	0	2899448	Any coordinate is null
step14	0	2899448	Any coordinate is empty
step15	56573	2842875	Trip distance less than 0.05 mile
step16	38509	2804366	Empty taxi travel time larger than 95 percentile of that for all trips (133 minutes) or empty travel distance is larger than 95 percentile of all trips (10.216 mile)
Total Records Deleted	261574		
% of Deletion	**8.53%**		

Preparing Taxi Data for Trip Generation Modeling

The dataset that has been adopted for trip generation modeling is the 10-month taxi GPS data with a date range between February 1, 2010 and November 28, 2010. The raw taxi data requires some filtering in order to remove false records in the data set. The deficiencies in the GPS location data are mostly due to satellite errors, receiver noise errors, coordinate transformation errors, and errors made by the driver (Zito et al., 1995). The taxi GPS data have been processed to minimize the influence of outliers. Some records are obviously false and have been eliminated:

1) Records that have total fare amount equal to zero or travel distance that is less than the straight-line distance between the origin and destination.
2) For other records, two or more criteria are used to determine whether to remove a data point (e.g., fare amount, distance, and travel time). Those removed records include: records with both zero travel time and zero trip distance; records with both zero travel time and payment type is 'Dispute' or 'No Charge'; records with both zero travel distance and payment type is 'Dispute' or 'No Charge'.
3) The taxi trip records which did not have valid locational information are also eliminated. For example, those records with zero or null pickup or drop-off coordinates are removed.

Ultimately, less than 2 percent of the original taxi records were eliminated through this filtering process.

The goal of the trip generation analysis is to identify which demographic, employment, and land use factors have the strongest effect on the number of taxi trips made. An additional goal of this study is to identify if the availability and accessibility of public transit is related to the use of taxis in a neighborhood when controlling these other factors. In order to conduct this analysis, the raw taxi data must be processed into a format that is compatible with other data sources. Since the spatial resolution of much of the demographic data is at the level of census tracts, this is the same level of spatial resolution that is used for aggregating the taxi data. For the taxi trip generation models developed in this study, census tracts are used as the geographic unit for the Transportation Analysis Zones (TAZs).

In addition to the spatial aggregation, the taxi data must also be aggregated by the hour of the day so that the trip generation model can account for temporal variations in demand. The process of aggregating pickup and drop-off records for this study is similar to the way that taxis were used as traffic probes by time of day in Yazici et al. (2012). The distribution of pickups (origins) and drop-offs (destinations) are considered separately because they are clustered differently in time and space. Thus separate models are developed to understand these two trip ends. Further models include both the aggregated pickups and drop-offs in one model.

This processed data can be visualized on maps of New York City in which each census tract is shaded based on the number of taxi trips that are observed starting or ending. Although the data set is split into 24 hours of the day, illustrative examples are shown for the afternoon peak at 5:00 p.m. (Figure 3.1) and late at night at 12:00 a.m. (Figure 3.2). The figures provide a visualization of where pickups and drop-offs are located at different times of day. All maps are constructed with the same scale so that they can be compared directly with one another. The figures show that the locations where demand is concentrated are mostly in Manhattan and downtown Brooklyn, but a

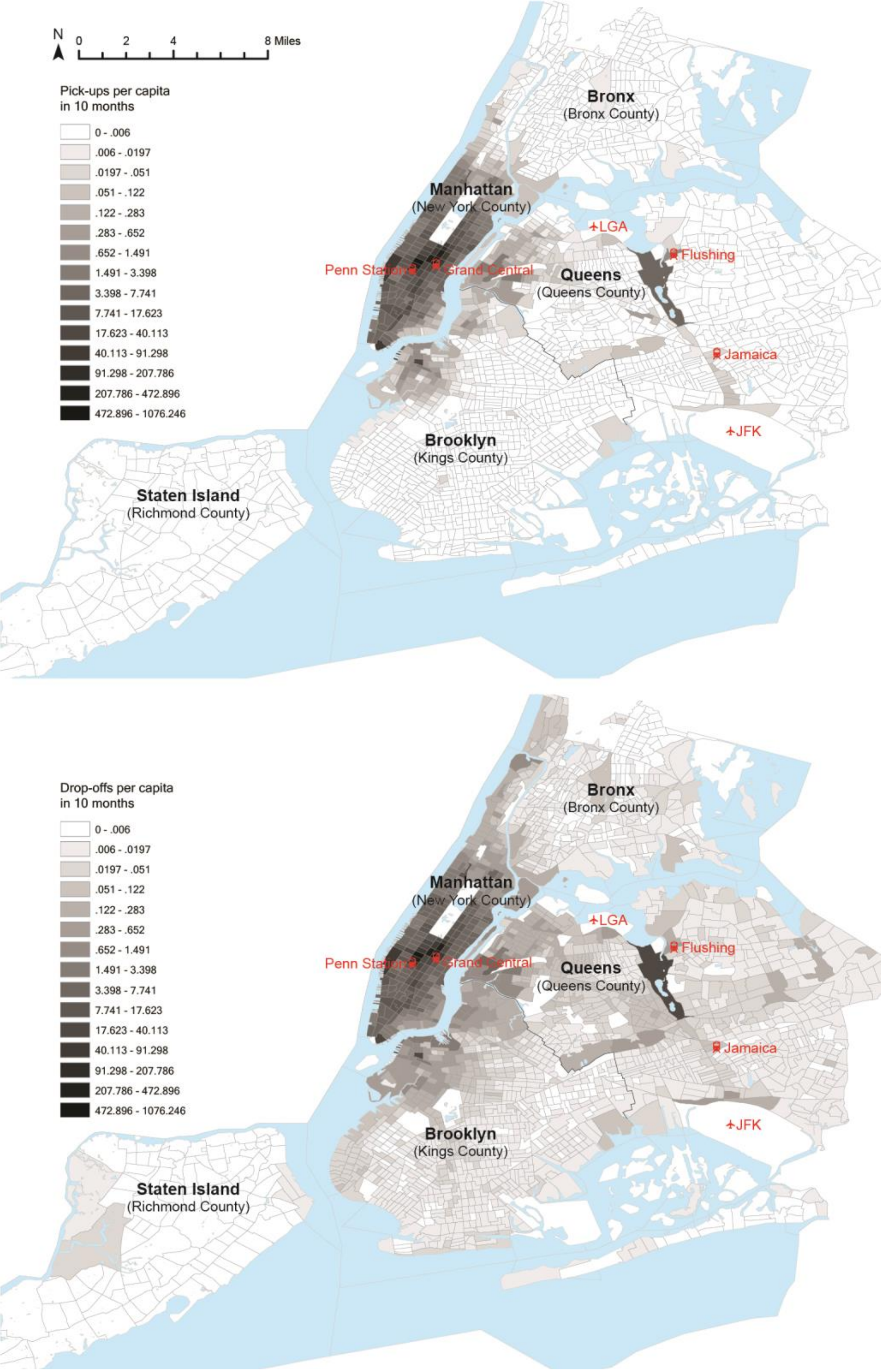


Figure 3.1 Taxi Pickups and Drop-Offs during 5:00 p.m. – 6:00 p.m.

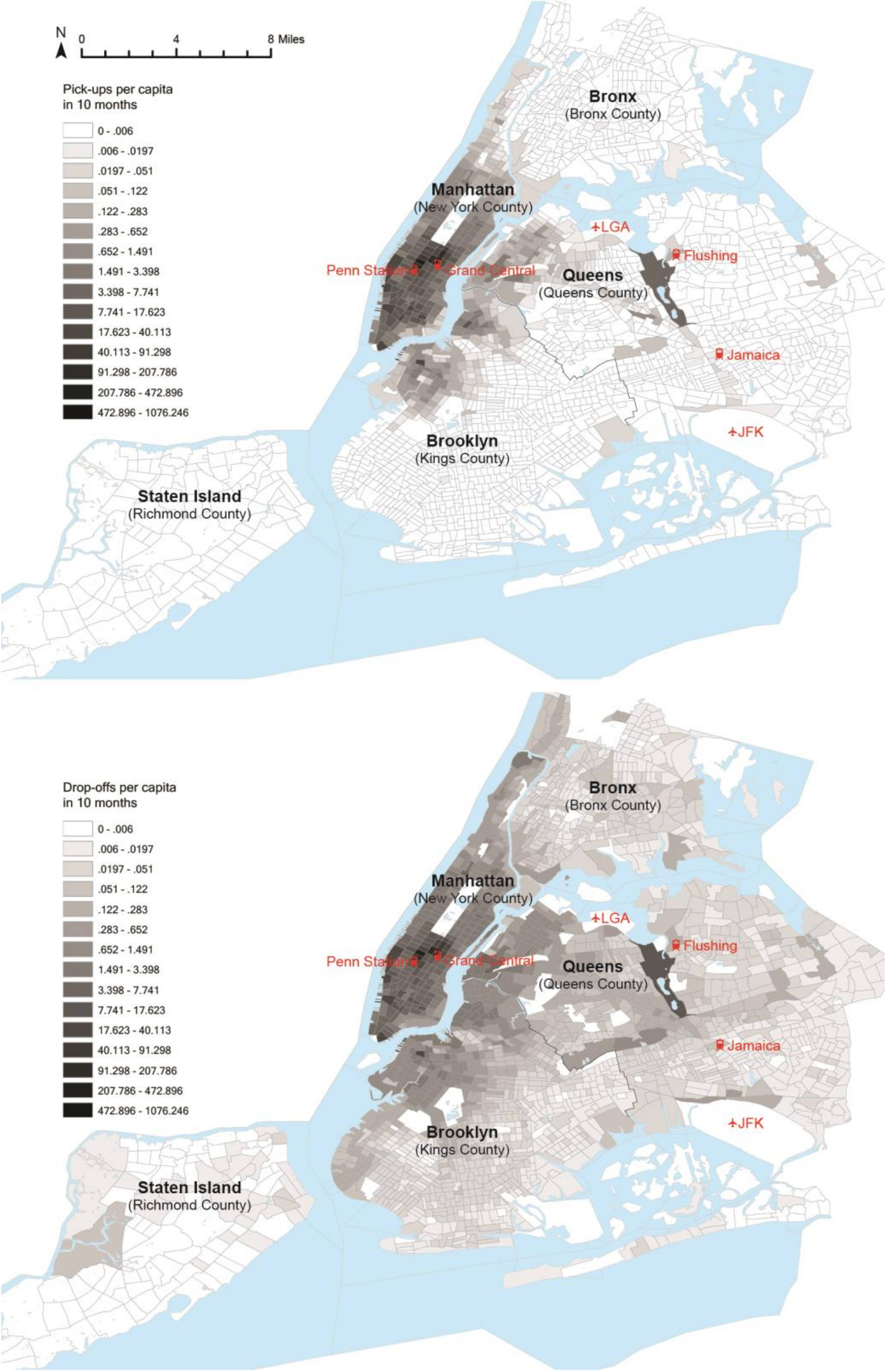


Figure 3.2 Taxi Pickups and Drop-Offs during 12:00 a.m. – 1:00 a.m.

more complete statistical analysis is necessary to quantify how this demand relates to characteristics of each census tract and the transit service that is available at each location.

Preparing Taxi Data for Mode Choice Analysis

In order to evaluate mode choice, the analysis focuses on a few specific origin-destination pairs involving travel between New York Pennsylvania Station (Penn Station) and the region's three main airports. The relevant taxi trips are extracted from the cleaned larger data set (the same one used for trip generation modeling) by identifying only those trips that have one trip end near Penn Station and the other trip end near an airport. All airport trips within 500 feet radius of the center of Penn Station, within a 1-mile radius of the center of EWR, and within the census tract of JFK (Census tract ID: 36081071600), and LGA (Census tract ID: 36081033100) are considered. As NYC taxis are not allowed to pick up passengers from EWR, the cabs must return empty so only trips from Penn Station to EWR are included in this study. JFK and LGA have taxi trips in both directions, so a total of five OD pairs are included in this study.

3.2 Google Transit Feed Specification (GTFS) Data

Having data on transit service over the same geographical area as the taxi data is necessary for determining how transit service affects taxi trip generation and how travelers choose between the two modes. The Metropolitan Transit Authority (MTA) operates extensive bus, subway, and commuter railroad services in New York City. Additionally, New Jersey Transit operates commuter rail services from Penn Station to New Jersey, including a connection to EWR. AirTrain services at JFK and EWR connect the airport terminals with the regional rail network.

In addition to the published routes and schedule information provided by the operating agencies, several web-based services make route and schedule information available in an electronic format. Examples include Google Maps, Bing Maps, and MapQuest. For this study, data from Google is utilized in two ways. First, station and schedule information is extracted in an electronic format for analysis of transit accessibility. Second, a specialized program has been developed to make use of the trip planning functionality of Google Maps API Transit Directions Service.

3.2.1 Google Feed Data

One of the goals of this study is to identify what role transit accessibility plays in determining the number of taxi trips that start or end in a census tract. Therefore, information is needed about the locations of transit stations and the frequency of service at these stations. This information is available in a standardized electronic format for public transportation schedules and associated geographic information called the General Transit Feed Specification (GTFS). The MTA developer data download website has a link to download the GTFS data for New York City: http://web.mta.info/developers/index.html.

Public transit agencies publish their route and schedule information in GTFS to Google so that developers can write applications using Google Maps to search for directions online. The GTFS “feeds” are a series of files that describe the locations of transit stations, the sequence of stations served by each route, and the set of times when vehicles depart from each station. Together, these make up a comprehensive description of all routes and schedules operated by an agency.

3.2.2 The Google Directions API

Google Maps API Directions Service, which offers free transit route guidance with a daily request limit, is used to obtain transit data. The information is gathered in XML format using a web-based JavaScript code. An application is developed that extracts one week of travel time and route information (including weekdays and weekends) based on schedules for the five origin-destination pairs every five minutes throughout the day. The routing information provided by Google is assumed to be the optimal transit option for the requested time and origin-destination pair since the web-based routing service compiles all available scheduling information for different transit modes and routes. The fare is estimated based on the optimal route. Data for approximately 2,016 transit trips have been collected for each origin-destination pair over a 10-month period. The transit travel duration of each trip includes waiting time, transfer time, and in-vehicle travel time. Each data point is also associated with an estimated fare based on the service utilized to complete the trip.

3.2.3 Data Preparation

Preparing Transit Data for Modeling Taxi Trip Generation

The raw data from GTFS can be directly combined to determine the number of scheduled vehicle departures per hour from each transit station. For example, Figure 3.3 shows a map of subway stations in NYC with each station shaded to represent the number of subway trains serving the station in the 5:00 p.m. to 6:00 p.m. hour on weekday afternoons. Additional analysis and calculations are required to convert the information on this map into a measure of transit accessibility for each census tract in the city. That process is part of the methodology for this study, and the details are described in Section 5.2.1. Although subway schedules are complete throughout the region, some data for bus routes is missing from Google's databases (notably, bus routes in Queens are not currently available in the GTFS format). Due to this limitation of data, the analysis in this study is based only on measuring subway accessibility in all 5 boroughs in New York City.

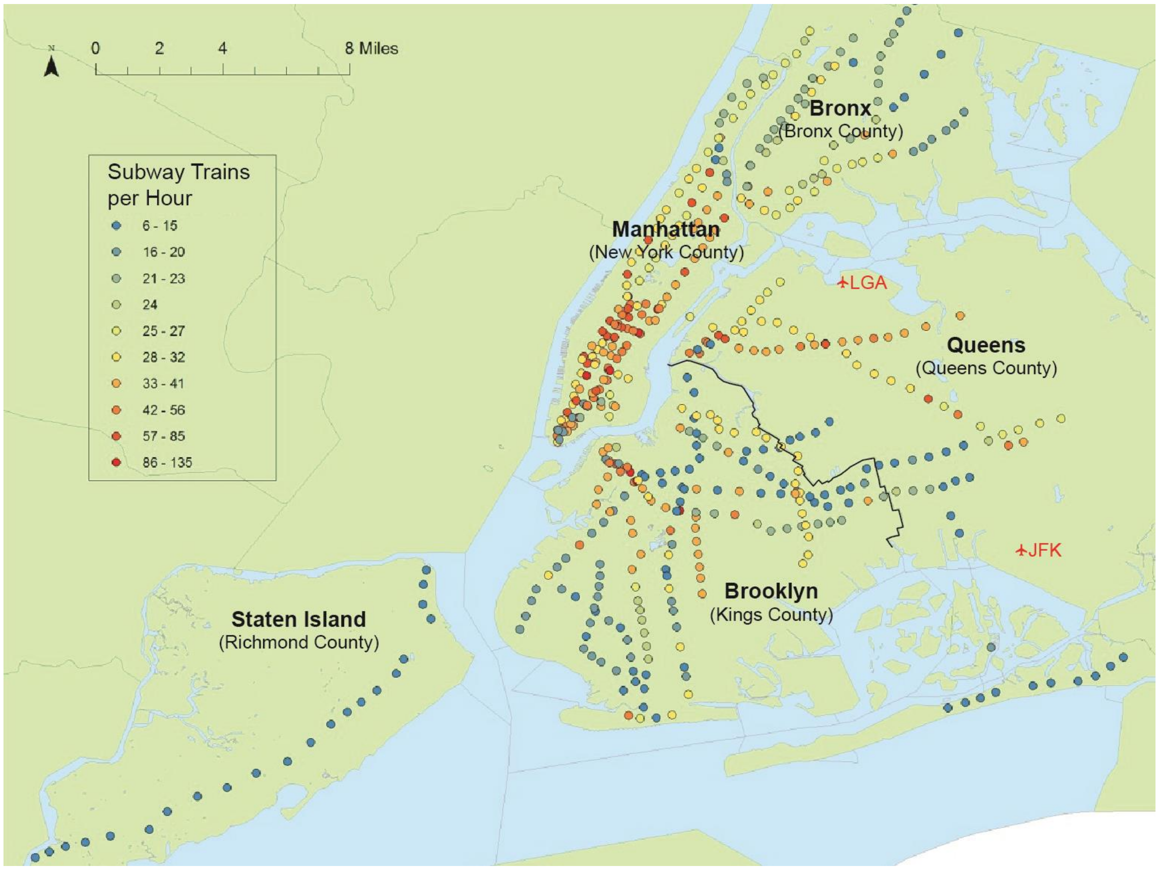


Figure 3.3 Frequency of Subway Service at Each Station in New York City between 5:00 p.m. – 6:00 p.m.

Preparing Transit Data for Mode Cost Modeling

Whereas trip generation models are constructed only based on the characteristics of the locations at the beginning and end of a trip, mode cost models require information about the complete trip itself. In order to compare taxi and transit service between Penn Station and the airports in the NYC region, specific information about travel times and fares are needed for each of the relevant origin-destination pairs. The accessibility through transit, taxi, or car for all airports is listed as follows (Port Authority of New York & New Jersey (PANYNJ) website):

1. JFK, located in Queens, NY, is accessible from AirTrain, buses, and car/taxis. AirTrain JFK connects to the Long Island Rail Road (LIRR) and the NYC subway and bus system at Jamaica and Howard Beach.
2. LGA, located in Queens, NY, is four miles from Manhattan and can be accessed by car or taxi. LGA does not have a direct rail link, but bus services do connect to the LIRR and subway.
3. EWR is located in Newark, NJ, and is accessible from Manhattan via the Holland and Lincoln Tunnels by car or taxi. AirTrain Newark provides access to New Jersey Transit trains into NYC Penn Station.

The travel time and monetary costs of each airport trip are summarized by hour of the day in the taxi data set, so comparable data is required for the competing transit trip. Without live data collection of travel times for passengers, the next best data source for assessing transit is to look to the schedules. The travel time and fare for using transit depends on the specific route selected (e.g., whether only subway and bus are used, or if commuter rail is used).

The Google Directions API has provided the estimated subway travel time and travel fare. The information was gathered in XML format using a web-based JavaScript code. An application was developed that extracts one week of travel time and route information (including weekdays and weekends) based on schedules for the five OD pairs every 5 minutes throughout the day. The routing information provided by Google is assumed to be the optimal transit option for the requested time and OD pair since the web-based routing service compiles all available scheduling information for different transit modes and routes. As the process of acquiring data is done automatically and programmed using Java Script code, we do not have faulty raw data to eliminate.

3.3 Demographics and other Data Sources

In addition to characteristics of the taxi and transit modes themselves, there are characteristics of the places where taxi trips start and end that are likely to have an effect on the magnitude of taxi demand. As described in Chapter 2, the literature on trip generation models shows that population, demographic characteristics of the population, employment, land use, and other characteristics of a transportation analysis zone can all have important explanatory power for predicting the number of trips that each zone generates (Corpuz, 2007; O'Neill and Brown, 2001; Kumar and Levinson, 1992). Much of this data is collected and made available by government entities such as the United States Census Bureau. Since many of the relevant population characteristics are aggregated at the level of census tracts, this is a logical scale for analyzing taxi demand.

3.3.1 Demographics

Demographics are quantified information about a given population in a region, usually referring to the Decennial Census, which is conducted every 10 years, and other surveys of individuals and households administered by the Census Bureau, such as the American Community Survey (ACS) and the American Housing Survey (AHS) (U.S. Census Bureau). A census tract is a geographic region usually within the limits of cities, towns, and counties, defined in the purpose of taking a Decennial Census. They can be further divided into smaller geographic regions including census block groups and census blocks. This study selected census tract as the scale of the study because the most complete data is available at this level. The tables and maps of census data are accessible from the *American FactFinder* tool[3].

The population data often comes from Decennial Census data. It includes population data categorized by different races and ethnicities (e.g., hispanic, white, black) and by different age groups (e.g., 0-5 years, 6-10 years, etc.). Information is also available on gender, housing cost, home value, poverty, employment status, and marital status. However, there is no detailed education background or income information.

3.3.2 Socioeconomics

The socioeconomic data are the economic and social information of a given population in a region. The education and income data are acquired from averaged 5-year American Community Survey (ACS) data from 2007-2011 (U.S. Census Bureau, 2012). Although ACS produces population, demographic and housing unit estimates, it comes from the Census Bureau's Population Estimates Program that produces and disseminates the official estimates of the population and housing unit by sampling part of the population instead of collecting data for the whole population.

The employment data is also important in this study because the number of jobs is a good indicator of people’s economic and social activities. Where there are jobs, there is a need to use transportation to go to work and participate in economic activities. Besides the total number of jobs, the data has detailed information for each job category. The

[3] The data are all accessible from U.S. Census Bureau website:
http://www.census.gov/main/www/cen2000.html

data is accessible from 2009 & 2010 Workplace Area Characteristic (WAC) available at U.S. Census website.[4]

3.3.3 Land Use Data

Land use data is also an indicator of people's activities and a relevant factor for modeling trip generation. The data is acquired from PLUTO,[5] an extensive land use and geographic data, and MapPLUTO, an extensive land use and geographic data at the level of tax lots available in ESRI AcrGIS shape format and dbase files. PLUTO is available from the NYC Department of City Planning (DCP). It includes 11 land use categories in the scale of tax lot, such as family buildings, commercial and office buildings, industrial and manufacturing, transportation and utility, parking, and so on.

3.3.4 Other geographic data

There are a number of geographic data (ESRI shapefiles) used to perform analysis at different levels of aggregation. Those data includes county, census block, and census tract, which are available from TIGER/line shapefiles in U.S. census website.[6] All analysis is based on the census tract data as it is a relatively small scale. In the trip generation results, the New York City community district data[7] is only used to show the neigborhoods at a community level. This data is the same as Public Use Microdata Areas (PUMAs) system, and they've been given the names of community districts in NYC.

[4] The WAC data is downloadable from: http://lehd.ces.census.gov/data/

[5] The PLUTO data is downloadable from: http://www.nyc.gov/html/dcp/html/bytes/applbyte.shtml

[6] TIGER/line shapefiles: https://www.census.gov/geo/maps-data/data/tiger-line.html

[7] New York City Community District Data is available from: http://www.nyc.gov/html/dcp/html/bytes/districts_download_metadata.shtml

3.3.5 Data Preparation

The datasets for demographics and socioeconomics span the year 2009 and 2010, it is necessary to include explanatory data for both of these two years. The sources of data for the explanatory factors considered in this study include:

1) demographic data for each census tract available from the U.S. Census 2000, and 2010, including total population, population categorized by age, and population categorized by race;
2) socioeconomic data available from the American Community Survey 5-year (2007-2011) estimate of education and income;
3) employment data by census tract, including categorization by age, earnings, type, race, ethnicity, educational attainment, and sex, which is available for NYC from 2009, 2010 Workplace Area Characteristic (WAC) data available from the U.S. Census Bureau;
4) PLUTO and MapPLUTO in ESRI ArcGIS shape format and dbase table format) data available from NYC Department of City Planning (DCP) which include 11 land use categories in the scale of tax lot; and
5) geographic data including relevant shapefiles (e.g., rivers, roads, county, census tract) and land area.

The population density and employment density (Figure 3.4) in 2010 is calculated for all 2,167 census tracts in NYC. PLUTO data provides a more refined level of information in Figure 3.5, but it can be aggregated into census tracts. Figure 3.4 shows that the population density and employment density are concentrated in Manhattan. Some census tracts consisting of cemeteries, parks, or islands do not have employment associated with them, so the WAC employment data covers 2,143 census tracts. Census tracts with variables that are lacking certain required information are excluded from the linear model analysis; e.g., where there is no population or employment. Ultimately, 116 out of 2,167 census tracts (5 percent) were omitted from the analysis, because there is insufficient population or employment in those few regions to obtain a useful data point. With all of the demographic, employment, and land use data aggregated by census tract, the data set is prepared with a large set of variables that is used to develop models for taxi trip generation.

The expected results are that there is a relationship between taxi demand and land use because the use of taxis is closely related to people's activities, and land use could reflect when and where people engage in certain activities. For example, residential areas may have a concentration of taxi trip pickups in the morning and drop-offs in the afternoon peak for travel to and from work, while commercial areas may have taxi trips at all times of the day.

One of the difficulties in this study is to account for the reliability and sensitivity of land use data when aggregated to the level of a census tract rather than at the level of the individual tax lot. The relationship between land use categories and people's activities depends on how land use is categorized according to building classes, and the same activities could be related to several categories. Therefore, there might be some difficulties in creating explanatory factors by aggregating the land use data before having an idea of how they are related to taxi trips. Some variables are lacking information for certain census tracts, and those census tracts are excluded from the linear model analysis.

In the end, we omitted about 5% (118 census tracts) out of 2,167 census tracts for NYC when we use Land Use instead of employment data in the trip generation analysis.

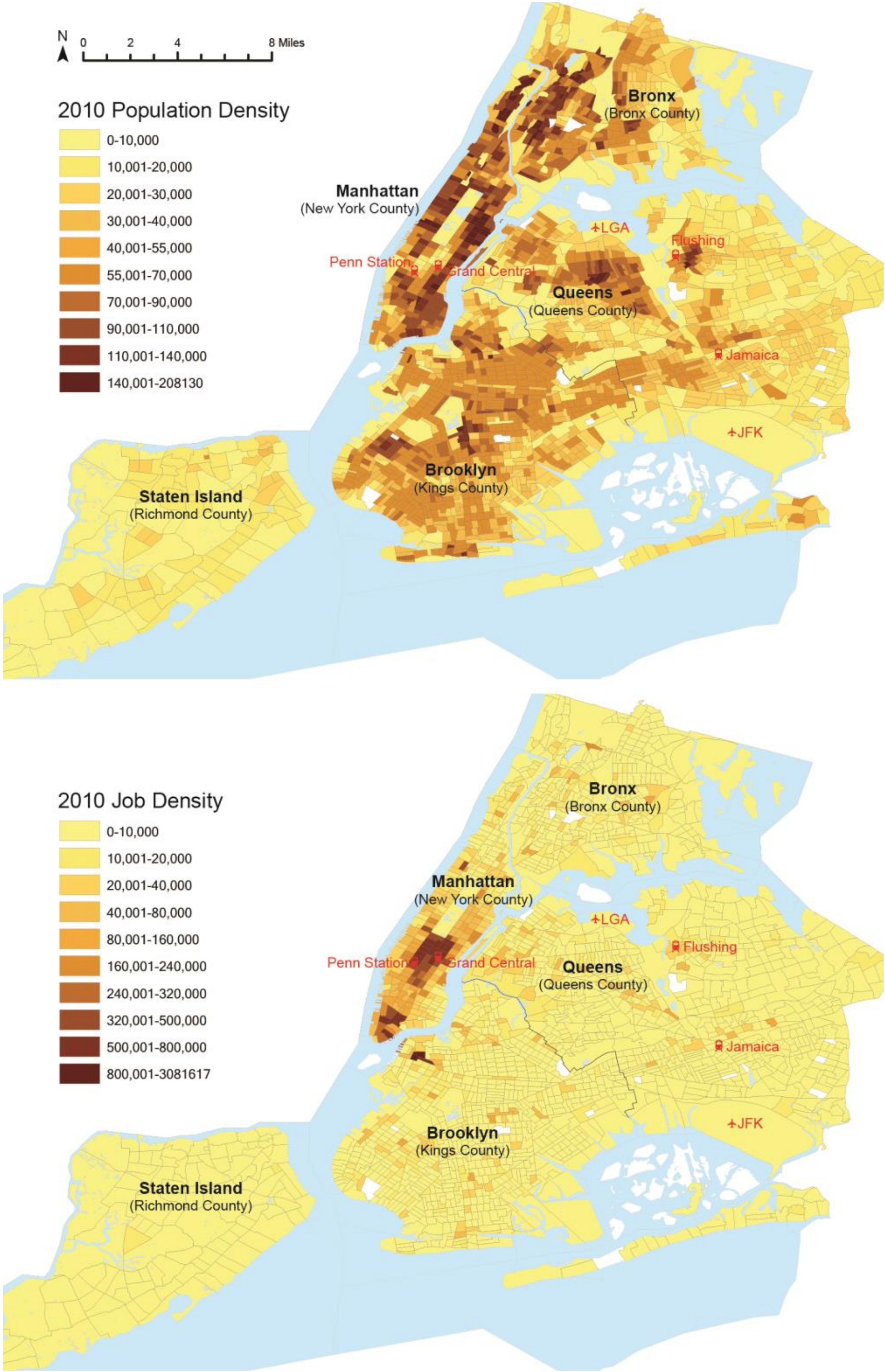


Figure 3.4 Year 2010 Population and Job Density (per square mile).

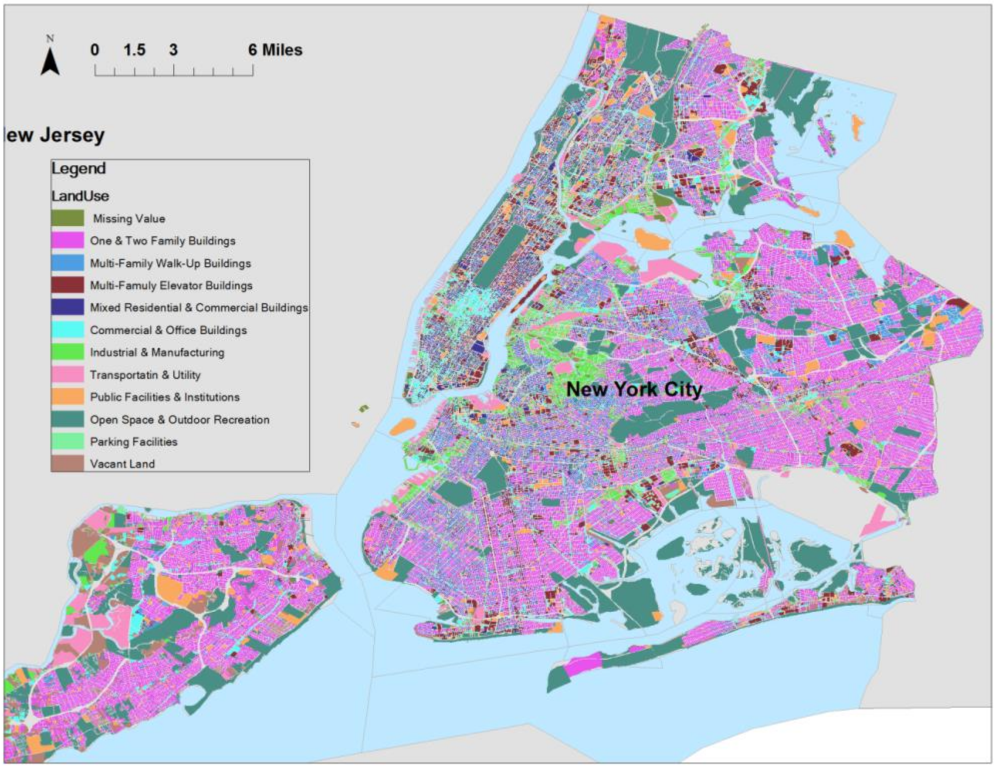

N
0 1.5 3 6 Miles
ew Jersey
Legend
LandUse
Missing Value
One & Two Family Buildings
Multi-Family Walk-Up Buildings
Multi-Famuly Elevator Buildings
Mixed Residential & Commercial Buildings
Commercial & Office Buildings
Industrial & Manufacturing
Transportatin & Utility
Public Facilities & Institutions
Open Space & Outdoor Recreation
Parking Facilities
Vacant Land
New York City

Figure 3.5 Year 2010 Land Use Data in New York City.

CHAPTER 4

TAXI SUPPLY

Taxi supply refers to the availability of taxi services in the taxi market. The taxi supply is a major indicator of how much taxi service has been made available. This can be measured by number of medallions in operation each day, the number of taxi shifts each day, the number of hours of operation each day, or the number of empty taxis on the streets. Therefore the definition of supply is introduced in the beginning of this chapter. Then the temporal and spatial characteristics of NYC taxi supply are discussed, followed by introduction of a novel methodology for modeling the efficiency of the customer search by vacant taxis. A summary of findings is provided at the end of this chapter.

4.1 Quantifying Taxi Supply

Taxi demand and supply are economic concepts related to the taxi market. The number of taxi pickups is aggregated by space and time to represent the taxi demand, because the number of pickups is an indicator of travel demand on when and where people use taxis. Taxi supply is a little more difficult to quantify, because the observations of vacant taxis are only available when a customer is dropped off and the taxi becomes available for another customer and when a customer is actually found and picked up.

Historically, without the availability of big taxi data, a researcher could only use summary statistics to represent the taxi supply, typically, the number of registered taxis in a city (Eckert, 1979; Frankena & Pautler, 1984; Gaunt, 1995). This has to do with the regulation of taxi market. Regulators from cities that have heavily restricted entry into taxi market (e.g., New York City, Los Angeles, Chicago, Boston, Miami, Houston, San Antonio, Buffalo, Albany, Salt Lake City, and San Francisco) must decide how many medallions are needed to meet the demand of taxi services. Cities like Denver and Philadelphia have preserved open entry into taxi market, so the taxi supply fluctuates reflecting the taxi demand in their market.

In New York City, the number of taxi medallions is fixed, thus the variation of the number of taxi trips completed every day relates less to number of taxi medallions and more to when and where taxi drivers choose to work. The number of taxi shifts completed and the hours of operations each day are representative of supply of taxis in operation. However, a taxi medallion can be owned by private taxi drivers and taxi agents who rent out their cars. A shift is the period when only one driver occupies the taxi whether it is owned by the driver or another entity. Specifically, the hours of operation per shift suggest how long each taxi driver works in each shift. Usually there is an upper limit of 12 hours per shift. The objective of this supply analysis is to identify what factors influence the behavior of a taxi driver, and where taxi drivers tend to search for customers. The time series analysis is going to show the variation of taxi supply with time and help identify factors that may be related to the temporal fluctuation of those values.

To prepare for the trip generation model in the next chapter, we define the taxi supply as the number of drop-offs in a neighborhood during a period of time. The reason is that the number of drop-offs, in a relatively short period of time, for example, each hour, is an indicator of how many empty taxis become available in that neighborhood. Those taxis that just make a drop-off immediately become available for hire. The drivers of vacant taxis may search nearby areas in the same neighborhood, or they may drive further if their previous drop-off location is somewhere that they believe they are not likely to find a customer. The empty driving distance and empty driving time reveal NYC taxi drivers' favorite neighborhoods or specific hot spots for finding their next customers. Details on the analysis of spatial dimensions and the efficiency of the vacant cab customer search are provided after the time series analysis.

4.2 Time Series Profile

The purpose of investigating on the variation in taxi use over time is to understand 1) the temporal variation of taxi trips; and 2) the potential factors that influence the temporal variation of taxi trips. The temporal variation is examined by looking at the taxi trips aggregated by day, month, and day of the week, as well as some of the potential variables that may influence the variation of the taxi trips, which is introduced in the following paragraph.

There are some possible factors that are related to the number of hours of operation per taxi shift. Some studies have shown the influence of weather on the taxi travel time and variability (Yazici, 2013a) and taxi ridership (Kamga, 2013). Weather is also likely to influence a taxi driver's behavior. For example, if it is snowing, the roads become slippery, a taxi driver may decide to stay home and not work on that day or reduce work hours on that day. Also temperature may be a significant reason why a taxi driver does not work or reduces working hours in his shift. For instance, if the temperature is too low or too high, a taxi driver may decide to serve fewer customers or change shifts earlier. A number of taxis are used by people who are going to work or taking off from their job. If it is a national holiday, a taxi driver may not get enough customers to fill up his hours in the shift, or they may just decide to give themselves a break or a holiday.

The taxi supply is shown in the temporal dimension in this section, we also investigate the relationship between taxi trips and other factors such as weather and holidays. The hours of operation are used as the response variable representing the taxi trips with respect to a taxi driver. A time series analysis is used on the hours of taxi operations using the 12-month taxi GPS dataset from January 1, 2009 to December 31, 2009.

4.2.1 Variables

The variation in taxi use over time is studied to understand what factors affect the taxi supply. Variables representing taxi supply are aggregated by day to create the time series. Five different time series were analyzed for the period between January 1, 2009 and December 31, 2009: the number of trips each day, the number of shifts each day, the average number of trips per shift each day, the total hours of operation each day, and hours of operation per shift each day. The hour of operation describes the time period from the first pickup in a shift to the last drop-off in the same shift.

The potential factors that influence the number of trips, shifts, and operating hours each day include:

1. The total precipitation to the nearest hundredth of an inch (WTR), and snow to the nearest tenth of an inch (SNW) that were measured at the Central Park Station control by the National Climatic Data Center (NCDC) from December 1, 2008 to November 30, 2010. This local climatological data provided by National Weather Service is available from the National Oceanic and Atmospheric Administration (www.noaa.gov).
2. The total number of medallions is considered in case the number changes during our study period. If it does not change during the whole study period, it can be excluded from the analysis.
3. Holiday, which is built as a dummy variable (i.e., indicator variable equal to 1 for Federal Holidays and equal to 0 otherwise). All Holidays between December 2008 and December 2010 are listed in Table 4.1 to construct this indicator variable.
4. Day of week variable, indicating to which day of the week each date belongs in order to understand the weekly pattern.
5. Average temperature (AVG) which is acquired from NCDC, but it is only included in the analysis of the 14 month dataset (12/1/2008 – 1/24/2010) along with the hours of operation analysis. Therefore, the following time series will only be analyzed using this 1 year sub-set in 2009.

4.2.2 Profile

The relationship between the above-mentioned factors and taxi supply is illustrated in Figure 4.1. Time series of total operation hours each day and operation hours per shift each day, and factors (Holiday, WTR, and SNOW) are plotted.

Table 4.1 The Federal Holidays from Dec. 2008 to Dec. 2010.

Date	Holiday
1/1/2009	New Year's Day
1/19/2009	Birthday of Martin Luther King, Jr.
2/16/2009	Washington's Birthday
5/25/2009	Memorial Day
7/3/2009	Independence Day
9/7/2009	Labor Day
10/12/2009	Columbus Day
11/11/2009	Veterans Day
11/26/2009	Thanksgiving Day
12/25/2009	Christmas Day

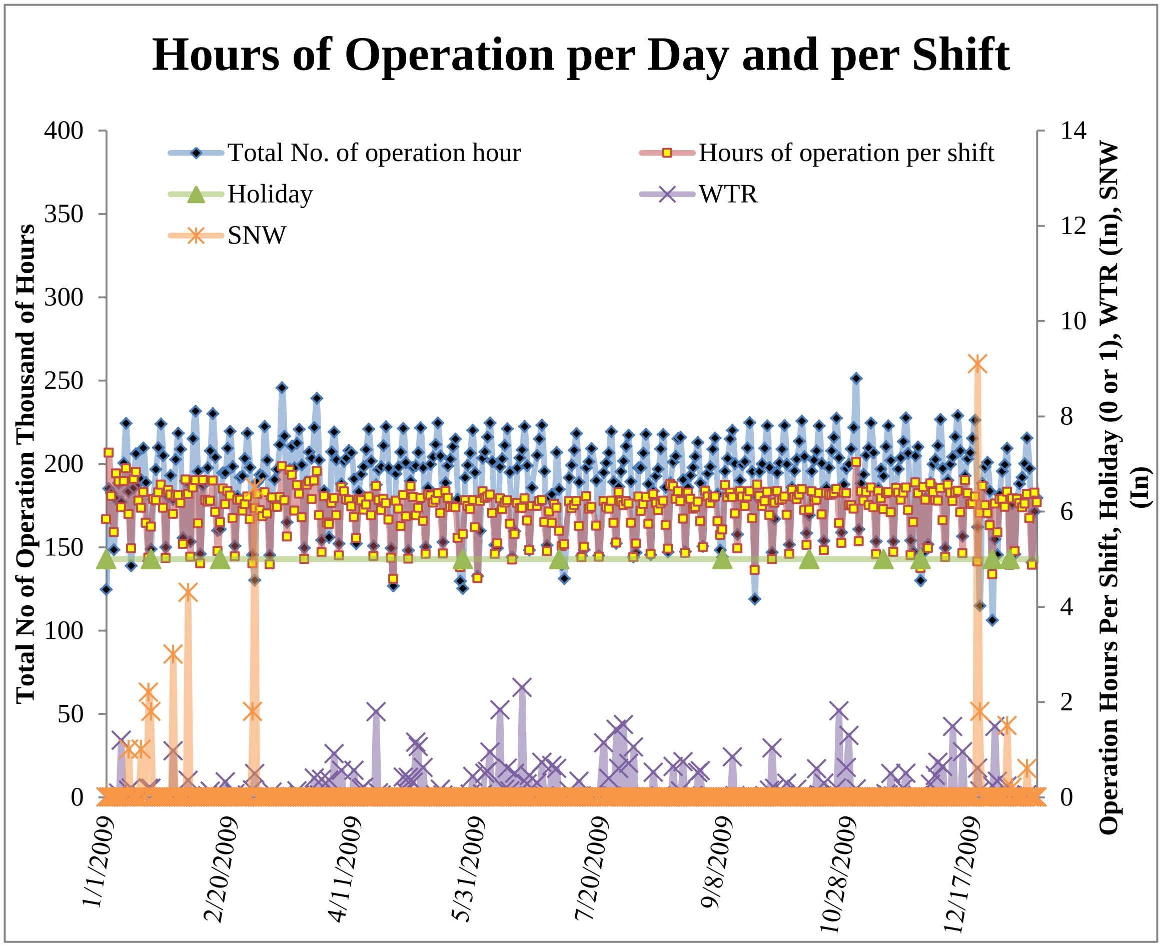


Figure 4.1 Hours of Operation per Day and per Shift V.S. Holiday, Precipitation, and Snow.

Figure 4.1 shows a relationship between the factors and the time series (total hours of operation each day or hour operation per shift each day). On federal holidays, the taxi operating hours were significantly reduced. On some days of heavy snow, the taxi operating hours were reduced as well. Holidays seem to have the larger effects on the number of operation hours each day compared to heavy snow days. No variation in taxi operating hours was observed for days with measureable total precipitation (WTR>0). There are cyclic patterns in each time series every week. Compared to weekends, weekdays have relatively higher total operating hours. There is no obvious monthly variation in the hours of operation observed from this figure.

In order to verify the patterns that exhibited in Figure 4.1, we have made boxplots of the weekly and monthly average of these five time series (Figure 4.2) as well as the averages against snow precipitation (SNW), holiday (dummy variable), and temperature (AVG). The central rectangle of the boxplot spans the first quartile to the third quartile. The segment inside the rectangle shows the median. IQR stands for the length of the box. The upper "whisker" shows the smaller value between the maximum and 1.5 times IQR above the third quartile, while the lower "whisker" is the larger value between the minimum and 1.5 times IQR below the first quartile. Observations fall outside the whisker are considered outliers. The overall averages for all five time series are shown by the red lines. The average temperature is grouped into 10 intervals: [12.9,20], (20, 27], (27.2,34.3], (34.3,41.4], (41.4,48.5], (48.5,55.6], (55.6,62.7], (62.7,69.8], (69.8,77], (77,84.1], and the average of each group is calculated separately.

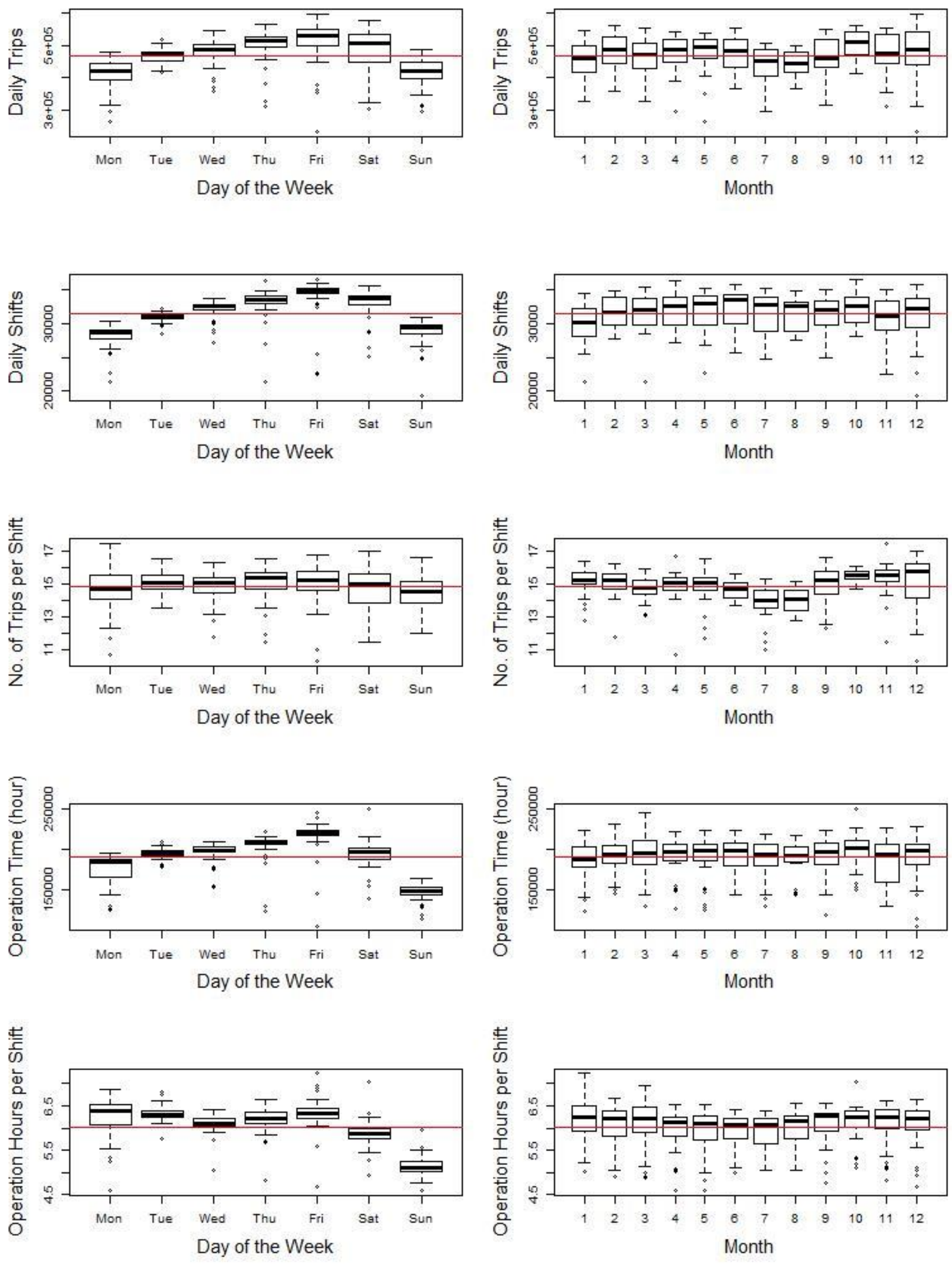


Figure 4.2 The Weekly and Monthly Average of Each Time Series (the red line represents the overall average of each time series).

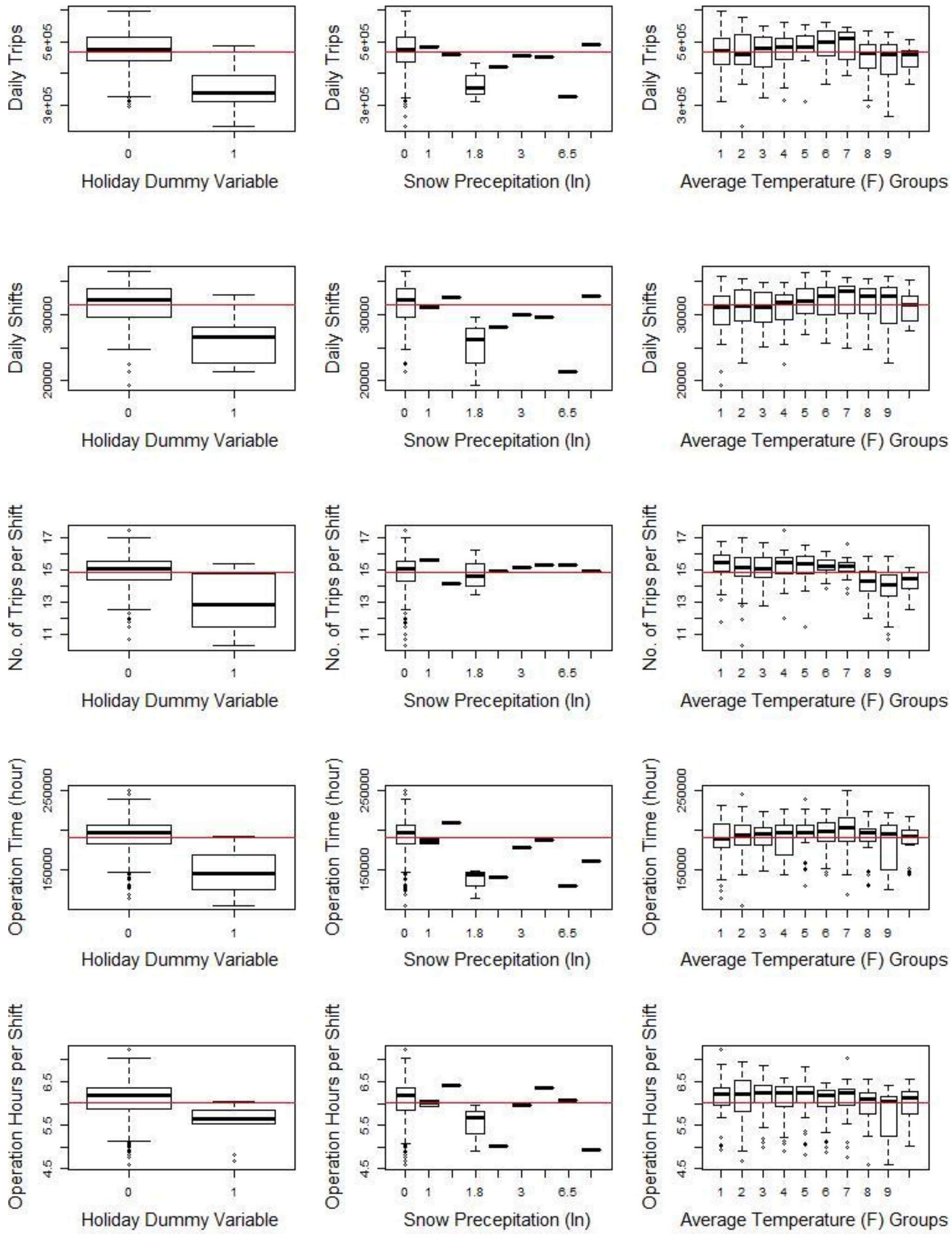


Figure 4.3 The Averages of Five Time Series Aggregated by Snow Precipitation, Holiday, and Average Temperature (the red line represents the overall average of each time series).

The day of week is one of the most influential factors on observed taxi trips. Figure 4.2 indicates that Monday and Sunday have relatively fewer taxi trips compared to Tuesday through Saturday. Similar patterns are observed in daily shifts and operating hours. On Mondays, the total taxi trips, shifts, and operating hours are below the overall average while the taxi trips and operating hours per shift are close or above the overall average. Part of the Monday trips include trips starting from midnight (12 A.M. Monday), which may be thought of as a weekend night; therefore, the average number of taxi trips are skewed to below the overall average taxi trips in 2009. However, the above average number of trips per shift and operating hours per shift suggest that taxi drivers actually work more efficiently on Mondays. In other words, weekdays have relatively more taxi trips compared to weekends. Figure 4.2 also verifies that a slight variation exists among the monthly averages of all five time series. This means that the taxi trips are relatively stable during the whole year, and it makes sense to aggregate taxi trips by hour as described in the following chapters.

Holidays have a significant influence on taxi availability and supply as well. The averages aggregated by holiday for all five time series in Figure 4.3 suggest that there are many more taxi trips, shifts, and operating hours during holidays than non-holidays. The day with snow of 6.5 inches (March 2, 2009) reduced taxi trips, while a heavier snow of 9.1 inches (December 19, 2009) reduced operating hours but not the taxi trips. This tells us that snow influences either taxi trips or operating hours. No trend or pattern was observed from the monthly average of all five time series.

To summarize, non-holidays and weekdays are likely to have more taxi trips, shifts, and operating hours compared to holidays and weekends. This makes sense, because a lot of activities in NYC are produced by people who work there, so fewer taxi trips are demanded during weekends or holidays. As a result, taxi drivers are less likely to work at all on holidays. Those who work are likely to have fewer customers during their shift and work for fewer hours. While heavy snow might influence taxi supply, the month of the year and average temperature have no effect on the taxi supply.

4.3 Supply for Empty Taxis

This study aims to understand the places where taxi drivers tend to find their next pickup after dropping off a customer. The study considers different drop-off locations using a one week sample from a 10-month taxi GPS dataset in NYC that has been used in Yang and Gonzales (2014). The data has been aggregated by 2167 census tracts in NYC so that the number of taxi pickups and drop-offs can be counted for each census tract. An Origin-Destination (O-D) matrix for vacant taxi movements is built to include all the movements of vacant taxis from passenger drop-offs to their next passenger pickups. Each row stands for each origin census tract (previous drop-off location) and each column (next pickup location) stands for each destination census tract. Thus the size of the O-D matrix is 2167 by 2167.

A unique procedure is developed to analyze the empty time and empty distance traveled while taxi drivers are searching for their next customers. An O-D trip matrix for empty cabs is created and two optimized O-D matrices are calculated using two different techniques: first, to minimize the distance traveled, and then, maximize the individual taxi drivers' revenue. It is also possible to identify where the taxi supply satisfies taxi demand and where there is not enough taxi service available based on user's preference on either one of the scenarios. This analysis will be addressed in the trip generation analysis from Chapter 5.

4.3.1 Preliminary Data Analysis and Visualization

The database of taxi trips has complete information of 147 million taxi trips made between February 1, 2010 and November 28, 2010, including temporal and spatial information acquired by GPS (taxi pickup and drop-off date, time, location), fare (including tolls, tip, total fare paid), and distance travelled. The taxi trip counts for each pickup and drop-off location are aggregated by census tract. A one-week sample dataset has been extracted from the 10-month data dated from June 28th, 2010 to July 4th, 2010. One complete week of data is chosen to work with to capture the pattern for both weekdays and weekends. July 5th, 2010 was treated as Independence holiday for pay and leave purpose instead of July 4th, 2010. Therefore, July 4th, 2010 is only considered as a weekend. This sub-dataset has complete information required for the analysis of empty time and empty distance traveled in searching for the next pickup.

Data Summaries

The processing of the data in order to build the model was discussed in Chapter 3. In this part, additional details of data treatment are introduced with results to show the complexity of the customer search analysis. As the customer search model needs more refined data than simple aggregations, the detailed steps are provided for the data processing combined with the summary statistics as a part of preliminary data analysis. The third step of the data treatment is the visualization of spatial distribution of empty taxis travel time and travel distance.

Step 1: Elimination of Error Records and Duplicates

In this taxi dataset, the medallion field is an aggregated quantity that may represent many different taxis, and each shift number represents a shift ranging from 8 hours to 12 hours in length. The average number of medallions in operation each day is very similar across days of week as shown Figure 4.4a. Thursday and Friday have relatively higher number of shifts (Figure 4.4b) and average number of shifts per medallion (Figure 4.4c). There is an average of 37 shifts per medallion, indicating that each medallion in the dataset represents more than one taxi and is an aggregation of the actual medallions issued. Within each shift, records of pickups and drop offs are recorded by time. The data first needs to be organized and cleaned by ordering all records according to medallion identification (first) and shift number (second) because some of the records are not in the right order by time and some of them are duplicates. Records that show a trip time and trip distance of zero are also eliminated from the data set.

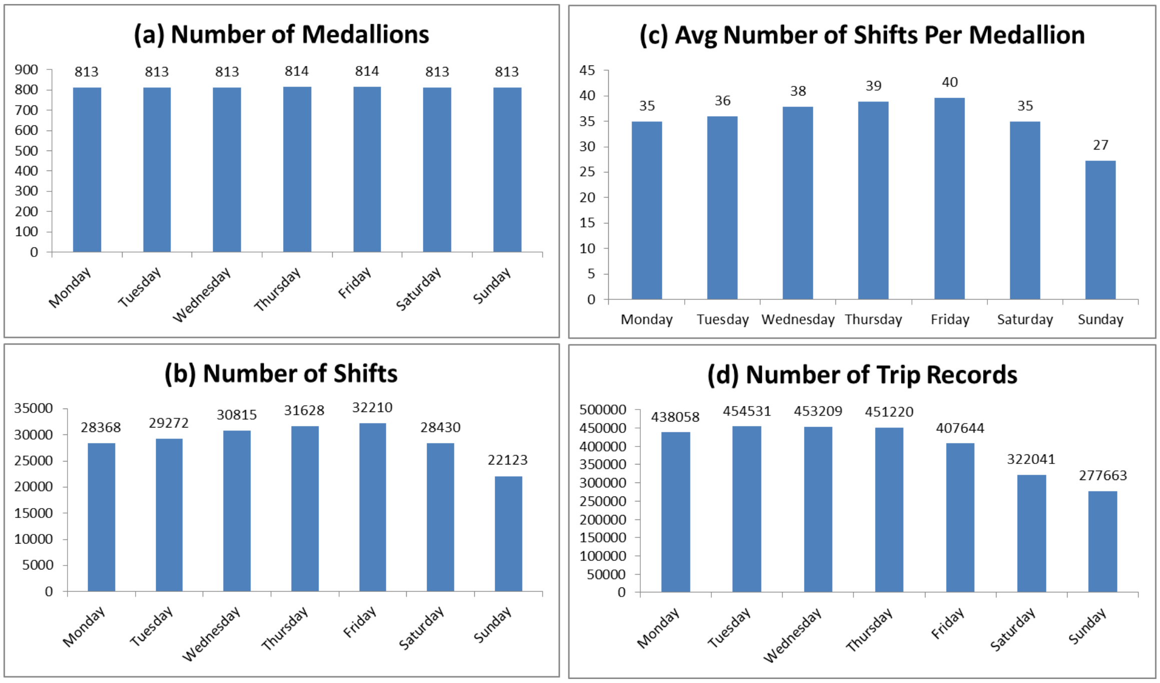


Figure 4.4 Summary of Medallions and Shifts for NYC GPS Taxi Data during June 28th, 2010 (Monday) to July 4th, 2010 (Sunday).

Step 2: Calculation of Empty Time and Empty Distance

As soon as the data is ranked by time, we are able to calculate the duration and straight-line distance between the previous drop-off location and subsequent pickup location for each individual record.

Step 3: Aggregations by Time or Location or Both

The data is composed of 2,804,366 individual taxi records after cleaning with trip pickup and drop-off information, so the data has been aggregated by NYC census tract (total 2167 census tracts) and the number of trips between each pair of census tracts is counted to build the O-D matrix for empty taxis. In order to compare the observed data with modeled data, we included only the records that have both a pickup and a drop-off in five boroughs of NYC. The calculation of empty time comes from Eq 4.1.

$$\text{operation time} = \text{empty time} + \text{occupied time} \qquad \textbf{Eq 4.1}$$

The empty time and distance are aggregated by hour of the day (Figure 4.5 and Figure 4.6), indicating that the average empty time and empty distance is aggregated for each census tract at each hour by averaging all the records that originated from that census tract. The data can be mapped to show the geo-spatial distribution of the average empty time and average empty distance for the taxi trips at each hour and at each census tract as shown in Figure 4.5 and Figure 4.6. It appears that the total empty time (empty time = operation time - occupied time) is less in morning and afternoon peaks than the rest of the day (Figure 4.5 & Figure 4.6), because peak travel demand keeps taxis relatively more occupied during the rush. The occupied time and the empty time follow closely with each other across the day except from 3 p.m. to 7 p.m. (Figure 4.5), but the occupied VMT seems to be much larger than the empty VMT (Figure 4.6) indicating that taxi drivers may drive much slower in searching for a customer than when having a customer on-board.

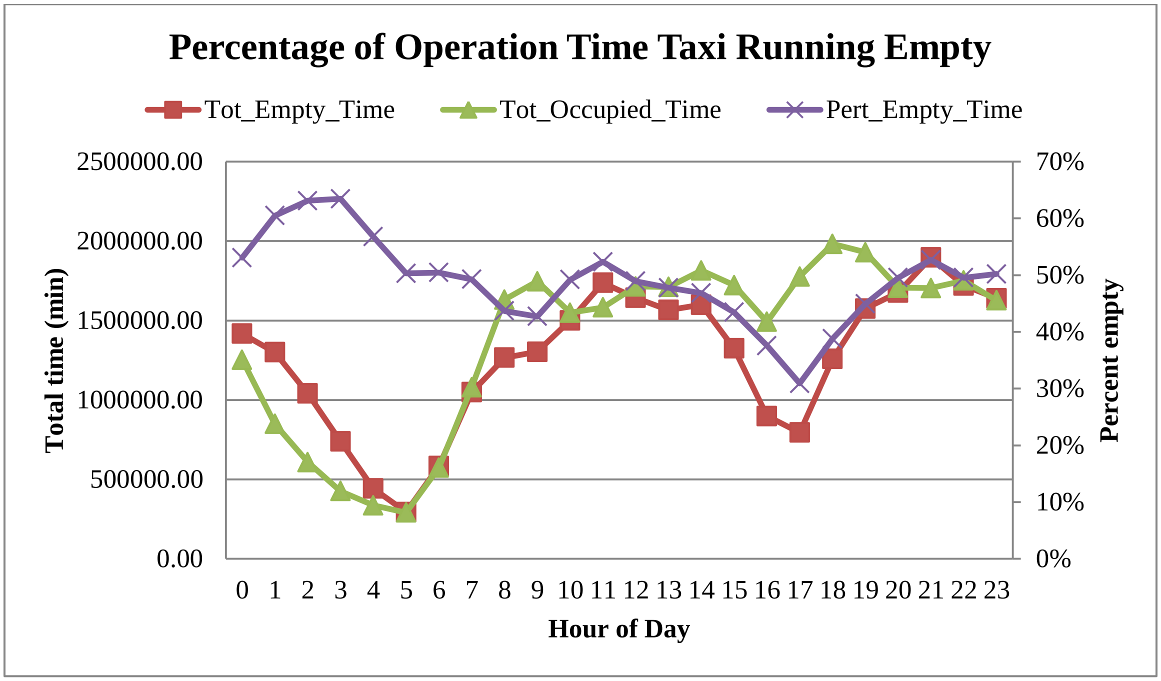


Figure 4.5 Plot of Total Empty Time, Total Occupied Time, and the Percent of Time Taxis Running Empty.

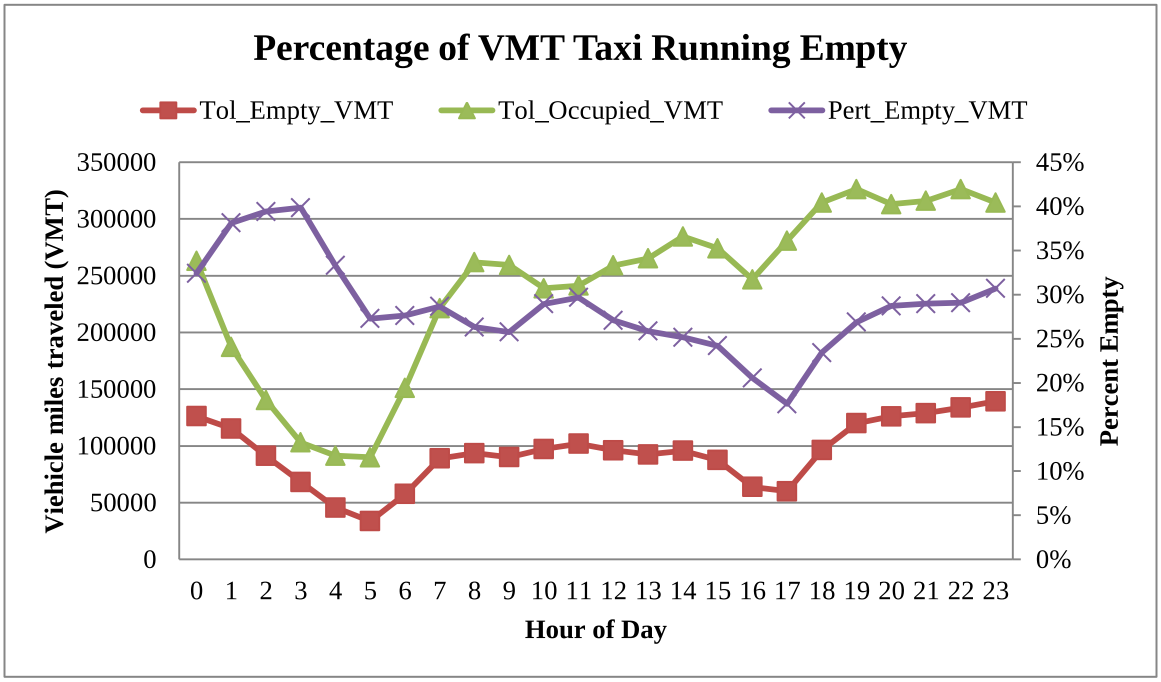


Figure 4.6 Plot of Total Empty Vehicle Miles Traveled (VMT), Total Occupied VMT, and the Percent of VMT Taxis Running Empty.

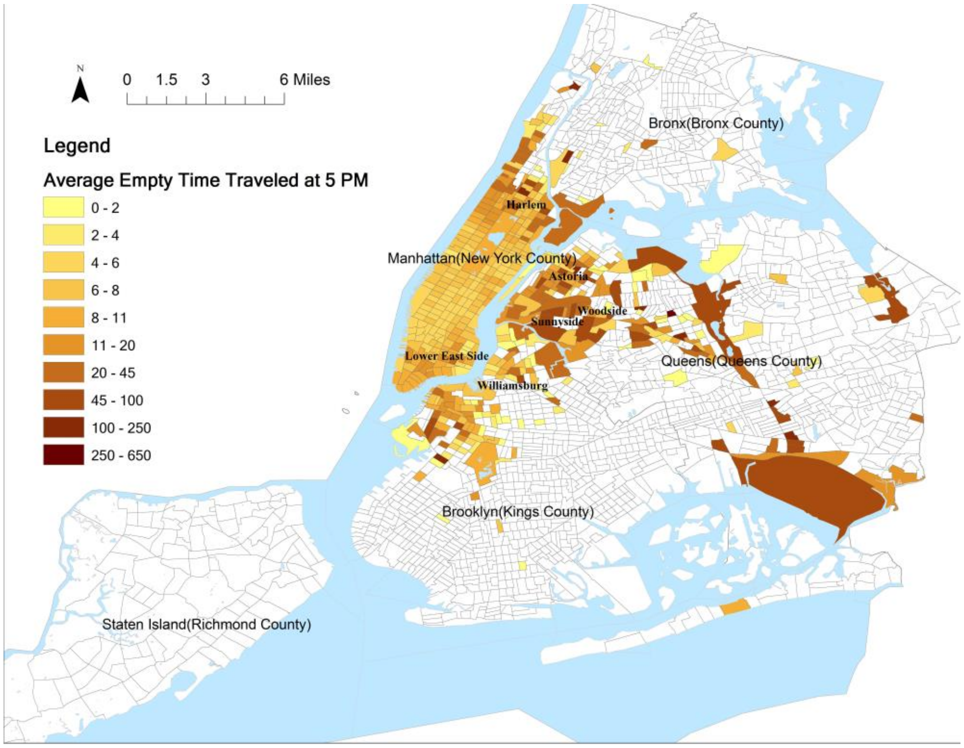


Figure 4.7 The Average Empty Time Traveled (in minutes) Before a Pickup for Each Census Tract at 5 PM.

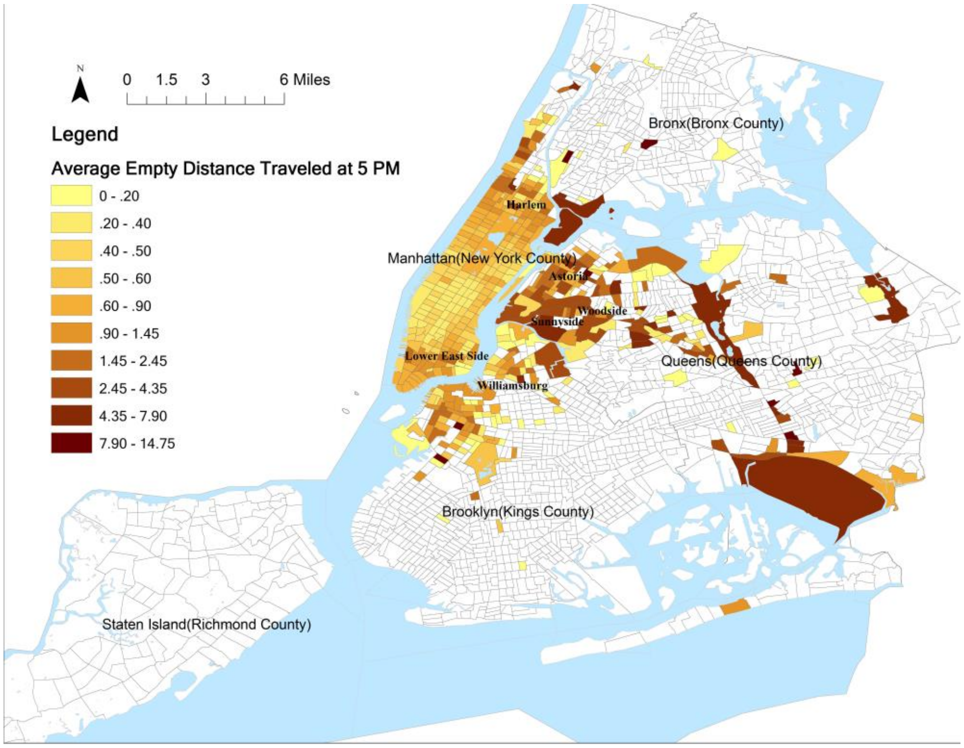


Figure 4.8 The Average Empty Distance Traveled (in miles) Before a Pickup for Each Census Tract at 5 PM.

Step 4: Extract partial data for the taxi efficiency analysis. Based on results from Figure 4.7 and Figure 4.8, Manhattan has relatively less empty time and empty distance than the other four counties in New York City, which is consistent with the higher levels of demand in Manhattan at 5 PM. Taxis that have drop-offs outside of Manhattan are likely to destine for hot spots (e.g. airports) or popular neighborhoods in order to maximize revenue. Otherwise it is difficult to find a customer while cruising in areas with a low density of demand. In order to better understand areas that taxi drivers tend to go and their operational patterns, all taxi trips which have the previous drop-offs outside Manhattan are extracted to study the taxi efficiency (about 7% of all trips in the one-week data) in the following customer search model.

4.4 Efficiency of Customer Search

The second objective of modeling the taxi customer search is to analyze the efficiency of a taxi driver in search of their next customer. The number of pickups and drop-offs for two modeled scenarios are developed to compare with the observed data. The first scenario is from a system's perspective by minimizing the total miles traveled by each trip in the O-D matrix using a Linear Programming (LP) technique. The second scenario is from a taxi driver's perspective by maximizing the revenue earned by the driver.

In scenario 1, we assume the system works with centralized control with complete information on the locations of every taxicab so the system can re-distribute the cabs to where taxis are demanded with minimum total distance traveled. In scenario 2, the taxi drivers are assumed to have perfect information about the popular locations with higher demand for taxis, which is estimated using the average taxi demand from historical 1-

week taxi data. This study is an off-line optimization of re-distributing empty taxis. Our study is based completely on historical data without predicting the future. This is different from an online optimization that predicts the distribution of vacant taxis without knowing anything about the future. Like the emerging crowdsourcing apps, it is probably better to use real time taxi data to know the taxi demand at an exact time and location, so that the taxi demand can be better served.

Both scenarios are modeled under ideal boundary conditions under these assumptions. Scenario 1 is expected to achieve better system performance overall by reducing the total distance traveled by taxis while serving customers' needs. Scenario 2 is an example of what could happen if all drivers are provided with information about where the most customers are located but are not coordinated in a way that ensures efficient system-wide performance. If every cab driver seeks the profit-maximizing route, the cabs will cluster in the few places that have the most demand, while in the places that taxis do not go, there will be many customers and no taxis at all. This would create a severe imbalance problem. In reality, yellow taxis in NYC are mostly street hail cabs without centralized control. Also, taxi drivers are not likely to have perfect information to re-distribute themselves to when and where they can earn the maximum revenue.

4.4.1 Models and Results

Scenario 1: System Optimization using Linear Programming (LP) Algorithm

Taxis need to be reallocated across regions, because pickups and drop-offs are not necessarily spatially balanced. Suppose every time a taxi drops off a passenger in region i, it drives to pick up a passenger in region j. The number of empty cabs traveling from i to j is X_{ij}. Cabs that stay in the same region as their previous drop-off are associated with

i = j. The average straight-line distance from a point in i to a point j is denoted d_{ij}, and if i = j, this is assumed to be approximately ½ of the diameter of the region. Furthermore the data can be summarized by the total number of drop-offs in region i (D_i = $\sum_j X_{ij}$) and the total number of pickup in region j (P_j = $\sum_i X_{ij}$).

The O-D matrix is constructed as follows:

$$\begin{pmatrix} X_{11} & \cdots & X_{1n} \\ \vdots & \ddots & \vdots \\ X_{n1} & \cdots & X_{nn} \end{pmatrix} \qquad \textbf{Eq 4.2}$$

and the mathematical program to minimize the total distance traveled by all empty taxis can be formulated as a linear program (LP) as follows:

$$\min \sum_{i}^{n} \sum_{j}^{n} X_{ij} d_{ij} \qquad \textbf{Eq 4.3}$$

$$\text{subject to } \sum_j X_{ij} = D_i \quad \forall i \qquad \textbf{Eq 4.4}$$

$$\sum_i X_{ij} = P_j \quad \forall j \qquad \textbf{Eq 4.5}$$

$$X_{ij} \geq 0 \quad \forall i, j \qquad \textbf{Eq 4.6}$$

where n is the total number of census tracts considered. The objective function of LP algorithm is to minimize the total travel distance traveled by all empty taxicab trips. The total number of trips can be calculated by adding the counts in the O-D matrix together. The constraints are to ensure that flows of taxis are positive, and that the number of taxicabs leaving each census tract and arriving in each census tract is consistent with the total numbers of observed drop-offs and pickups, respectively. As there are $n = 2{,}167$ census tracts, there are 4,286 equations for constraints according to Eq 4.4 and Eq 4.5, and 2,167^2 inequalities for the constraints (Eq 4.6). This is a classic linear programming formulation of the transportation problem. The number of equations is too large to solve using MATLAB or Excel solver, so GAMS 2.0 is used to successfully solve this problem.

Scenario 2: Taxi Drivers' Optimization by Maximizing the Revenue

Now, rather than seeking to minimize the total distance traveled by all empty taxicabs, suppose that each taxi driver will seek to maximize his or her own revenue. This analysis is performed assuming that the drivers have perfect information about where the demand is located and what characteristics the potential users have. The expected revenue can be calculated dividing the total fare to the time spent for each trip (including time driving, time looking for a fare using demand per hour, and how long the average trip takes).

$$R = \frac{\text{Fare}}{T} = \frac{\text{Fare}}{T_{\text{drive}} + T_{\text{search}} + T_{\text{trip}}} = \frac{\text{Fare}}{1/(\frac{\text{demand}}{\text{hour}})} = \text{Fare}\left(\frac{\text{demand}}{\text{hour}}\right) \qquad \textbf{Eq 4.7}$$

where R is the revenue in U.S. dollar, T is the total time in hours spent for each trip, estimated by the average demand per hour, T_{drive} is the time driving to a location of interest, T_{search} is the time spent looking for a customer, and T_{trip} is the time that a trip takes.

The expected revenue per time is denoted by α, and this can be calculated by:

$$\alpha = \frac{sum(\text{Fare})}{sum\left(T_{\text{drive}} + T_{\text{search}} + T_{\text{trip}}\right)} = \frac{sum(\text{Fare})}{sum\left(T_{\text{empty}} + T_{\text{trip}}\right)} \ \forall \text{all taxi records} \qquad \textbf{Eq 4.8}$$

The sum of T_{drive} and T_{search} can be represented by the time that a taxi runs empty, which is calculated from the dataset before. The average value of time for a taxi driver is estimated to be \$19.66/hour using the extracted partial taxi data with only the records had the previous drop-off trips outside Manhattan.

Some assumptions are made to simplify the analysis. It is assumed that all taxi drivers are rational and well-informed about the demand everywhere and about the empty travel time needed to reach the popular spots in NYC. Then it is reasonable to assume that R is the same for all taxi drivers from i to j, because they will make similar decisions to choose the best destination j to maximize their revenue. Maximizing revenue is the same as minimizing the time spent for each trip.

A behavior parameter, k, is introduced, to represent the census tract that a taxi driver will most likely to choose. From the measured data, k is set to be the census tract where the maximum taxi trips (of all pickups) are headed compared to number of taxi pickups in other census tracts after a drop-off. For example, after a taxi makes a drop-off at census tract $i = 1621$ (Census tract GEOID= 36081033100) where LaGuardia airport (LGA) is located, census tract $j = 1621$ also gets the largest number of taxi pickups (15,795 pickups out of 31,888 pickups). This indicates that taxi drivers who make a drop-off near LGA airport are likely to stay at the airport to pick up their next customer because they are aware that there is a higher probability to find a pickup at the airport instead of circulating the surrounding neighborhoods or spending time going back to Manhattan empty.

The results show that there are some census tracts for which multiple census tracts j correspond to the maximum number of taxi pickups after dropping off a passenger at census tract i. Therefore, the number of census tracts with tied maximum number of pickups, k, is defined in Scenario 2.

Scenario 2 maximizes the revenue by evenly distributing all trips from i to end in the census tracts with the maximum numbers of observed trip counts. An example is shown in Table 4.2 to explain the calculation. If there is only one census tract with maximum pickups (optimal location for taxi drivers if they want to maximize revenue), it is easy to just choose all trips end in that single one census tract. If there are two or more census tracts with maximum pickups, all taxis will be distributed evenly to all of those census tracts. The distance calculation is shown as an example in Table 4.2.

Table 4.2 Example Showing How to Calculate Total distance Based on k.

O-D table	Destined in j=									Total
	1	2	3	4	5	6	7	8	9	
Trips originated i=1	2	3	4	10	10	10	5	0	9	53
d_{ij}	50	67	300	200	100	60	120	112	100	
If k=1						√				
Total distance						60*53				3180
If k=3				√	√	√				
Total distance				200*53/3	100*53/3	60*53/3				6360

Comparison between Scenario 1 and Scenario 2

The summary of the results from the scenarios is listed in Table 4.3. As shown in Table 4.3, after redistributing the O-D counts based on each scenario, the number of fields in the O-D table that have a value greater than zero is significantly reduced. If the sum of either each row, or each column, of the O-D matrix is larger than zero, meaning that the census tract has at least one pickup or drop-off trip. There are 2143 census tracts (out of a total of 2167 census tracts) that have information on either pickups or drop-offs in both the observed data and LP results. This indicates that there are a number of taxi drivers who find their next pickups immediately after dropping off a passenger outside Manhattan, which makes the taxi systematically efficient (minimizing the total distance traveled for all taxis). However, the number of census tracts with more than one pickup or drop-off for scenario 2 is 357, which is much less than 2143, revealing that only a partial set of the census tracts are popular to cab drivers, because cab drivers have incentives to choose the census tract with maximum revenue for their next customer.

Table 4.3 Results from the Observation and Modeled Scenarios.

	Observation	Scenario 1	Scenario 2
Method	Aggregation from original data	Linear Programming (LP)	Maximize the revenue by maximizing the number of trips per unit of time
Number of O-D fields that have value>0	57617	3569	3322
Number of census tracts that have at least 1 drop-off or pickup	2143	2143	357
Avg (O-D matrix)	3	61.09	65
Var (O-D matrix)	5,462.60	322,029.71	364,767.68
Total number of trips/week	218022	218022	218022
Total distance for all trips (km)/week	970,021	645,130	521,019

The linear programming algorithm (Scenario 1) is designed to minimize the total travel distance in the system by keeping the total trips that are originated in i and total trips that are destined for j equal to that of the observed drop-offs and pickups, respectively. However, the algorithm for Scenario 2 is based on maximizing the revenue, and the cab drivers are likely to choose destination census tracts k based on popularity. Therefore, the total number of empty trip originated in each census tract i (D_i) equals to the observed data but the total number of trips destined in each census tract j (P_j) does not have to equal to the observed pickups (Table 4.3). But the sum of D_i and the sum of P_j

should always equal to 218,022 trips. The O-D matrix of Scenario 2 has the highest variance than the rest of scenarios and the observed O-D data has the least variance (Table 4.3).

The total distance in meters is estimated by calculating the straight line distance from the centroid of census tract i to the centroid of census tract j for each scenario. The approximation is very useful to get an understanding of which model is more appropriate. However, the estimation is not 100% accurate, because the straight-line distance is always shorter than the actual trip distance and the distance will be zero if $i = j$. As predicted, Scenario 1 has a much lower distance travelled than the observed data. Scenario 2 has relatively less distance travelled than the observed data (Table 4.3). The reason why Scenario 1 has larger distance traveled than Scenario 2 is because the LP algorithm constrains that the sum of P_j should also equal to that of the observation.

If the O-D matrix is aggregated into the census tracts, both the observed data and LP should have the same aggregated number of pickups and drop-offs for each census tract. However, there are differences between the distribution of the counts for observed data, the LP, and Scenarios 2. Those aggregated trip counts are mapped into ArcGIS (Figure 4.9) to visualize the differences of taxi drivers' choices. All three maps (Figure 4.9a-c) show that census tracts with LGA airport and JFK airport located have relatively higher supply of taxi services than the rest of the NYC, including Manhattan, which indicates that taxi drivers strongly favor hotspots such as airports. This also reveals that taxi drivers are still likely to come back to Manhattan to get the following pickups after they drop off a passenger outside Manhattan.

The aggregated pickups from Scenario 2 (Figure 4.9c) seem to follow the trend of the observed pickups except that most of the areas outside Manhattan do not have pickups. This shows that even though we assume that every taxi driver goes to the most popular census tract for the following pickups after dropping off a customer outside Manhattan, we still end up with so many popular places outside Manhattan except the airports, for example, Williamsburg, western Queens, and Flushing. Some drivers are willing to drive a long distance to get back to Manhattan, especially lower Manhattan, to get the next pickups.

As introduced in the beginning of this section, a major limitation of scenario 2 is that taxi drivers would never have such perfect information on the popular locations, where they can maximize revenue. If so, all the popular areas would be overserved as they are targeted by all taxicabs; and the rest of the areas would lose the level of taxi service. The result is a scenario with severe imbalance of taxi demand and supply that will not be repeated. Yellow taxis will potentially lose customers to crowdsourcing services such as Uber and public transportations such as subways and buses. In reality, taxi drivers re-distribute themselves based on their past experiences. Once taxi drivers realize that they cannot all make profit in the same locations, there will be a process of redistribution that will tend toward something like what they do today: drivers go to places that they think will be busy enough to get them a customer but not so busy that they will have to compete with too many other cabs. Therefore, they tend to behave somewhere between scenario 2 (with complete information) and a state with repeatable information that can be predicted for any day, e.g., taxi drivers tend to go to Manhattan because they know that Manhattan has more demand than the Outer Boroughs.

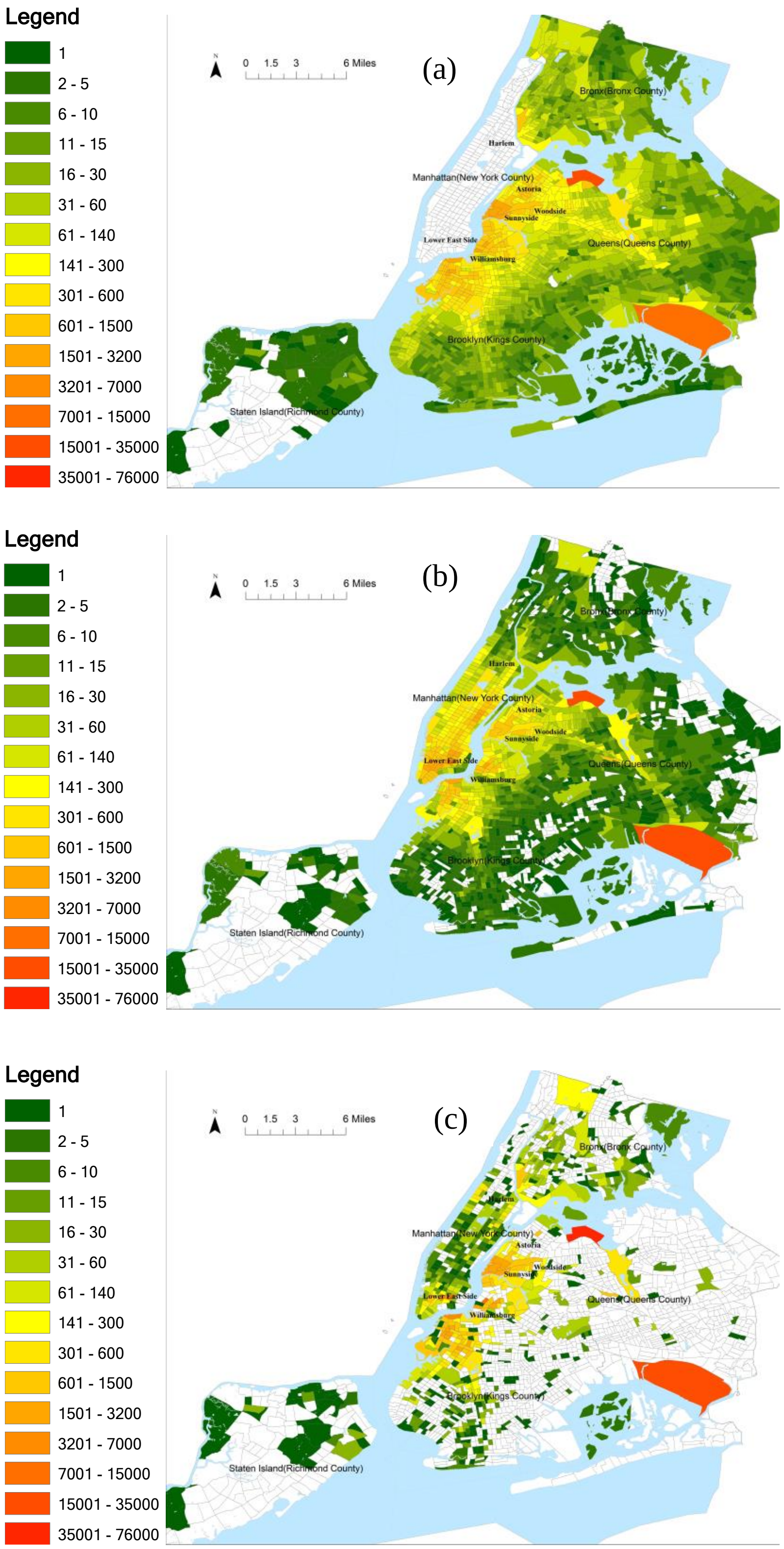


Figure 4.9 Model Output. (a) Distribution of Observed (Scenario 1 and Scenario 2) Drop-off Trips Outside Manhattan; (b) Distribution of Observed (Scenario 1) Pickup Trips; (c) Distribution of (Scenario 2) Pickup Trips.

4.5 Summary of Findings

There are three parts of analysis performed in this chapter: the analysis of variation in taxi use over time, the spatial dimension analysis, and efficiency of empty taxi customer search analysis. This chapter helps prepare the data for NYC taxi supply, and its temporal and spatial analysis provides evidence that is necessary to build a trip generation model later.

The analysis of variation in taxi use over time indicates that there are strong weekly patterns in the five time series representing taxi supply, e.g., total trips each day and total operating hours each day. The day of the week and holiday are found to be most important factors influencing taxi supply. Heavy snow is a less important factor. Precipitation and month of the year are not influential factors. The average temperature is not significantly related to taxi supply. The number of taxi medallions remains the same for the time period of this study, thus it is not included in the analysis of variation in taxi use over time.

The analysis of the spatial dimension prepares data for the efficiency study of the customer search model. The spatial dimension analysis also identifies popular neighborhoods and points of interest for empty taxi drivers, while efficiency analysis helps evaluate the efficient empty taxi driver in the search of the next customer after dropping off a customer outside Manhattan. Both analyses are important components in a customer search model.

This study demonstrates a data processing and modeling method to show the locations of interest for taxi drivers in New York City using a one-week sample of taxi GPS data. In the summary of the taxi data, it was found that the number of medallions, shifts, and trips during weekdays (especially Wednesday and Thursday) are much higher than on the weekends. This relationship between the number of medallions and number of shifts facilitates calculation of the distance and time traveled while the taxi is empty. The empty time and empty distance are much less during the peak hours than the rest of the day. The percentage of time that a taxi is empty ranges from 30% to 65%, and the percentage of vehicle miles traveled (VMT) that a taxi is empty ranges from 16% to 40%. This finding reveals that taxi drivers might be driving slowly while circulating empty. The smaller percentage of empty travel distance may also be due to the fact that a lot of taxicabs are not moving while waiting at the airports or cabstands.

As indicated from Table 4.3, two scenarios are designed from the systematic perspective (Scenario 1) and the taxi driver's perspective (Scenario 2), to model the taxi drivers' operational behaviors and compare their efficiency with the observed data. Scenario 1 optimizes the best combinations of demand and supply at each census tract by minimizing the total empty distance traveled for all taxis (at system level) in NYC at this one-week period. Scenario 2 calculates the demand and supply for each census tract by maximizing the revenue of each taxi driver (individual level). Also, Scenario 2 has less total distance travelled compared to Scenario 1 and the observation, which implies that the taxi system can be made more efficient if the taxi drivers are well-informed.

The contribution of this study includes: 1) novel methods in estimating the trips that have drop-offs in a less popular region (e.g. in this study we use outside Manhattan); 2) quantifying the efficiency of taxis by total distance traveled; 3) comparison of scenarios from a system perspective and individual's perspective; 4) visualization and comparison of the taxi drivers' places of interests providing empirical information on when and where the demand is for taxi drivers and when and where taxicabs are available for people who use taxi service in NYC.

The study can also be made more detailed by:

1) estimating a distance as close as the actual trip distance instead of using straight line distance;
2) building more complicated models for individual drivers by distribute the taxi drivers based on the probability that any census tract is chosen by a taxi driver instead of just selecting the most popular census tract to pick up an passenger;
3) constructing a complete list of places of interests based on the pickups and drop-offs, but this could be a separate study;
4) conducting potential causality analysis: do taxi drivers seek out customers or vice versa, or do they seek out one another simultaneously.

CHAPTER 5

TAXI TRIP GENERATION

Taxis play a significant role as a transportation mode in New York City (NYC). Taxis make 175 million trips annually (Bloomberg et al., 2014). The Taxi and Limousine Commission (TLC) regulates all NYC taxis and issues medallions and sets fare rules, although cab drivers choose where to circulate to pick up passengers. NYC yellow taxis have been facing new challenges since ride-share services such as Uber and Lyft have entered the NYC taxi market. To effectively plan and manage the taxi fleet, including both yellow taxis and the emerging ride-share taxi services, one must understand what factors drive yellow taxi demand; how taxi use is related to the availability of public transit; and how these patterns vary over space and time. A trip generation model that relates taxi demand to observable characteristics of a neighborhood (e.g., demographics, employment, and transit accessibility) is useful for planners and policy makers to manage taxi services effectively. Also, trip generation modeling can help identify neighborhoods and hours with insufficient (underserved) yellow taxi demand, which are potential areas and times that can be served by ride-share services. In this chapter, a case study of NYC is used to examine taxi trip generation using various modeling approaches.

5.1 Variables

The dependent variable, taxi demand, is the number of customer pickups aggregated by each hour of the day because the trip generation during each hour varies throughout the day (Yazici, 2012). Thus one model is developed for each hour to understand the temporal variation. Since the Transportation Analysis Zones (TAZ) in this study are the census tracts, all of the data is grouped by census tract so that the dependent variable and independent variables are aggregated at the same spatial resolution.

The sources of data for the independent factors considered in this study include:

- Transit Level of Service (LOS) based on NYC subway schedules, which are available from Google Transit Feed Data in the format of General Transit Feed Specification (GTFS);
- Demographics data for each census tract is available from the U.S. census 2010, including total population, population categorized by age, percent of population by age groups, and population categorized by race;
- Socioeconomic data is available from the American Community Survey (ACS) 5-year estimate of education and income;
- Employment data by census tract, including the number of jobs categorized by age groups, earnings, type, race, ethnicity, educational attainment, and sex is available for NYC from 2010 Workplace Area Characteristic (WAC) data from the U.S. Census Bureau; and
- PLUTO (extensive land use and geographic data) and MapPLUTO (Extensive land use and geographic data at the tax lot level in ESRI ArcGIS shape format

and dbase table format) data available from the NYC Department of City Planning (DCP) which includes 11 land use categories in the scale of tax lot.

With the availability of various data sources, it is possible to create a number of independent variables. Most of the independent variables are built by aggregating data to the level of census tract. Two novel variables are created: Transit Access Time (TAT) and the number of taxicab drop-offs in each census tract during each hour (DrpOff). TAT is the estimated walking and waiting time to get to the next transit services. DrpOff is added as an additional independent variable representing the supply of taxicabs that is immediately available at each location and time. DrpOff is created by aggregating the number of taxicab drop-offs by census tract and by time of day, similar to the creation of the taxi demand variable whereas the TAT is more difficult to calculate. Therefore, its calculation is explained in detail in the following sub-section.

5.1.1 Transit Access Time (TAT)

Transit level of services (LOS) and accessibility must be quantified in order to be used as an independent variable to model taxi trips. A new measure is developed that combines the estimated walking time a person must spend to access the nearest station (transit accessibility) and the estimated time that person will wait for transit service (transit LOS). This measure is the Transit Access Time (TAT), and it represents the minimum expected time for a person at a specific location and time-of-day to walk to, wait for, and board a transit vehicle. For a walking speed of 3.1 mph (5.0 kph) (Browning, 2005), the transit access time in minutes is:

$$\mathrm{TAT} = \frac{60D}{v_w} + \frac{60}{f} \qquad \textbf{Eq 5.1}$$

where f is the frequency of subway dispatches per hour at the nearest station, D is the distance to the nearest station (mi), and v_w is the walking speed (mph).

The minimum TAT is calculated at each location by the following steps. First, the transit schedule in GTFS provides the number of transit departures (i.e., frequency) in each hour at each station. The waiting time depends on the frequency based on the second term of Equation 5.1, and it is calculated separately for each hour of the day to account for variations in the schedule. Then, a fine grid is imposed on the study area with cells measuring 250×250 meters (820 feet), which is small enough that the walking time to cross each cell is less than 1 minute. Each cell is characterized by the location of its centroid, and a TAT will be calculated for each cell. A modified K-nearest neighbor algorithm is implemented by calculating the minimum TAT from the k nearest transit stations by screening distance and waiting time to all transit stations from the centroid.

People are assumed to be well-informed about transit schedules and to choose the nearest station that minimizes the sum of their walking and waiting time. Thus, the TAT is a metric of transit accessibility that is independent of specific origin-destination demand patterns. For simplicity, the method looks only at the closest access from each location (cell centroid) to the nearest subway departure, in space and time, anywhere in the system. The minimum TAT is calculated for each cell in NYC at each hour of the day, and this is used to quantify transit accessibility in the city with a spatial resolution of 250 meters (820 feet) and temporal resolution of one hour.

Each census tract is composed of more than one cell (each cell measuring 250×250 meters). Once the minimum TAT for each census tract is determined, it is necessary to calculate TAT by averaging the values across the cells included within the census tract. This provides a better TAT measure than simply calculating from census tract centroids, because a large census tract may have a centroid near a transit station but lots of land that has relatively low accessibility. The TAT is calculated for different times of the day for each census tract using only the subway data in this study, because the complete GTFS bus schedule data is incomplete (e.g., bus data for Queens are not available).

Visualizing TAT, DrpOff, and Taxi Demand

In order to visualize New York City with both information of TAT and taxi demand, I attempt to include all of them in Figure 5.1. Because the distribution of activities in NYC changes with time, taxi trips are separated by the hour of the day. Figure 5.1 shows the TAT for subways at 12:00 a.m. (midnight) and 5:00 p.m. (afternoon peak) along with both taxi pickups and drop-offs per capita in the same hours. The map of TAT shows that there is greater transit accessibility in Manhattan and along the subway routes than in other parts of the city, which is expected given the spatial coverage of the subway network. The transit accessibility is also generally greater at 5 p.m. than at 12 a.m., because services operate more frequently during the peak hours than late at night. Figure 5.1 suggests that the pickups and drop-offs per capita are higher where the TAT is lower (i.e., transit is more accessible), which suggests a negative correlation between TAT and taxi use.

There are also differences between the rates of taxi pickups per capita at 5 p.m. and at 12 a.m. For example, there are more taxi pickups at Jamaica at 5 p.m. than that at 12 a.m., which could result from people getting off the subway at Jamaica and then taking a taxi to complete a trip from work to home. In some areas of lower Manhattan there are more pickups at 12 a.m. than at 5 p.m., which indicates concentrations of nightlife.

The drop-offs per capita show large differences between 5 p.m. and 12 a.m. as well. For example, there are more drop-offs per capita at some popular locations such as Penn Station, Grand Central Station, and Flushing at 5 p.m. than at 12 a.m., which is consistent with commuters using these busy transit hubs. Although the total amount of travel activity in the city is lower at midnight than at 5p.m., many areas of the Outer Boroughs actually see a greater rate of drop-offs in the late night hours. This suggests that people use taxis more often to travel to outlying neighborhoods when it is dark and transit services are less frequent. There appears to be a consistent trend at all times of day that pickups are more concentrated around transit hubs and central areas, whereas drop-offs are more dispersed around the city. Clearly, trip making behavior by taxis is asymmetric.

5.1.2 Research Hypotheses

As indicated from literature, taxi demand is often closely related to the characteristics of people (demographics and socioeconomics), and to the level of services of the public transportation (subway accessibility and walkability) in the area as well (O’Neill and Brown, 2011; Racca and Tatledge, 2004; Kumar and Levinson, 1992; Schaller, 2005). Therefore, six theoretically important variables are considered in the following sub-sections: TAT, taxi drop-offs (DrpOff), income, education, population, and total jobs.

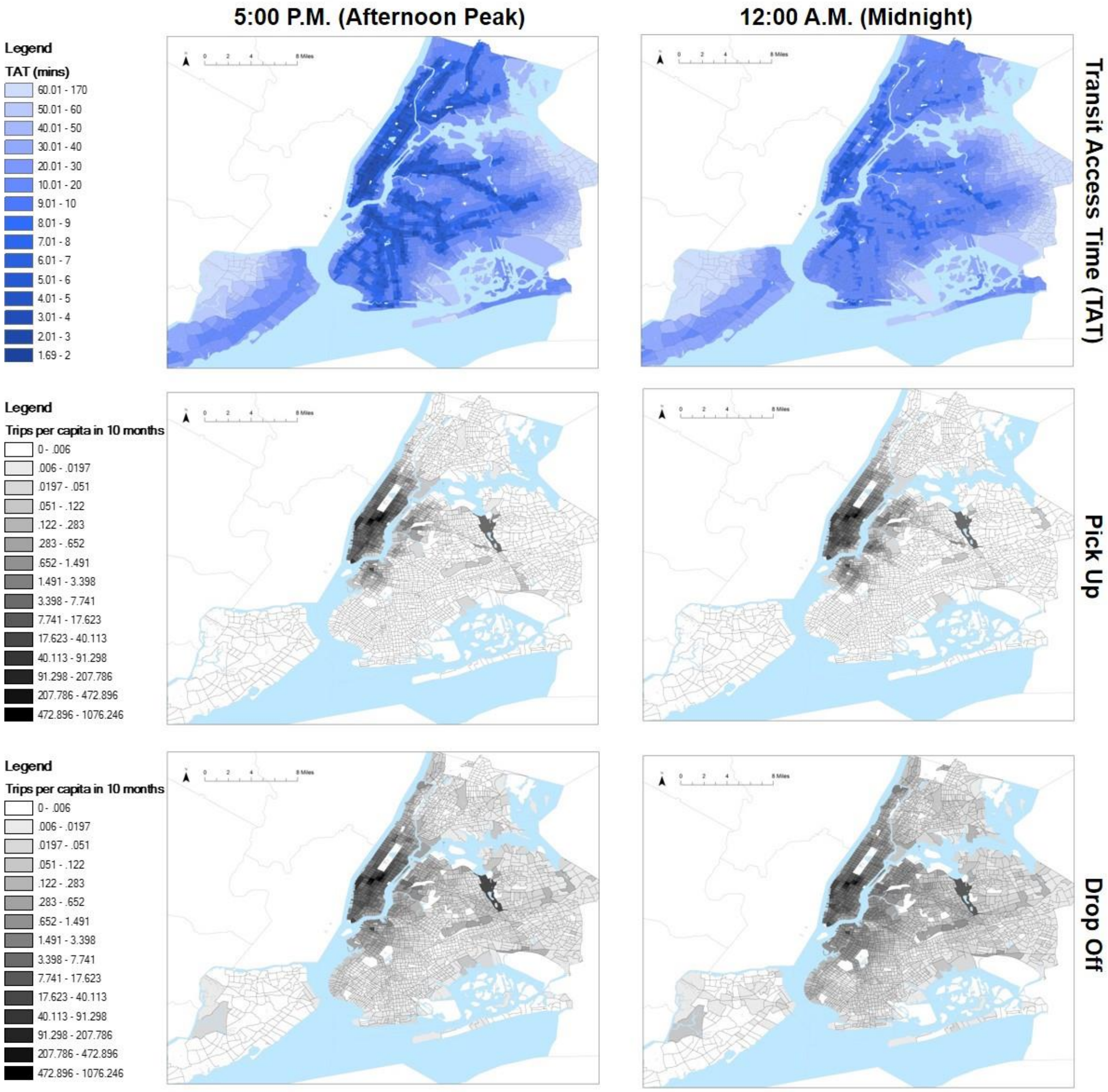


Figure 5.1 TAT and Pickup and Drop-off Taxi Demand per Capita at 5 p.m. and 12 a.m.

TAT

As explained in detail from the previous section, transit access time (TAT) is correlated to taxi demand both in theory and from the maps. While areas with easier transit accessibility tend to generate more taxi trips, such as Central Manhattan, those areas with less transit accessibility tend to generate fewer taxi trips, such as most of the outer boroughs.

Taxi Drop-off

The DrpOff variable indicates the immediate availability of taxis at each census tract and each hour. Conceptually, the number of taxi drop-offs represents the “supply” of empty taxicabs that are immediately available and can be hailed from the streets. An area with more empty taxis is likely to generate more taxi trips. For example, people may go to those places to hail a cab knowing that there are enough vacant taxicabs available, especially at popular areas such as a transit hub. It is also evident from Figure 5.1 that DrpOff per capita is very closely related to taxi demand (the number of pickups per capita).

Income and Education

In theory, taxi travel is an economic behavior/decision. Those people with relatively higher average income (e.g., higher per capital income) are likely to afford taking taxis on a frequent basis. Education attainment is correlated with income, and since well-educated people are likely to be wealthier, they tend to generate taxi trips as well, making both income and education important factors.

Population

Also, neighborhoods with larger populations are likely to generate more taxi trips. Figure 5.2 shows that the population density is concentrated in Manhattan, where there is a higher level of taxi demand, and conversely, a smaller population density is observed in the Outer Boroughs, where there is a lower level of taxi demand. Therefore, a positive correlation between the population and taxi demand is expected.

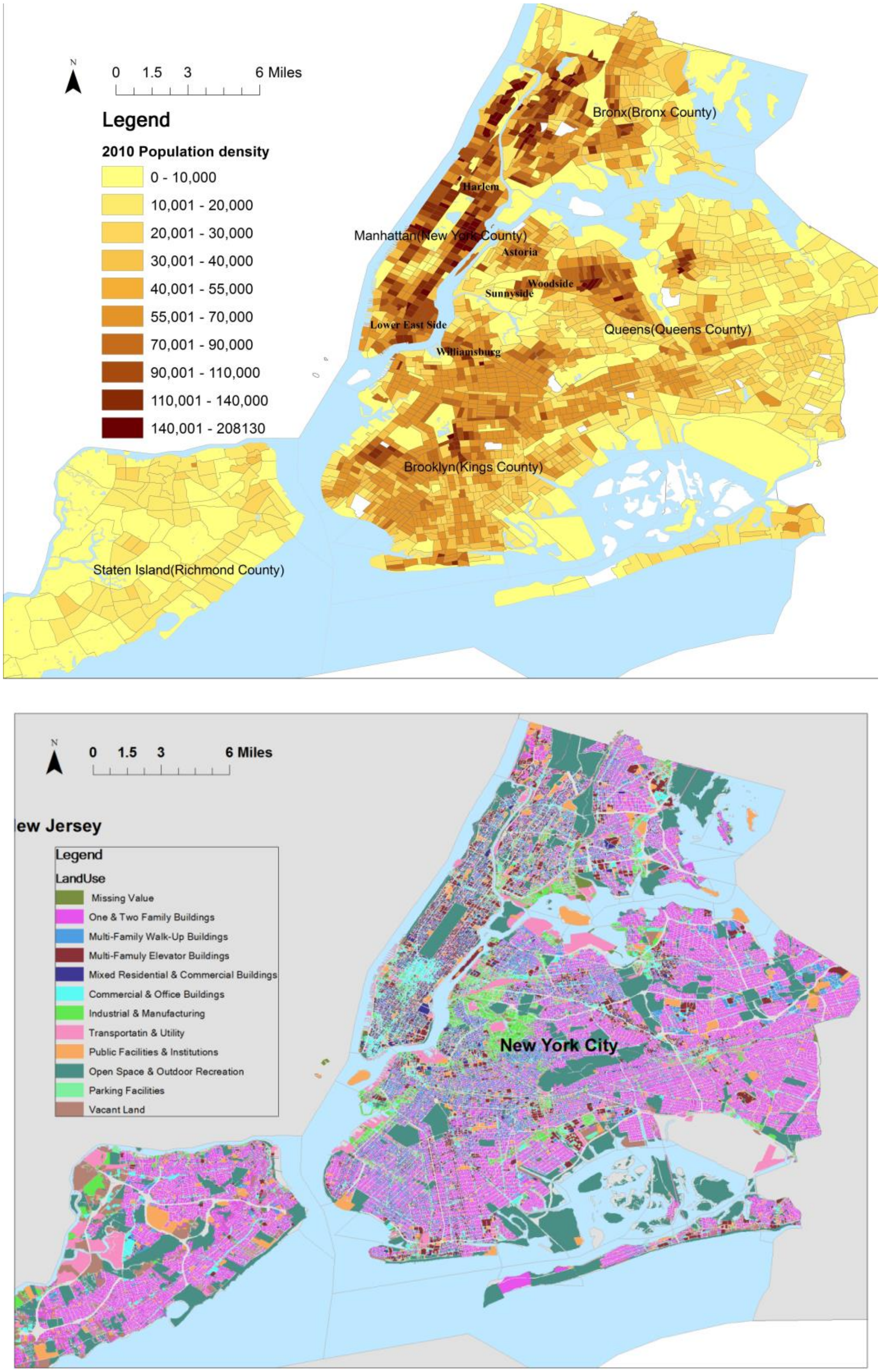


Figure 5.2 2010 Population Density and Land Use Data in New York City.

Percentage of the Population by Age Groups

The percentages of population are calculated for four age groups: ages below 15 (PerAge0-14), ages between 15 and 34 (PerAge15-34), ages between 35 and 64 (PerAge13-64), and ages 65 and above (PerAge65). The age groups are aggregated so that within each age group, people have relatively homogeneous behaviors. For example, ages below 15 are not likely to travel on their own by taxi.

Some of those age groups are likely to influence the generation of taxi trips. The NYC TLC In-Taxi Passenger Surveys (Bloomberg et al., 2014) found that, "over two-thirds of taxi passengers are 35 or under, with 35% of passengers reporting that they are younger than 20 years old, while another 35% report being between 21 and 35 years old." Although those numbers are over-represented because of the limitation of the survey, it indicates that the major taxi passengers are likely to be younger than 35 years old. Those aged between 15 and 34 have higher activity levels, so they are likely to generate more taxi trips. Those aged 65 and above may have more limited mobility, and they may demand more taxis for activities. Ages between 35 to 64 and ages above 64 are unlikely to have the same level of activities compared to that of the younger people, therefore, those two age groups tend to generate fewer taxi trips compared to younger people.

Although we believe that individuals from each age group are likely to generate different levels of taxi trips, we don’t know whether the variable, the percentage of population in each age group for all census tracts, would exhibit any relationship with taxi demand. Therefore, the population percentage of each age group is included as a control variable.

Total Jobs and Job Categories

Figure 5.2 and Figure 3.4 shows that the total job density, similar to population density, is also highly concentrated in Manhattan. Similar to the pattern of taxi demand, there are fewer jobs available in the Outer Boroughs; therefore, total jobs (TotJob) might be correlated to taxi demand as well. The categorization of the number of jobs, such as jobs by age, earnings, types, race, ethnicity, education attainment, and sex could also be potential factors that affect taxi demand. In theory, taxi travel is often closely related to the travel purpose; therefore, areas with different type of jobs might have higher correlation with taxi demand compared to other categorization of jobs and may demand a different level of taxi services at different times of the day. Also, the jobs by types could reflect when and where people would like to do a certain type of activity. For example, places with many jobs in high-technology information and finance attract taxi trips to visiting in the morning; places with food and accommodation attract taxi trips to eat in the afternoon and evening; places with retail jobs attract taxi trips to shopping mostly in the afternoon and evening; and places with night club jobs attract taxi trips to parties at night.

Some census tracts consisting of cemeteries, parks, or islands do not have employment associated with them, so the WAC employment data covers 2,143 census tracts. Census tracts that have variables that are lacking certain required information are excluded from the linear model analysis. Ultimately, 116 out of 2,167 census tracts (5%) were omitted from the analysis if the employment variables are included in the model, because there was insufficient population or employment in those few regions to create a useful data point.

Land Use Categories

The PLUTO data in tax lot level is plotted in Figure 5.2. Similar to population density, Figure 5.2 shows that a mixture of different land use types are concentrated in Manhattan and its surrounding neighborhoods in the Outer Boroughs, where there is a high level of taxi demand. Mixed residential and commercial buildings (LU04) exist in most of those neighborhoods as well, so they might be related to taxi trip generation. Commercial and office buildings (LU05) are concentrated in Lower and Midtown Manhattan, but the rest of Manhattan is fairly mixed with different land use types. One and two family housing (LU01) are common in the other four boroughs. The west part of Queens has some industrial and manufacturing land uses. Figure 5.1 shows that most pickups are concentrated in Manhattan, northern Brooklyn, and the west and north sides of Queens, while the drop-offs are more spread over the boroughs of Manhattan, Brooklyn, Queens, and Bronx.

A land use variable is the ratio of areas between a land use type and the census tract. To be consistent, the percentage of land use type is labeled the same way as the land use types. Commercial and mixed land use types, in theory, are likely to be related to where people travel for social and economic activities; therefore, LU04 and LU05 are hypothesized to be closely related to the generation of taxi trips.

One of the difficulties in this study is the reliability and sensitivity of aggregating land use data to a larger scale (census tract) rather than a small scale (tax lot). Also, the relationship between land use categories and people's activity depends on how land use is categorized according to building classes. For example, same activities could be related to several different land use categories. Therefore, there might be some difficulties in creating independent factors by aggregating the land use categories before having an idea of how they are related to taxi trips. Some variables are lacking information for certain census tracts, and those census tracts are excluded from the linear model analysis. Ultimately, we omitted about 5% (118 census tracts) out of 2,167 census tracts for NYC when land use categories are included in the model.

5.1.3 Initial Screening of the Variables

In the previous section, the theoretical relationships between the potential factors and taxi demand provide evidence for theoretically important factors. In this section, an initial screening for all variables is performed based on the relationship with taxi pickup trips by examining correlations. Additional information is also used to assist variable screening by mapping those variables in the previous section. As shown in Table 5.1, among all 88 independent variables, six major variables and three factor categories have been determined to affect taxi demand based on the literature and the initial screening of the correlation coefficient table (Table 5.2).

Table 5.1 List of Independent Variables in Trip Generation Model.

Factor group	Factors or factor category	No. of variables	Initial Screening	Theoretical Major Factors	Selected Categorical Factors
TAT	Transit Access Time (TAT) at specific hour*	1	√	√	—
Drop-off	The number of drop-off trips aggregated by hour and census tract (DrpOff)	1	√	√	—
Population	Total population (Pop)*	1	√	√	—
	Population by race	8	√	—	—
	Population by age	14	√	—	—
Age	Percentage of Population by age group*	4	√	—	√ (2 out of 4)
Education	Percentage education higher than high school	1	√	—	—
	Percentage education higher than Bachelor (EduBac)*	1	√	√	—
Income	Median household income*	1	√	—	—
	Mean household income	1	√	—	—
	Median family income	1	√	—	—
	Mean family income*	1	√	—	—
	Per capita income (CapInc)*	1	√	√	—
Employment	Total jobs (TotJob)*	1	√	√	—
	Jobs by age	3	√	—	—
	Jobs by earnings	3	√	—	—
	Jobs by types*	20	√	—	√ (6 out of 20)
	Jobs by race	6	√	—	—
	Jobs by ethnicity	2	√	—	—
	Jobs by education attainment	4	√	—	—
	Jobs by sex*	2	√	—	—
Land Use	Land Use category*	11	√	—	√ (2 out of 11)
Total No. of variables		88	88	6	35

*****influential factors or factor category identified that are correlated with taxi demand
√ factor considered in the model
— factors omitted from the model

Table 5.2 List of Correlation Coefficient (R) between Independent Variables and Dependent Variables.

Correlation Coefficient, R	Pickup_hr0 (12 A.M.)	Pickup_hr7 (7 A.M.)	Pickup_hr17 (5 P.M.)
TAT_hr0	-0.16	-0.16	-0.17
TAT_hr7	-0.15	-0.15	-0.16
TAT_hr17	-0.16	-0.15	-0.16
DrpOff_hr0 (12 A.M.)	0.95	0.83	0.85
DrpOff_hr7 (7 A.M.)	0.63	0.73	0.84
DrpOff_hr17 (5 P.M.)	0.85	0.93	0.98
Pop	0.16	0.23	0.16
PerAge0-14	-0.41	-0.37	-0.40
PerAge15-34	0.31	0.23	0.22
EduBac	0.48	0.51	0.54
CapInc	0.57	0.65	0.71
TotJob	0.47	0.46	0.56
JobRet	0.44	0.42	0.56
JobInf	0.55	0.48	0.63
JobFin	0.36	0.36	0.48
JobRea	0.42	0.45	0.55
JobPro	0.58	0.58	0.71
JobFod	0.77	0.63	0.80
LU04	0.32	0.33	0.30
LU05	0.48	0.41	0.55

This table only includes independent variables that have correlation coefficient larger than 0.3 for at least one hour.

‘JobRet’ Number of jobs in Retail Trade; JobInf' Number of Jobs in Information; ‘JobFin’ Number of jobs in Finance and Insurance; ‘JobRea’ Number of jobs in Real Estate and Rental and Enterprises; ‘JobPro’ Number of jobs in Professional, Scientific, and Enterprises; ‘JobFod’ Number of jobs in Accommodation and Food Services; ‘LU04’ Mixed Residential & Commercial Buildings; ‘LU05’ Commercial & Office Buildings

In theory, TAT, drop-off, population, education, income, total jobs are related to taxi demand; therefore, they are considered as controlling variables regardless of the value of the correlation coefficient. Population (Pop) is selected because it is a theoretically important factor. If population is included in the model, population by race and population by age have to be left out because they are highly correlated with total population (R ranges from 0.7 to 0.9). Instead, the percentage of the population for different age groups is considered. Percentage of population with an education attainment higher than a bachelor degree (EduBac) is selected because it has higher correlation with taxi demand compared to the percentage of population with an education attainment higher than high school. Similarly, per capita income (CapInc) is selected because it is more correlated to the number of taxi pickups compared to other income variables. Total job (TotJob) is a theoretically important factor and it is highly correlated with taxi demand, so it is also selected in the initial screening process. 'DrpOff ' is highly correlated with taxi demand ((R= 0.7-0.9). It is expected that TAT would show a positive relationship with taxi demand while 'DrpOff' would show a negative relationship.

The three categorical factors including percentage of the population by age groups, jobs by type, and land use categories, are selected to be put in the model because some of those categories are highly correlated to taxi demand. For those categorical variables, usually a correlation coefficient, R, is considered to be 'high' if it is larger than 0.5; however, in this dissertation, to prevent the elimination of potential influential variables, an R value larger than 0.3 is considered in the initial screening process as shown in Table 5.2.

Two age categories,'PerAge0-14' and ' PerAge15-34', are selected after the initial screening. First, they have relatively higher correlation coefficient (R>0.3) with the dependent variable compared to the other two age groups. Second, it is evident that those two age groups are major taxi passengers. The predicted relationship between 'PerAge15-34' and taxi demand is positive, while the relationship between 'PerAge0-14' and taxi demand is negative. Ages between 15 and 34 are likely to have a good job with stable income and to generate more taxi trips for commuting (between work and home) and social activities. They are most likely to be able to afford travelling by taxi as well. However, areas with higher percentages of children are correlated to lower incomes and lower education levels, thus, areas with higher percentage of age group 'PerAge0-14' are likely to generate less taxi demand.

Those two age variables are included in the experiments of model trials as soon as the appropriate modeling method is determined in later sections. In the final model, 'PerAge0-14' will be dropped, most probably, because it is highly correlated (negatively) with important theoretical factors such as income and education (R< - 0.5), which are preferred to be included in the model.

Among all 7 different categorization methods for jobs (Table 5.1), jobs by type is selected because, first, almost all job categories by type have relatively higher correlation (R=0.5~0.7) with taxi demand compared to the other 6 categorization systems; and, second, jobs by type is a better indication of people's travel purposes. Jobs by type variables are highly correlated to TotJob, so in order to prevent multi-collinearity, they cannot be included in the same model.

There are six job groups that are highly correlated to taxi demand (R>0.3):

- 'JobRet', the number of jobs in Retail Trade;
- 'JobInf', the number of Jobs in Information;
- 'JobFin', the number of jobs in Finance and Insurance;
- 'JobRea', the number of jobs in Real Estate and Rental and Enterprises;
- 'JobPro', the number of jobs in Professional, Scientific, and Enterprises;
- 'JobFod', the number of jobs in Accommodation and Food Services.

Among those job categories, the number of jobs in accommodation and food services (JobFod) have the highest correlation (R=0.7 ~0.8) with taxi demand. All job categories are expected to be positively related to taxi demand. Based on the above information, total jobs and some job categories will be selected to include in the following count regression model.

Among all land use categories, commercial and mixed land used related factor, LU04 and LU05, have relatively higher correlation with taxi demand compared to other land use categories. Therefore, they are selected in the initial screening process. As illustrated from the spatial distribution of each land use type above, it is expected that LU04 and LU05 are positively related to taxi demand.

Summary of Expected Results

In summary, I expect to observe significant relationships between the taxi demand and the population and transit accessibility. The mapping of TAT and taxi demand provides a visualization of their relationship and helps provide insights about why such a relationship exists. With the hourly data for TAT, taxi pickups, taxi drop-offs, and all other demographic and socioeconomic information, visual inspection of the maps is interesting but insufficient for determining the quantitative relationship between the independent variables and the taxi demand. To achieve this objective, a count regression model is introduced in the next section.

It is expected that the age groups are significantly related to taxi demand especially two groups: the PerAge0-14 and the PerAge15-34. I expect there to be a relationship between taxi demand and employment or land use, because taking taxis is a transportation method that is closely related to people’s activities, and both employment and land use could in some sense reflect when and where people would like to have their activities. For example, places with lots of financial jobs attract people who work there, resulting in trips peaking in the morning and in the afternoon while places with night club jobs attract people who like night life. Similarly, residential areas may have the most taxi trips in the morning and afternoon to travel to and from airports, while commercial areas may have taxi trips throughout the day. We expect to see some categories, such as mixed residential or commercial areas, are more influential in generating taxi trips than other land use categories.

5.2 Count Regression Model

There are two important empirical contributions of this study. The first is the development of a novel transit accessibility measure based on the time to access and wait for transit. This requires processing raw transit schedule information to determine how much time it takes per person at a specific location and time of day to access the public transit system, which has provided in the first section ‘variables’ in this chapter. The second is the development of a hybrid cross-classification/regression model for estimating taxi trip generation. The taxi data is cross-classified by pickup and drop-off and aggregated by hour of the day. Within each classification, a count regression model is estimated to identify the factors that influence taxi demand (the number of pickups). The number of taxi drop-offs is used as an independent variable representing taxi supply. We will first determine the modeling method and then select the appropriate variables for the model.

5.2.1 Methodology

Linear models have been broadly applied in trip generation (Bein-Edigbe and Rahman, 2010; Mousavi, 2012). Linear regression models are inadequate for count data because the dependent variable is a count of random events, which cannot be negative. As a result, the models that are built and compared in this study are count models that are specifically developed to represent count processes.

The count regression model is a Generalized Linear Model (GLM). The idea behind count regression modeling is similar to linear regression: to explore the relationship between the log (dependent variable) and independent variables with the assumption that this relationship is linear as follows

$$Y = e^{\sum_{i=1}^{n} \beta_i X_i + \beta_0 + \varepsilon} \qquad \textbf{Eq 5.2}$$

where Y is the number of taxi trips generated in a TAZ (dependent variable), X_i is one of n independent variables, β_0 is the intercept, β_i is the coefficient corresponding to X_i, and ε is the error representing the difference between the modeled and observed number of taxi trips.

The dependent variable in the model is the number of pickups generated in each census tract by hour of the day from the 10-month taxi GPS data in NYC. The full list of theoretically important independent variables considered in the initial model is shown in Table 5.1. Maximum likelihood estimation (MLE) method is used to estimate the model coefficients for each independent variable. The goal is to select a set of independent variables that results in low model error and in which each independent variable has a statistically significant coefficient. But before model selection, it is necessary to determine which type of count regression model is appropriate.

Modeling for Over-dispersed Data

There are three types of count regression models that are commonly used: Poisson, Quasi-Poisson, and negative binomial. First, the Poisson regression model is introduced, following the reference of Ismail and Jemain (2007). In order to account for the varying effects that each of the independent variables have on the dependent variable at different times of day, a separate model is estimated for the data in each hour.

Let Y_i be the independent Poisson random variable for the count of taxicab trips in census tract $i = 1 \ldots 2126$. The mean and variance of a Poisson distribution has a relationship as follows:

$$\mu_i = E(Y_i) = var(Y_i)$$ **Eq 5.3**

The Poisson process is based on the assumption that the mean of the independent random variables is equal to its variance. Therefore, Poisson regression is only appropriate if the observed count data has an equal mean and variance. If the observed variance for the count data exceeds the mean, the data is said to be overdispersed, and an alternative model specification using either quasi-likelihood estimation or negative binomial regression may be used instead. Both methods use a generalized linear model framework. The approach using quasi-likelihood estimation, which is similar to the Poisson-like assumptions except the relationship between mean and variance is not the same. The mean and variance of the random variable Y_i are:

$$E(Y_i) = \mu_i$$ **Eq 5.4**

$$Var(Y_i) = \theta\mu_i$$ **Eq 5.5**

where θ is a dispersion parameter; when $\theta = 1$, this model becomes a Poisson model.

The quasi model formulation leaves the parameters in a natural state and allows standard model diagnostics without losing efficient fitting algorithms.

The negative binomial model is characterized by a quadratic relationship between the variance and mean of the dependent variable (Hoef, 2007). For the negative binomial model, the mean of Y_i is still μ_i as given by Eq 5.3, but the variance is:

$$Var(Y_i) = \mu_i + {\mu_i}^2\alpha^{-1} = \mu_i + {\mu_i}^2\theta \qquad \textbf{Eq 5.6}$$

where α is the inverse of the overdispersion parameter, θ. The mean and the variance are equal if $\theta = 0$, so the Poisson distribution is also a special case of the negative binomial distribution. Values of $\theta > 0$ indicate that the variance exceeds the mean, and the observed distribution is overdispersed.

Table 5.3 Mean and Variance of Dependent Variable, Number of Pickups by Census Tract at each Hour.

Hour of day	Mean value of pickup counts for all census tracts	Variance of pickup counts for all census tracts
0	2446	100,804,416
1	1827	63,070,167
2	1373	41,751,337
3	994	24,875,367
4	711	10,338,265
5	558	4,547,384
6	1202	25,984,332
7	2213	79,611,311
8	2862	128,381,917
9	2948	137,735,882
10	2801	122,508,279
11	2857	131,214,089
12	3063	153,234,156
13	3030	149,228,639
14	3133	161,516,405
15	2976	143,190,381
16	2586	106,107,666
17	3115	151,429,667
18	3759	224,925,568
19	3913	244,880,173
20	3615	209,472,196
21	3469	195,600,486
22	3360	186,065,870
23	3017	147,131,917
All Hours	61,828	57,071,324,807

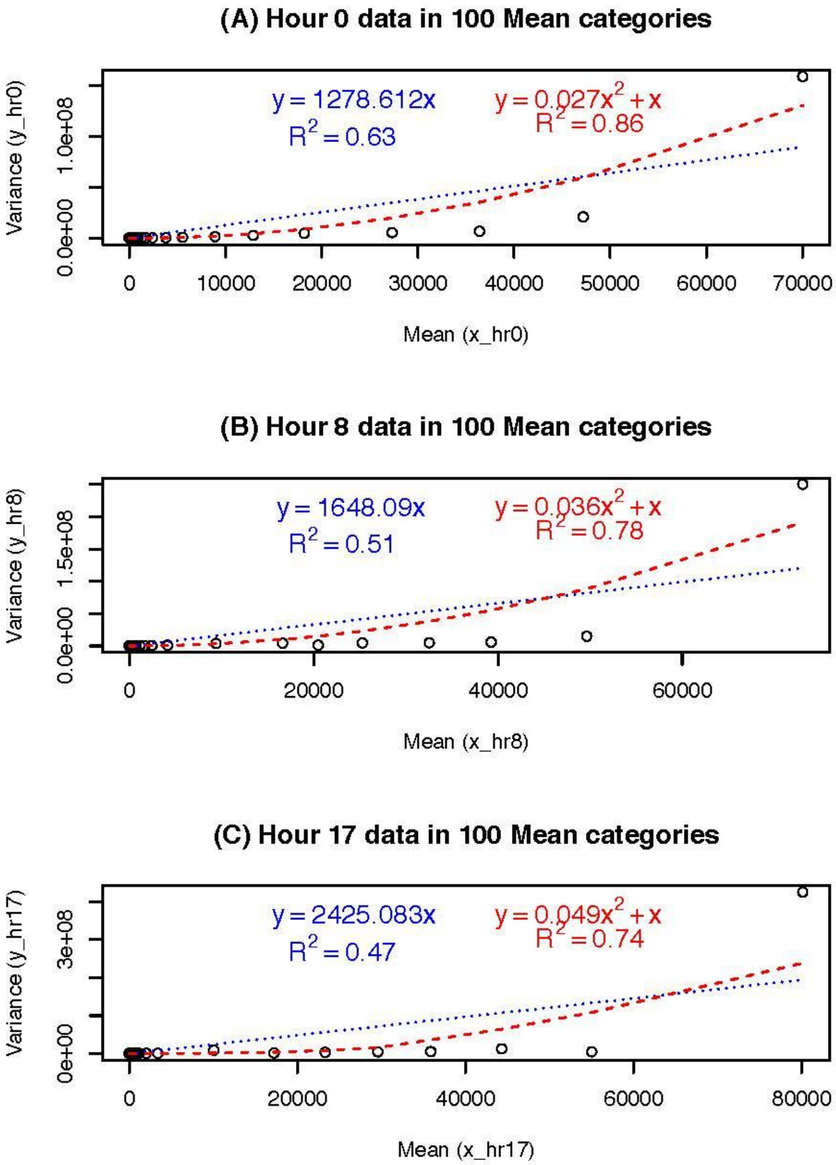
Table 5.3 presents a summary by hour of the day for the mean and variance of the total number of taxicab pickups per census tract in the ten-month data sample. The variance is much larger than the mean of the dependent variables. Therefore, it is not appropriate to use Poisson model. Instead, models for overdispered data should be considered.

Selecting Quasi-Poisson Distribution or Negative Binomial Distribution

Two types of models are generally used to account for the overdispersion of the count data. This sub-section introduces the comparison between those two modeling techniques: quasi-Poisson and negative binomial (NB). Both are estimated using maximum likelihood estimation (MLE) method, referred to as NB-MLE. In order to select the most appropriate model specification for the count regression, one must compare the mean and variance of the taxicab pickup counts, which are the dependent variable for the proposed model.

The variance of taxicab pickups per census tract in the ten-month dataset greatly exceeds the mean, as shown in Table 5.3, which provides an indication that the data is overdispersed. This pattern holds whether all counts from all hours of the day are considered together or the records are broken down by hour of the day. The implication is that the count model for the regression should be appropriate for overdispersed data. To choose between the quasi-Poisson distribution and the negative binomial distribution, it is necessary to look at how the mean and variance appear to be related. Since a goal of this study is to consider how the effect of independent variables changes with the hour of the day, a separate model is estimated for each hour, and the comparison of mean and variance must be considered within each hourly aggregation. Figure 5.3 presents separate plots comparing count mean and variance for three representative hours: hour 0 is 12:00 A.M. – 1:00 A.M. (midnight); hour 8 is 8:00 A.M. – 9:00 A.M. (morning peak); and hour 17 is 5:00 P.M. – 6:00 P.M. (evening peak).

In order to choose the distribution that most appropriately represents the dependent variable, the data within each hourly aggregation is divided into 100 subsets using the quantiles of the taxicab pickup counts. The first category includes taxicab pickups for census tracts in which counts fall between the 0 quantile and 0.01 quantile, the second category includes census tract data in the range of the 0.01 quantile and 0.02 quantile, and so on. Within each quantile category, the mean and variance of each subset of the data is calculated and plotted in Figure 5.3. A linear function of the form shown in Eq 5.5 estimates θ to match the data to the assumed relationship for a quasi-Poisson regression model. A quadratic function of the form shown in Eq 5.6 estimates α or θ to match the data to the assumed relationship for a negative binomial regression model. The linear equation for the quasi-Poisson model is shown in Figure 5.3 with the blue dotted line, while the quadratic equation for the negative binomial model is shown with the red dashed line. The goodness of fit parameter, R^2, is used to identify which specification fits the data better. A value of R^2 closer to 1 indicates a better fit. It can be seen in the examples for hours 0, 8, and 17 (Figure 5.3) that the quadratic function provides a better fit for relating the variance and mean, indicating that the negative binomial distribution is more appropriate for the counts of taxicab pickups.



(A) Hour 0 data in 100 Mean categories

(B) Hour 8 data in 100 Mean categories

(C) Hour 17 data in 100 Mean categories

Figure 5.3 Plot of the Variance vs. Mean of the Aggregated Hourly Taxicab Pick-up Counts for (a) Hour 0 (midnight), (b) Hour 8 (morning peak), and (c) Hour 17 (Evening Peak).

5.2.2 Model Selection

There are many methodological and statistical criteria for selecting important variables.

First, six theoretically important variables and three age groups are selected that are highly correlated with taxi demand in the variables section in the initial screening process.

Check Multi-collinearity

Next, the correlation between each pair of independent variables is checked to understand the correlation among the independent variables themselves. This information helps to prevent multi-collinearity by not including independent variables that are highly correlated with each other in the same model.

The correlation coefficient between each pair of independent variables is presented in Table 5.4. Important correlations among independent variables are summarized:

1. Age categories are highly correlated with themselves, but they are not correlated with TotJob, job categories, or Land Use categories, except that LU01 (One & Two Family Buildings) is correlated with PerAge15-34.
2. Job categories are highly correlated with themselves and TotJob; therefore, only one of those variables can be included in the model at a time.
3. Land use categories are correlated with themselves except LU04 and LU05 (R=0.11 between LU04 and LU05). Land use categories are also highly correlated with employment categories; therefore, they should not be included in the same model. LU04 is not correlated with TotJob (R=0.1), thus they can exist in the same model.
4. CapInc and EduBac are highly correlated with each other with R=0.8, thus they cannot be included in the same model.

Table 5.4 Correlation Coefficient (R) between Each Pair of Independent Variables.

R	End_hr0	End_hr7	End_hr17	Pop	PerAge15-34	EduBac	CapInc	TAT_hr0	TAT_hr7	TAT_hr17	TotJob	JobRet	JobInf	JobFin	JobRea	JobPro	JobFod	LU04	LU05
End_hr0	1	0.61	0.88	0.25	0.36	0.56	0.63	-0.20	-0.18	-0.19	0.46	0.40	0.49	0.34	0.43	0.58	0.70	0.39	0.42
End_hr7	0.61	1	0.80	0.03	0.18	0.36	0.49	-0.13	-0.12	-0.12	0.60	0.52	0.65	0.71	0.56	0.78	0.78	0.14	0.63
End_hr17	0.88	0.80	1	0.19	0.22	0.55	0.71	-0.17	-0.16	-0.16	0.53	0.54	0.61	0.42	0.51	0.67	0.78	0.32	0.51
Pop	0.25	0.03	0.19	1	0.08	0.08	0.09	-0.13	-0.13	-0.13	0.01	0.05	-0.01	-0.02	0.08	-0.02	0.07	0.37	-0.03
PerAge15-34	0.36	0.18	0.22	0.08	1	0.16	0.02	-0.33	-0.32	-0.32	0.26	0.14	0.15	0.18	0.10	0.22	0.18	0.28	0.21
EduBac	0.63	0.49	0.71	0.09	0.02	0.82	1	-0.05	-0.04	-0.05	0.37	0.39	0.35	0.28	0.43	0.40	0.47	0.27	0.31
CapInc	-0.20	-0.13	-0.17	-0.13	-0.33	-0.05	-0.05	1	1	1	-0.11	-0.09	-0.09	-0.07	-0.10	-0.11	-0.12	-0.34	-0.15
TAT_hr0	-0.18	-0.12	-0.16	-0.13	-0.32	-0.04	-0.04	1	1	1	-0.11	-0.09	-0.09	-0.07	-0.09	-0.1	-0.11	-0.33	-0.15
TAT_hr7	-0.19	-0.12	-0.16	-0.13	-0.32	-0.04	-0.05	1	1	1	-0.11	-0.09	-0.09	-0.07	-0.09	-0.1	-0.11	-0.33	-0.15
TAT_hr17	0.46	0.60	0.53	0.01	0.26	0.30	0.37	-0.11	-0.11	-0.11	1	0.47	0.56	0.57	0.50	0.67	0.60	0.10	0.60
TotJob	0.40	0.52	0.54	0.05	0.14	0.27	0.39	-0.09	-0.09	-0.09	0.47	1	0.43	0.32	0.44	0.51	0.53	0.13	0.65
JobRet	0.49	0.65	0.61	-0.01	0.15	0.26	0.35	-0.09	-0.09	-0.09	0.56	0.43	1	0.54	0.44	0.73	0.76	0.08	0.57
JobInf	0.34	0.71	0.42	-0.02	0.18	0.20	0.28	-0.07	-0.07	-0.07	0.57	0.32	0.54	1	0.45	0.69	0.56	0.04	0.51
JobFin	0.43	0.56	0.51	0.08	0.10	0.29	0.43	-0.1	-0.09	-0.09	0.50	0.44	0.44	0.45	1	0.52	0.49	0.14	0.43
JobRea	0.58	0.78	0.67	-0.02	0.22	0.30	0.40	-0.11	-0.1	-0.1	0.67	0.51	0.73	0.69	0.52	1	0.76	0.08	0.65
JobPro	0.70	0.78	0.78	0.07	0.18	0.36	0.47	-0.12	-0.11	-0.11	0.60	0.53	0.76	0.56	0.49	0.76	1	0.19	0.65
JobFod	-0.34	-0.22	-0.30	-0.36	-0.38	-0.16	-0.17	0.44	0.43	0.43	-0.22	-0.25	-0.17	-0.13	-0.19	-0.18	-0.24	-0.44	-0.31
LU04	0.39	0.14	0.32	0.37	0.28	0.25	0.27	-0.34	-0.33	-0.33	0.10	0.13	0.08	0.04	0.14	0.08	0.19	1	0.11
LU05	0.42	0.63	0.51	-0.03	0.21	0.19	0.31	-0.15	-0.15	-0.15	0.60	0.65	0.57	0.51	0.43	0.65	0.65	0.11	1

Red cells contain absolute value of R>0.4,
End_hr0, End_hr7, and End_hr17 are the number of drop-offs at 12 AM, 7 AM, and 5 PM, respectively.

The Overdispersion Parameter

The overdispersion parameter for negative binomial model represents the quadratic relationship between the mean and variance of the data. The statistical significance (usually at a level of 0.05) for the overdispersion parameter is checked to make sure that negative binomial is appropriate for models at different hours.

Model Selection Criteria

Then, several model selecting criteria are used to select among models including:

1. Akaike Information Criterion (AIC)

The Akaike Information Criterion (AIC) is a measure of the relative quality of statistical models by trading off the complexity and goodness of fit. A smaller AIC value represents a better model. The measure is then used to ensure that the model is not overfitted to the data.

2. Log Likelihood (LL)

Likelihood is a probability that is used to describe a function of a parameter given a set of data points (outcomes). Likelihood function is important in statistical inference, and it is commonly used to estimate a parameter from a set of statistics. Log likelihood is usually a negative value, and a larger LL value indicates a better model.

3. Sum of Model Deviances (Deviance)

The sum of squared deviance residuals, G^2, is a measure of model fit which is used for count regressions. The sum of model deviances is calculated as (Washington et al., 2003):

$$G^2 = 2\sum_{i=1}^{n} y_i \ln\frac{y_i}{\hat{\mu}_i} \qquad \textbf{Eq 5.7}$$

If the model fits the data perfectly, then $G^2 = 0$, because $\hat{\mu}_i = y_i$ for every census tract i. For a count model, such as the Poisson or negative binomial, the observed values are always integers. However, the model produces values of $\hat{\mu}_i$ that are continuous, so it is very unlikely to achieve zero sum of squared deviance residuals. Nevertheless, the value of G^2 provides a useful measure of the error in the model.

Finally, based on those statistics, models have been selected that have relatively better statistics (e.g., smaller AIC, larger LL, and smaller deviance) and produce statistically significant coefficients. However, the residuals of those NB-MLE models are not randomly distributed in space; neighborhoods are likely to have residuals with the same sign and close magnitude. Therefore, a spatial model is used by incorporating spatial random effects into the negative binomial model. The spatial auto-correlation is tested for all dependent variables and independent variables in the next sub-section.

Moran's I test

A common statistic used to measure the spatial autocorrelation is called Moran's I. A number of statistical tools can be used to perform Moran's I test: R, SAS, WinBUGS, and CrimeStat. The Moran's I test has been performed to see whether the spatial correlation (of each individual variable) exists using R "ape" package. To calculate Moran's I in R, a matrix of inverse distance weights is required. The straight line distance calculation is based on spherical coordinates.

Table 5.5 lists the Moran's I test results from R. The Moran's I statistics test the null hypothesis that there is no spatial autocorrelation (Levin et al., 2010). An observed I value that is high indicates more spatial autocorrelation than an observed I value that is low. The expected value of I is not equal to zero but is given by $I_0 = -1/(n-1)$, so it is the same for all independent variables. If the observed value of I is significantly greater than I_0, then values of X are positively autocorrelated, whereas if I < I_0 is observed, this will indicate negative autocorrelation. The two-sided p-value in Table 5.5, if less than 0.05, indicates significance for all the variables listed. Table 5.5 shows that all dependent variables and all theoretical independent variables have statistically significant spatial autocorrelation. The rest of the categorical variables are also found to have spatial autocorrelation using the same method. There are autocorrelated spatial models that can incorporate an additional spatial parameter into the negative binomial count model that performed in the previous section, which reduces the overfitting problems of the count regression model parameters.

Table 5.5 Moran's I Test Results.

R variables	*Short Name*	*Observed I*	*SD*	*p-value*
Independent Variables	**Pop**	0.0589	0.0009	0.0E-04
	EduBac	0.1611	0.0009	0.0E-04
	CapInc	0.1707	0.0009	0.0E-04
	TAT_hr0 (12 A.M.)	0.1793	0.0009	0.0E-04
	TAT_hr7 (7 A.M.)	0.1764	0.0009	0.0E-04
	TAT_hr17 (5 P.M.)	0.1774	0.0009	0.0E-04
	DrpOff_hr0 (12 A.M.)	0.1943	0.0009	0.0E-04
	DrpOff_hr7 (7 A.M.)	0.1263	0.0009	0.0E-04
	DrpOff_hr17 (5 P.M.)	0.1944	0.0009	0.0E-04
	TotJob	0.0711	0.0008	0.0E-04
Dependent Variables	**Pickup_hr0 (12 A.M.)**	0.1768	0.0009	0.0E-04
	Pickup_hr7 (7 A.M.)	0.1675	0.0009	0.0E-04
	Pickup_hr17 (5 P.M.)	0.1900	0.0009	0.0E-04

5.3 Spatial Model

5.3.1 Methodology

We now consider the potential problem of spatial autocorrelation, because many neighborhood areas tend to have similar data on the dependent variable (Taxi demand) and independent variables (demographics, socioeconomics, and Transit Access Time) among adjacent census tracts. The residuals from neighborhood census tracts are also likely to be similar. Therefore, a spatial model, Poisson-Gamma-Conditional Autoregressive (CAR) model, is introduced in this section.

CrimeStat IV is used to model Poisson-Gamma-CAR (Levine et al., 2010). It is able to perform spatial analysis, including: spatial distribution, spatial autocorrelation, distance analysis, 'hot spot" analysis, interpolation, space-time analysis. As the software is designed for crime events, which are rare events, compared to potentially high frequency event, such as taxi pickups, the data structure differs a little. However, the tool can be used to model trip generation by correctly specifying the independent and dependent variables.

MCMC Algorithm

The spatial autocorrelation is programmed in the Poisson-Gamma-Conditional Autoregressive (CAR) model with Markov Chain Monte Carlo (MCMC) algorithm instead of the regular Poisson model and regular negative binomial model, which uses Maximum Likelihood Estimation (MLE) method. The maximum likelihood algorithm works efficiently with functions from the single-parameter exponential family, but MCMC works better with complex (mixed-function) models. However, the MCMC method takes time to estimate the model parameters (Levine et al., 2010). For example, the average time to estimate an NB model with MLE using the default settings in CrimeStat, is less than one minute, while that for a MCMC-CAR model is about 1 hour.

MCMC is a Bayesian algorithm used in estimating the coefficients and parameters in the spatial model. There are five steps of MCMC algorithm (Levine et al., 2010):

1. Functional model specification and model parameters setup;
2. Set up likelihood function and prior distribution of each parameter;
3. Define joint posterior distribution for all unknown parameters, which is simply multiplying the likelihood and the prior as expressed in the following equations:

$$P\ (\theta|X) = \frac{P(X|\theta) \times P(\theta)}{P(X)} \qquad \textbf{Eq 5.8}$$

$$Then, P\ (\theta|X) \propto P(X|\theta) \times P(\theta) \qquad \textbf{Eq 5.9}$$

4. The Markov Chain draws samples from joint posterior distribution repetitively, and parameters are sampled sequentially from the full conditional distribution holding existing parameter constant; and
5. Estimate all coefficients based on the summary statistics results from samples drawn in the previous step. This should also explain why each run of the model may have different outputs.

This Poisson-gamma model is actually a negative binomial model, defined in CrimeStat IV as:

$$y_i|\ \lambda_i \sim Poisson\ (\lambda_i\ ) \qquad \textbf{Eq 5.10}$$

with the mean of Poisson-Gamma-CAR

$$\lambda_i\ = exp({X_i}^T\beta + \varepsilon_i + \phi_i) \qquad \textbf{Eq 5.11}$$

where ε_{i} is the same error from regular Poisson-Gamma model, and ϕ_i is the spatial random effect for each observation i.

CAR model Settings

Based on the theoretical information of CAR model, the models are set up to include the intercept because it is possible that the six major independent variables are not enough to explain all the variability of the dependent variable.

G-R stat is one important criteria of the CAR model that the NB-MLE model does not use. The G-R statistic is a ratio of MC error on MC error standard deviation, which should be below 1.05-1.20. It is necessary to ensure that the G-R stat is below 1.2, i.e., the coefficients for independent variables are believed to be stable and converged. The number of iterations is defaulted to be 25,000, and the burn-in is set to be 5,000, indicating that the first 5,000 iterations are thrown away among all 25,000 iterations. Often, increasing the number of iterations can help achieve the goal for the G-R stat to be smaller than 1.2. The number of burn-in has a slight effect on the computation time, but is not as significant as the influence of iterations. It is recommended to use 150,000 iterations with 30,000 burn-ins based on the convergence of coefficients from some of the initial experiments using different iterations, e.g., 50,000 iterations w/ default 5,000 burn-ins and 100,000 iterations w/ 20,000 burn-ins.

Estimation of Alpha (α)

The neighborhood component, Alpha (α), needs to be specified in order to run the CAR model. Alpha (α) is the exponent for the distance decay function in the spatial model, and it is a component in the computation of the spatial parameter, Phi (ø). It is necessary to determine a distance decay function for Alpha (α) so that a weight can be applied to the values of nearby records (Levine et al., 2010). The weight is defined as follows:

$$Weight = exp(-\alpha * dd_{ij}) \quad \textbf{Eq 5.12}$$

$$Alpha = \ln(Weight)\ /NN(distance)) \quad \textbf{Eq 5.13}$$

where:

α= the absolute value of Alpha (α);

dd_{ij}= the distance between observations in specified units, e.g. miles, meters; and

$NN(distance)$= the nearest neighbor distance in specified distance units, (e.g. miles, meters), and the average NN(distance) for this data set is 0.249 miles.

It is also necessary to examine the Moran Correlogram to see whether the distance decays in the dependent variable (the hourly number of pickups) is very 'sharp'. Eq 5.12 is used to fit the Moran'I points (blue dots) by plotting a red curve in Figure 5.4. The equation of the fitted curve is $Moran's\ I = 0.185 + exp(-0.335 * dd_{ij})$. Note that a constant, 0.185, is added to Eq 5.13 to give the curve a vertical shift. However, this does not change the estimation of Alpha (α) to be -0.335, which is close to the suggested value for shallow decay.

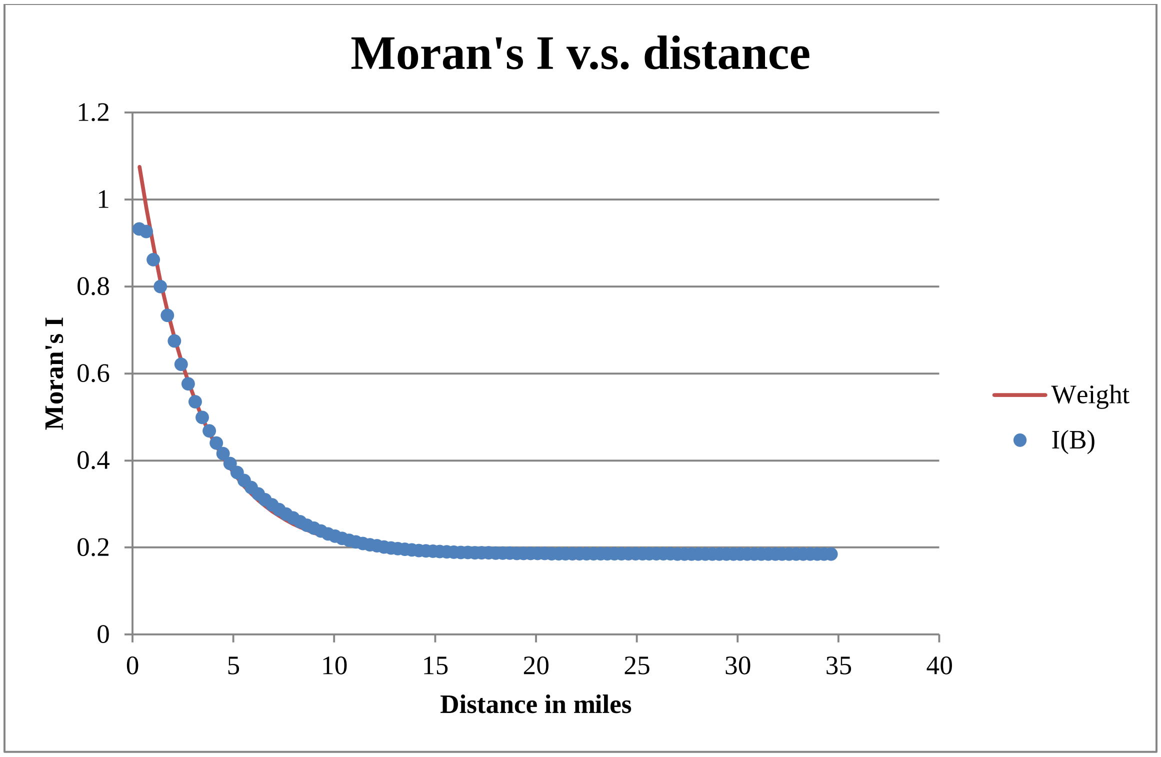


Figure 5.4 Moran's I v.s. Distance with Fitted Curve to Determine Alpha (α).

Experiments of Model Trials

A total of more than 100 CAR models are estimated for 12 A.M (hour 0), 7 A.M. (hour 7), and 5 P.M. (hour 17) using combinations of six theoretical major factors and three factor categories (Table 5.6). Similar to the variables selection in NB-MLE models, these major theoretical variables are included in almost all models. DrpOff, Pop, and TAT are selected for all models. In all the 20 models listed in Table 5.6, between CapInc and EduBac, only one can be included in each model. TotJob is included to indicate the level of social and economic activities, and it cannot be included with any of the job categories in the same model. For the other variables, the same rule as the NB-MLE model is used to prevent multi-collinearity, to make sure that no pairs of variables are highly correlated with each other. For example, land use categories cannot be included in the same model with employment categories. Based on this rule, 20 combinations (i.e., 20 models per hour) for any hour that can be tested are listed in Table 5.6.

Next, the same likelihood statistics from the NB-MLE model selection used are applied to the CAR model to select variables. Also, the model coefficients are examined to see whether they are statistically significant and their implications are also checked. If the sign and magnitude of the coefficients of all variables make sense, then statistics, including AIC, LL, deviance, overdispersion parameter, are used to select a better model, i.e., a combination of variables that make the model behave relatively better than the rest of the models.

Using 7 A.M. as an example, 20 models are examined for the 7 A.M. slot as shown in Table 5.6. Their corresponding statistics (Table 5.7) and coefficients (Table 5.8) are calculated for each model. The model experiments are composed of five steps based on those statistics at this hour. The first two models (the first step) only include theoretically important variables, and the model with EduBac has a smaller AIC compared to the model that contains CapInc. EduBac also has a lower correlation with other major variables or variable categories compared to CapInc; therefore, EduBac is the preferred variable in the following steps.

Table 5.6 The Experiments of Model Trials Template.

Steps	Model	Major Variables						Age Groups		Land Use Groups		Employment Groups					
		DrpOff	Pop	TAT	EduBac	CapInc	TotJob	PerAge 0-14	PerAge 15-34	LU 04	LU 05	Job Ret	Job Inf	Job Fin	Job Rea	Job Pro	Job Fod
Step1: only include major variables	1	√	√	√	√		√										
	2	√	√	√		√	√										
Step 2: choose the best age category	3	√	√	√	√		√	√									
	4	√	√	√	√		√		√								
	5	√	√	√	√		√	√	√								
Step 3: best age category + land use groups	6	√	√	√	√				√	√							
	7	√	√	√	√				√		√						
	8	√	√	√	√				√	√	√						
Step 4: best age category + employment groups	9	√	√	√	√				√			√					
	10	√	√	√	√				√				√				
	11	√	√	√	√				√					√			
	12	√	√	√	√				√						√		
	13	√	√	√	√				√							√	
	14	√	√	√	√				√								√
Step 5: Remove the age category	15	√	√	√	√						√						
	16	√	√	√	√					√	√						
	17	√	√	√	√		√			√							
	18	√	√	√	√												√
	19	√	√	√		√				√	√						
	20	√	√	√		√	√			√							

‘JobRet’ Number of jobs in Retail Trade; ‘JobInf' Number of Jobs in Information; ‘JobFin’ Number of jobs in Finance and Insurance; ‘JobRea’ Number of jobs in Real Estate and Rental and Enterprises; ‘JobPro’ Number of jobs in Professional, Scientific, and Enterprises; ‘JobFod’ Number of jobs in Accommodation and Food Services; ‘LU01’ One & Two family Buildings; ‘LU04’ Mixed Residential & Commercial Buildings; ‘LU05’ Commercial & Office Buildings.

The second step examines age categories. PerAge15-34 is preferred based on the statistics, and is similarly used in the following step. Two land use groups (LU04, LU05) and three job categories (JobRet, JobFin, and JobFod) have significant coefficients and G-R stat below 1.2. This makes sense: the best land use group is LU05 (commercial and office building), while the best employment group is JobFod (number of jobs in food and accommodation) based on the other model statistics in the third and fourth step. Some variables become insignificant (p-value<0.05) when they are included with PerAge15-34 in the same model. Therefore, this age category is removed while the best land use category or job category is tested in the fifth step. Also, CapInc is tested interchangeably with EduBac in some of those well-behaved models, such as model 17 and model 20 in Table 5.6.

Following the guidelines, one model works relatively better in the experiments at each steps, which is labeled '*' in Table 5.7. All those labeled models exhibit statistically significant coefficients for all of their independent variables with the satisfied G-R stat (G-R stat below 1.2). Eliminated models include those with coefficients that do not converge to equilibrium; insignificant coefficients; and relatively smaller AIC values. Finally, model 14 in Table 5.7 is selected as the best model among all 20 models at hour 7. The same methods apply to the model selection for the other two hours: hour 0 and hour 17. The best CAR model for each hour is presented in the next sub-section.

Table 5.7 The Model Statistics of Model Trials at 7 A.M.

Steps	Hour 7 Model	Model Statistics					
		All G-R stat<1.2?	All coef. sig. at a level of 0.05?	AIC	LL	Deviance	P-value of Overdispersion Parameter
Step 1	1*	yes	yes	469,099	-234,542	459,598	0.001
	2	yes	yes	738,030	-369,007	728,299	0.001
Step 2	3	no	no	2,165,422	-1,082,702	2,156,385	0.001
	4*	no	no	3,558,549	-1,779,265	3,549,031	0.001
	5	no	no	1,625,446	-812,713	1,615,781	0.001
Step 3	6	yes	yes	558,627	-279,305	549,322	0.001
	7*	yes	yes	417,462	-208,722	407,874	0.001
	8	yes	yes	438,067	-219,024	428,589	0.001
Step 4	9	no	yes	540,265	-270,124	530,909	0.001
	10	yes	no	523,945	-261,963	514,540	0.001
	11	no	yes	580,010	-289,996	570,694	0.001
	12	no	no	600,660	-300,321	591,367	0.001
	13	yes	no	468,735	-234,359	459,111	0.001
	14*	yes	yes	312,667	-156,325	302,741	0.001
Step 5	15	yes	yes	515,718	-257,851	506,356	0.001
	16	yes	yes	345,086	-172,534	335,333	0.001
	17	yes	yes	687,434	-343,708	678,415	0.001
	18*	yes	yes	335,820	-167,902	325,911	0.001
	19	yes	yes	1,953,887	-976,935	1,945,229	0.001
	20	yes	yes	1,459,012	-729,497	1,450,096	0.001

‘*’ represents the best model selected for hour 7 at each step (except the second step) based on statistics such as p-value of coefficients, G-R stat, AIC, LL, and overdispersion parameter. In the second step, since PerAge0-14 is highly correlated with education and income (R<-0.5), it cannot be included with EduBac and CapInc in the same model, therefore model 3 and 5 are not selected.

Table 5.8 The Model Coefficients of Model Trials at 7 A.M.

	Hour 7	Major Variables						Age Groups		Land Use Groups		Employment Groups					
Steps	Model	DrpOff	Pop	TAT	EduBac	CapInc	TotJob	PerAge0-14	PerAge15-34	LU04	LU05	JobRet	JobInf	JobFin	JobRea	JobPro	JobFod
Step 1	1*	5.0E-5	3.9E-4	-0.116	0.091		4.0E-5										
	2	4.0E-5	3.7E-4	-0.125		7.0E-5	4.0E-5										
Step 2	3	8.0E-5	5.8E-4	0.000#	0.000^		5.0E-5	-0.257									
	4*	1.1E-6	5.7E-4	0.000^	0.000		4.0E-5		0.164								
	5	8.0E-5	5.5E-4	0.000^	0.000^		3.0E-5	-0.229	0.125								
Step 3	6	6.0E-5	3.0E-4	-0.084	0.082				0.087	0.054							
	7*	4.0E-5	4.0E-4	-0.093	0.088				0.090		0.041						
	8	4.0E-5	3.2E-4	-0.082	0.082				0.080	0.054	0.041						
Step 4	9	5.0E-5	3.9E-4	-0.094	0.086				0.093			9.1E-4					
	10	6.0E-5	3.9E-4	-0.097	0.087				0.098				9.0E-5^				
	11	8.0E-5	3.8E-4	-0.096	0.086				0.099					-2.0E-4			
	12	6.0E-5	3.9E-4	-0.097	0.087				0.095						2.0E-4^		
	13	7.0E-5	3.8E-4	-0.098	0.088				0.095							-1E-4^	
	14*	4.0E-5	3.8E-4	-0.097	0.086				0.094								9.5E-4
Step 5	15	4.0E-5	4.1E-4	-0.114	0.092						0.051						
	16	4.0E-5	3.1E-4	-0.095	0.086					0.063	0.050						
	17*	5.0E-5	2.8E-4	-0.100	0.083		4.0E-5			0.065							
	18	4.0E-5	3.8E-4	-0.115	0.090												1.1E-3
	19	3.0E-5	2.8E-4	-0.102		6.0E-5				0.068	0.038						
	20	4.0E-5	2.6E-4	-0.104		6.0E-5	5.0E-5			0.071							

Cells without labels represent that the coefficient is statistically significant at a level of 0.05 (p-value<0.05).
Cells labelled “#” represent that the coefficient is statistically significant between the levels of 0.05 and 0.1 (0.05<p-value<0.1).
Cells labelled “^”represent that the coefficient is not statistically significant at a level of 0.1 (p-value>0.1).

5.3.2 Results

CAR Model Output

Based on the variable selection procedures in the previous section, a separate model has been estimated for the best variable combination of each of the three representative hours of the day: 12 A.M. (midnight), 7 A.M. (morning peak), and 5 P.M. (afternoon peak). All three models have statistically significant coefficients and all coefficients converge (G-R stat smaller than 1.2). The model with the smallest AIC value is selected for each hour. Model 17 (DrpOff, Pop, TAT, CapInc, TotJob and LU04) is chosen for hour 0; model 14 (DrpOff, Pop, TAT, EduBac, JobFod) is chosen for hour 7; and model 17 (DrpOff, Pop, TAT, EduBac, TotJob and LU04) is chosen for hour 17. Table 5.9 displays the coefficients of the independent variables for CAR models at all three hours.

The overdispersion parameter theta, θ (Eq 5.6), is significant at each hour for all CAR models, which suggests that the model is accounting for the overdispersion. All coefficients are significant for the CAR models at all three hours of the day. At all hours, the coefficients of the same variable for both models have the same sign and are close in magnitude.

The spatial parameter (Phi, ø) is a function of three parameters including global component (Rho), local component (Tauphi, τ_{ϕ}), and neighborhood component (Alpha, α). They are all significant in all three hour models, indicating that CAR model is appropriate in accounting for the spatial autocorrelation.

Table 5.9 CAR Model Output for 12 A.M., 7 A.M., and 5 P.M.

CAR Model	Hour	12 AM	7 AM	5 PM
Overdispersion Parameter	**Theta (θ)**	0.9413	0.1393	0.0632
Model Statistics	**AIC**	38,996	312,667	443,148
	LL	-21,849	-156,325	-221,565
	Deviance	30,902	302,741	433,940
Coefficients (*95% Credible Interval*)	**DrpOff**	1.61E-04	3.70E-05	8.20E-05
	(95% CI)	*(1.38E-04,1.86E-04)*	*(2.20E-05,5.3E-05)*	*(7.00E-05,9.50E-05)*
	Pop	1.23E-04	3.82E-04	1.73E-04
	(95% CI)	*(0.62E-04,1.86E-04)*	*(3.14E-04,4.41E-04)*	*(1.09E-04,2.39E-04)*
	TAT	-1.05E-01	-9.67E-02	-9.08E-02
	(95% CI)	*(-1.18E-01,-0.92E-01)*	*(-11.04E-02,-8.31E-02)*	*(-10.33E-02,-7.84E-02)*
	EduBac	—	8.58E-02	7.43E-02
	(95% CI)		*(7.92E-02,9.22E-02)*	*(6.72E-02,8.16E-02)*
	CapInc	4.10E-05	—	—
	(95% CI)	*(3.50E-05,4.70E-05)*		
	TotJob	3.50E-05	—	4.50E-05
	(95% CI)	*(1.5E-05,5.8E-05)*		*(2.8E-05,6.3E-05)*
	PerAge15-34	—	9.37E-02	—
	(95% CI)		*(7.42E-02,11.01E-02)*	
	LU04	5.74E-02	—	5.81E-02
	(95% CI)	*(4.36E-02,7.14E-02)*		*(4.54E-02,7.12E-02)*
	JobFod	—	9.48E-04	—
	(95% CI)		*(5.00E-04,13.84E-04)*	
Spatial Parameters	**Phi (ø)**	-0.2431	-0.2527	-0.2793
	Rho (ρ)	0.1003	0.2394	0.2418
	Tauphi (τ_{ϕ})	0.0004	0.0005	0.0005
	Alpha (α)	-0.3350	-0.3350	-0.3350

All coefficients and parameters are significant at the level of 0.05 in the above table.
'—' indicates variables that are not included in the model.

The 95% credible interval (CI) of a coefficient means that there is a 95% probability that the mean coefficient falls into this interval. Although the statistics from previous model trials suggest that one model behaves slightly better than other models, the mean estimated coefficient is generally close in magnitude for different CAR models at the same hour, i.e., the coefficient of one variable in one model might fall in the 95% CI of the same coefficient in another model. For example, the coefficient of population

(Pop) in model 2 (Table 5.8) at 7 A.M. has a mean value of 0.00037, which falls in the 95% CI (0.000336, 0.000453) of the coefficient of model 1 at 7 A.M. (the coefficient of Pop in model 1 is 0.00039). However, the CAR coefficients for one hour does not necessarily fall into the 95% credible interval of the same coefficients for another hour, indicating that the influence of population changes at different hours. For instance, as shown in Table 5.9, the coefficient of Pop at 12 A.M. does not fall in the 95% CI for any of the other two hours. Similar patterns are observed for other variables as well.

Implication of Model Coefficients

At almost all hours, all the coefficients of CAR models are statistically significant at a level of 0.05 (Table 5.8). The signs of model coefficients are consistent in different CAR models. They are also consistent with the predicted signs discussed in the beginning of this chapter. The relationships (arithmetical signs and statistical significance) between independent variables and taxi demand are as hypothesized in the beginning of this chapter. The statistically significant variables, drop-offs and population, are positively related to taxi demand (Table 5.8 and Table 5.9); areas with more available empty cabs and population tend to have higher levels of taxi demand. As shown in Table 5.8 and Table 5.9, the coefficient of transit access time (TAT) is negative at all three hours. In other words, places with better transit accessibility have more taxi pickups, which is consistent with the expected relationships. Socioeconomic factors like income, education, and total jobs are positively related to the number of taxi pickups as expected, since, in theory, income and education represent the affordability of the population that live in the areas. TotJob is an indicator of the level of economic activities, and economic activities drive taxi demand.

Some categorical factors, especially those selected in the best CAR model at each hour, PerAge14-35, LU04 (Mixed residential and commercial buildings), and JobFod (Jobs in Food and Accommodation), have the stronger influence on taxi demand compared to other categorical factors in their respective groups. They are all positively associated with taxi demand. This implies that areas with higher percentages of ages between 14 and 35, higher percentages of mixed and commercial land use, and more jobs in food and accommodation are generating a statistically significant number of taxi pickups at the level of 0.05. This is consistent with the expected results.

Some variations are observed among the coefficients of the same variable from models at different hours. This suggests that intuitively, the implication of the coefficients for the same variable may not be slightly different among different models. For example, the number of drop-offs is more influential at 12 A.M. than that at 7 A.M. and 5 P.M., which makes sense because the number of empty taxis circulating on the street at midnight is much smaller than at day time; people having night activities may go to those places with higher level of empty cabs to hail a cab to get back home. The results also indicate that the midnight taxi demand is more associated with good transit accessibility (shorter TAT) compared to that at day time. It is possible that some of the midnight trips are being made to or from transit facilities, enabling taxi services to complement transit much more at midnight compared to day time. It is also possible that places with good transit services are more desirable for taxi use for other reasons. For example, it might be easier to hail a cab on streets near subways in NYC under which the busiest subway lines run at midnight.

The Pop, EduBac, TotJob, and LU04 have a higher influence on taxi pickups in the day time compared to midnight, which indicates that extra taxi demand during morning peak and afternoon peak in NYC is likely caused by people going to and from work or work-related activities. Since all model coefficients are theoretically consistent as expected and the implications of temporal variation of the coefficients make sense, the model forecasting must be tested in the next step.

Mapping the Predicted Taxi Demand

Once the best model is selected at each hour, the mean predicted taxi demand, the upper bound and lower bound of the 95% credible interval of the predicted taxi demand, are discussed to understand CAR model forecasting. The mean predicted taxi demand is calculated using the mean coefficients of all variables (including intercept) following Eq 5.2. The lower bound and upper bound are calculated using the 95% credible interval of the coefficients for each variable based on the same equation. Essentially, the CAR model is a type of NB model, so the linear relationship exists between the logarithm of the dependent variable and the independent variables.

The actual and predicted taxi demand are mapped along with the upper bound and lower bound of the 95% credible interval of the predicted taxi demand using the best model selected at each hour (12 A.M, Figure 5.5 - Figure 5.6; 7 A.M. Figure 5.7 - Figure 5.8; 5 P.M., Figure 5.9 - Figure 5.10). The same color scale has been used for all four maps at each figure so they can be compared fairly. The number of census tracts that are within 95% credible interval (CI) for each model is also summarized in Table 5.10.

The overall predictability for taxi demand at hour 7 is better compared to the other two hours. For example, the percent of census tracts in NYC that are within the 95% CI is 47%, 75%, and 56% for hour 0, hour 7 and hour 17, respectively (Table 5.10). Table 5.10 also indicates that the model predictability for Manhattan is weaker compared to NYC (5 Boroughs). This is probably because there is a relatively higher percentage of census tracts with over-estimated (extreme predicted values) taxi demand in Manhattan compared to overall NYC.

Table 5.10 The Number of Census Tracts of Actual Demand within 95% CI of CAR Model.

			Actual Demand within 95% CI		
		Total	12 A.M. Model	7 A.M. Model	5 P.M. Model
NYC (5 Boroughs)	No of Census Tracts	2124	1003	1601	1187
	%	100%	47.22%	75.38%	55.89%
Manhattan	No of Census Tracts	285	154	160	147
	%	100%	54.04%	56.14%	51.58%

The overall predictability of the CAR model is not very good, i.e. less than 60% of the census tracts are within the 95% CI of the predicted taxi demand. The CAR model exhibits a better ability to predict the taxi demand at most of the Outer Boroughs, while having a weaker ability to predict taxicab demand in Central Manhattan as shown in all figures (Figure 5.5 - Figure 5.10). Taxi demand is under-estimated in the neighborhood areas near Manhattan, such as Williamsburg and Astoria, i.e., the predicted demand is much less compared to the raw taxi demand. However, the raw taxi trip demand at each hour is between the lower and upper bound of the 95% predicted credible interval. Airports are eliminated from the model because of the lack of employment and population. That is why they are not shown in the predicted taxi demand or the lower and upper bound for taxi demand.

Manhattan is zoomed in so that the detailed variation of the 95% credible interval (CI) of the predicted taxi demand can be visualized and analyzed (Figure 5.6, Figure 5.8, and Figure 5.10). At 12 A.M., transit hubs such as Penn Station are showing the same level (color) of raw taxi demand as the predicted taxi demand (Figure 5.6), and it is within the 95% CI of the predicted taxi demand. Lower demand areas also show similar level of taxi demand between the raw taxi demand and the 95% CI of predicted taxi demand, such as the Lower East Side of Manhattan and Harlem. In the same hour, while most of the central Manhattan areas are over-estimated, some neighborhoods have the number of predicted pickups smaller than the upper 95% CI, for example, Lincoln Square.

Similar patterns for Harlem, the Lower East Side of Manhattan, and Lincoln Square are observed for CAR model at 7 A.M. (Figure 5.8), except that the lower bound of the 95% CI exhibits a smaller level of taxi demand compared to raw taxi demand at 7 A.M. Also, the upper bound of the predicted 95% CI has higher variability than that of hour 0 model. This is probably why there is a greater number of census tracts that fall in the 95% CI for hour 7 than that of the rest two hours (Hour 0 and hour 17). Hour 17 model prediction exhibits similar patterns as hour 0 model prediction (Figure 5.10).

The CAR model predictions are not very good. However, the maps (Figure 5.5 - Figure 5.10) indicate that the predicted taxi demand has values that are close in magnitude compared to the actual taxi demand in most census tracts. This means that if a smaller CI is selected, for example 90%, the percentage of census tracts within the CI will increase compared to that of the 95% CI. We don’t know the reason for the low predictability of the models. Maybe the scale of spatial resolution (census tract level) is too small, or maybe a different level of aggregation is needed for different neighborhoods. Next, the sign of residuals (the difference between the actual taxi demand and the predicted taxi demand) is studied instead of the magnitude of the residuals.

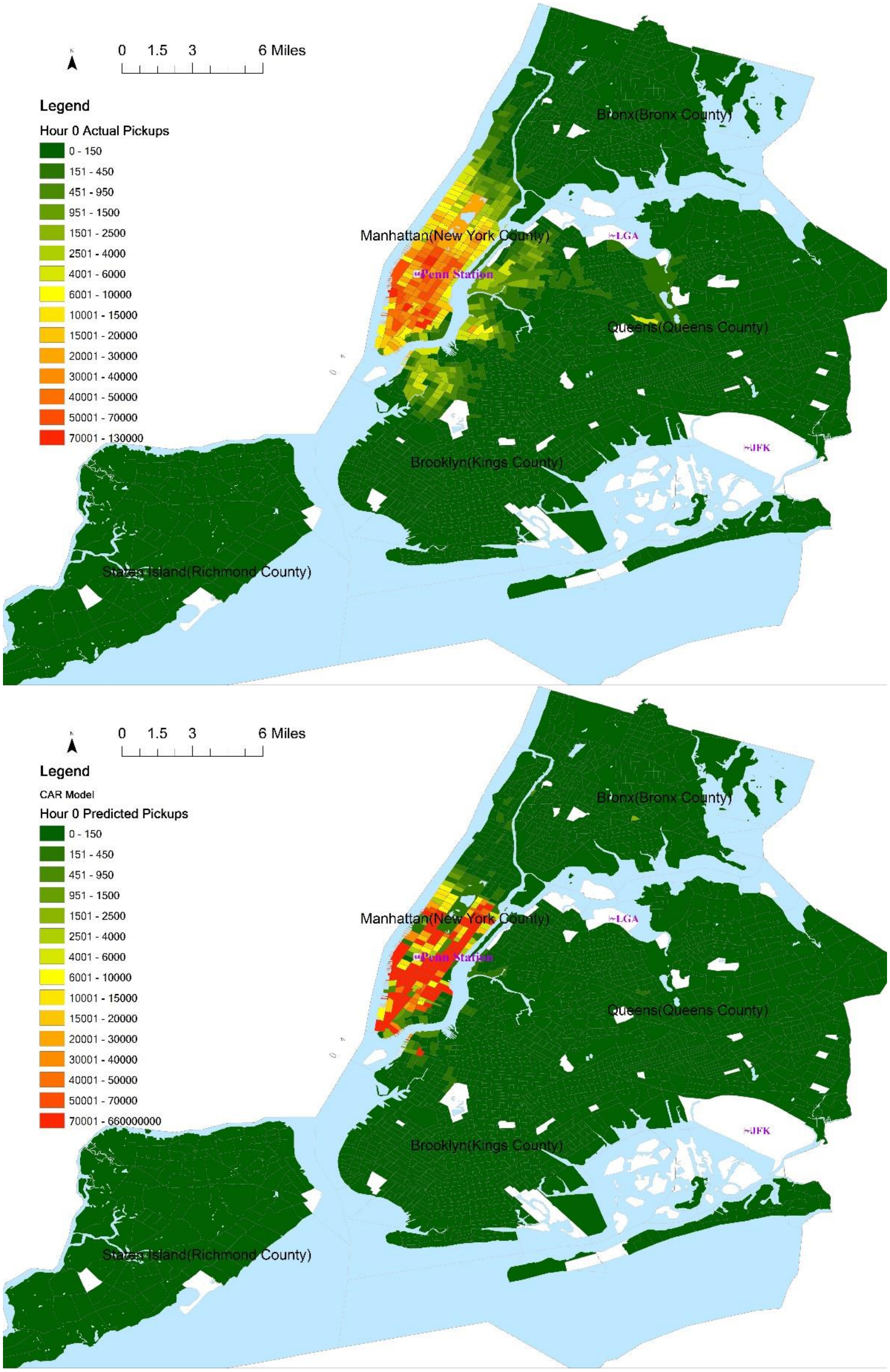


Figure 5.5 The Actual and the Predicted Taxi Demand at 12 A.M.

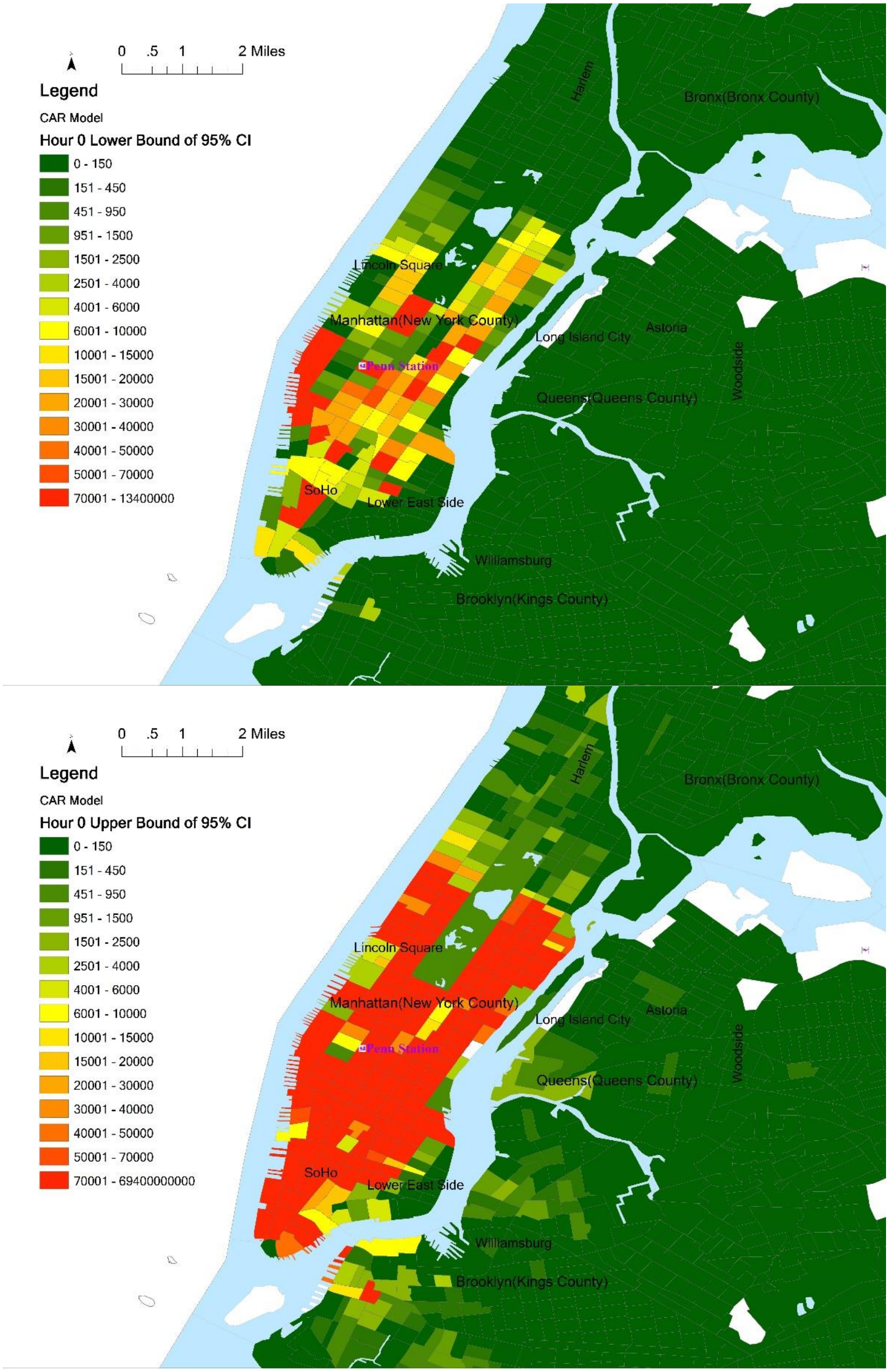


Figure 5.6 The Lower and Upper Bound for the 95% CI of the Predicted Taxi Demand at 12 A.M.

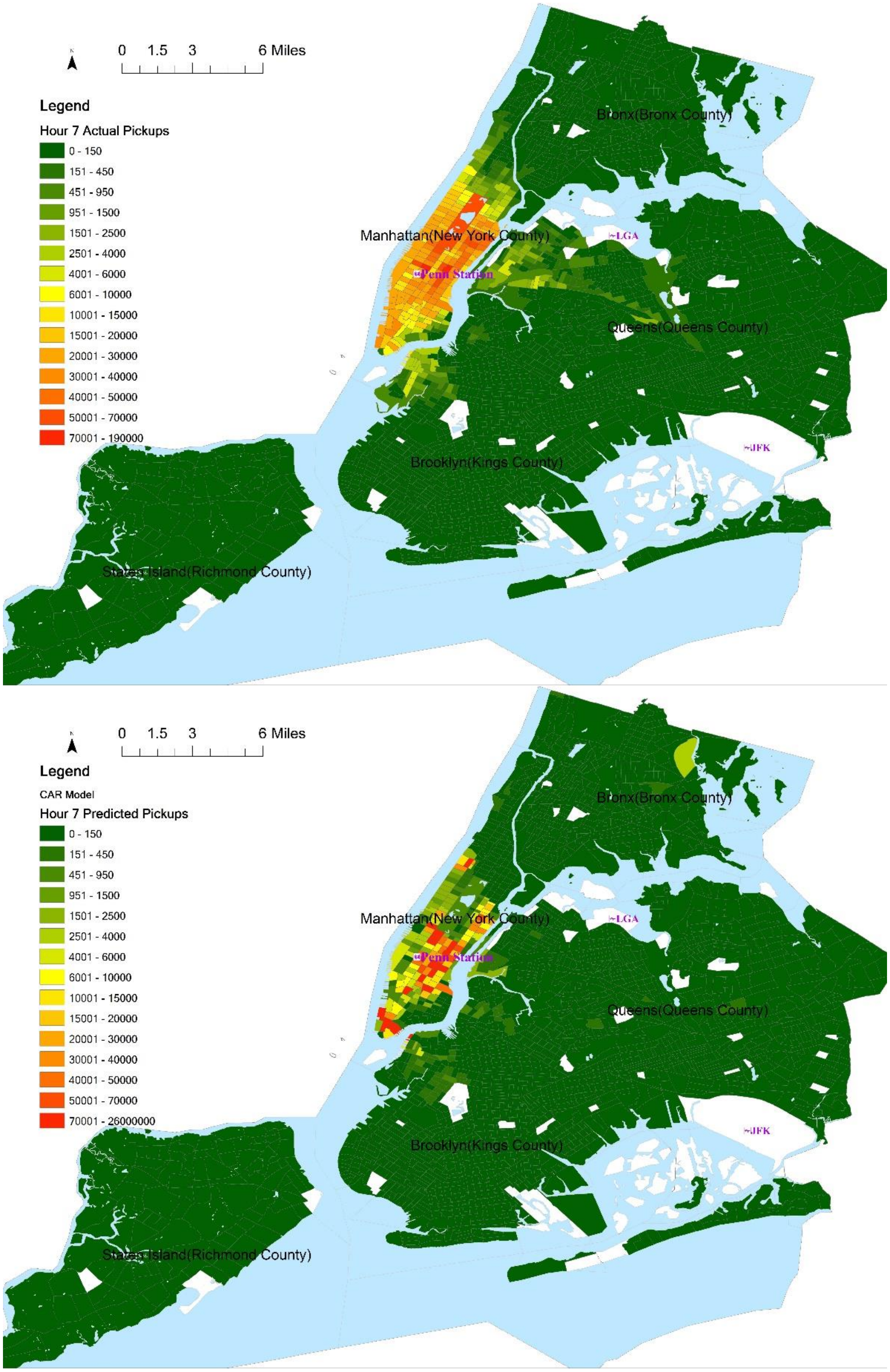


Figure 5.7 The Actual and the Predicted Taxi Demand at 7 A.M.

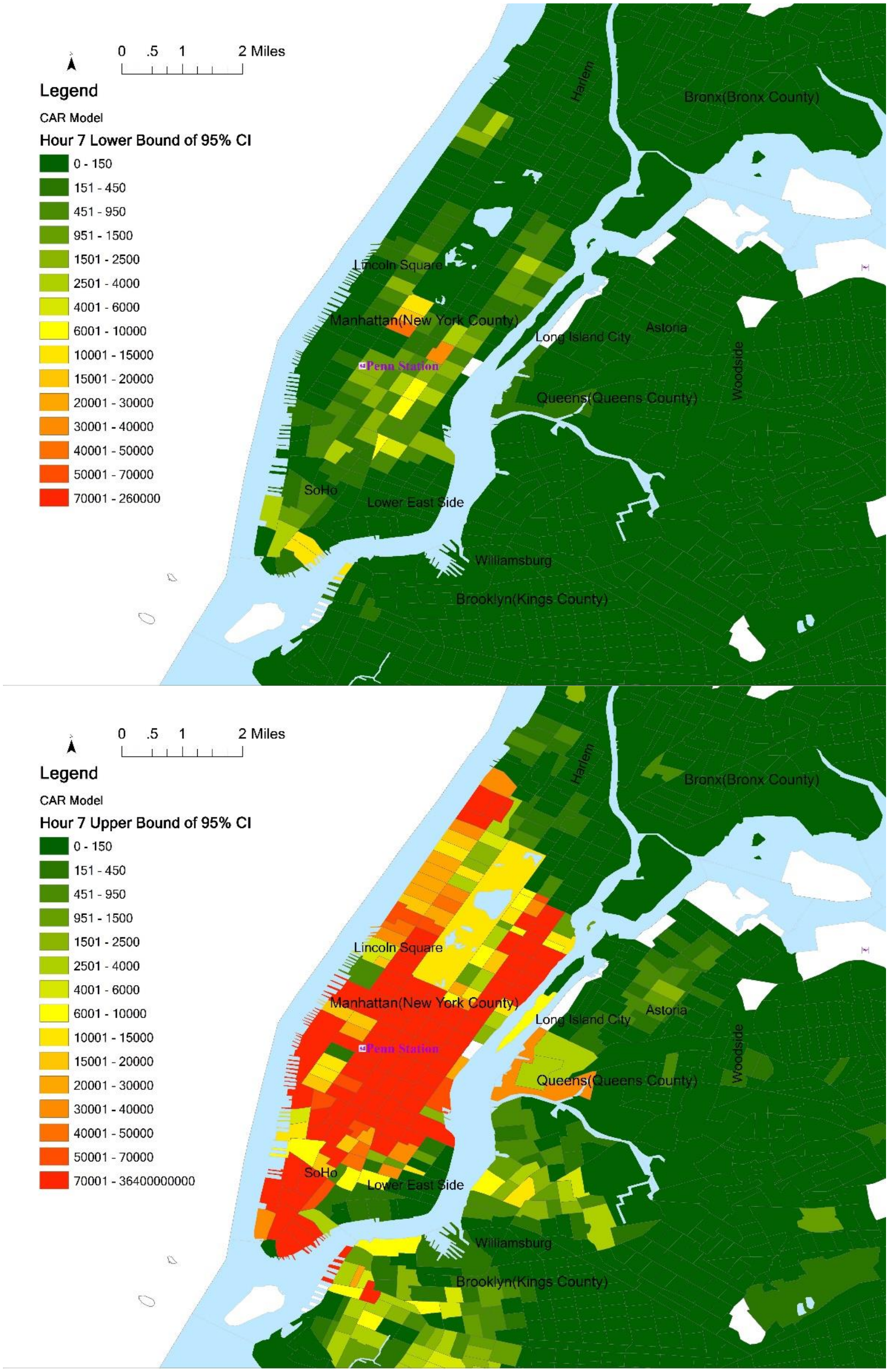


Figure 5.8 The Lower and Upper Bound for the 95% CI of the Predicted Taxi Demand at 7 A.M.

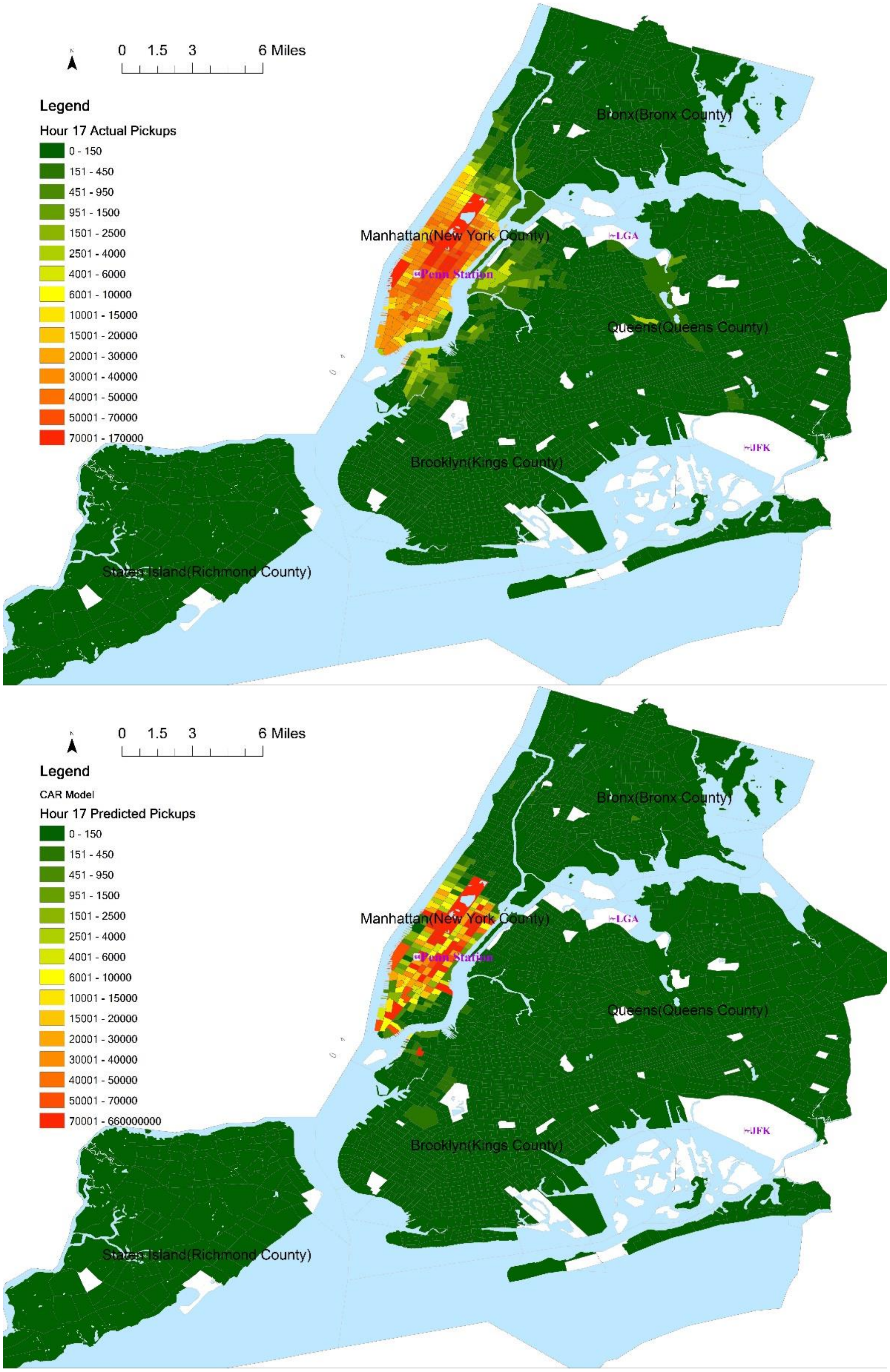


Figure 5.9 The Actual and the Predicted Taxi Demand at 5 P.M.

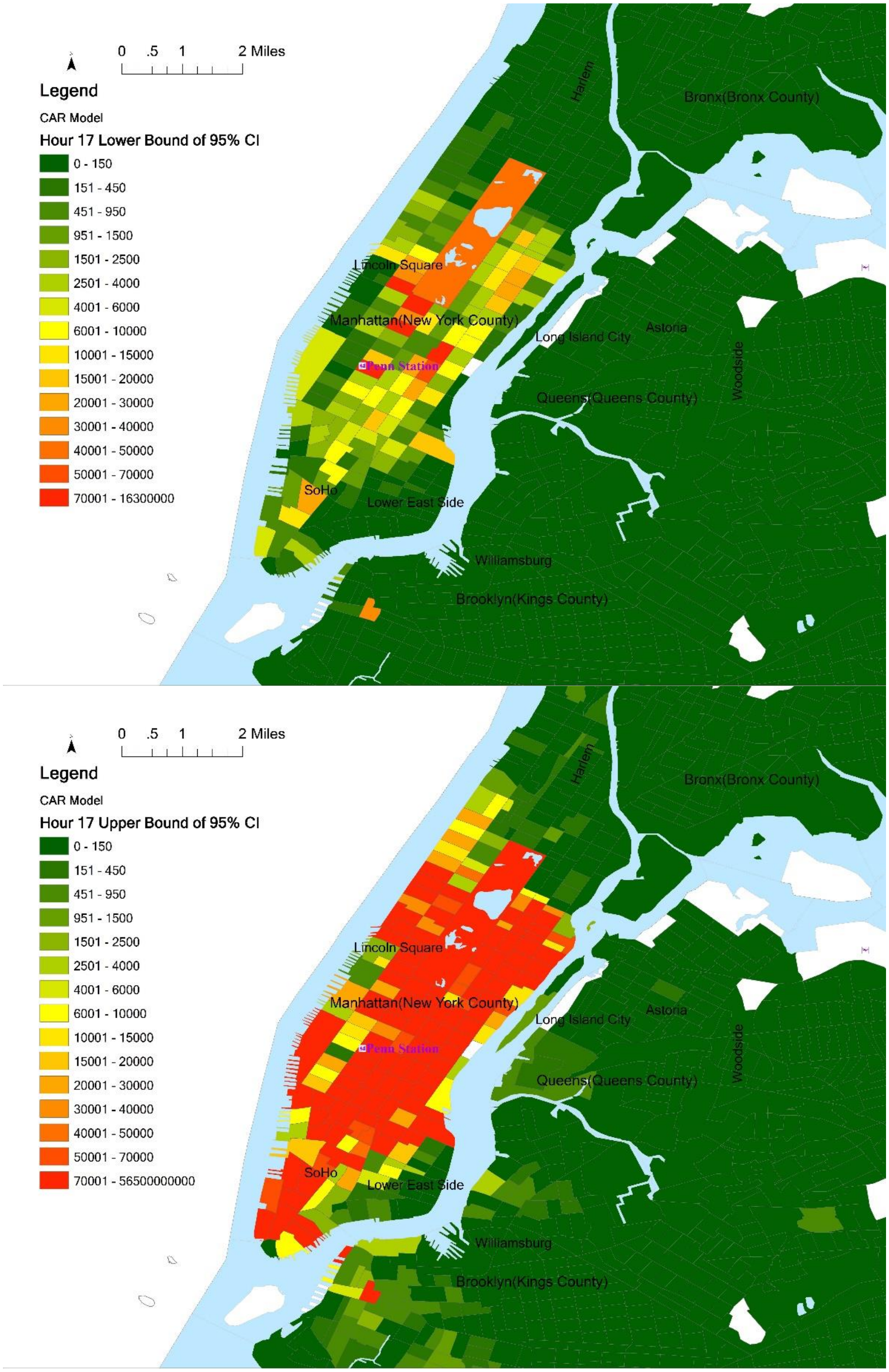


Figure 5.10 The Lower and Upper Bound for the 95% CI of the Predicted Taxi Demand at 5 P.M.

5.4 Interpretation of Model Residuals and Imbalance

In the beginning of this dissertation, three questions have been brought out concerning the demand and supply of NYC taxis, which have been answered step by step during each chapter. First, it is confirmed that there is a variability and imbalance between taxi demand and supply across different times of day and at different locations. Secondly, the six major factors related to the distribution of taxi demand and supply have been identified. Lastly, the imbalance between taxi demand and supply has been identified and quantified using trip generation modeling using CAR models. However, there is no literature related to the identification and quantification of the imbalance. Therefore, it is very important to understand whether the results of the CAR models make sense and whether they are insightful for policy makers in taxi planning and management.

5.4.1 Model Validation

One way to validate the models is to use cross-validation methods, which is conventionally used in validating a model. This requires first separating the daily taxi data randomly into two parts: a training dataset and a testing dataset. Then, a model is built using the training data and the coefficients from the model are tested with the test dataset to see if the model is correct. However, this analysis is more focused on identifying the important factors that affect taxi demand and whether the influence is positive or negative. Acknowledge that the model is not a perfect fit, we are more interested in the relative differences between actual demand and modeled demand and how regions compare with each other. Therefore, the cross-validation approach is not implemented in this dissertation.

A second way to confirm whether the model implies correct information is to see whether the model results qualitatively match known issues related to the taxi demand and supply in NYC. For example, the Green Taxi licenses have been introduced to serve a recognized need for additional taxi service to be provided outside of the most densely served parts of Manhattan and the airports. Thus the second approach is used in the following interpretation and validation process.

5.4.2 Comparison between Model Residuals and Imbalance

The imbalance and residuals from the CAR model must be compared to understand the model results. The imbalance is the difference between the observed taxi demand (pickups) and supply (drop-offs). A positive imbalance means that there are more pickups than drop-offs, so empty taxis are actually entering the neighborhood. A negative imbalance means that there are more drop-offs than pickups, so empty taxis are actually leaving the neighborhood. The imbalance has been calculated and mapped as shown in Chapter 1.

The residual is the difference between the observed number of pickups and the modeled number of pickups, based on demographic characteristics and the supply of empty taxis as measured by the number of drop offs. A positive residual means that controlling for demographic characteristics, there is more observed demand than the number of drop-offs would suggest. This means, that in order to provide a level of empty taxi availability relative to the demand that is comparable with the city-wide average, empty taxis should be entering the neighborhood. A negative residual means that the model overestimates demand, so there are more empty taxis in the neighborhood relative to the demand than the city-wide average.

The magnitudes of the imbalance and residuals are not as important as their signs in order to understand patterns of supply and demand in NYC. The four possible combinations of relationships between the residuals and imbalance are summarized based on their signs in the following Table 5.11.

Table 5.11 The interpretation of Residuals and Imbalance between Demand and Supply.

			Residuals		
			+ Not enough drop-offs	- Too many drop-offs	
Imbalance Between Demand and Supply	+	More pickups than drop-offs	Balanced High Demand (BHD) Area	Overserved Area	Taxi Actually Entering
	-	More drop-offs than pickups	Underserved Area	Balanced Low Demand (BLD) Area	Taxi Actually Leaving
			Taxi Should Be Entering	Taxi Should Be Leaving	

All neighborhoods are classified into four categories as shown in Table 5.11 and Figure 5.11-Figure 5.13. Those four categories include: 1) Balanced High Demand (BHD); 2) Overserved; 3) Underserved; 4) Balanced Low Demand (BLD).

As concluded from the model selection procedures, the interpretation will be based on the best CAR model for 12 A.M., 7 A.M., and 5 P.M. Also, community district data is used to gain information of the communities or the neighborhoods that have similar pattern, and those interested community districts are labeled on the maps. Since imbalance and model output are aggregated by hour, the comparison will also be based on each hour.

If the imbalance is positive in an area at a certain hour, there are more pickup than drop-off taxi trips in that neighborhood in this hour, which means that during the 10-

month overall, taxis are always entering the area to get the extra pickups at this hour; i.e., it may indicate that after dropping off passengers in other neighborhoods, taxis tend to come to this neighborhood to get the next customer (Table 5.11). Vice versa, if the imbalance is negative, the taxi drivers tend to exit the area for their next pickup (Table 5.11). The imbalance implies the actual status of demand and supply in NYC.

The residual represents the difference between the actual number of pickups and the modeled pickups. The coefficients for the variable “supply” are always positive, if one assumes that the model is correct, then all coefficients for all variables are correctly calculating the generated taxi pickups. If we assume that the coefficients for demographic, TAT, and socioeconomic variables are correct (i.e. controlling for those factors), and knowing that the coefficient on drop-offs are always positive, one can adjust to the number of drop-offs that make the modeled demand closer to actual demand leading residuals to be zero. Therefore, if the residual is “+”, we can adjust the modeled pickups by increasing the number of drop-offs. The positive residual implies that there are not enough taxi drop-offs, and the taxis should be entering the area to satisfy the extra demand that is not captured by the model. On the contrary, if the residual is “-”, meaning that there are too many taxi drop-offs in the area, then, the taxis should be exiting the area to reduce the over-explained demand by the model.

Now that we understand the meaning of positive and negative signs for CAR residuals and imbalance, the following discussion provides a summary of neighborhoods that represent each of the four categories (Table 5.12) and detailed interpretation of the four categories. Information in Table 5.12 is based on the mapped four categories from Figure 5.11-Figure 5.13.

Table 5.12 Neighborhoods Representing Each of the Four Categories.

Categories	Details	Representing Neighborhoods
Balanced High Demand	Residual (+) Imbalance (+)	Relatively small number of areas in Manhattan and the Outer Boroughs
Overserved	Residual (+) Imbalance (-)	Midtown Manhattan
Underserved	Residual (-) Imbalance (+)	Lower East Side of Manhattan, Harlem, Woodside, Sunnyside, and Williamsburg
Balanced Low Demand	Residual (-) Imbalance (-)	Most residential neighborhood in the Outer Boroughs at all three hours

If both residuals and imbalance in an area are positive or negative at the same time, it means that the model results agree with the observations, suggesting that the models correctly represent the relationships between demand and all the factors. It also means that the actual demand and supply relationship have already taken care of what the model predicts the taxi should do. For example, if imbalance is positive, the demand is more than the supply, and the empty taxis are actually entering the area; if the residual is positive, the model tells us the same thing: the taxis should be entering the area. If both are positive, the model agrees with the data, meaning that the high demand is met by empty taxis moving into the neighborhood to serve the customers.

Similarly, if both imbalance and residuals are negative, the supply is always more than the demand, indicating that empty taxis are exiting the area, and the model implies that the empty taxis should indeed exit the area, meaning that taxis are leaving low demand areas where they are not needed. What this means is that the area may not have significantly different characteristics that generate taxi demand, so even though people need to take a taxi to go there, which makes empty cabs available after making drop-offs, the empty cabs are more likely to leave the area immediately instead of circling in the area for their next customer. These areas include, for example, Soho and Lincoln Square at Midnight; most residential neighborhoods in the Outer Boroughs at all three hours (Figure 5.11-Figure 5.13).

The interesting part of this interpretation is that the residuals and the imbalance for some areas have different signs, indicating that the models provide useful information. If the residual is positive while the imbalance is negative (underserved areas), the real pickups are not enough to attract empty taxis, but the demand model suggests there should be more empty taxis to provide a comparable level of service. This indicates a potential latent demand: demand that would exist if there were more available taxis to serve them. Those areas are usually urban in character but lie outside of the areas of densest taxi demand where empty taxis tend to congregate. Such areas include the Lower East Side of Manhattan, Harlem, Woodside, Sunnyside, and Williamsburg at 12 A.M. (Figure 5.11). The green taxi was introduced in 2012 to provide more taxis to complement the supply of yellow taxis and meet the demands of the neighborhoods in north of Manhattan and the Outer Boroughs. This is consistent with the model findings.

If the residual is negative while the imbalance is positive, as seen in Midtown Manhattan at all hours (Figure 5.11-Figure 5.13), the taxis become over-supplied because there are too many drop-offs, and empty taxis should leave, but instead they are entering. A possible reason for this result is that taxi drivers believe that Midtown Manhattan has a greater potential customer base, and so they tend to circulate in Midtown Manhattan providing surplus taxicab supply. Efficient policies can effectively provide information for yellow taxi drivers, such as the installation of centralized information system by updating demand geographically (on a map). Such policies would aid in distributing taxis to areas other than Midtown Manhattan.

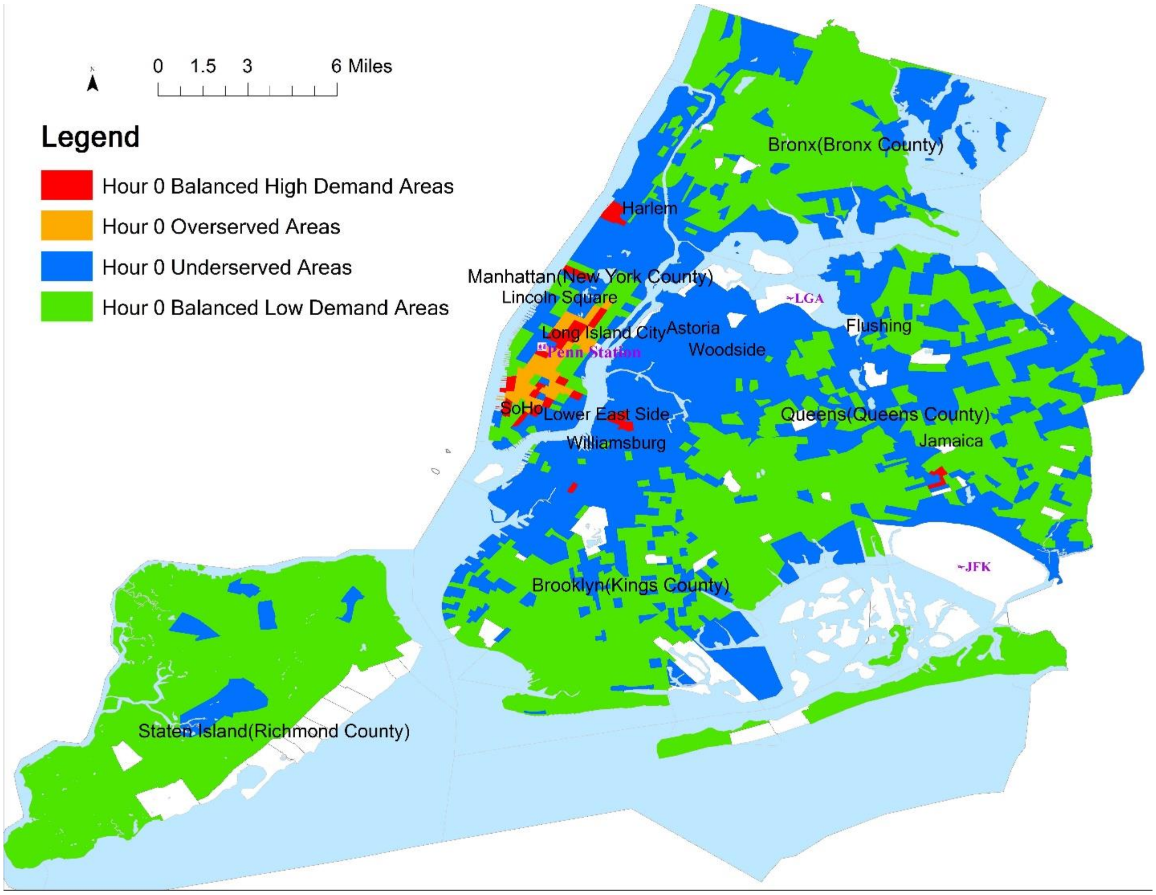


Figure 5.11 Four Categories Based on CAR Residuals and Imbalance between Demand and Supply at 12 A.M.

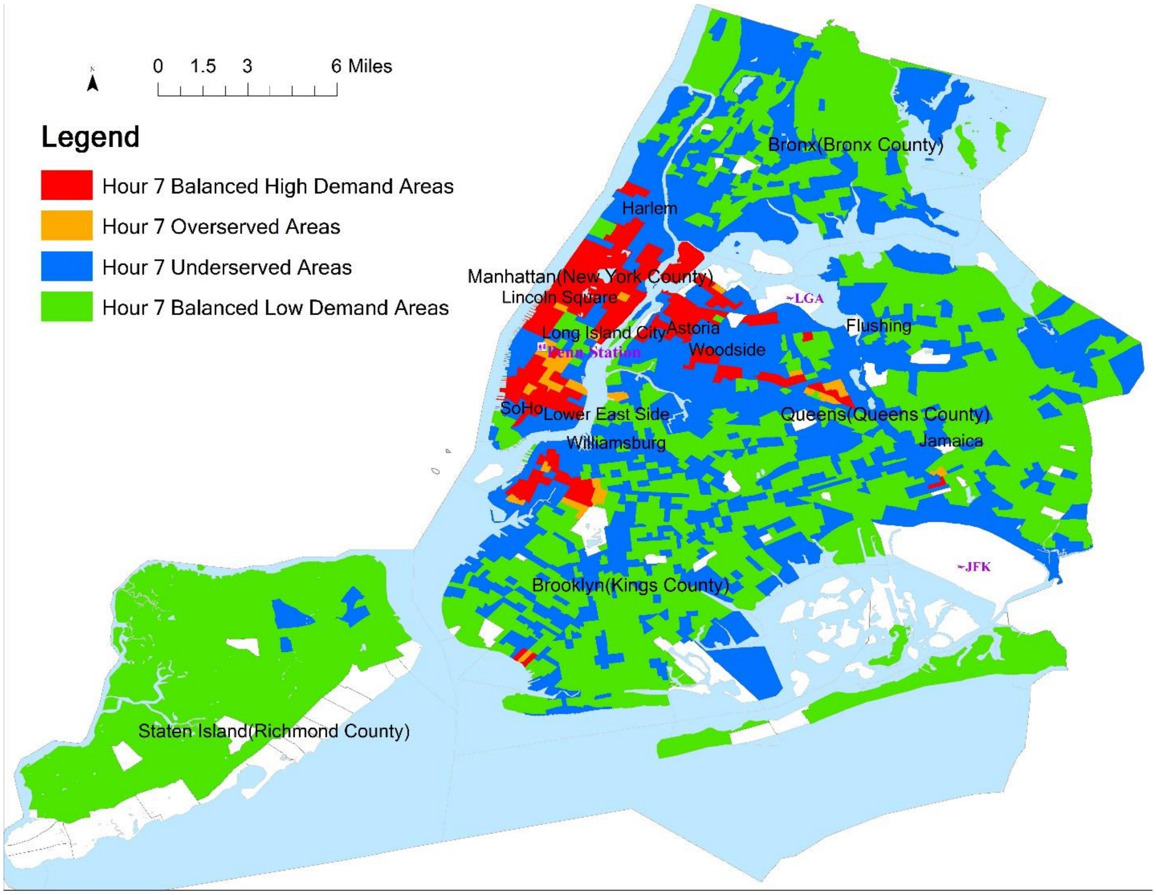


Figure 5.12 Four Categories Based on CAR Residuals and Imbalance between Demand and Supply at 7 A.M.

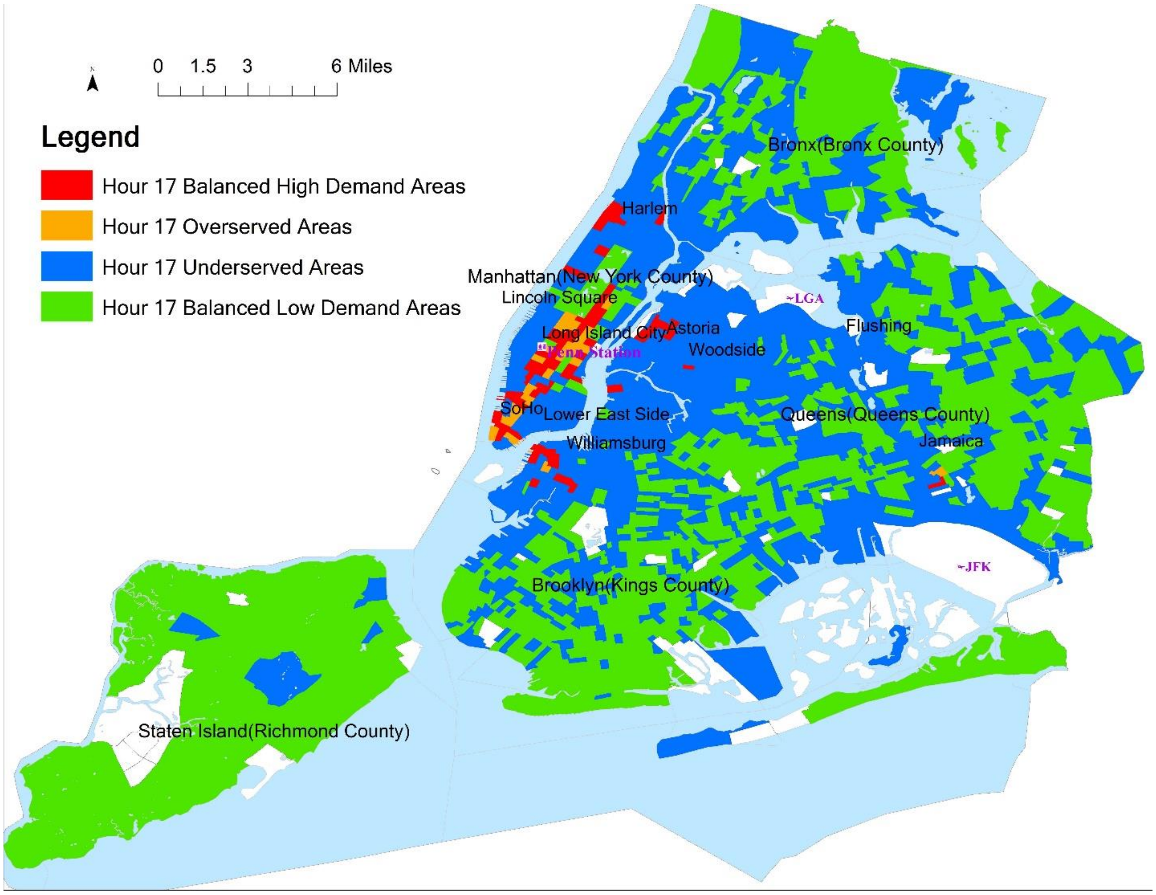


Figure 5.13 Four Categories Based on CAR Residuals and Imbalance between Demand and Supply at 5 P.M.

Of more interest are the underserved and overserved areas. The insights provided in those areas are essential for improving and planning future taxi demand. The populations and total jobs of census tracts that belong to each of the four categories are calculated for both NYC and Manhattan in Table 5.13. According to the table, at 12 A.M., there are 37% of the population and 28% jobs are underserved in New York City (all 5 boroughs), while 49% of the population and 17% of jobs in Manhattan are underserved. This indicates a large potential market for ride-share services like Uber and Lyft. This may also suggest that the yellow medallion would make more revenue if they can be better distributed with more information (like centralized coordination) regarding where and when those population and jobs demand taxi services.

Table 5.13 The Population and Total Jobs that Belong to Census Tracts for Each of the Four Categories.

	Hour	Class Code	Class Name	Pop by Categ.	Pop Pert. (%)	Jobs by Categ.	Job Pert. (%)	Ratio (pop/jobs)
NYC	12 A.M.	1	BHD	179198	2%	581881	16%	0.31
		2	Overserved	288774	4%	901977	25%	0.32
		3	Underserved	2961898	36%	1031460	28%	2.87
		4	BLD	4712102	58%	1138090	31%	4.14
	7 A.M.	1	BHD	1218461	15%	742290	20%	1.64
		2	Overserved	122913	2%	217955	6%	0.56
		3	Underserved	2606556	32%	1148969	31%	2.27
		4	BLD	4126220	51%	1539445	42%	2.68
	5 P.M.	1	BHD	347983	4%	952267	26%	0.37
		2	Overserved	99886	1%	559477	15%	0.18
		3	Underserved	3638582	45%	1079693	30%	3.37
		4	BLD	4035429	50%	1059788	29%	3.81
Manhattan	12 A.M.	1	BHD	149707	9%	561211	26%	0.27
		2	Overserved	288774	18%	901977	42%	0.32
		3	Underserved	775592	49%	366687	17%	2.12
		4	BLD	371798	23%	298606	14%	1.25
	7 A.M.	1	BHD	850930	54%	578464	27%	1.47
		2	Overserved	122913	8%	217955	10%	0.56
		3	Underserved	421523	27%	590906	28%	0.71
		4	BLD	190505	12%	741156	35%	0.26
	5 P.M.	1	BHD	271353	17%	811498	38%	0.33
		2	Overserved	99886	6%	559477	26%	0.18
		3	Underserved	936856	59%	454314	21%	2.06
		4	BLD	277776	18%	303192	14%	0.92

To help understand which factors have bigger influence on taxi demand, the ratio of population to jobs is calculated for all census tracts belong to each category. All overserved areas are located in Manhattan at all three hours; therefore, the ratio is the same for NYC and Manhattan. The underserved areas tend to have higher ratio of population to jobs than overserved areas, which is consistent with the underserved areas shown in Figure 5.11, such as the Lower East Side of Manhattan. Conversely, overserved areas have a lower ratio of population to jobs than the underserved areas in both NYC and Manhattan, which makes sense because taxis are more easily driven by the level of economic activities (jobs) than just the number of people living in the area. The same pattern is observed at the other two hours (7 A.M. and 5 P.M.). The balanced low areas have the highest ratio of population to jobs, since those neighborhoods are mostly one and two family residential land use.

The ratio of population to jobs indicates whether an area is underserved or overserved. This empirical statistic (the ratio) for NYC and Manhattan in Table 5.13 can be potentially used to estimate which category an area belongs to according to the actual ratio of population versus jobs in that area.

5.5 Summary of Findings

Using large-scale taxi GPS data in NYC, this study demonstrates the temporal and spatial variation of taxi demand and the relationship between taxi demand and transit accessibility and other demographic and socioeconomic factors. A novel approach for calculating the minimum transit access time (TAT) uses transit LOS and K-NN algorithm, and a unique independent variable is the number of taxi drop-offs (DrpOff). The negative binomial model is shown to be appropriate for the overdispersed taxi count data. Moran's I tests on both dependent and independent variables indicate that spatial autocorrelation exists in the data. Therefore, a Poisson-Gamma-Conditional Autoregressive (CAR) model with Markov Chain Monte Carlo (MCMC) algorithm was selected for the trip generation modeling approach.

The theoretical understanding of the relationships and correlation between variables was first applied in an initial screening process. Then, a number of model selection criteria — significance of overdispersion parameter, significance of coefficients, satisfied G-R stat, AIC, LL, deviance — are used to select among different combinations of theoretically important factors and categorical factors. Models are estimated for data at three time intervals: 12 A.M., 7 A.M., and 5 P.M. A best model was estimated for each time interval based on the best statistical fit. The CAR residuals and imbalance between demand and supply are used to classify NYC census tracts into four categories: balanced high demand (BHD) areas, overserved areas, underserved areas, and balanced low demand areas (BLD).

This study reveals six key findings:

1. Six important independent variables influence taxi trips: population (Pop), education (EduBac), drop-offs (DrpOff), income (CapInc), TAT, and employment (TotJob). The important factors in the three final CAR models are mixed residential and commercial land use (LU04) and the number of jobs at accommodation and food services (JobFod), when controlling for population with ages between 14 and 35.
2. The influence of these factors on taxi pickups confirms the hypotheses. All independent variables are positively correlated with taxi demand except TAT, which makes sense because higher accessibility (smaller TAT) leads to more taxi demand.
3. The influence of these factors on taxi pickups varies by time. Drop-off and TAT have a stronger impact on the taxi demand at 12 A.M. compared to 7 A.M. and 5 P.M., while population, education, and employment have a stronger influence on taxi demand at daytime hours than at midnight.
4. The overall predictability of the models is not very good. The percent of census tracts in NYC that has the actual taxi demand falling in the 95% CI of

the predicted demand is 47%, 75%, and 56% for hour 0, hour 7 and hour 17, respectively (Table 5.10). This indicates that the 7 A.M. model has better predictability compared to that of 12 A.M. and 5 P.M. The model predictability for Manhattan is weaker compared to NYC (5 Boroughs).

5. Midtown Manhattan has a sufficient supply of taxis, while most residential areas in the Outer boroughs do not generate much demand for taxis. Some neighborhoods are identified as underserved areas: the Lower East Side of Manhattan, Harlem, and Williamsburg.
6. Some community districts have a change in the pattern of demand from one hour to the next, indicating a directional imbalance between different times of day; i.e., Harlem, Astoria, Sunnyside, and Woodside.

To sum up, large datasets such as the records of taxicab trips in NYC present some challenges, because the raw data is too big to be analyzed directly through conventional methods. By processing the data and developing models that relate the taxicab data to other sources of demographic information at different times of day, it is possible to gain useful insights about the role that taxicabs play in the broader transportation system. This work not only provides a visualization of the raw data, but also the predicted taxi demand, indicating the imbalance between demand and supply as well as areas with insufficient taxi supply (underserved) and surplus taxi supply (overserved). More importantly, these insights can be used to plan and improve the transportation system to meet the needs of users. For example, the underserved area may not be better served with conventional taxis; there may not be enough demand to warrant the level of available cabs they would need for good service. This demand may be better served by for-hire services that use technology to facilitate the customer-cab matching process, such as Uber and Lyft. Maybe conventional yellow cabs could fill this need by adopting these technologies as well.

This study provides many useful insights for taxi planning; however, there is still some interesting information that is limited from the trip generation modeling. For example, the methods used in this study are inconclusive regarding the relationship between taxis and transit is competitive or complementary. Also, it is interesting to know whether the taxis are operating at their full capacity or earning more revenue at the overserved areas, i.e., are they running empty more or less than the taxis circulating in areas that belong to other categories; similarly, are they earning more or less revenue compared taxis in other areas. Additionally, it is unknown whether the low predictability of the CAR model is due to the scale of spatial aggregation. It is recommended that the future work include:

1. Collect similar ridership data from transit and conduct survey on taxi passenger's travel purpose. This may provide information to understand when and where taxis are complimentary or competitive to transit.
2. Study the empty travel time of taxis and the fare of taxi trips along with the trip generation model, to understand when and where taxis have higher occupied rate or create higher revenue.
3. Expand to larger spatial aggregation scales, such as census block, and maybe use different scales at different areas. For example, taxi demand in most of the Outer Boroughs is relatively homogenous, so a larger spatial scale may be

warranted, while a smaller spatial scale may be sufficient for Manhattan because of its highly varied taxi pickups and neighborhood characteristics.

CHAPTER 6

EVALUATION OF DATA

6.1 The Advantages and Major Deficiencies of Big Data

The primary objective of this study is to investigate the taxi supply and forecast taxi demand using trip generation modeling techniques with taxi GPS data collected over a two-year span by TLC in New York City. Relevant data such as transit and demographic data have been used to conduct the proposed analyses: an analysis of favorable neighborhoods for vacant taxis and trip generation modeling. This data is used to explore other possibilities, such as mode cost comparison between taxis and transit. This chapter also discusses the advantages and major deficiencies of using big data.

This dissertation uses three main bodies of data. The first is a complete collection of GPS taxi data for every taxi trip made in New York City within a two-year period. Second, detailed transit schedules for the same geographic regions are acquired using Google Transit Feed Data (GTFS), and the directional services from one point to another point are acquired using Google Maps Direction API. Finally, this transportation data is supplemented with demographic, employment, and land use data, which is expected to include key characteristics of the locations that are associated with the highest rates of taxi use. Chapter 3 describes these data sources and briefly discusses their limitations.

This study makes three major contributions in filling the gaps in data and knowledge related to taxi demand and providing empirical evidence regarding taxi planning and regulation. The foremost contribution is that this study uniquely utilizes a detailed and complete GPS taxi data for developing a customer search model and a taxi trip generation model. The customer search model provides insights about the estimation of vacant taxi travel time, travel distance, and efficiency in re-allocating the empty taxis. The trip generation model identifies several important factors that are associated with taxi demand. The second contribution is that the study aggregates the data in refined spatial and temporal scales, which provides detailed insights on when and where the taxi demand is predicted to be insufficient or surplus. This has not existed in any previous literature. The third contribution of the study is the implementation and combination of novel technologies (e.g., GPS data, Google Maps, Google transit feed data, GIS), and various statistical and economic methodologies, which provide a systematic way to analyze and forecast taxi demand in NYC.

Up until this point, a number of topics in taxi demand and supply have been investigated by using this taxi data, demographic data, socioeconomic data, and Google transit data. The big data has shown to be very useful because it has provided:

1. Complete, continuous, and detailed information. This allows us to aggregate the taxi pickups and drop-offs with good coverage of different locations in NYC at different times. For example, it includes every day in each of the datasets (a 10-month one and a 14-month one), it covers all times of day, and

it includes all the vehicles' information. Additionally, the complete data provides a possibility to understand when and where there are fewer taxi pickups or drop-offs. For instance, if an area does not have even a single taxi trip in a month, then it would not be able to be identified using a one week of data.

2. More refined spatial and temporal data that is able to show some statistical significance. Big taxi data allows the aggregation of data in a scale as refined as possible; for example, to a scale of census tract, to a scale of census block and so on. This aggregated data is then able to show some statistical significance.
3. Rough estimation of travel time and travel distance from one place to another. By geographically and temporally keeping track of the number of the trips for certain origin and destination (O-D) pairs, we are able to estimate the average travel time and travel distance at different times of day and at different parts of New York City.

However, the data has a number of limitations:

1. Although the taxi data spans a large period of time for the entire New York City region, the data is not traceable and therefore it cannot be used for routing and network modeling. Each taxi trip contains only two geo-locations, the origin and the destination of the trip. Therefore, after the taxi data is aggregated into certain origin-destination (O-D) pairs (this could be a very small place, such as a train station, or a building), it returns a very small number of taxi trips. However, traceable data contains more trip information

depending on how frequently the trip location and time is recorded when taxis are occupied or empty. For example, if the frequency of recording is every 5 seconds, a 10 minute taxi trip will have at least 120 sub-trip geo-locations. Those geo-locations are very useful to build network and routing models.

2. This taxi data is not sufficient to build a traffic realization model in a refined time scale for a similar reason that has stated above: there is not enough resolution of taxi movements to reflect the realized traffic network, especially at locations where taxi drivers travel less frequently, for example, the Outer Boroughs.
3. Similarly, this taxi data can only be used to estimate travel time for popular OD pairs; however, for areas that taxis do not travel to very often, there is no sufficient data to estimate the travel time.

Because there is no aggregated transit data in such a refined scale, other sources of transit data must be identified, such as Google transit feed data, as well as routing data using The Google Directions API services. This chapter illustrates the advantages and limitations of big GPS data sources. It is important because the models are built based on the data. The following sub-section of this chapter will show a mode cost model between taxi and transit.

6.2 Case Study: Mode Cost Comparison

Transportation planning for airport ground access has attracted increasing attention in recent years from both planners and researchers. Approximately 65% of airport trips are made by private vehicles in the US and Europe (Humphreys and Ison, 2005). The remaining 35% of trips depend on alternative airport access modes. For example, 37% of the passengers leaving John F. Kennedy International Airport (JFK) are estimated to use taxis (Convey et al. 2012). Several studies have been conducted to measure the effectiveness of airport ground transportation services other than private automobiles such as taxi cabs, buses, or trains, based on the speed and reliability of travel times for these alternative modes (Shriner, and Hoel, 1999; Gosling, 2006). Another challenging issue is to understand air passengers' airport access mode choice. Trips to and from airports should be treated separately in regional behavioral travel models due to their unique characteristics (e.g., different value of time for passengers) and a number of factors that influence mode choice such as trip purpose (e.g., business or leisure) and time-of-day (Gupta, 2008). The trip cost of using an alternative mode relative to driving has been shown to be one of the major factors affecting the mode choice of passengers along with factors such as the frequency of service and luggage capacity (Akar, 2013).

The generalized cost of travel for different modes of transportation varies significantly depending on several factors related to traffic conditions and public transportation schedules. The use of public transportation might save money, but in return people give up the convenience of a door-to-door ride directly to or from their home or work place. The willingness to pay for more reliable transportation service for airport access is estimated to be considerably higher than for regular daily commutes since scheduling constraints are more binding (Gupta et al., 2008). Estimated values of time for airport trips have also been found to be significantly different than estimated values for overall network travelers (Gupta, 2008). Moreover, shorter travel times are valued more for business trips compared to non-business trips to avoid the risk of missing the flight (Koster, 2011).

This study is motivated by the importance of realistically comparing travel costs of alternative modes for airport ground access using revealed data sources. The majority of studies in the literature calculate the transit trip costs using previously determined schedules for certain types of transit services, which cannot address all of the available transit options for airport ground access in cities with highly complex transit networks such as New York City (NYC). Different transit options for different times-of-day should be considered in the analysis, because the speed and reliability of each mode varies over the course of a day. While some of the existing methodologies consider waiting times as a part of total travel time, rough estimates for possible transfer times at train or bus stations can lead to biased results. It has also been demonstrated that about 40% of the population in the US uses smart phones, and there is a great potential for these devices to disseminate reliable real-time transit travel information (Gould, 2013). Taxi travel times, on the other hand, are usually calculated assuming that the vehicle travels on the shortest path to the airport. Traffic conditions on the network by time of day have not been addressed adequately in previous studies although traffic congestion affects both the travel time and taxi fare.

In this study, a novel methodology is introduced utilizing big data to compare travel times and fares for NYC airport ground access by taxi and transit. The analysis is based on historical taxi global positioning system (GPS) data recordings in NYC and transit schedule information from a web-based application that we developed using Google Maps Developer Application Programming Interface (API). The transit travel time data enables us to compare the travel times for a transit passenger who uses his/her smart phone to plan an airport trip. Taxi travel times, on the other hand, are the average values by hour of the day of the observed trips that are extracted from the NYC taxi GPS data. Therefore the comparison in this study aims to evaluate the travel options available to a well-informed passenger, who has perfect knowledge about the expected taxi fare and travel time.

6.2.1 Methodology

The objective of this study is to develop a data-oriented method to compare the generalized cost for different non-driving modes for airport access and to understand whether transit or taxi yields a better utility at different times of the day. While the results of this study may be useful to individuals making travel choices, the method proposed in this study can also help policymakers understand the factors that affect mode choice in order to plan airport ground access.

As web services and information technology become more advanced, it is easier for people to acquire complete information about travel by transit and taxi. Assuming that passengers make travel decisions based on money costs and travel time, the relative attractiveness of one mode over the other may change as transit schedules, fares, and taxi travel times vary for different times of day and days of the week. The relevant information can be obtained from Google Transit and a large taxi GPS data.

The total generalized cost for an individual trip in units of dollars can be computed for each mode i at time j is denoted by TC_{ij} and calculated as follows (Zhang, 2008; Morgul, 2011):

$$\mathrm{TC_{ij}} = \alpha \times \mathrm{T_{ij}} + \frac{\mathrm{F_{ij}}}{\mathrm{n}} \qquad \textbf{Eq 6.1}$$

where α is the passenger's value of time ($/hr), $\mathrm{T_{ij}}$ is the average travel time for the trip (hours), $\mathrm{F_{ij}}$ is the average fare paid for the trip ($), and n is the number of passengers sharing a taxi cab; for transit $\mathrm{n} = 1$. The total generalized cost can also be expressed in units of hours by dividing the $\mathrm{TC_{ij}}$ by α. The utility of each travel by each mode in hour j is based on the generalized cost:

$$\mathrm{U_{tran,j}} = \mathrm{b} - \beta \times \mathrm{TC_{tran,j}} = \mathrm{b} - \alpha\beta \times \mathrm{T_{tran,j}} - \beta \times \mathrm{F_{tran,j}} \qquad \textbf{Eq 6.2}$$

$$\mathrm{U_{taxi,j}} = \mathrm{b} - \beta \times \mathrm{TC_{taxi,j}} = \mathrm{b} - \alpha\beta \times \mathrm{T_{taxi,j}} - \frac{\beta \times \mathrm{F_{taxi,j}}}{\mathrm{n}} \qquad \textbf{Eq 6.3}$$

where b is the benefit for each individual of completing a trip to or from the airport, and β is the equivalent utility of a dollar. For an airport trip, the benefit is assumed to be the same for both choices as long as the OD pair is fixed. The choice between two modes, such as transit and taxi, is typically modeled with a binary logit model based on the difference of utilities between the choices (Train, 2009). The probability that an individual will choose transit over taxi is:

$$P_{trans,j} = \frac{e^{U_{trans,j}}}{e^{U_{trans,j}} + e^{U_{taxi,j}}} = \frac{e^{(-\beta \times TC_{tran,j})}}{e^{(-\beta \times TC_{tran,j})} + e^{(-\beta \times TC_{taxi,j})}} \qquad \textbf{Eq 6.4}$$

and the probability of choosing taxi over transit is:

$$P_{taxi,j} = 1 - P_{trans,j} \qquad \textbf{Eq 6.5}$$

The number of passengers choosing mode i is the product of $P_{i,j}$ and the total travel demand. In the following sections, transit and taxi trips in NYC are compared based on their travel time, total cost and the corresponding choice probability.

6.2.2 NYC Airport Taxi and Transit Cost Comparison at Different Times of Day

The two main types of public transportation services for airport access in NYC are transit (including train, AirTrain, subway, and bus) and taxi are compared for trips between Penn Station and the three main airports in NYC (Figure 1): JFK, Newark Liberty International Airport (EWR) and LaGuardia Airport (LGA). These constitute the largest airport system in the United States. Penn Station is selected as the non-airport trip end of interest for this study, because it is a major hub of transit and taxi activity. About 18% of taxi trips from Penn Station that leave the city are to EWR, and about 1% of all taxi trips to/from Penn Station are from/to LGA and JFK.

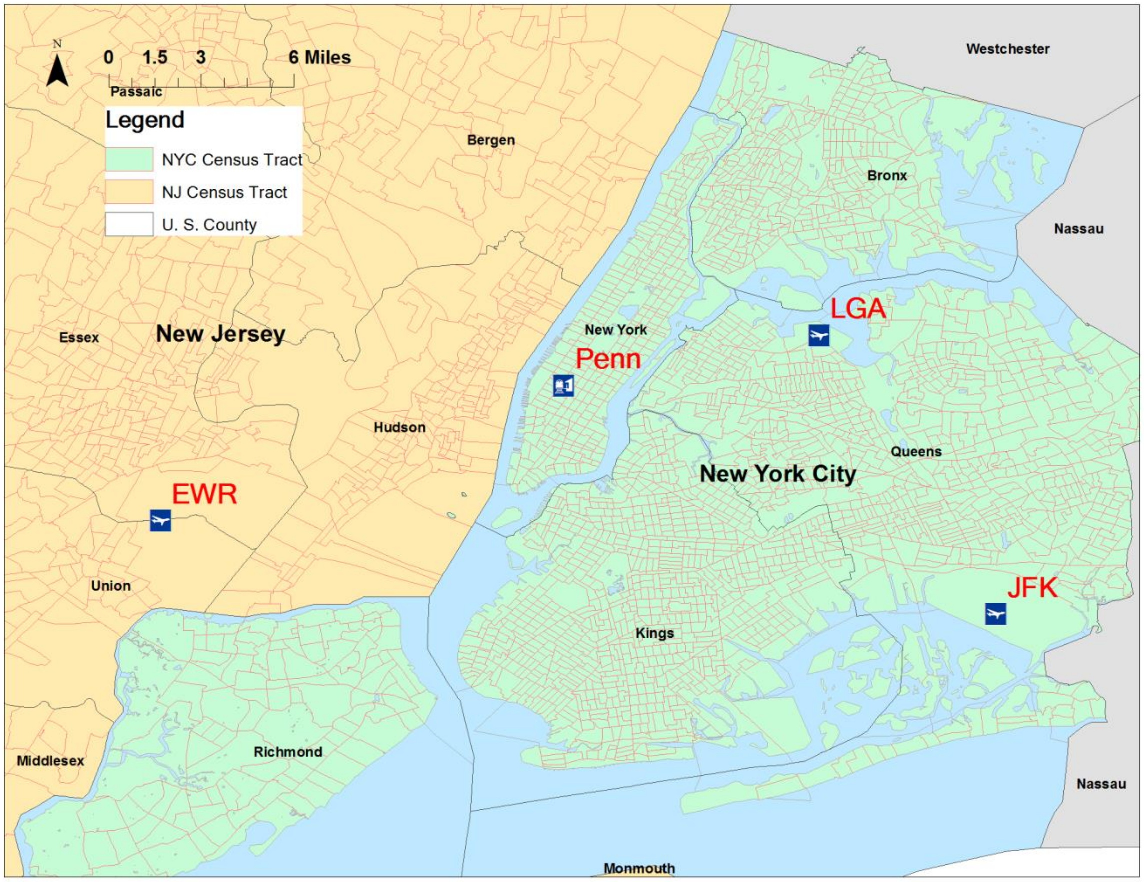


Figure 6.1 Location of NYC Penn Station and three airports.

Taxi Data

Taxi GPS data for NYC is available for every trip in a 10-month period (February 1, 2010 to November 28, 2010). The data consists of 147 million taxi trips across the city, from which trips between Penn Station and the three airports are extracted. NYC taxi trip data have also been used in studies of travel time reliability for inner-city traffic (Yazici, 2012) and commercial vehicle delivery time estimation (Morgul, 2013).

The trip information in this taxi dataset such as fare, travel time, and trip distance is used to estimate the cost of each transportation mode. Five origin-destination (O-D) pairs between Penn Station and the three airports are examined in this case study. Information regarding each of the 5 OD pairs is summarized in the following Table 6.1 .

Table 6.1 The Taxi Trips Extracted from 10-month GPS Data and 1-week Transit Trips.

	Taxi					**Transit**	
OD pair	**No. of Obs**	**Passenger No**	**Total Amount ($)**	**Trip Time (min)**	**Trip Distance (mi)**	**Trip Time (min)**	**Fare ($)**
		Mean ***(SD)***	**Mean** ***(SD)***	**Mean** ***(SD)***	**Mean** ***(SD)***	**Mean** ***(SD)***	**Mean** ***(SD)***
Penn-JFK	5624	1.81 *(1.29)*	52.34 *(5.26)*	47.79 *(17.57)*	17.24 *(1.89)*	52.51 *(10.34)*	12.47 *(1.69)*
JFK-Penn	2691	1.87 *(1.27)*	51.69 *(4.33)*	45.11 *(12.74)*	17.72 *(1.79)*	60.96 *(10.03)*	12.65 *(1.37)*
Penn-LGA	9697	1.63 *(1.21)*	34.85 *(5.64)*	31.43 *(10.30)*	10.23 *(1.38)*	60.85 *(6.32)*	4.92 *(3.57)*
LGA-Penn	3630	1.65 *(1.22)*	35.07 *(5.58)*	32.06 *(9.35)*	10.21 *(1.71)*	61.48 *(7.57)*	6.91 *(3.23)*
Penn-EWR	1445	1.80 *(1.28)*	67.30 *(10.48)*	32.09 *(9.49)*	17.08 *(2.05)*	58.87 *(19.90)*	17.25 *(9.54)*

SD: Standard Deviation

The total fare is a flat rate between most of Manhattan and JFK or EWR airports plus any tolls, tips, and surcharges. The fare between Manhattan and LGA has some variability because trips are charged the normal metered rate. The travel time has largest variability among all four variables, which indicates variability of traffic conditions. The trip distance is mostly stable for all airport trips, but the slight variability indicates that various alternative routes might be taken for the same OD pairs.

Transit Data

As mentioned in Chapter 3, the Google Directions API Service was used to obtain transit travel time, and it offers free transit route guidance with a daily request limit. The fare is estimated based on the optimal route. Data for approximately 2,016 transit trips have been collected for each OD pair. The transit travel duration of each trip includes waiting time, transfer time, and in-vehicle travel time.

The transit travel time distributions for all 5 OD pairs are shown in Figure 6.2, indicating that travel time in weekdays are consistent, which could result from the fact that the transit schedule is similar for all weekdays, but it is necessary to analyze Saturday and Sunday separately. The average travel times for transit on weekends is higher because service headways are longer, and average travel times for taxis on weekends is lower because traffic is less congested. The day of the week affects the costs that travelers face on each mode.

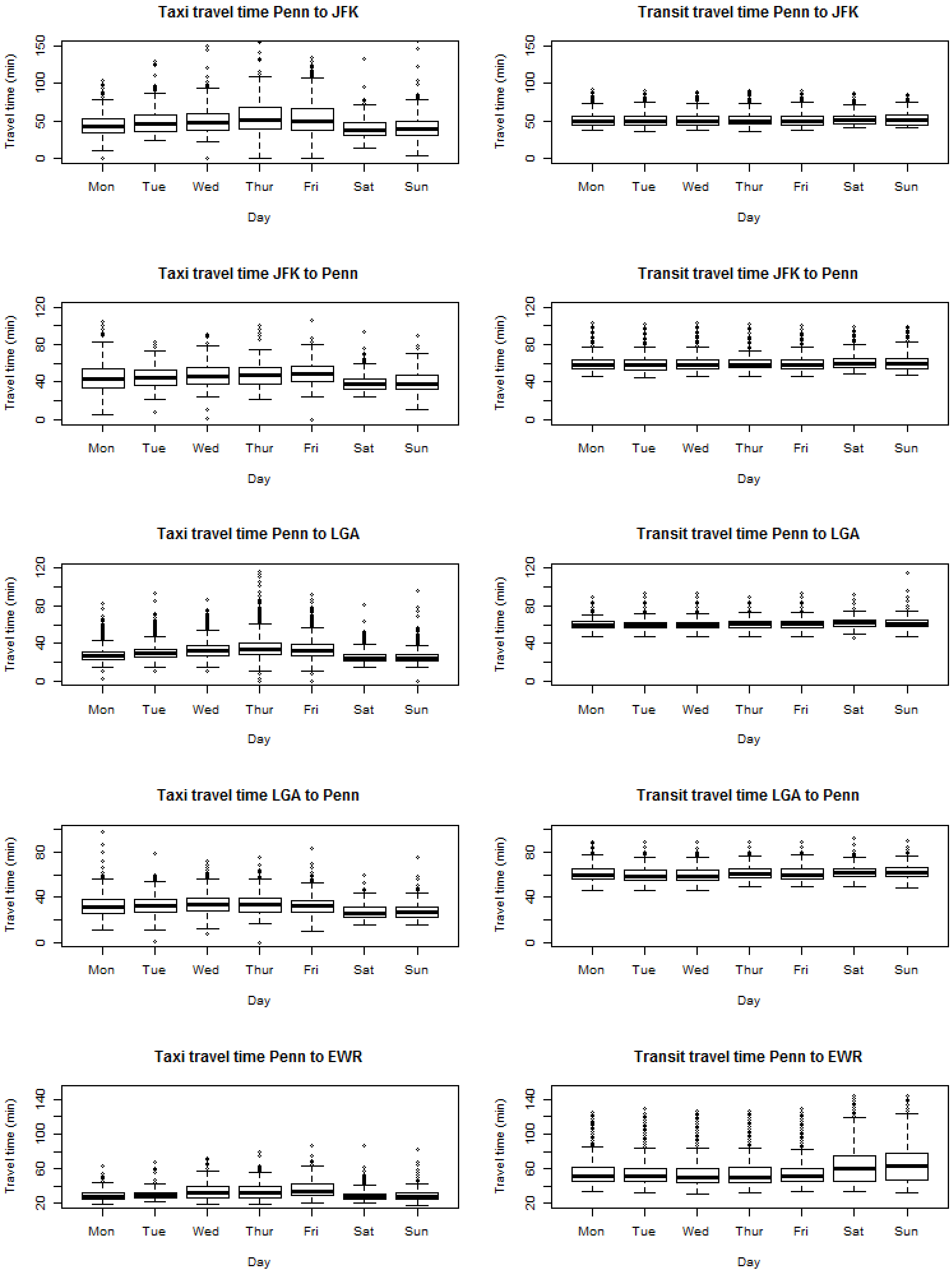


Figure 6.2 Boxplot of Transit Travel Time by Day of the Week for 5 OD-Pairs.

Determination of Value of Time, α

The value of time for airport trips varies considerably from person to person, and it can be considered as a continuous random variable that is distributed across the user population (Jiang, 2013). The value of time for business trips can be higher than for leisure trips (Harvey, 1986; Gupta, 2008). The distribution of values of time for airport trips is also likely to differ from that of other trip purposes, which makes estimation of this value difficult.

The UK Department of Transport suggests £47.95 (equivalent $76) as the value of time for a taxi/minicab passenger and £39.65 ($63) for a rail passenger in 2010 (UKDOT website). For trips to/from Penn Station, the income in Manhattan is used as a reference value (Harvey, 1986). The 5-year American Community Survey estimate of per capita income in Manhattan is $61,290 (2007-2011 American Community Survey), which is $29.5/hr if working a full time job with 40 hours per week. Gupta et al. (2008) considers a higher value of time for airport trips because travelers may be willing to pay more to avoid missing their flight. The authors suggested $42/hr for leisure trips and $63/hr for business trips. Since we do not know the value of time for passengers that made airport trips in NYC, a preliminary value of time $40/hr is used to represent everyone and anytime based on above references. In the sensitivity analysis that follows, wide ranging of values of time are considered.

Calibration of Coefficient β

Eq 6.2 and Eq 6.3 display the relationship between cost and utility, and the β coefficient plays an important role in determining the outcome of the binary logit model. The β value is the marginal utility of total cost. In order to estimate the probability of choosing one travel mode, it is necessary to determine β, which can be roughly estimated by comparing the total number of transit and taxi trips. The transit ridership data from all airports are essential to estimate β, however, it is very difficult to acquire those data for two major reasons. First, it is impossible to track how many subway passengers travel between Penn Station and the airports, even if we can get the MetroCard data on how many people enter and exit subway stations every day; we cannot tell where they are headed because passengers only swipe their MetroCard when entering the station, and there is no record of where they leave the system. Traceable ridership data would be more useful if information on every single passenger were available or a survey were conducted regarding transit passenger's travel behavior. Second, it is possible to get the total ridership of AirTrain to estimate the usage of transit to get to/from the airports, but only JFK is served by AirTrain.

In order to illustrate this methodology, JFK-AirTrain ridership information and taxi GPS data are used to estimate a single β. One value of β is used for all three airport trips because it is likely that the average marginal utility of total cost is similar for passengers using each of the airports. Furthermore, data are not available to estimate specific β values for LGA and EWR.

Paid ridership of the JFK AirTrain was 5.3 million passengers in 2010 (2010 Airport Traffic Report), which accounts for nearly all of the transit trips to and from JFK. In the same time period, there were 3.386 million taxi trips to and from JFK, extrapolated from the complete 10-month records of taxi GPS data. Based on the trip counts above, 39% of non-driving trips were made by taxi and 61% were made by transit to get to and from JFK.

Without more detailed transit ridership data, the overall mode share for all trips to and from JFK during 2010 is considered to be the same as mode share for trips between Penn Station and JFK. The logit model is calibrated by selecting the β value that makes the model estimates over the course of the day match this observed mode share. As an example, Figure 6.4 shows the relationship between β and the aggregated probability of taking a taxi, based on the total generalized cost in each hour j between Penn Station and JFK (including both directions: Penn-JFK and JFK-Penn), and $n = 1$, $\alpha = \$40$/hour. At each hour, the relationship between number of taxi trips ($n_{taxi,j}$) and the estimated number of transit users ($\hat{n}_{transit,j}$) is:

$$\frac{n_{taxi,j}}{P_{taxi,j}} - n_{taxi,j} = \hat{n}_{transit,j} \qquad \textbf{Eq 6.6}$$

$$P_{taxi,j} = \frac{\sum_j n_{taxi,j}}{\sum_j \hat{n}_{transit,j} + \sum_j n_{taxi,j}} \qquad \textbf{Eq 6.7}$$

When $\beta = 0.012$, the expected probability of people choosing taxi to the airport is 0.39, which matches the data from 2010. This value of β is applied to the cost data for all the airport trips in order to estimate the mode share by taxi and transit.

Assumptions

In this part of the analysis, the main challenges are the data collection and data processing. Some assumptions made in order to perform the comparison of transit versus taxi use are as follows:

1) Taxi fares are calculated per person, so the total fare is divided by the number of passengers (Eq 6.1), but the travel time is experienced by each passenger regardless of the group size.

2) There are always a sufficient number of taxis available at each airport for passengers to hail, so no time is spent for walking and waiting for a taxi at any airport.
3) The origin and destination of the transit trips are very close to the transit stops, so the walking distance is negligible.
4) The average trip duration in an hour of the day is assumed to represent the travel time at that hour of the day for both train and taxi, provided that there are a large number of trips per hour.
5) All passengers are able to buy the tickets before boarding transit to avoid additional fees. MTA subway and bus riders pay a flat rate of $2.50 per trip. Discount fares such as senior citizen tickets, weekly tickets, or monthly tickets are not considered.
6) Travelers to all airports are assumed to have the same average utility preferences, so β is the same for all trips.

The deficiencies in GPS data, mostly due to satellite errors, receiver noise errors, coordinate transformation errors, and errors made by the driver need to be filtered (Zito, 1995). The taxi GPS data is processed to minimize the influence of outliers. Some false records are eliminated, for example, records that have total fare amount equal to zero or travel distance is less than the straight-line distance between the origin and destination. Sometimes more than two criteria are used to determine whether to remove a data point (e.g., fare amount, distance, and travel time). Ultimately, less than 2% of the original taxi records were eliminated through this filtering process. This data selection procedure requires familiarity with information such as fare amount, distance between Penn and each airport, travel time, and the rate codes.

Transit data collected from Google every 5 minutes for all three airports are kept without filtering because Google estimates are already based on clean schedule data. The transit fare is calculated according to the routes that Google Maps provides at different times of the day.

6.2.3 Results

On weekdays, the average travel time for taxis is less than that of transit at all times. When considering both the time and money spent on the trip, the total cost indicates that even if the passenger is travelling alone, taxi has a cost advantage only in the middle of the night (12 a.m. to 6 a.m.). The taxi travel times vary significantly, and the longest travel times are usually observed during the morning peak (6 a.m. to 10 a.m.) and afternoon peak (2 p.m. to 6 p.m.) as shown in Figure 6.3.

In order to consider the variability of travel costs, we also calculate the standard error (SE) for the total cost of trips (Figure 6.3). Some taxi data has few observations at midnight, which results in relatively higher SE and a wider 95% confidence interval for the mean values at each hour (approximately equal to mean ± 1.96×SE). On the other hand, the transit data show relatively small variance within each of the 24 hours. The main difference in transit travel times arises from the waiting times and transfer times for the next available train or bus. The transit travel time and cost are less variable than taxi travel time and cost, which depend on the traffic condition at different times of day and from day to day (Tam, 2011).

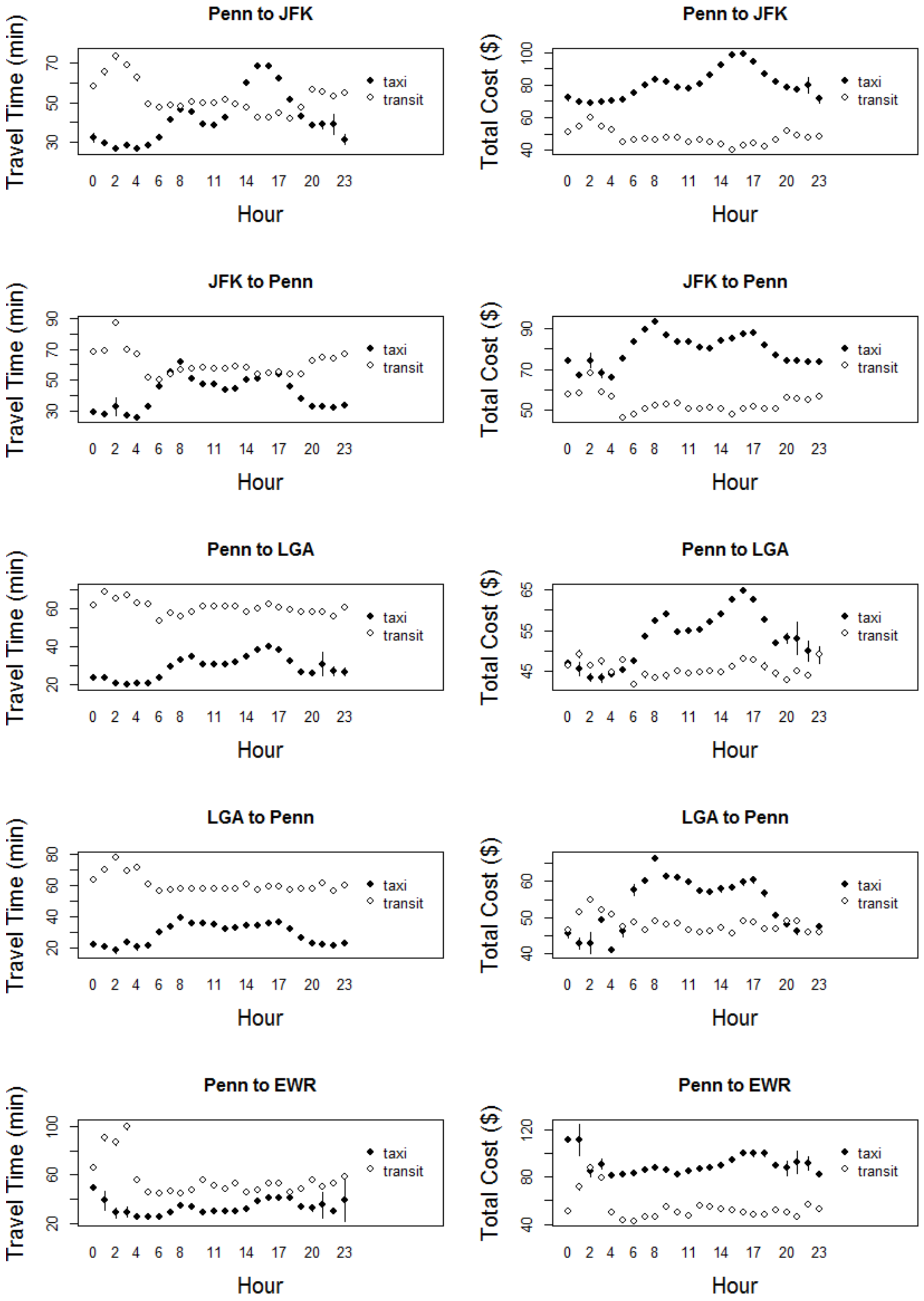


Figure 6.3 Weekday Taxi and Transit Travel Time and Cost Comparison Mean and Mean +/- Standard Error.

This analysis is limited to trips between Penn Station and the three airports. The probability of choosing a taxi at each time of day is calculated using the binary logit model, and the results are plotted in Figure 6.4. However, instead of relying on the probability calculation based on the assumptions and limited ridership data used to estimate β, we estimate the mode share on the cost comparison in Figure 6.3. Based on the difference of total cost for taxi and transit, the mode share can be expected to change for different times of the day. For example, transit tends to be more competitive during rush hours when traffic congestion makes taxi trips slower. On the other hand, taxis are more competitive in late night hours when transit headways are long.

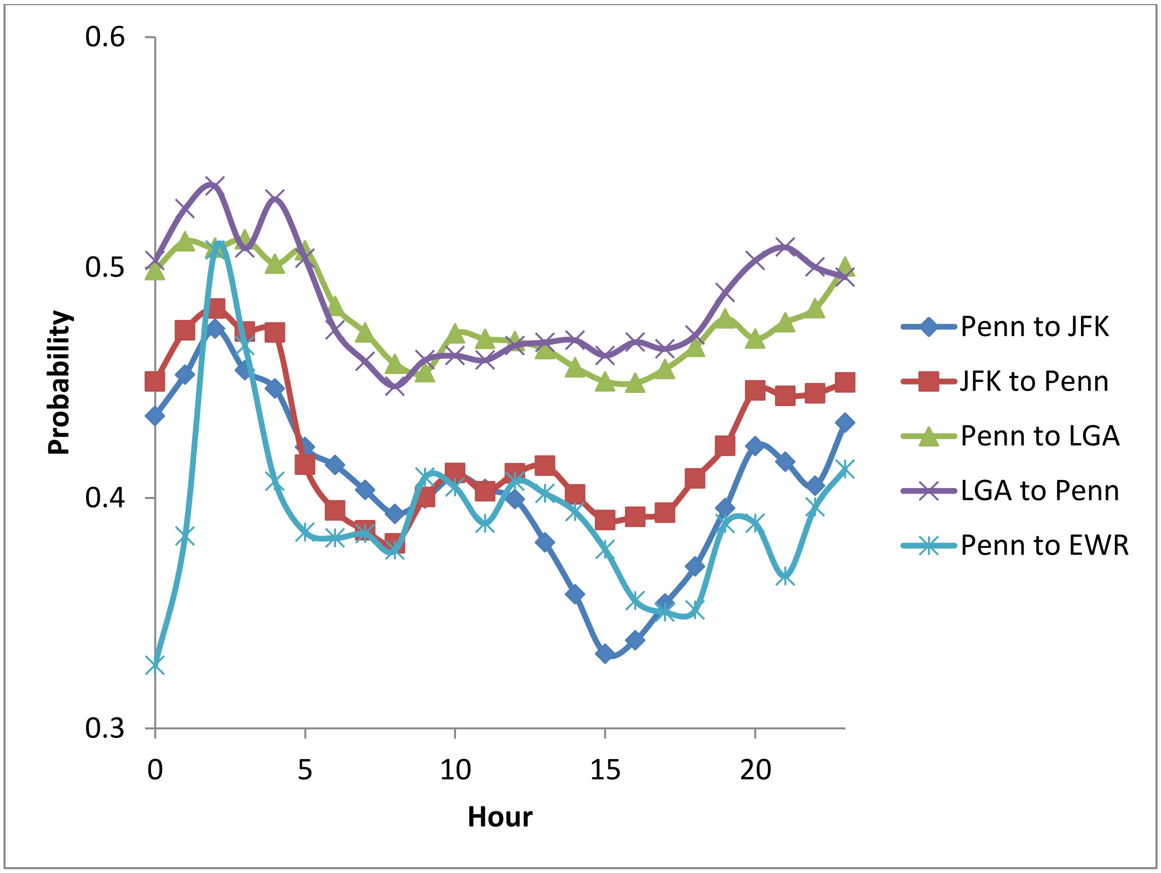


Figure 6.4 Weekday Probability of Choosing Taxi for All Airports.

Since these OD pairs are just a partial set of all trips that go to and from the airports, the analysis only reflects the costs at those locations. Different locations may have a totally different trend based on the travel time and cost. Time and money are not the only things that people consider when making travel choices, but the literature suggests that these are the most important factors. It is possible that some people use transit for almost all activities without even considering taxi, or others take taxis to the airport without ever considering the transit trip.

Some factors that likely influence mode choice other than travel time and fare are that taxis provide a more personalized door-to-door service with additional benefits such as assistance with luggage if traveling to/from airports. Some of this value is captured in the tips that are included in the taxi data and the total fare paid, which includes tip. In reality people may experience an addition penalty for using transit because they need to walk a certain distance to get transit service. These additional benefits or penalties are omitted from the analysis in order to focus on the effects of money cost and travel time on the competitiveness of each mode.

6.2.4 Sensitivity Analysis

The total cost is also influenced by the value of time and the number of passengers traveling together. Table 6.1 shows that on average there are 1.6 to 1.8 passengers taking taxis together to go to or from each airport according to the GPS data. A sensitivity analysis is performed to investigate the effects of both the value of time and the number of passengers in the group on the probability of an individual’s travel mode choice in detail.

The average travel time and fare from all records for taxi and transit are considered as travel time and fare for each OD pair. The sensitivity analysis considers variation of the value of time, α (in range \$10/hr to \$70/hr), and passenger count, n (in the range 1 to 3), to see how much influence these factors have on total cost as shown in Figure 6.5 a-f. If the value of time is fixed, changing the number of passengers only affects the taxi fare per person because the transit fare per person is always fixed. The slope in Figure 6.5 for each mode is the travel time, and the intercept is the fare per person according to Eq 6.1. Intersections of taxi cost and transit cost are found for all OD pairs except trips from Penn to JFK (Figure 6.5 a). The intersection indicates a value of time when the cost of taxi and transit are the same for different numbers of passengers. This value of time at the intersection is a tipping point above which passengers are willing to pay extra fare for the faster mode. For example, the transit cost for JFK-Penn intersects with taxi cost at \$57/hr for n=2, indicating that the total cost of taxi is higher when value of time is less than \$57/hr, because the slope for transit exceeds the slope for taxi (Figure 6.5 b). This means that if two passengers are travelling as a group, it is better to choose transit if the value of time is lower than \$57/hr, otherwise it is more cost-effective to choose taxi.

On average, there is no way that a trip from Penn Station to JFK will be less costly by taxi in the assumed range of values of time and number of passengers (Figure 6.5 a). For the reverse direction, JFK-Penn, taxis do become competitive for sufficiently high values of time and passenger occupancies (Figure 6.5 b).

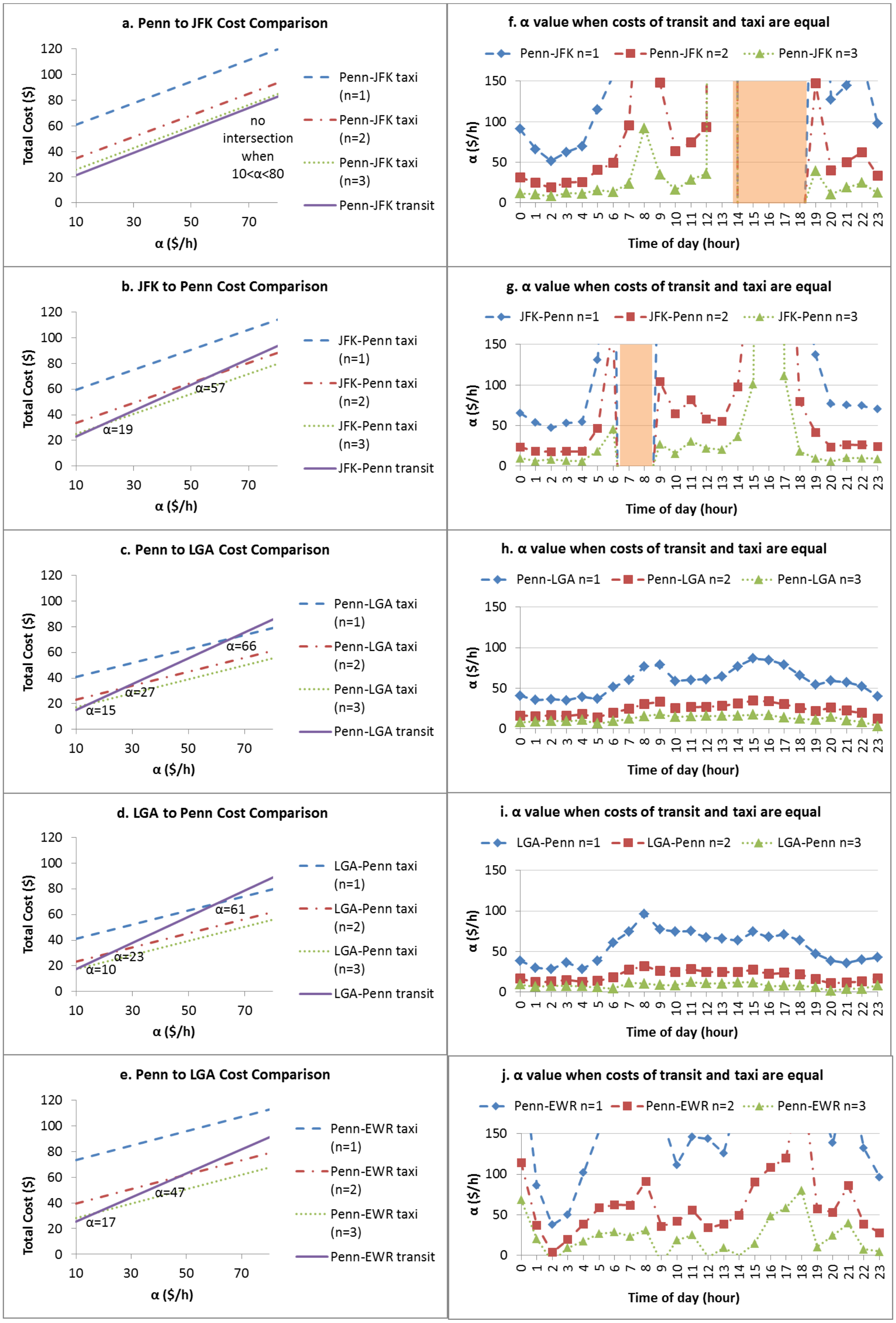


Figure 6.5 Sensitivity Analysis on Value of Time (α) and Passenger Count (n).

For trips to and from LGA (Figure 6.5 c-d), if traveling alone, the taxi costs are higher than transit costs when the value of time is less than $66/hr (Penn-LGA) or $61/hr (LGA-Penn), however, if traveling with more than two people, threshold is $27/hr. This relatively low value is reasonable because there is not direct transit service between Penn Station and LGA, i.e. taxi is more competitive with relative short distance.

The trip cost from Penn to EWR seems similar to JFK- Penn, except that the intersection points differ slightly. Transit costs more if the value of time exceeds $47/hr (*n*=2) or $17/hr (*n*=3). Considering $63/hr as the value of time for business trips in NYC and $42/h as the value of time for leisure trips (Gupta, 2008), it is likely that a business trip will use a taxi for JFK-Penn or Penn- EWR trips if traveling with more than 2 people, but a leisure trip will use taxi only if travelling with at least 3 people.

In order to account for the effect on the variation of travel time throughout a day, the threshold value of time within each hour at which passengers will switch their preferred mode is plotted (Figure 6.5 f-j). For most cases, travel times are longer by transit than by taxi (i.e., the slope for transit exceeds the slope for taxi), so values of time greater than the threshold are associated with more cost-effective taxi service, and values of time less than the threshold are associated with more cost-effective transit service. For a couple of time periods, transit is actually faster than taxi because traffic congestion has such a severe effect on taxi travel times, and the interpretation switches, so in the shaded areas of Figure 5 f and 5 g, all trips are more cost effectively served by transit, regardless of the value of time. During these times, transit is faster and cheaper than taxi.

The relatively low tipping point values for LGA compared to EWR and JFK show that taxi is more competitive than transit for that airport, appealing to a wider range of values of time. There is also a pattern at all airports that taxi is more competitive in the early hours of the morning (around 2 a.m.) when transit service is also less frequent. These results have policy implications, because they show how airports differ in the competitiveness of ground access modes, and how this changes by time of day.

Results for EWR (Figure 6.5 j) suggest that transit is more competitive from Penn Station to EWR, but for midnight trips taxis have a lower total cost than transit since the frequency of service is lower, which results in longer waiting times. However, because of the relatively long distance between Penn and EWR, it is possible that taxi is more likely to be chosen based on factors like convenience and comfort, which are not considered in this study.

6.3 Summary of Findings

This chapter aims to illustrate the advantages and limitations of the taxi data by showing that the taxi data is able to estimate travel time and travel distance for certain popular O-D pairs. There are two main limitations of the data: first, it is not sufficient to build a network model; and second, there is no survey data on travel preference of passengers available to provide behavior information. Extra Google transit data is required to work with the big taxi data to produce some useful insights.

This case study presents a methodology to compare the total cost for two modes of transportation (transit and taxi) using taxi GPS data and high-resolution transit schedule information. Trips between NYC Penn Station and three New York area airports (JFK, LGA and EWR) at different times of day are used to illustrate the methods. As shown in the analysis of total mode cost, transit is found to be more cost-effective than taxi for most times of the day if passengers are traveling alone and value time at

$40/hour, except during some midnight periods when transit service has long headways that contributes a significant amount of time to waiting or transfers.

The sensitivity analysis suggests that people are more likely to choose taxi to travel from Penn Station for airport trips if: 1) they have high value of time, 2) they are traveling with a large group of people, or 3) they are traveling in late night hours. It is also found that if people are traveling for business trips taxis become a less costly choice for airport access. For both JFK and EWR, because of the long distance, the taxi fare is very high, making transit a more competitive mode most of the time (especially when n=1) even though taxis have an advantage in travel time. However, LGA airport is closer to Penn Station, and the relatively low taxi fare and low travel time make taxi a more competitive choice for that OD pair.

The results show that the total cost or travel time for taxis always has a morning peak (between 6 a.m. to 10 a.m.) and an afternoon peak (between 1 p.m. to 6 p.m.). The taxi data provides an indication of traffic conditions in NYC (Yazici, 2012), so the use of this data to calculate the travel cost could incorporate both the temporal and spatial effects of traffic congestion in the city. However, this study is limited to the temporal analysis of the 5 most popular OD pairs for the airports, which all include trips to and from Penn Station. This can create bias if used to estimate total costs for the entire city. Future applications could be expanded to consider the spatial dimension as well by include multiple OD pairs distributed all over the city.

There are other factors that affect the choice of mode for trips to and from the airport, such as convenience and comfort, which are not considered in this study because they cannot be easily measured and quantified. With additional data about the number of passengers using each mode by time of day, it may be possible to gain some insights into the effect of these less tangible factors by comparing the expected mode shares from the utility functions in this dissertation with the observed mode shares.

In sum, there are potential data sources that can be used for calibrating and validating our model:

1. Traceable GPS taxi data can provide more data points for greater spatial and temporal coverage for the estimation of taxi travel time from point A to point B, which helps expand 5 OD pairs in our study to almost any other origin and destination pair at any time of day in NYC. As soon as this travel time network model is built, it can be used to estimate not only travel time for the mode cost comparison, but also the traffic congestion in the network.
2. A passenger preference survey is also useful to understand why passengers choose one travel mode over another. This will help to measure and estimate variables that are difficult to quantify, such as convenience and comfort. Also, a survey can provide demographic information about those who take taxis and those who take transit.
3. Transit ridership data is needed to estimate the equivalent utility of a dollar, β, which is important to estimate mode share. The ridership information could come from data such as train tickets and MetroCards, although some technology changes would be required to match the locations that passengers enter the subway with the locations that they exit.
4. Data for an individual transit (subway + bus) traveler, similar to GPS traceable data for taxis, might be available in the future if traveling in transit

becomes completely digital (e.g., using one metro card for all transit services). Then, this data would be very helpful to estimate β as well as travel time in real time.

This mode cost study used as an example of a practical method to estimate the travel cost including both time and money. As information and resources like travel time and fare are increasingly accessible, it is possible to design a smartphone app or a small computer program at the transit ticket vending machine to estimate total cost using this methodology. This information along with the choice model can be used to understand the factors that affect the aggregate mode choice decisions of the public. This will be useful for transportation planners and policy makers to improve the quality of travel options available to people traveling to and from airports.

CHAPTER 7

CONCLUSIONS AND FUTURE WORK

This dissertation has sought to address three primary issues on NYC taxi demand and supply modeling using large-scale taxi GPS data in NYC. These are: 1) the locational distribution of taxis and the imbalance of taxi demand and supply; 2) the temporal distribution of taxis and the imbalance of demand and supply; and 3) the directional distribution of empty taxis in search of their next customers. These issues have been successfully addressed using various modeling approaches as follows: the customer search model addresses where and how efficiently empty cab drivers pick up their next customer; the geo-spatial mapping and trip generation model show how taxi demand is generated in space and time, and the factors associated with taxi demand; and the mode cost comparison analyzes the difference in costs between taxis and other modes of transportation.

A large volume of relevant data has been collected and organized to perform a variety of analyses with New York City taxi GPS data. The study uses three major types of data. The first is a complete collection of GPS taxi data for every taxi trip made in NYC within a two-year period. The data is aggregated into refined spatial and temporal dimensions representing the taxi data into a detailed number of pickups and drop-offs at each census tract for each hour. Second, Google Transit Feed Specification (GTFS) and the Google Directions API are used to represent a detailed transit schedule for the same geographic region. Transit data is also aggregated for each census tract and each hour. Finally, these transportation data are supplemented with demographic, employment, and land use data to include key characteristics of the locations that are related to the highest rates of taxi use.

The following three paragraphs present some of the key findings from this dissertation:

First, taxi supply is quantified to estimate trip generation models. There exists a weekly oscillation of taxi trips, shifts, and operating hours. Holidays and heavy snowfall are important factors that affect the daily number of taxi trips and shifts. A customer search model identifies that vacant taxi drivers would most commonly travel to areas like Manhattan and its nearby neighborhoods for their next customers. Taxi drivers prefer the two major airports (LGA and JFK), where they are assured to find customers. Two scenarios are considered to optimize the efficiency of taxi drivers who made a drop off in the Outer Boroughs using a one-week dataset. Scenario 1 minimizes the total empty distance traveled for all taxis (at system level) in NYC, and Scenario 2 maximizes the revenue of each taxi driver (individual level). It was discovered that Scenario 2 had a smaller distance travelled than that of Scenario 1. This implies that there can be a more efficient taxi system if the taxi drivers are better-informed. However, Scenario 2 would not work in reality, because it is unlikely that all taxi drivers can circulate in the same areas and actually maximize their own profit.

Second, the Poisson-Gamma-Conditional Autoregressive (CAR) model with Markov Chain Monte Carlo (MCMC) algorithm was selected as the appropriate modeling technique for the taxi trips. The time of day modeling of trip generation identifies six important independent variables that influence taxi trips: population, education, taxi supply, income, TAT, and employment. All variables indicate positive correlations with taxi demand except TAT, because locations with lower access time tend to attract more taxis. Mixed residential and commercial land use (LU04), and the number of jobs at accommodation and food services (JobFod) are also identified to be important factors in the 12 AM, 7 AM, and 5 PM models, when controlling for population between ages 14 and 35. The overall predictability of CAR model is not very good, possibly because the spatial aggregation of the taxi data is too small. The residual analysis indicates that Midtown Manhattan is overserved, while most residential areas in the Outer Boroughs are not generating much taxi demand. The Lower East Side of Manhattan, Harlem, and Williamsburg are underserved areas. A higher percentage of population and jobs is underserved in Manhattan compared to New York City (all 5 boroughs). The underserved areas indicate a potential market for crowdsourcing taxi services like Uber and Lyft.

Third, the mode cost analysis is an example of a practical methodology for estimating the travel cost based on time and money, using Google Directions API and taxi trips between New York City Penn Station and three New York area airports (JFK, LGA and EWR) at different times of day. The cost comparison between taxi and transit suggests that transit is more cost-effective than taxi for most times of the day if passengers are traveling alone and the value of time is $40/hour. However, taxi is more cost-effective during some midnight periods when transit service has long headways; this contributes a significant amount of time to wait or transfer.

The major contribution of this study is to fill the gaps of insufficient empirical evidence in taxi planning and regulation. First, this study uniquely utilizes a large-scale GPS taxi dataset to model the taxi demand in New York City. Second, this study models taxi trip generation, trip distribution, and mode cost at refined spatial and temporal scales. The models are aggregated to each census tract according to time of day, which has not been investigated in any previous studies. The third contribution of the study is the implementation and integration of novel technologies (e.g., GPS data, Google Map, Google transit feed data, GIS) with various statistical and economic methodologies, which provide a systematic method for analyzing and forecasting taxi demand in NYC.

Big data has served the purpose of successfully building a taxi customer search model, trip generation, and mode cost comparison. However, there are some limitations of the data and the models that are built using this data:

1. This taxi data is capable of estimating travel time and travel distance for certain popular O-D pairs, but not sufficient to build a network model to estimate travel time for any O-D pairs, because the limited data points provided by only the start and end point of each trip.
2. We have identified areas underserved and overserved by the taxi system from the trip generation model; however, we don’t know whether the taxis are operating efficiently in the overserved areas.
3. No data from taxi crowdsourcing companies such as Uber, Lyft, or Sidecar are available to compare their demand to yellow taxi demand, which would give

us information on whether it is worthwhile to distribute the crowdsourcing services to the underserved areas.

4. The mode choice model is limited to mode cost model because of the lack of transit ridership data.
5. The mode cost comparison has to rely on a number of assumptions because we don't have data to estimate both taxi and transit passenger demographics, their value of time, their origins and destinations, and how they choose one mode over another.

The major purpose of the study is to understand how effectively empty taxis redistribute themselves; to identify factors related to taxi demand and supply; and to understand the imbalance between demand and supply based on data-driven models. Some improvements can be made to expand this study to future work:

- Online optimization model can be used to estimate the performance improved by minimizing the total travel distance or by maximizing the revenue of taxis. This model is different from off-line optimization models in that the online optimization model has the ability to reflect the locations of demand in real time, and it is updated by adding new information to the model.
- Studying whether there are any other factors that are important in generating taxi trips in New York City. For example, certain explanatory variables can be considered to account for the activities or popularity of a census tract or to account for the movements of empty taxicabs when locating potential customers.

- Expand the development of CAR models for other hours and using spatial aggregation larger than census tract. Future work can focus on completing the spatial autocorrelation model for all hours of the day or examining the CAR model using different aggregations for different areas, or both.
- Utilizing a Bayesian perspective rather than using conventional frequentist method to identify factors related to taxi demand and supply. This opens other forecasting opportunities, such as estimating the posterior travel demand using the prior travel demand information.
- Analysis can be performed to improve the understanding of the efficiency of taxi cabs in an under or overserved area. For example, we can calculate the empty travel time for the whole database for each trip to estimate the occupancy rate of all taxi cabs.
- If possible, collect taxi trip data from crowdsourcing companies like Uber and Lyft to estimate a separate trip generation model. This could be useful to compare crowdsourcing services with the yellow taxis because their competition has existed since Uber entered NYC taxi market in 2011. There are more Uber cars than yellow cabs now in NYC. The demand model of Uber cars will help to understand what factors are related to generating Uber demand and when and where Uber cars are over or underserved. This information can help policy makers figure out whether Uber can be regulated to serve the areas that are underserved by yellow cabs, and limited to serve areas that are overserved by yellow taxis.

- Another challenge related to app-based for-hire services such as Uber is that they are competing with NYC's black car livery services as well. The regular for-hire industry is likely to feel unrelenting competition from crowdsourcing app services. Therefore, future research can also focus on comparing the black livery services and Uber, if data is available.
- Instead of solely using the airports and NYC Penn Station, an application of the mode cost analysis methodology can include more areas (O-D pairs) in the New York City area.
- Introduce new transit data sources such as ridership data or traceable transit data to get more information on the transit travel demand, which will help get a real time estimation of β to compare mode share between taxi and transit.
- A revealed preference survey for both taxi and transit can be performed to discover the demographics of the passengers of each mode, the value of time, and the passengers' preferred transportation mode. This is also useful in determining when and where taxis are supplementing or competing against public transit.
- Investigating the value of time as a distribution of different taxi riders and trip purposes, such as airport employees and the regular air passengers.
- Performing additional analysis in two steps following trip generation in travel demand modeling: trip distribution and trip assignment.

REFERENCES

Abelson, P. (2010). The high Cost of Taxi Regulation, with Special Reference to Sydney. Australian National University. Retrieved from http://press.anu.edu.au/apps/bookworm/view/Agenda,+Volume+17,+Number+2,+2010/6691/abelson.xhtml

Akar, G. (2013). Ground Access to Airports, Case Study: Port Colombus International Airport. In *Journal of Airport Management*, Vol.30, pp. 25-31. doi:10.1016/j.jairtraman.2013.04.002

Alhussein, S.N. (2011). Analysis of Ground Access Modes Choice King Khaled International Airport, Riyadh, Saudi Arabia. In *Journal of Transport Geography,* Vol. 19, pp. 1361-1367. doi:10.1016/j.jtrangeo.2011.07.007

An, S., Hu, X., & Wang, J. (2011). Urban Taxis and Air Pollution: A Case Study in Harbin, China. In *Journal of Transport Geography*, Vol. 10, 2011, pp. 960-967. doi:10.1016/j.jtrangeo.2010.12.005

Austin, D., and Zegras, C. (2012) Taxicabs as public transportation in Boston, Massachusetts. In *Transportation Research Record*, No. 2277, pp. 65-74.

Bacache-Beauvallet, M., & Janin, L. (2012). Taxicab License Value and Market Regulation. In *Transport Policy,* Vol. 19, pp. 57-62. doi:10.1016/j.tranpol.2011.08.001

Bai, R., Li, J., Atkin, J. A. D., & Kendall, G. (2013). A Novel Approach to Independent Taxi Scheduling Problem Based on Stable Matching. In *Journal of the Operational Research Society,* Vol. 65, pp. 1501-1510. doi:10.1057/jors.2013.96

Baik, H., Trani, A. A., Hinze, N., Swingle, H., Ashiabor, S., & Seshadri, A. (2008). Forecasting Model for Air Taxi, Commercial Airline, and Automobile Demand in the United States. In *Transportation Research Record*, Vol. 2052, pp. 9-20.

Balan, R. K., Khoa, N. X., & Jiang, L. (2011). Real-time Trip Information Service for a Large Taxi Fleet. Proceedings from MobiSys '11: The 9th international conference on Mobile systems, applications, and services, pp. 99-102. doi:10.1145/1999995.2000006

Baptista, P., Ribau, J., Bravo, J., Dilva, C., Adcock, P., & Kells, A. (2011). Fuel Cell Hybrid Taxi Life Cycle Analysis. In *Energy policy*. Vol., 39, pp. 4683-4691. doi:10.1016/j.enpol.2011.06.064

Barlett, A., & Yilmaz, Y. (2011). *Taxicab Medallions- A Review of Experiences in Other Cities*. Government of the District of Columbia, Office of The Chief Financial Officer, Office of Revenue Analysis Briefing Note.

Barrett, S. D. (2003). Regulatory Capture, Property Rights and Taxi Deregulation: A Case Study. In *Institute of Economic Affairs*, Vol. 23, pp. 34-40. doi: 10.1111/j.1468-0270.2003.00441.x

Beesley, M.E. (1973). Regulation of Taxis. In *The Economic Journal*. Vol. 83(329), pp. 150-172.

Ben-Edigbe J., & Rahman, R. (2010). Multivariate School Travel Demand Regression Based on Trip Attraction. In *World Academy of Science, Engineering and Technology*, Vol. 42, pp. 1169-1173.

Bergantino, A. S., Villemeur, E. B., & Longobardi, E. (2007). *The Taxi Market: Failures and Regulation*. Retrieved from http://www.academia.edu/2332547/The_taxi_market_failures_and_regulation

Biggar, D. (2011). *Why and How Should We Regulate Taxis? Prepared for the Victorian Taxi Inquiry Roundtable 2*. Retrieved from http://www.taxiindustryinquiry.vic.gov.au/__data/assets/pdf_file/0007/57733/Darryl-Biggar-roundtable-paper.pdf

Bloomberg, M. R., Yassky, D. (2014). *2014 Taxicab Fact Book*. City of New York. Retrieved from http://www.nyc.gov/html/tlc/downloads/pdf/2014_taxicab_fact_book.pdf

Browning, R., Baker, E., Herron, J., & Kram, R. (2006). Effects of Obesity and Sex on the Energetic Cost and Preferred Speed of Walking. In *Journal of Applied Physiology*, Vol. 100, pp. 390-398. doi: 10.1152/japplphysiol.00767.2005

Cairns, R. D., & Liston-Heyes, C. (1996). Competition and Regulation in Taxi Industry. In *Journal of Public Economics*, Vol. 59, pp.1-15. doi:10.1016/0047-2727(94)01495-7

Cetin, T., & Eryigit, K.Y. (2011). Estimating the Effect of Entry Regulation in the Istanbul Taxicab market. In *Transportation Research Part A*, Vol. 45, pp. 476-484. doi:10.1016/j.tra.2011.03.002

Cetin, T., & Eryigit, K.Y. (2013). The Economic Effects of Government Regulation: Evidence from New York. In *Transport Policy*, Vol. 25, pp.169-177. doi:10.1016/j.tranpol.2012.11.011

Chadwick, E. (1859). Results of Different Principles of Legislation and Administration in Europe: of Competition for Field, as Compared with Competition within the Field, of Service. In *Journal of the Statistical Society of London*, Vol. 22, pp. 381-420.

Chang, Y. C. (2013). Factors Affecting Airport Access Mode Choice for elderly air passengers. In *Transportation Research Part E*, Vol. 57, pp. 105-112. doi:10.1016/j.tre.2013.01.010

Chatman, D. G. (2013). Does TOD Need the T? In *Journal of the American Planning Association*, Vol. 79, pp. 17-31. doi:10.1080/01944363.2013.791008

Chen, X, Ender, P. B., Mitchell, M., & Wells, C. (2003). *Stata Web Books Regression with Stata: Chapter 2 - Regression Diagnostics*. UCLA Institute for Digital Research and Education. Retrieved from http://www.ats.ucla.edu/stat/stata/webbooks/reg/chapter2/statareg2.htm Accessed on 11/1/2013.

City of New York (2015). *List of Current Medallions* [EXCEL File: current_medallions]. Retrieved from www.nyc.gov/html/tlc/downloads/excel/current_medallions.xls

City of New York (2015). *List of Current Street Hail Livery (SHL) Permit Holders* [EXCEL file: current_shl_permits]. Retrieved from www.nyc.gov/html/tlc/downloads/excel/current_shl_permits.xls

Coffman, R. B. (1977). The Economic Reasons for Price and Entry Regulation of Taxicabs (Comment and Rejoinder). In *Journal of Transport Economics and Policy*, Vol. 11(3), pp. 288-304.

Conway, A., Kamga, C., Yazici, A., & Singhal, A. (2012). Challenges in Managing Centralized Taxi Dispatching at High-Volume Airports: Case Study of John F. Kennedy International Airport, New York City. In *Transportation Research Record: Journal of Transportation Research Board*, No. 2300, pp.89-90.

Corpuz, G. (2007). Public Transport or Private Vehicle: Factors that Impact on Mode Choice. In *Sydney, N.S.W. Transport Data Centre,* 30th Australasian Transportation Research Forum.

Cross, H. New York Taxis – Getting around New York City in a Taxi, Retrieved from http://gonyc.about.com/a/taxi.htm. Accessed on February 19, 2014

Daganzo, C.F. (1978). An Approximate Analytic Model of Many-to-many Demand Responsive Transportation Systems. In *Transportation Research*, Vol 12, pp. 325-333.

Darr, A., & Lewin, A. C. (2008). The Implementation of Democratic Justice Regimes in the Israeli Taxi Sector: Econimic Imperative or Ethnic Origin. In *Journal of Socio-Economics*, Vol. 37, pp. 2072-2079.

Davies, A. (2008). Exhaustive Regression an Exploration of Regression-Based Data Mining Techniques Using Super Computation. In *Research Program on Forecasting*, The George Washington University, RPF Working Paper No. 2008-008. Retrieved from http://www.gwu.edu/~forcpgm/2008-008.pdf Accessed on 7/27/2013.

De Neufville, R. (2006). Planning Airport Access in an Era of Low-Cost Airlines. In *Journal of the American Planning Association*, Vol.72, pp. 347-356.

Dempsey, P. S. (1996). Taxi Industry Regulation, Deregulation & Reregulation: the Paradox of Market Failure. In *Transportation Law Journal*, Vol. 24(73), pp. 73-120.

Douglas, G. W. (1972). Price Regulation and Optimal Service Standards The Taxicab Industry. In *Journal of Transport Economics and Policy*, Vol. 6(2), pp. 116-127.

Du, X., Fu, L., Ge, W., Zhang, S., & Wang, H. (2011). Exposure of Taxi Drivers and Office Workers to Traffic-related Pollutants in Beijing: A note. In *Transportation Research Part D: Transport and Environment*, Vol. 16, pp. 78-81. doi: 10.1016/j.trd.2010.08.002

Eckert, R. D. (1973). On the Incentives of Regulators: The case of taxicabs. In *Public Choice*, Vol. 14(1), pp. 83-99.

Elliott, G., Gargano, A., & Timmermann, A. (2012). Complete Subset Regressions. In *Journal of Econometrics*, Vol. 177, pp. 357-373. doi:10.1016/j.jeconom.2013.04.017

Ewing, R., Schroeer, W., & Greene, W. (2004). School Location and Student Travel Analysis of Factors Affecting Mode Choice. In *Transportation Reserach Record*, Vol. 1895, pp. 55-63.

Fernández L., J. E., De Cea Ch, J., & Briones M., J. (2006). A Diagrammatic Analysis of the Market for Cruising Taxis. In *Transportation Research Part E: Logistics and Transportation Review*, Vol. 42, pp. 498-526. doi:10.1016/j.tre.2005.05.001

Flath, D. (2006). Taxicab Regulation in Japan. In *Journal of Japanese International Economics*, Vol. 20, pp. 288-304.

Fleiter, J. J., Gao, L., Qiu, C., & Shi, K. (2009). Availability, functionality, and use of seat belts in Beijing taxis prior to the 2008 Beijing Olympic Games. In *Accident Analysis and Prevention*, Vol. 41, pp. 342-344. doi: 10.1016/j.aap.2008.12.007

Flores-Guri, D. (2003). An Economic Analysis of Regulated Taxicab. In *Review of Industrial Organization*, Vol. 23, pp. 255-266.

Fox, J. (2008). *Applied Regression Analysis and Generalized Linear Models*. SAGE Publications, Inc, 2nd edition.

Frankena, M. W., & Pautler, P. A. (1984). An Economic Analysis of Taxicab Regulation, Bureau of Economics Staff Report. Federal Trade Commission (FTC). Retrieved from http://www.ftc.gov/sites/default/files/documents/reports/economic-analysis-taxicab-regulation/233832.pdf

Funderburg, R. G., Nixon, H., Boarnet, M. G., & Ferguson, G. (2010). New Highways and Land Use Change: Results From a Quasi-Experimental Design. In *Transportation Research Part A*, Vol. 44, pp. 76-98. doi:10.1016/j.tra.2009.11.003

Gao, H. O., & Kitirattragarn, V. (2008). Taxi Owners' Buying Preferences of Hybrid-Electric Vehicles and Their Implications for Emissions in New York City. In *Transportation Research Part A*, Vol. 42, pp. 1064-1073. doi:10.1016/j.tra.2008.03.002

Garber, N. J, & Hoel, L. A. (2014). *Traffic and Highway Engineering* (5th edition). Cengage Learning, January.

Gaunt, C. (1995). The impact of taxi deregulation on small urban areas: some New Zealand evidence Clive. In *Transport Policy*, Vol. 2(4), pp. 257-262.

Gaunt, C., & Black, T. (1996). The Economic Cost of Taxicab Regulation: The Case of Brisbane. In *Economic Analysis & Policy*, Vol. 26(1), pp. 45-58.

Gebeyehu, M., & Takano, S. (2007). Diagnostic Evaluation of Public Transportation Mode Choice in Addis Ababa. In *Journal of Public Transportaion*, Vol. 10(4), pp. 27-50.

Gebeyehu, M., & Takano, S. (2008). Modeling the Relationship between Traveler's Level of Satisfaction and Their Mode Choice Behavior Using Ordinal Models, In Journal of Trasnportation Research Forum, Vol. 47, pp. 103-118.

Gelhausen, M. C. (2011). Modeling the Effects of Capacity Constraints on Air travellers' Airport Choice. In *Journal of Air Transport Management*, Vol. 17, pp. 116-119. doi:10.1016/j.jairtraman.2010.11.004

Gerrard, M. J. (1974). Comparison of Taxi and Dial-a-Bus Services. In *Transportation Science*, Vol. 8(2), pp. 85-101.

Girardin, F., & Blat, J. (2010). The Co-evolution of Taxi Drivers and Their In-car Navigation Systems. In *Pervasive and Mobile Computing*, Vol. 6, 424-434. doi:10.1016/j.pmcj.2010.03.002

Glaister, S. (1987). Regulation through Output Related Profits Tax. In *Journal of Industrial Economics*, Vol. 35(3), pp. 281-296.

Gosling, D. G. (2006). Predictive Reliability of Airport Ground Access Mode Choice Models. In *Transportation Research Record: Journal of Transportation Research Board*, No. 1951, pp. 69-75.

Gould, J. (2013). *Transport Survey Methods: Best Practice for Decision Making*, Chapter 3: Cell Phone Enabled Travel Surveys, The Medium Moves the Message, pp. 51-70, Emerald Group Publishing.

Guo, H., Zou, S. C., Tsal, W.Y., & Blake, D.R. (2011). Emission Characteristics of Nonmethane Hydrocarbons From Private Cars and Taxis at Different Driving Speeds in Hong Kong. In *Atmospheric Environment*, Vol. 45, pp. 2711-2721. doi:10.1016/j.atmosenv.2011.02.053

Gupta, S., Vovsha, P., & Donnelly, R. (2008). Air Passenger Preferences for Choice of Airport and Ground Access Mode in the New York City Metropolitan Region. In *Transportation Research Record: Journal of Transportation Research Board*, No.2042, pp.3-11.

Häckner, J., & Nyberg, S. (1995). Deregulating Taxi Services: A Word of Caution. In *Journal of Transport Economics and Policy*, Vol. 29(2), pp. 195-207.

Harshbarger, R. (2015). 'Uber's New York Business is Illegal': Yellow Cabs File Lawsuit. NYPost. Retrieved from http://nypost.com/2015/03/31/ubers-new-york-business-is-illegal-yellow-cabs-file-lawsuit

Harvey, G. (1986). Study of Airport Access Mode Choice. In *Journal of Transportation Engineering*, Vol. 112(5), pp. 525-545.

Hess, S., & Polak, J. W. (2005). Mixed Logit Modelling of Airport Choice in Multi-airport Regions. In *Journal of Air Transport Management*, Vol. 11, pp. 59-68. doi:10.1016/j.jairtraman.2004.09.001

Hess, S., & Polak, J. W. (2006). Exploring the Potential for Cross-nesting Structures in Airport-choice Analysis: A Case-study of the Greater London Area. In *Transportation Research Part E*, Vol. 42, pp. 63-81. doi:10.1016/j.tre.2005.09.001

Hess, S., Adler, T., & Polak, J. W. (2007). Modeling Airport and Airline Choice Behaviour with the Use of Stated Preference Survey Data. In *Transportation Research Part E*, Vol. 43, pp. 221-233. doi:10.1016/j.tre.2006.10.002

Hu, X., Gao, S., Chiu, Y. C., Lin, & D. Y. (2012). Modeling Routing Behavior for Vacant Taxi Cabs in Urban Traffic Networks. *Proceedings of the 91st Annual meeting of the Transportation Research Board, National Research Council*, Washington, D.C., United States.

Humphreys, I., & Ison, S. (2005). Changing Airport Employee Travel Behavior: The Role of Airport Surface Access Strategies. In *Transport Policy*, Vol. 12, pp. 1-9. doi:10.1016/j.tranpol.2004.07.002

Hung, W.T. (2009). Taxation on Vehicle Fuels its Impacts on Switching to Cleaner Fuels. In *Energy Policy*, Vol. 34, pp. 2566-2571. doi:10.1016/j.enpol.2004.08.018

Ishii, J., Jun, S., & Dender, K. V. (2009). Air travel choices in multi-airport markets. In *Journal of Urban Economics*, Vol. 65, pp. 216-227. doi:10.1016/j.jue.2008.12.001

Jiang, L., & Mahmassani, H. S. (2013). Toll Pricing: Computational Tests on How to Capture Heterogeneity of User Preferences. In *Transportation Research Record*, Vol. 2343, pp. 105–115

Jo, W. K., & Yu, C. H. (2001). Public Bus and Taxicab Drivers' Work-Time Exposure to Aromatic Volatile Organic Compounds. In *Environmental Research Section A*, Vol. 86, pp. 66-72. doi:10.1006/enrs.2001.4257

Kamga, C., Yazici, M. A., & Singhai, A. (2013). Hailing in the Rain Temporal and Weather-Related Variations in Taxi Ridership and Taxi Demand-Supply Equilibrium. In *Transportation Research Board's 92nd Annual Meeting*.

Kang, C. H. (1998). *Taxi Deregulation: International Comparison* (MSc. Dissertation). The University of Leeds. Retrieved from http://www.taxi-library.org/kang0898.htm

Kim, H., Oh, J. S., & Jayakrishnan, R. (2005). Effect of Taxi Information System on Efficiency and Quality of Taxi Services. *Transportation Research Record*, Vol. 1903, pp. 96–104.

Kim, Y. J., & Hwang, H. (2008). Incremental Discount for Taxi Fare with Price-sensitive Demand, In *International Journal Production Economics*, Vol. 112, pp. 859-902.

Koehler, B. (2004). *Regulating Supply in Taxi Markets* (MSc Dissertation). City Univerisity, London.

Koster, P., Kroes, E., & Verhoef, E. (2011). Travel Time Variability and Airport Accessibility. In *Transportation Research Part B*, Vol.45, pp. 1545-1559. doi:10.1016/j.trb.2011.05.027

Kumar, A., & Levinson, D. (1992). Specifying, Estimating, and Validating a New Trip Generation Model: A Case Study of Montgomery County, Maryland. In *Transportation Research Record*, Vol. 1413, pp. 107-113.

Lam, L. T. (2004). Environmental Factors Associated with Crash-related Mortality and Injury among Taxi Drivers in New South Wales, Australia. In *Accident Analysis and Prevention*, Vol. 36, pp. 905-908. doi:10.1016/j.aap.2003.10.001

Lau, J., Hung, W. T., & Cheung, C. S. (2011). On-board Gaseous Emissions of LPG taxis and Estimation of Taxi Fleet Emissions. In *Science of the Total Environment*, Vol. 409, pp. 5292-5300. doi: 10.1016/j.scitotenv.2011.08.054

Lawler, R. (2013). Crowdsourcing An Alternate Name For ‘Ride Sharing’. Retrieved from http://techcrunch.com/2013/05/12/crowdsourcing-an-alternate-name-for-ride-sharing

Levine, N. (2010). CrimeStat Help Document, Chapter 14: Trip distribution. University of Michigan. Retrieved from https://www.icpsr.umich.edu/CrimeStat/files/CrimeStatChapter.14.pdf Accessed 1/30/2014

Levine, N. (2010). *CrimeStat*: A Spatial Statistics Program for the Analysis of Crime Incident Locations (v 3.3). Ned Levine & Associates, Houston, TX, and the National Institute of Justice, Washington, DC. July.

Levine, N., Lord, D., & Park, B. J. (2010). *CrimeStat Version 3.3 Update Notes: Part 2: Regression Modeling*. Retrieved from: https://www.icpsr.umich.edu/CrimeStat/files/CrimeStat3.3updatenotesPartII.pdf

Li, S. (2006). *Multi-Attribute Taxi Logistics Optimization* (M.S. Thesis). Massachusetts Institute of Technology.

Lin, Y., Li, W., Qiu, F., & Xu, H. (2012). Research on Optimization of Vehicle Routing Problem for Ride-sharing Taxi. In *Procedia- Social and Behavioral Science*, Vol. 43, pp. 494-502. doi:10.1016/j.sbspro.2012.04.122

Liu, B., Chen, C., Zhang, B., Bu, M., Bi, J., & Yu, Y. (2012). Fuel Use Pattern and Determinants of Taxi Drivers' Fuel Choice in Nanjing, China. In *Journal of Cleaner Production*, Vol. 33, pp. 60-66. doi:10.1016/j.jclepro.2012.05.016

Loos, N. (2006). *Value Creation in Leveraged Buyouts: Analysis of Factors Driving Private Equity Investment Performance*. German University Publishers (DUV).

Luken, B., L., & Garrow, L. (2011). Multiairport Choice Models for the New York Metropolitan Area: Application Based on Ticketing Data. In *Transportation Research Record: Journal of Transportation Research Board*, No. 2206, pp. 24-31.

Maa, P. (2005). *Taxicabs in US Cities and How Governments Act* (MMSS Senior Thesis). Northwestern University.

Mallows, C. L. (1975). Some Comments on C_P. In *Technometrics,* Vol. 15, pp. 661–675.

Mandel, B., Gaudry, M., & Rothengatter, W. (1997). A disaggregate Box-Cox Logit Mode Choice Model of iIntercity Passenger Travel in Germany and its Implications for High-speed rail Demand Forecasts. In *Annals of Regional Science*, Vol. 31, pp. 99-120.

Manini, p., Palma, G. D., Andreolo, R., Poli, D., Mozzoni, P., Folesani, G., Mutti, A., & Apostoli, P. (2006). Environmental and Biological Monitoring of Benzene Exposure in a

Cohort of Italian Taxi Drivers. In *Toxicology Letters*, Vol. 167, pp. 142-151. doi:10.1016/j.toxlet.2006.08.016

Mankiw, N. G. (1998). *Principles of Microeconomics*. Elsevier, Volume 1.

Marell, A, & Wsstin, K. (2002). The Effects of Taxicab Deregulation in Rural Areas of Sweden. In *Journal of Transport Geography,* Vol. 10, pp. 135-144. doi:10.1016/S0966-6923(02)00006-6

McLeod, A., & Xu, C. *R Help bestglm: Best Subset GLM*. Retrieved from http://cran.r-project.org/web/packages/bestglm/vignettes/bestglm.pdf Assessed on 7/26/2013.

Moore, A. T., & Balaker, T. (2006). Do Economists Reach a Conclusion on Taxi Deregulation? In *Economic Journal Watch*, Vol. 3, pp. 109-132.

Morgul, E. F., & Ozbay, K. (2011). Simulation-Based Evaluation of a Feedback Based Dynamic Congestion Pricing Strategy on Alternate Facilities. In *Transportation Research Board 90th Annual Meeting*, Washington, D.C., USA.

Morgul, E. F., Ozbay, K., & Iyer, S. (2013). Holguin-Veras, J. Commercial Vehicle Travel Time Estimation in Urban Networks Using GPS Data from Multiple Sources. In *Transportation Research Board 92nd Annual Meeting*, Washington, D.C., USA.

Mourato, S., Saynor, B., & Hart, D. (2004). Greening London's Black Cabs: A Study of Driver's Preferences for Fuel Cell Taxis. In *Energy Policy*, Vol. 32, pp. 685-695. doi:10.1016/S0301-4215(02)00335-X

Mousavi, A., Bunker, J., & Lee, B. 92012). A New Approach for Trip Generation Estimation for Traffic Impact Assessments. In *25th ARRB Conference – Shaping the future: Linking policy research and outcomes*, Perth, Australia.

NYC Department of City Planning, *PLUTO data*. Retrieved from http://www.nyc.gov/html/dcp/html/bytes/applbyte.shtml

NYC TLC (2015), *Current Licensee.*, New York City Taxi & Limousine Commission (TLC). Retrieved from http://www.nyc.gov/html/tlc/html/industry/current_licensees.shtml

NYC TLC (2015). *Understanding the For-Hire Vehicle Industry*, New York City Taxi & Limousine Commission (TLC). Retrieved from http://www.nyc.gov/html/tlc/downloads/pdf/fhv_base_fact_sheet.pdf

O’Neill, W. A., & Brown, E. (2001). Long-Distance Trip Generation Modeling Using ATS. In *Transportation Research E-circular Number E-C026. Journal of Transportation Research Board*. Washington D.C.

Ohazulike, A. E., Still, G., Kern, W., & Berkum, E. C. V. (2013). An Origin-destination Based Road Pricing Model for Static and Multi-period Traffic Assignment Problems. In *Transportation Research Part E*, Vol. 58, pp. 1-27. doi:10.1016/j.tre.2013.06.003

Orr, D. (1969). The Taxicab Problem: A Proposed Solution. In *Journal of Political Economy*, Vol. 77 (1), pp. 141-147.

Otsuka, K., & Murakami, N. (1989). Incentives and enforcement under contract The Taxicab in Kyoto. In *Journal of The Janpanese and International Economics,* Vol. 3, pp. 231-249. doi:10.1016/0889-1583(89)90020-8

PANYNJ (2011). *2010 Airport Traffic Report*. Port Authority of New York & New Jersey (PANYNJ). Retrieved from http://www.panynj.gov/airports/pdf-traffic/ATR2010.pdf Accessed on 7/22/2013

PANYNJ (2015). *Airports*. The Port Authority of New York & New Jersey (PANYNJ). Retrieved from http://www.panynj.gov/ Accessed on 7/10/2013

Papacostas, C. S., & Prevedouros, P. D. (2000). *Transportation Engineering and Planning* (3rd Edition). Prentice Hall, June.

Paradis, E. (2014). *Moran's Autocorrelation Coefficient in Comparative Methods,* R help Manual. Retrived from: http://cran.r-project.org/web/packages/ape/vignettes/MoranI.pdf

Pels, E., Nijkamp, P., & Rietveld, P. (2003). Access to and Competition Between Airports: A Case Study for The San Francisco Bay Area. In *Transportation Research Part A*, Vol. 37, pp. 71-83. doi:10.1016/S0965-8564(02)00007-1

Psasaki, V., & Abacoumkin, C. (2002). Access Mode Choice for Relocated Airports: The New Athens International Airport. In *Journal of Air Transportation Management*, Vol. 8, 2pp. 89-98. doi:10.1016/S0969-6997(01)00033-3

Pyrcz, M. J., & Deutsch, C. V. (2014). *Geostatistical Reservoir Modeling* (Illustrated Edition). Oxford University Press.

Qian, X., Zhang, X., & Ukkusuri, S. V. (2014). Characterizing Urban Dynamics Using Large Scale Taxicab Data, In *Transportation Research Board's 93rd Annual Meeting*, Paper No. 14-5301.

Racca, D., & Ratledge, E. C. (2004). *Project Report for Factors That Affect and/or Can Alter Mode Choice*. Prepared for Delaware Transportation Institute and The State of Delaware Department of Transportation. Retrieved from http://udspace.udel.edu/bitstream/handle/19716/1101/transitmodel.pdf?sequence=1

Rometsch, S., & Wolfstetter, E. (1993). The Taxicab Market An Elementary Model. In *Journal of Institutional and Theoretical Economics (JITE),* Vol. 149(3), pp. 531-546.

Routley, V., Ozanne-Smith, J., Qin, Y., & Wu, M. (2009). Taxi driver seat belt wearing in Nanjing, China. In *Journal of Safety Research*, Vol. 40, pp. 449-454. doi:10.1016/j.jsr.2009.10.004

Schaller, B (2007). Entry Controls in Taxi Regulation: Implications of US and Canadian Experience for Taxi Regulation and Deregulation. In *Transport Policy*, Vol. 14, pp. 490-506. doi:10.1016/j.tranpol.2007.04.010

Schaller, B. (1999). Elasticities for Taxicab Fares and Service Availability. In *Transportation*, Vol. 26, pp. 283-297.

Schaller, B. (2005). A Regression Model of the Number of Taxicabs in U.S. Cities. In *Journal of Public Transportation*, Vol. 8, pp. 63-78.

Schaller, B. (2006). *The New York City Taxicab Fact Book*. Retrieved from http://www.schallerconsult.com/taxi/taxifb.pdf Accessed on 11/1/2013.

Schmöcker, J. D., Quddus, M. A., Noland, R. B., & Bell, M. G. H. (2005). Estimating Trip Generation of Elderly and Disabled people. In *Transportation Research Record*, Vol. 1924, pp. 9-18.

Schroeter, J. R. (1983). A Model of Taxi Service under Fare Structure and Fleet Size Regulation. In *Journal of Economics*, Vol. 14(1), pp. 81-96.

Schwanen, T., & Mokhtarian, P. L. (2005). What Affects Commute Mode Choice: Neighborhood Physical Structure or Preferences toward Neighborhoods? In *Journal of Transport Geography*, Vol 13, pp. 83-99. doi:10.1016/j.jtrangeo.2004.11.001

Seibert, C. (2006). Taxi Deregulation and Transaction Costs. In *Institute of Economic Affairs*, Vol. 26, pp.71-73.

Seymour, D. (2009). *Taxi Deregulation*. Report for FCPP Policy Series No. 55.

Shreiber, C. (1975). The Economic Reasons for Price and Entry Regulation of Taxicabs. In *Journal of Transport Economics and Policy*, Vol. 9(3), pp. 268-279.

Shreiber, C. (1977). the economic reasons for price and entry regulation of taxicabs a rejoinder. In *Journal of Transport Economics and Policy*, Vol. 11(3), pp. 81-81.

Shreiber, C. (1981). The Economic Reasons for Price and Entry Regulation of Taxicabs: A Rejoinder. In *Journal of Transport Economics and Policy*, Vol. 15(1), pp. 81-83.

Shriner, H. W., & Hoel, L. A. (1999). Evaluating Improvements in Landside Access for Airports. In *Transportation Research Record: Journal of Transportation Research Board*, No. 1662, pp. 32-40.

Tam, M., Lam, W. H. K., & Lo, H. (2011). The Impact of Travel Time Reliability and Ground Service Quality on Airport Ground Access Mode Choice. In *Journal of Choice Modelling*, Vol. 4, pp. 49-69. doi:10.1016/S1755-5345(13)70057-5

Teal, R. F., & Berglund, M. (1987). The impacts of taxicab regulation in USA. In *Journal of Transport Economics and Policy*, Vol. 21(1), pp. 37-56.

Toner, J. P. (1992). *Regulation in the Taxi Industry, Institute of Transport Studies*. University of Leeds, Working Paper 381.

Toner, J. P. (2010). The Welfare Effects of Taxicab Regulation in English Towns. In *Economic Analysis & Policy*, Vol. 40(3), pp. 299-312. doi:10.1016/S0313-5926(10)50031-6

Train, K. (2009). *Discrete Choice Methods with Simulation* (2nd edition). Cambridge University Press.

Turvey, R. (1961). Some Economic Features of the London Cab Trade. In *The Economic Journal*, Vol. 71 (281), pp. 79-92.

U.S. Census Bureau (2012). *2007-2011 American Community Survey (ACS) 5-Year Estimates DP03: Selected Economic Characteristics*. Washington, DC: U.S. Census Bureau. Available from http://factfinder2.census.gov Accessed on 11/5/2013

U.S. Census Bureau. *Demographic Data*. Retrieved from U.S. Census Bureau Center. for Economic Studies (CES) website http://www.census.gov/ces/dataproducts/demographicdata.html

U.S. Census Bureau. *2009 & 2010 Workplace Area Characteristic (WAC)*. Available on-line at: http://lehd.ces.census.gov/data/

U.S. Office of Personnel Management. *2010 Federal Holidays*. Retrieved from: http://archive.opm.gov/operating_status_schedules/fedhol/2010.asp

UK Department of Transport (2013).*Values of Time and Vehicle Operating Costs TAG Unit 3.5.6*. UK Department of Transport, Transport Analysis Guidance (TAG), February. Retrieved from http://www.dft.gov.uk/webtag/documents/expert/unit3.5.6.php Accessed on 10/15/2013

Waheed, S., & Romero-Alston, L. (2008). *Taxi Drivers and the Cost of Moving the City*. A report by the Community Development Project of the Urban Justice Center.

Wang X. H., & Hofe, R. V. (2007). *Research Methods in Urban and Regional Planning*. Springer Berlin Heidelberg. doi:10.1007/978-3-540-49658-8

Wilbur Smith Associates. *Demographics & Travel Demand Model* (Chapter 3). Report from City of Laredo. Retrieved from http://www.ci.laredo.tx.us/city-planning/Departments/MPO/files/mtp/chapter_3.pdf Accessed on 1/30/2014.

Williams, D. J. (1980). The Economic Reasons for Price and Entry Regulation of Taxicabs: A Comment. In *Journal of Transport Economics and Policy*, Vol. 14(1), pp. 105-112.

Wong, K. I., Wong, S. C., & Yang, H. (2001). Modeling Urban Taxi Services in Congested Road Networks with Elastic Demand. *Transportation Research Part* B, Vol. 35, pp. 819-842. doi:10.1016/S0191-2615(00)00021-7

Wong, K. I., Wong, S. C., Bell, M. G. H., & Yang, H. (2005). Modeling the Bilateral Micro-searching Behavior for Urban Taxi Services using the Absorbing Markov Chain Approach. In *Journal of Advanced Transportation*, Vol 39, pp. 81-104. doi: 10.1002/atr.5670390107

Wong, K. I., Wong, S. C., Yang, H., & Tong, C. O. (2002). A Sensitivity-Based Solution Algorithm for the Network Model of Urban Taxi Services. In: *Taylor, M.A.P. (ed.),* Proceedings of the 15th International Symposium on Transportation and Traffic Theory. Elsevier Science, pp. 23-42.

Wong, K. I., Wong, S. C., Yang, H., & Wu, J. H. (2008). Modeling Urban Taxi Services with Multiple User Classes and Vehicle Modes. *Transportation Research Part B*, Vol. 42, pp. 985-1007. doi:10.1016/j.trb.2008.03.004

Wong, R. C. P., Szeto, W. Y., & Wong, S. C. (2014). A Cell-based Logit-opportunity Taxi Customer-search Model. *Transportation Research Part C*, Vol. 48, pp. 84-96. doi:10.1016/j.trc.2014.08.010

Yan, X., & Su, X. G. (2009). *Linear Regression Analysis: Theory and Computing* (1st edition. World Scientific Publishing Company.

Yang, H., & Wong, S. C. (1998). A Network Model of Urban Taxi Services. In *Transportation Research Part B*, Vol. 32, pp. 235-246. doi:10.1016/S0191-2615(97)00042-8

Yang, H., & Yang, T. (2011). Equilibrium Properties of Taxi Markets with Search Frictions. In *Transportation Research Part B*, Vol. 45, pp. 696-713. doi:10.1016/j.trb.2011.01.002

Yang, H., Fung, C. S., Wong, K. I., & Wong, S. C. (2010). Nonlinear Pricing of Taxi Services. In *Transportation Research Part A*, Vol. 44, pp. 337-348. doi:10.1016/j.tra.2010.03.004

Yang, H., Wong, K. I., & Wong, S. C. (2001). Modeling Urban Taxi Services in Road Networks: Progress, Problem and Prospect. *Journal of Advanced Transportation*, Vol. 35, pp. 237-258. doi: 10.1002/atr.5670350305

Yang, H., Ye, M., Tang, W. H. C., & Wong, S. C. (2005). A Multi-period Dynamic Model of Taxi Services with Endogenous Service Intensity. In *Operations Research*, Vol. 53, pp. 501-515. doi: 10.1287/opre.1040.0181

Yang, H., Ye, M., Tang, W. H., & Wong, S. C. (2005). Regulating Taxi Services in the Presence of Congestion Externality. In *Transportation Research Part A*, Vol. 39, pp. 17-40. doi:10.1016/j.tra.2004.05.004

Yazici, M. A., Kamga, C., & Mouskos, K. (2012). Analysis of Travel Time Reliability in New York City Based on Day-of-Week Time-of-Day Periods. In *Transportation Research Record*, Vol. 2308, pp. 83-95. doi: 10.3141/2308-09

Yazici, M. A., Kamga, C., & Mouskos, K. C. (2012). Analysis of Travel Time Reliability in New York City Based on Day-of-Week and Time-of-Day. In *Transportation Research Board's 91st Annual Meeting*.

Yazici, M. A., Kamga, C., & Singhai, A. (2013a). Weather's Impact on Travel Time and Travel Time Variability in New York City. In *Transportation Research Board's 92nd Annual Meeting*.

Yazici, M. A., Kamga, C., & Singhai, A. (2013b). A Big Data Driven Model for Taxi Drivers' Airport Pick-up Decisions in New York City. *IEEE International Conference on Big Data*.

Zhan, X., Hasan, S., Ukkusuri, S. V., & Kamga, C. (2013). Urban Link Travel Time Estimation Using Large-scale Taxi Data with Partial Information. In Transportation Research Part C, Vol. 33, pp. 37-49. doi:10.1016/j.trc.2013.04.001

Zhang, D., & Song, L. (2009). *Travel Mode Choice Behavior of Taxi Passengers*. Traffic Research Center, Beijing University of Technology.

Zhang, G., Wang, Y., Wei, H., & Yi, P. (2008). A Feedback-Based Dynamic Tolling Algorithm for High-Occupancy Toll Lane Operations. In *Transportation Research Board*, No. 2065, pp. 54-63.

Zito, R., D'Este, G., & Taylor, M. A. P. (1995). Global Positioning Systems in the Time Domain: How Useful a Tool for Intelligent Vehicle-highway System. In *Transportation Research Part C*, Vol. 3, pp. 193-209.

APPENDIX A

ACRONYMS

AIC – Akaike Information Criterion

BIC – Bayesian Information Criterion

CAR – Spatial Autocorrelation (Poisson-Gamma-CAR Model)

CP – Central Park

EWR – Newark Liberty International Airport

GCS – Grand Central Station

GIS – Geographic Information System

GPS – Global Positioning System

GTFS – Google Transit Feed Specification

JFK – John F. Kennedy International Airport

HL – High Line

LGA – LaGuardia Airport

LIRR – Long Island Rail Road

LOS – Level of Service

MCMC – Markov Chain Monte Carlo

MSP - Madison Square Park

MTA – Metropolitan Transit Authority

NYC – New York City

NB – Negative Binomial model

NBMC – Non-spatial Negative Binomial Model with MCMC Algorithm

OD – Origin-Destination (in the context of origin to destination pair)

PA – Port Authority Bus Terminal

PS – Penn Station

RC – Rockefeller Center

RSS – Residual Sum of Squares

SE – Standard Error

TAT – Transit Access Time

TAZ – Transportation Analysis Zone

TLC – Taxi and Limousine Commission

TS – Time Square

VIF – Variance Inflation Factor

WAC – Workplace Area Characteristic