2020

# Advanced Econometic Models for Modeling Flows: Application to Shared Economy

Bibhas Dey
*University of Central Florida*

# ADVANCED ECONOMETRIC MODELS FOR MODELING FLOWS: APPLICATION TO SHARED ECONOMY

by

BIBHAS KUMAR DEY
B.Sc. Bangladesh University of Engineering and Technology, 2014
M.S. University of Central Florida, 2018

A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in the Department of Civil, Environmental and Construction Engineering
in the College of Engineering and Computer Science
at the University of Central Florida
Orlando, Florida

Spring Term
2020

Major Professor: Naveen Eluru

**ABSTRACT**

Travel and tourism industry is undergoing transformation with the flourishing of online sharing economy marketplaces such as Bike Share services, Uber/Lyft (for taxi services), Eatwith (for community restaurants), and AirBnB (for accommodation). The current research effort contributes to literature on sharing economy service flow analysis by formulating and estiamting econometric approaches for analyzing frequency variables. The sharing economy alternatives investigated include: (a) accommodation service (AirBnB), (b) bikeshare service (Citi bike, NYC) and (c) ride hailing service (UBER/LYFT/Taxi). In the first part of the dissertation, we develop a copula based negative binomial count model framework to count AirBnB listings at census tract level to capture the snapshot of accommodation supply for tourists in NYC. In the second part, considering bike sharing as one of the transportation sharing systems, the dissertation identifies two choice dimensions for capturing the bike share system demand: (1) station level demand and (2) how bike flows from an origin station are distributed across the network. In the third part of the dissertation on ride sharing systems, we identify two choice dimensions: a demand component that estimates origin level transportation newtwork company (TNC) demand at the taxi zone level and (2) a distribution component that analyzes how these trips from an origin are distributed across the region. A linear mixed model is considered to estimate station or taxi zone level demand while a multiple discrete continuous extreme value (MDCEV) model to analyze flows distribution is employed. In the final part of this dissertation, we develop an innovative joint econometric model system to examine two components of the rapid ride share market transformation: (a) the increase in ride hailing demand and (b) the shift from traditional taxi services to TNC services. The first component is analyzed adopting a negative binomial (NB) count model while the second component is analyzed using a multinomial fractional split (MNLFS) model.

## ACKNOWLEDGMENTS

I would like to convey my heartiest gratitude to my honorable supervisor Dr. Naveen Eluru for his excellent supervision and being a constant support in this dissertation.

I would also like to gratefully acknowledge Dr. Sabreena Anowar for providing her overall support on the papers.

## TABLE OF CONTENTS

## LIST OF FIGURES

## LIST OF TABLES

## CHAPTER 1: INTRODUCTION

### 1.1 Sharing Economy System

The sharing economy is an economic model often defined as a peer-to-peer (P2P) based activity of acquiring, providing or sharing access to goods and services that are facilitated by a community based on-line platform. Sharing has become a powerful force of market participation. By sharing access to extra bedrooms, back seats of cars, special camera or cooking equipment, and their own time and skills, urban dwellers have earned extra money and joined a community of like-minded sellers and consumers. In effect, websites and mobile phone applications have allowed such individuals to start up the tiniest of businesses to leverage the value of assets that would otherwise serve only their own personal uses.

The basic of sharing economy comes from concept of peer to peer (P2P) system. A peer-to-peer (P2P) economy is a decentralized model whereby two individuals interact to buy or sell goods and services directly with each other, without an intermediary third-party, or without the use of a company of business.

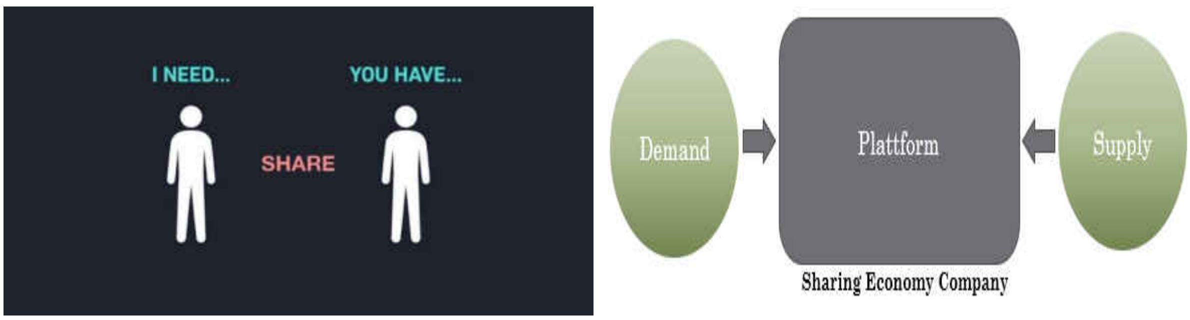



**Figure 1.1: Fundamental Concept of Sharing Economy System**

Sharing economy system consists of three parties while one party (seeker) requests some service that asset can be shared by other party (owner) and the third party make the deal possible via an online platform for service fee (platform) (Botsman & Rogers, 2011). The whole cycle of sharing economy illustrated in Figure 1.2.

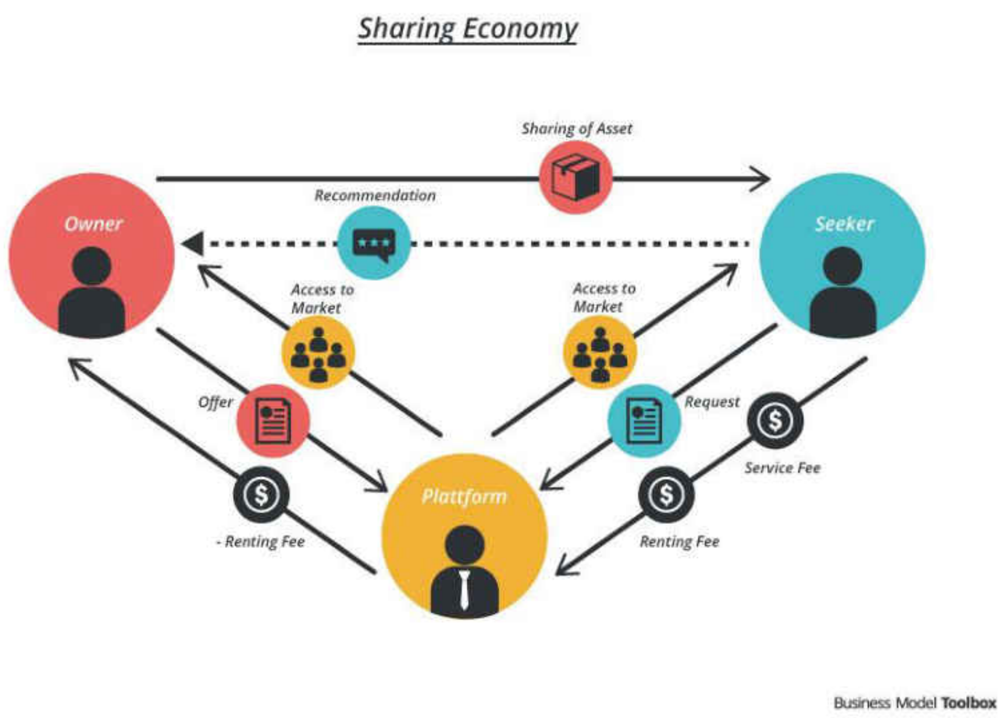



**Figure 1.2: Working Process of Sharing Economy System**

The concept and practice of a “sharing economy” and “collaborative consumption” suggest making use of market intelligence to foster a more collaborative and sustainable society. Prominent examples are bike- and carsharing schemes as well as web-based peer-to-peer platforms covering a broad range of activities from renting rooms to food business. Online peer-to-peer (P2P) marketplaces are growing at a rapid rate, especially in travel and tourism services (Pizam, 2014). Early marketplaces of this kind, such as eBay and Craigslist, have been associated with the trade of traditional retail items (Sundararajan, 2014). Recently, a new type of P2P commerce, mainly associated with the supply of services and commonly known as the “sharing economy,” has emerged (Botsman & Rogers, 2011). Sharing economy marketplaces have flourished particularly within the field of travel and tourism, in which locals supply services to tourists. Travel and tourism industry is undergoing transformation with the flourishing of online sharing economy marketplaces such as Bike Share services, Uber (for taxi services), Eatwith (for community restaurants), and AirBnB (for accommodation). In this study, we selected accommodation service (AirBnB), bikeshare service (Citi bike, NYC) and rideshare service (UBER/LYFT/Taxi).

#### 1.1.1 Accommodation Services (AirBnB)

The shared housing market place AirBnB with its large inventory and wide reach across the globe is redefining the hospitality sector. AirBnB is unique in its design as it does not own any properties but provides a platform for ordinary people (sellers) to rent their residences (entire house/apartment or a room) to tourists (consumers) (Botsman & Rogers, 2011). AirBnB accommodation system is quite easy to use: a consumer searches for an entire home or private (or shared) room based on their travel dates, destination on the AirBnB website (www.AirBnB.com). The user is provided with a list of housing alternatives based on the user preferences. The success and wide adoption of the system is based on available review information and background check procedures for renters and tourists. AirBnB charges a service fee for each transaction. Initiated in 2008, popularity of this sharing hospitality platform has rapidly grown with over 200 million guests having stayed in about 3 million listings in more than 65,000 cities and 191 countries (AirBnB, 2017). In fact, since 2016, over 100 million people have enjoyed the accommodation through AirBnB while over 1 million new listings worldwide have been added to the market place.

### 1.1.2 Transportation Field

#### 1.1.2.1 Bikeshare

Transportation field is undergoing a transformative change in response to several technological innovations in recent years. A product of these technological transformations is the adoption of shared mobility systems such as bikesharing (such as CitiBike in New York City), car sharing (such as Zipcar or Car2Go), ridesourcing (such as Uber and Lyft) and ride-splitting (such as dynamic carpooling in urban regions). As highlighted in a recent Transit Cooperative Research Program report (Feigon & Murphy, 2016), understanding shared mobility adoption and usage provides an unprecedented opportunity to address existing mobility shortcomings in urban regions. In fact, public transit agencies and transportation planning agencies can enhance mobility and accessibility by incorporating these shared mobility alternatives within their planning frameworks. Among the shared mobility alternatives, bike sharing offers a sustainable transportation alternative in urban core regions and could be an effective solution to the last mile problem (Jäppinen, Toivonen, & Salonen, 2013).

About 1000 cities around the world have a bikeshare system in operation or in consideration for development (Meddin & DeMaio, 2016). As reported by Richter, 2018 (Anowar, Eluru, & Hatzopoulou, 2017), the number of public use bicycles in the world have nearly quadrupled between 2013 and 2016. Further, a recent national association of city transportation officials (NACTO) report highlighted that of the 88 million trips made by bike share users in US between 2010-2016, 28 million were trips from 2016 only (Dey, Anowar, Eluru, & Hatzopoulou, 2018b).

#### 1.1.2.2 Transportation Network Company (TNC)

Ride hailing services have been available as a mode of transportation since the early 17$^{th}$ century in the form of horse-drawn hackney carriages in Europe. With the advent of the automobile, taxis for hire have been the most common ride hailing transportation alternative. However, ride hailing has undergone a rapid transformation in the recent few years in response to the transformative technological changes including smart mobile availability, ease of hailing a ride using mobile applications, integration of seamless payment systems and real-time driver and user reviews. In fact, the convenience offered by transport networking companies (TNC) such as Uber, Lyft, and Via has allowed for a tremendous growth in ride hailing demand. For example, in New York City, the average daily trips by taxi (yellow taxi) was varying between 400 thousand and 500 thousand for the years 2010 and 2014 (Metcalfe & Warburg, 2012). However, since 2014, with the advent TNC services in the city, the total number of trips have increased. Specifically in 2018, the daily trips have increased to more than a million trips with traditional taxi accounting for nearly 300 thousand trips, and TNC services accounting for 700 thousand trips. These trends are not specific to New York City. A recent report analyzing reimbursed travel in the US has found that the share of Uber and Lyft has increased from 8% to 72.5% within 2014-2018 at the cost of taxi and rental car business share (Silver & Fischer-Baum, 2016). The prevalence of TNC services is also not restricted to US. Uber operates in over 60 countries, while Didi Express in China, Ola in India currently capture a large share of the ride hailing market in these countries. The immense growth in market share and the spread of these services across the world illustrate how the ride hailing market has undergone a rapid transformation in a short time frame.

## 1.2 Empirical Motivation

### 1.2.1 <u>Accommodation Services (AirBnB)</u>

The growth of AirBnB impacts transportation and urban systems along two major directions. *First*, AirBnB provides a unique snapshot of the hospitality industry and can serve as a surrogate for the health of tourism industry in the region. The number of available listings on

AirBnB can serve as a proxy for tourist interest in the region. AirBnB provides renters with an opportunity to immediately respond to tourist demand by allowing for a simple listing process (without any substantial capital costs). In the event of a drop in tourist demand, renters on the website remove their listing. On the other hand, traditional hospitality industry with hotels respond to tourist demand slowly due to the large capital costs involved in increasing capacity. In addition, the traditional hospitality sector cannot dismantle their infrastructure as easily in response to the reduced tourist demand. Thus, with its ease of adding a listing, the AirBnB listings provide a unique snapshot of the health of tourism industry. *Second*, an analysis of AirBnB listings will allow transportation and urban regional professionals examine the demand arising from these tourists on transportation and urban infrastructure. Cities such as New York that receive significant expenditures from tourists can provide improved services by enhancing infrastructure in response to emerging tourist locations.

The first part of the research effort is focused on meeting these three dimensions. *First*, by developing a model framework to count AirBnB listings at census tract level to capture the snapshot of accommodation supply for tourist in NYC. *Second*, capture the unobserved heterogeneity in the model together with correlation between those matrices. *Finally*, based on the estimation results, a policy analysis is also conducted to illustrate how listings count is influenced by various exogenous attributes.

### 1.2.2 Transportation Field

#### 1.2.2.1 Bikeshare Destination Flows

As bike sharing is an emerging transportation mode, the current approaches being employed for analyzing system usage and performance measure are still in their infancy. In the 2nd task of our research, we focus our attention on developing a research framework to contribute to our understanding of bikeshare origin destination flows. In this study, we propose an enhanced framework to estimate usage dimensions of bike sharing at a system level.

To be sure, several earlier research efforts have explored approaches to model system level usage (Faghih-Imani & Eluru, 2015; Faghih-Imani, Eluru, El-Geneidy, Rabbat, & Haq, 2014; Rixey, 2013; Zhao, Deng, & Song, 2014). These research studies examine the impact of bicycling infrastructure, land use and built environment, public transportation infrastructure, temporal and meteorological attributes on bikeshare system usage (defined as station level arrivals and departures). These models can be viewed as analogous to the trip generation and trip attraction models in the traditional trip based modeling approach. While these models provide important insights on variables affecting bikeshare usage, they do not provide any information on the system level flows between the stations. To elaborate, the approaches provide trip end information without the trip distribution relationship. To address this shortcoming, recent research has developed destination choice models at an individual trip level (El-Assi, Mahmoud, & Habib, 2017; Faghih-Imani & Eluru, 2015, 2017b). In these studies, for every individual trip the choice of destination given the origin station is analyzed using a random utility based approach. The models developed at an individual trip level can be employed to obtain aggregate estimates of trip distribution (analogous to the gravity model). However, such an aggregation approach is purely a statistical construct and lacks behavioral support.

In this second task, we remedy this drawback, by developing a model framework for bikeshare system usage as well as origin destination flows. Towards this end, we characterize system demand as origin level demand (number of trips) and allocate these trips to various destination stations (number of trips from an origin to destination) in the system. For the first variable, a linear mixed model is developed while the second variable is analyzed using a multiple discrete continuous model system that implicitly recognizes that the total arrivals across stations should add up to the total number of trips leaving the origin.

1.2.2.2 TNC Destination Flows

The rapid transformation of the ride hailing market coupled with emerging shared mobility service expansions (such as Carshare, Bikeshare, and Scooter share) offers an unprecedented opportunity to address the existing mobility shortcomings in urban regions (as highlighted in a recent TCRP report (Feigon & Murphy, 2016). In fact, public transit and transportation planning agencies can enhance mobility and accessibility in a region by incorporating these shared transportation alternatives within their planning frameworks to provide holistic mobility options in denser urban regions. Specifically, dense urban regions with well-connected public transit systems can strategically target reducing the reliance on private automobile ownership (and use) by incorporating ride-hailing alternatives in trip planning tools. Further, by examining the spatio-temporal ride hailing data, transit agencies and shared mobility platforms can identify urban pockets with service needs to provide last mile connectivity. Towards understanding these patterns it would be beneficial to understand TNC demand and its spatial distribution in the region.

The current research effort (3rd task), contributes to this goal by developing quantitative models of TNC demand and flow distribution patterns. The study develops (1) a demand component that estimates origin level TNC demand at the taxi zone level and (2) a distribution component that analyzes how these trips from an origin are distributed across the region. The former component is analyzed using linear mixed models and the latter component is analyzed using a multiple discrete continuous model system. The model components are developed using a comprehensive set of independent variables including aggregate trip attributes, transportation infrastructure variables, land use and built environment variables, weather attributes, and temporal attributes. The model estimates are validated using a hold out sample. Further, a policy exercise is conducted to illustrate how the proposed model system can be utilized for evaluating the impact of changes to independent variables.

#### 1.2.2.3 TNC Transformation

The TNC service induced transformation can be viewed as constituting two major components. The first component is the overall increase in ride-hailing demand possibly drawing from population of individuals driving, using public transit and even inducing newer travel. The second component of the transformation is the shift in the share of traditional taxi service demand toward TNC services (Gerte, Konduri, Ravishanker, Mondal, & Eluru, 2019). In a short time frame, in NYC, TNC services have increased their market share from 0 to nearly 70% by the end of 2018. While preliminary research has begun to explore the reasons for the transformation, it is safe to assume economists and social scientists will continue to examine the transformation for several years into the future.

The proposed study contributes to our understanding of this transformation by examining the NYC data from a fine spatial and temporal resolution by adopting an innovative joint econometric model system. The study examines two components of the transformation (a) the increase in ride hailing demand and (b) the shift from traditional taxi services to TNC services. The first component – taxi zone ride hailing demand - is analyzed adopting a negative binomial count model. The second component - share of traditional and TNC services demand - is analyzed using a multinomial fractional split model. The two model components are stitched together in a joint framework that allows for the influence of repeated observations as well as for the presence of common unobserved factors affecting the two components. The study employs trip level data from the NYC Taxi and Limousine Commission from January 2015 through December 2018 for the analysis. The data is aggregated by taxi zone for every month in the study period and analyzed by ride hailing alternatives: yellow taxi, green taxi and TNC services (including Uber, Lyft, Juno and Via).

## 1.3 Methodological Perspectives

### 1.3.1 Count Approach

While observed variables can be included in the univariate models, the consideration of the influence of unobserved factors requires a panel multivariate or joint modeling approach. Earlier research efforts on modeling count variables have developed simulation oriented multivariate models that stitches together the various dimensions within a maximum simulated or Bayesian approach (see (Yasmin & Eluru, 2018)) for an extensive literature review). Alternatively, bivariate copula framework that treats the variable dimensions as a joint distribution have also been developed (see (Nashad, Yasmin, Eluru, Lee, & Abdel-Aty, 2016)). The first approach allows for accommodating unobserved attributes affecting the joint distribution as well the individual count components. The copula approach only allows for the influence of unobserved factors on the joint distribution within a closed form framework.

In our proposed research (1st task), we build on these two model structures to accommodate for repeated measures by developing a unified framework that accommodates for dependency within a joint copula framework while also allowing for random parameters within each count model. To the best of the authors' knowledge, this is the first attempt to employ such a unified framework for examining count events.

### 1.3.2 Approach for Destination Flows

Station level demand is a continuous variable and can be easily analyzed using linear regression models and their advanced variants. On the other hand, the second choice variable is quite different. Specifically, for an origin station with a predefined demand, the choice process involves identifying the flows to all destination stations in the system. There are two major challenges associated with it. First, the destinations for bike flows from an origin are likely to involve multiple alternatives (as opposed to a single chosen alternative). Second, the potential universal alternative set includes all stations in the bikeshare system. The multiple discrete continuous approaches that follow Kuhn-Tucker (KT) approaches developed in literature can be adapted to address this choice dimension. KT demand systems have been used in outdoor recreational demand studies (Phaneuf, Kling, & Herriges, 2000; von Haefen, 2004; von Haefen & Phaneuf, 2005), individual activity participation and time-use studies (Bhat, 2005; Nurul Habib & Miller, 2009; Pinjari & Bhat, 2010; Pinjari, Bhat, & Hensher, 2009; Rajagopalan, Pinjari, & Bhat, 2009), household vehicle ownership and usage forecasting (Ahn, Jeong, & Kim, 2008; Bhat, Sen, & Eluru, 2009; Fang, 2008) and household travel expenditure analyses (Ferdous, Pinjari, Bhat, & Pendyala, 2010; Rajagopalan & Srinivasan, 2008). Of these approaches, for our current choice context, Bhat (Bhat, 2008) offers a flexible alternative that can be adapted to our choice dimension.

The second task of the analysis focused on examination of bikeshare demand patterns and distribution patterns on a weekly basis while 3rd task focused on TNC distribution for daily peak hour. The processed data provides station or zonal level origin demand and the corresponding flow patterns from the origin to all destinations across the system. The second choice dimension has huge number of destination alternatives in our analysis. To the best of the authors' knowledge this is the largest number of alternatives considered in a KT system in literature.

#### 1.3.3 Approach for Demand Transformation

In the final task, the share of traditional and TNC services demand - is analyzed using a multinomial fractional split model. As the data for the two components is obtained for the same spatial record, there are several common unobserved factors influencing the two variables. The database generated also has multiple data points for each spatial unit. Thus, a joint econometric model that accommodates for repeated measures (panel) and common unobserved factors across the two dependent variables is developed. Specifically, we build on the cross-sectional joint negative binomial and multinomial fractional split model developed in Bhowmik et al. (Bhowmik, Yasmin, & Eluru, 2018) for a different empirical context.

### 1.4 Objectives of the Dissertation

The <u>first objective</u> is focused on examination of the evolution of AirBnB listings at a census tract level by listing type – entire home or private/shared room. The dependent variable is defined as the number of listings in the census tract by listing type. Given that each census tract has two dependent variables with multiple repeated observations for each CT, observed and unobserved factors affect these variables. While observed variables can be included in the univariate models, the consideration of the influence of unobserved factors requires a panel multivariate or joint modeling approach. Earlier research efforts on modeling count variables have developed simulation oriented multivariate models that stitches together the various dimensions within a maximum simulated or Bayesian approach (see (Yasmin & Eluru, 2018)) for an extensive literature review). Alternatively, bivariate copula framework that treats the variable dimensions as a joint distribution have also been developed (see (Nashad et al., 2016)). The first approach allows for accommodating unobserved attributes affecting the joint distribution as well the individual count components. The copula approach only allows for the influence of unobserved factors on the joint distribution within a closed form framework. In our proposed research, we build on these two model structures to accommodate for repeated measures by developing a unified framework that accommodates for dependency within a joint copula framework while also allowing for random parameters within each count model. To the best of the authors' knowledge, this is the first attempt to employ such a unified framework for examining count events.

The second objective of our research is to contribute to the research on bikeshare systems by examining system level demand and its distribution. To elaborate, our emphasis is on understanding bikeshare demand at the stations and the flow of these bikes to their corresponding destinations. The framework should provide system operators an estimate of system demand at a station level and how these bikes are distributed across the bikeshare system. We identify two choice dimensions: (1) station level demand and (2) how bike flows from an origin station are distributed across the network. For our analysis, we examine demand patterns and distribution patterns on a weekly basis. The processed data provides station level weekly origin demand and the corresponding flow patterns from the origin to all destinations across the system. The second choice dimension has 573 destination alternatives in our analysis. To the best of the authors' knowledge this is the largest number of alternatives considered in a KT system in literature. The model estimation results for the proposed model offers intuitive results. The proposed model was also validated using a hold-out sample and prediction exercise is undertaken.

The third objective of our dissertation is to develop TNC demand based planning models that can be integrated within existing frameworks or used to augment the outputs from existing demand frameworks. With this primary objective, the current study makes the following contributions. First, the current study develops a TNC demand model at the Taxi zone level for the morning peak hour (represented as pickups in the data). The demand variable is continuous in nature and a linear mixed model framework is employed to analyze the data. Second, conditional on the origin taxi zone demand, we develop a distribution model to determine TNC flows from the origin to all destinations in the study region. There are two major challenges associated with modeling the TNC flow distribution. First, the destinations for TNC flows from an origin are likely to involve multiple alternatives (as opposed to a single chosen alternative). Second, the potential universal alternative set includes all taxi zones in the system. The multiple discrete continuous approaches that follow Kuhn-Tucker (KT) approaches developed in literature can be adapted to address this choice dimension. In a recent study, Dey et al. (Dey, Anowar, & Eluru, 2019) developed a similar framework for studying bicycle sharing system flows. The data for our analysis from January 2018 through December 2018 is drawn from NYC Taxi & Limousine Commission (NYTLC). The data provides taxi zonal level daily origin demand and the corresponding flow patterns from the origin to all destinations across the system. The two model components were developed using a host of independent variables including trip attribute, transportation infrastructure variables, land use and built environment variables, weather attributes, and temporal attributes. The model estimation results for the proposed model offers intuitive results. The proposed model was also validated using a hold-out sample and prediction exercise is undertaken.

In the final objective of the dissertation, the study contributes to our understanding of the ongoing transformation of ride hailing market by examining the NYC data from a fine spatial and temporal resolution using an innovative joint econometric model. Specifically, as opposed to considering the transformation at a regional scale and in a 4 year period, we examine taxi zone based demand data from NYC for each month and explore the reasons contributing to (a) the increase in ride hailing demand and (b) the shift from traditional taxi services to TNC services. The first component – taxi zone ride hailing demand - is analyzed adopting a negative binomial count model. The second component - share of traditional and TNC services demand - is analyzed using a multinomial fractional split model. As the data for the two components is obtained for the same spatial record, there are several common unobserved factors influencing the two variables. The database generated also has multiple data points for each spatial unit. Thus, a joint econometric model that accommodates for repeated measures (panel) and common unobserved factors across the two dependent variables is developed. Specifically, we build on the cross-sectional joint negative binomial and multinomial fractional split model developed in Bhowmik et al. (Bhowmik, Yasmin, & Eluru, 2018) for a different empirical context.

### 1.5 Outline of the Dissertation

The remainder of the research proposal is divided into four chapters that shows how each chapter position the current research effort within the larger context of the literature. Within chapter three and four, a quick review of the current research effort along the with econometric framework adopted in the study are also discussed.

Chapter two provides a brief review of previous relevant researches and a detailed discussion on different approaches employed for demand modeling in sharing economy literature. The chapter is divided into two parts discussing the earlier studies regarding various scope for two sharing economy system as AirBnB and Bikeshare. Various dimension such as history, new scope, demand, pros and cons of those service systems are discussed in this chapter. Information on the study unit, methodological framework, estimation technique, dependent variables and the number of dimensions employed in these studies are discussed in a systematic format. Further, the limitation of the earlier frameworks used for analysis are also identified.

Chapter three contributes to objective one by comparing the performance of the simulation-based framework with closed form copula-based frameworks. For this study purpose, a copula based negative binomial count model system is developed so that implicitly recognizes the total AirBnB listings. Given these afore-mentioned implications, the proposed research conducts a comprehensive analysis of AirBnB listings in New York City region drawing on data from January 2015 to September 2017. We analysis the evolution of AirBnB listings at a census tract level by listing type – entire home or private/shared room. The dependent variable is defined as the number of listings in the census tract by listing type. Given that each census tract has two dependent variables with multiple repeated observations for each CT, observed and unobserved factors affect these variables. Within the copula framework, we estimate models for four copula structures: (1) FGM, (2) Frank, (3) Gumbel, (4) Clayton and (5) Joe. The model frameworks are compared based on statistical fit and a host of comparison metrics for estimation sample and hold-out sample. Finally, the applicability of the model for most tourism zone identification is illustrated by generating plots by AirBnB types in the NYC region.

Chapter four contributes to objective two by proposing a model framework that considered a large number of alternatives in a KT system in literature. The data for our analysis is drawn from New York City bikeshare system (CitiBike). Six months of bikeshare usage data from January 2017 through June 2017 was downloaded from CitiBike website and processed to obtain weekly bikeshare usage patterns. For our analysis, we examine demand patterns and distribution patterns on a weekly basis. The processed data provides station level weekly origin demand and the corresponding flow patterns from the origin to all destinations across the system. The second choice dimension has 573 destination alternatives in our analysis. The proposed model was also validated using a hold-out sample and prediction exercise is undertaken.

Chapter five contributes to this goal by developing quantitative models of TNC demand and flow distribution patterns. Using data from the NYC Taxi and Limousine commission, we conduct a comprehensive analysis of morning peak hour ride hailing data from Uber, Lyft, Juno and Via from 2018. The study develops (1) a demand component that estimates origin level TNC demand at the taxi zone level and (2) a distribution component that analyzes how these trips from an origin are distributed across the region. The former component is analyzed using linear mixed models and the latter component is analyzed using a multiple discrete continuous model system. The model components are developed using a comprehensive set of independent variables including aggregate trip attributes, transportation infrastructure variables, land use and built environment variables, weather attributes, and temporal attributes. The model estimates are validated using a hold out sample. Further, a policy exercise is conducted to illustrate how the proposed model system can be utilized for evaluating the impact of changes to independent variables.

Chapter six contributes to our understanding of this transformation by examining the NYC data from a fine spatial and temporal resolution by adopting an innovative joint econometric model system. The study examines two components of the transformation (a) the increase in ride hailing demand and (b) the shift from traditional taxi services to TNC services. The first component – taxi zone ride hailing demand - is analyzed adopting a negative binomial count model. The second component - share of traditional and TNC services demand - is analyzed using a multinomial fractional split model. The two model components are stitched together in a joint framework that allows for the influence of repeated observations as well as for the presence of common unobserved factors affecting the two components. The study employs trip level data from the NYC Taxi and Limousine Commission from January 2015 through December 2018 for the analysis. The data is aggregated by taxi zone for every month in the study period and analyzed by ride hailing alternatives: yellow taxi, green taxi and TNC services (including Uber, Lyft, Juno and Via).

Chapter seven concludes the dissertation by summarizing the findings, and identifies directions for future research.

# CHAPTER 2: LITERATURE REVIEW

In this chapter, we provide a review of relevant literature for the various shared market places examined in the dissertation. The chapter is organized to match the four research objectives described in Chapter 1 as follows.

1. Sharing accommodation literature:
2. Literature on Bikeshare Destination Flows: Earlier studies regarding various bikeshare demand and destination flows are summarized in this section.
3. Literature on TNC Destination Flows: Earlier studies regarding TNC demand and destination flows are summarized in this section.
4. Literature on Ride Hailing Transformation: Earlier studies regarding various ride hailing services demand are summarized in this section.

## 2.1 Earlier Research of Sharing Accommodation

Tourism is a burgeoning global industry contributing to economic activity. A major component of the economic activity is accounted by the hospitality industry with accommodations having a significant role. While it is not possible to review the entire spectrum of literature covering the accommodation industry, we focus our attention on the sharing accommodation literature encompassing accommodation websites such as AirBnB Vacation Rentals by Owners (VRBO) and HomeAway. Specifically, we review literature on sharing accommodation along three main streams: a) studies investigating evaluation of sharing accommodation systems , b) studies investigating the various qualitative characteristics of shared accommodation systems and c) studies exploring the quantitative aspects of shared accommodation systems and examining their relationship with traditional hotel system. Table 1 provides a summary of the reviewed studies along the three streams. The table provides information on the study area, data source, determinants examined and analysis methodology. Sharing economy listings analyzed in the literature span many urban cities of USA (such as New York, Los Angeles, San Francisco, Washington D.C., Boston, Dallas, Houston), Canada (various urban regions), Europe (such as Paris, London, Stockholm), Korea (such as Seoul, Busan, and Jeju) and India (various urban regions). The reader would note that a majority of the shared accommodation research examines AirBnB accommodation underlining the growing relevance of AirBnB in the shared accommodation industry.

The *first group of studies* focused on evaluation of shared accommodation systems such as AirBnB from different perspectives, including the theoretical and practical aspects of emergence of AirBnB as sharing economy system (see firs panel of Table 1). Multiple studies focused on the definition of shared accommodation systems, how these services have evolved over time, investigated the challenges and opportunities presented by real-time services and highlighted various opportunities for the future (Proserpio & Tellis, 2017, D. Guttentag, 2015, Zervas et al., 2015a, Oskam & Boswijk, 2016, Wang et al., 2018, Adamiak, 2018). Several studies analyzed future research scope of shared accommodation on tourism. These studies investigated shared economy's significant impact on tourism and found that policy making needs to be adaptive considering new aged sharing economy system (Edelman & Geradin, 2015, Juul 2015).

The *second group of studies* explored various qualitative characteristics and conducted quantitative analysis of shared accommodation systems. While qualitative studies typically rely on online reviews, photos, questionnaire surveys (mail, telephone, face-to-face, online, on-site) and data from field experiments quantitative studies used web script to download listings data for further analysis. According to Ert et al (2016), host's photo in AirBnB's website play an important role in increasing the probability of gaining guest's trust towards booking AirBnB. Several studies explored AirBnB service quality by conducting text analysis using online reviews. Based on these analysis, the authors evaluated how AirBnB experience contrasts with their home (Zhu, Cheng, Wang, Ma, & Jiang, 2019) and the trust issues experienced (Sthapit & Björk, 2019). The research regarding guest reviews also offer useful inputs for future guest's decisions to book AirBnB (Brochado, Troilo, & Aditya, 2017).

Another set of studies explored the influence of AirBnB on the neighborhood home rent/price increases, and income of middle class families (Sperling, 2015, Jiao & Bai, 2020). Sperling (2015) investigated the socio-economic conditions of a neighborhood after the emergence of AirBnB listings and concluded that income stagnation of middle-class family can potentially be overcome by hosting on AirBnB platform. (Jiao & Bai, 2020) explored how demographics, socioeconomics and transportation might affect AirBnB listings and found that neighborhoods with good transit service, short distances to the city center and household income has the positive association with AirBnB listings. (Jordan & Moore, 2018) investigated positive and negative impact of AirBnB in the economic, environmental, and sociocultural realms using thematic analysis of interview data of AirBnB, Vacation Rentals by Owner (VRBO), and HomeAway users. Several studies investigated the negative issues associated with shared accommodation systems (such as AirBnB) including racial discrimination and illegal listings (B. Edelman, Luca, & Svirsky, 2017, Fradkin, Grewal, & Holtz, 2018).

In recent literature, impact on AirBnB pricing owing to distinct neighborhood and listings characteristics is one of the often investigated dimensions (Deboosere, Kerrigan, Wachsmuth, & El-Geneidy, 2019; Gibbs, Guttentag, Gretzel, Morton, & Goodwill, 2018; Rodríguez-Pérez de Arenaza, Hierro, & Patiño, 2019; Tong & Gunter, 2020; Wyman, Mothorpe, & McLeod, 2020; Barron et al., 2018; Lee et al., 2016). AirBnB and VRBO listings price rate and revenue was investigated to illustrate the host's preference to replace long-term renters with short-term visitors to generate more revenue considering neighborhood characteristics such as transit accessibility to jobs, employment rate, population density, median income (Deboosere et al., 2019; Wyman et al., 2020; Rodríguez-Pérez de Arenaza et al., 2019; Barron et al., 2018; Lee et al., 2016) and listings characteristics overall rating, size, reviews, host attributes, site and property attributes, amenities and services, rental rules and distance from the CBD etc. (Tong & Gunter, 2020; Gibbs et al., 2018, Wang & Nicolau, 2017). In recent times, a group of studies examine the AirBnB supply impact on whole hospitality, tourism or leisure business considering the revenue and employment opportunity it brings with and found the positive correlation (Dogru, Mody, Suess, McGinley, & Line, 2020; Quattrone, Proserpio, Quercia, Capra, & Musolesi, 2016; Vinogradov, Leick, & Kivedal, 2020).

The *third group of studies* is comprised of research conducting comparative analysis of sharing accommodation system such as AirBnB, VRBO, HomeAway with traditional accommodation services (such as hotels and suites). A large portion of these studies using AirBnB and hotel listings data (such as listings, price, revenue) provided by or downloaded through automated scripts from AirBnB and hotel management. Studies in this group investigate the new age AirBnB demand considering relationship between AirBnB services with traditional hotel system. (Young, Corsun, & Xie, 2017) investigate travelers' preferences for VRBO relative hotels using an online survey in Denver, Colorado and found that factors like price, location, party size, dwelling size and trip length influence travelers to choose VRBO over hotel.

Few studies investigated location factors such points of interest, transport convenience, the surrounding environment impact on AirBnB listings and hotel supply (Sans & Quaglieri, 2016; Yang & Mao, 2020). Another set of studies consider supply of AirBnB listings impact on hotel performance such as revenue, prices and occupancy rates and found negative association (Neeser et al., 2015; Zervas et al., 2017; Dogru, Hanks, Mody, Suess, & Sirakaya-Turk, 2020; Dogru, Mody, Line, et al., 2020) while few literature discovered quite strange result that price have no effect on AirBnB and hotel supply so that AirBnB can be substitutes hotel (Gunter, Önder, & Zekan, 2020; Choi et al., 2015). *Finally*, the most commonly employed

.

**Table 2.1: Summary of Existing AirBnB Studies**

| Category | Study | Dimension | Country | Data | Method | Determinants |
|---|---|---|---|---|---|---|
| **Evaluation of Sharing Economy and Characteristics** | (Proserpio & Tellis, 2017) | Sharing Economy System | N/A | N/A | Literature Review | Evaluation |
| | (D. Guttentag, 2015) | Emergence of AirBnB | N/A | N/A | Literature Review | Accommodation and Tourism |
| | (Zervas, Proserpio, & Byers, 2015) | Emergence of AirBnB | Worldwide | AirBnB Listings and TripAdvisor | Statistical Distributions | Ratings |
| | (Oskam & Boswijk, 2016) | Emergence of AirBnB | N/A | N/A | Literature Review | Economic Benefits and Tourism |
| | (C. Wang, Komanduri, Viswanathan, Rossi, & West, 2018) | Visitor Demand | Los Angeles, USA | Hotel Occupancy and AirBnB Listings and Review | Statistical Comparison and Text Mining | Occupancy Rate |
| | (Juul, 2015) | Impact on Tourism | Europe | N/A | Literature Review | Impact on Tourism |
| | (B. G. Edelman & Geradin, 2015) | Policy Making | N/A | N/A | Literature Review | Rules and Regulation |
| | (Quattrone et al., 2016) | Impact on Tourism | London | Airbbnb Listings, Hotel and Census | Ordinary Least Squares (OLS) | Number of AirBnB and Hotel |
| | (Adamiak, 2018) | AirBnB Mapping | Europe | AirBnB Listings and TripAdvisor Hotel | Frequency and Comparison | Listings Capacity |
| **Qualitative and Quantitative Analysis** | (D. Lee, 2016) | Impact of AirBnB on Rent | Los Angeles, USA | AirBnB and Zillow | Regression | Price/Rent |
| | (Barron, Kung, & Proserpio, 2018) | Impact on Hotel Revenue | USA | AirBnB and Zillow | Regression | Price |

| Category | Study | Dimension | Country | Data | Method | Determinants |
|---|---|---|---|---|---|---|
| | (D. Wang & Nicolau, 2017) | Impact of AirBnB on Rent | 33 Citites in Europe, US, Canada and Australia | AirBnB Listings | Ordinary least squares and Quantile regression analysis | Host Attributes, Site and Property Attributes, Amenities and Services, Rental Rules and Online Review Ratings |
| | (Sperling, 2015) | Economic Impact | Portland, Los Angeles, New York City, San Francisco and Boston, USA | Income | Statistical Analysis | Income |
| | (Deboosere et al., 2019) | Economic Impact | New York City (NYC) | Listings Price | Hedonic Regression | Average Price of AirBnB per Night and Revenue |
| | (Gibbs et al., 2018) | Economic Impact | Canada | Listings Price | Hedonic Pricing Model | Price Rate of AirBnB |
| | (Tong & Gunter, 2020) | Economic Impact | Barcelona, Madrid, and Seville | Listings Price | Hedonic Pricing Model, Weighted Least Squares | Price Rate of AirBnB |

| Category | Study | Dimension | Country | Data | Method | Determinants |
|---|---|---|---|---|---|---|
| | | | | | (WLS) and Quantile Regression | |
| | (Wyman et al., 2020) | Economic Impact | Isle of Palms, SC, USA | Home Sales | Regression | Price Rate of AirBnB |
| | (Rodríguez-Pérez de Arenaza et al., 2019) | Impact on Residential Rental Price | Coast of Andalusia, Spain | Listings and Residential Rent Price | Regression | Residential Rental Price Rate |
| | (B. Edelman, Luca, & Svirsky, 2017) | Racial Discrimination | Baltimore, Dallas, Los Angeles, St. Louis, and Washington, D.C. | Field Experiment | Text Mining, Regression | Booking Confirmation |
| | (Brochado, Troilo, & Aditya, 2017) | Customer Experience and Preferences | India, Portugal and USA | SP Survey | Text Analysis | Stay, Host, Place, Location, Apartment, Room and City |
| | (Jordan & Moore, 2018) | Impact of AirBnB, Vacation Rentals By Owner (VRBO), and HomeAway | Oahu, Hawaii, USA | Interview | Thematic Analysis | Perception |

| Category | Study | Dimension | Country | Data | Method | Determinants |
|---|---|---|---|---|---|---|
| | (Jiao & Bai, 2020) | Emergence of AirBnB | USA | AirBnB Listings | Regression | AirBnB Listings Count |
| | (Zhu et al., 2019) | Emergence of AirBnB | New York, Los Angeles and Chicago (USA) | Online Reviews | Semantics Perspective | AirBnB Experience |
| | (Sthapit & Björk, 2019) | Trust and Reputation | -- | Online Reviews | Grounded Theory Research Design | Customer Service |
| | (Ert, Fleischer, & Magen, 2016) | Trust and Reputation | Stockholm, Sweden | Field Experiment | Hedonic Regression, Mixed Logit | Price |
| | (Fradkin, Grewal, & Holtz, 2018) | Trust and Reputation | AirBnB Reviews | Field Experiment | Logistic Regression | Online Review Analysis |
| | (Quattrone et al., 2016) | Impact on Tourism | London | AirBnB Listings, Hotel and Census | Ordinary Least Squares (OLS) | AirBnB and Hotel |
| | (Vinogradov et al., 2020) | Impact on Tourism and Rental Markets | Norway | AirBnB Listings | Agent Based Model | AirBnB Listings Supply |
| | (Dogru, Mody, Suess, et al., 2020) | Impact on Tourism | USA | AirBnB Listings | Regression | Employment |
| **Comparison with Hotel** | (D. A. Guttentag & Smith, 2017) | Price and Performance | Canada | SP Survey | T-test | Price and preferences |
| | (Young et al., 2017) | VRBO Performance | Denver, Colorado, USA | Email Survey | Statistical Analysis | Traveler's Preferences |

| Category | Study | Dimension | Country | Data | Method | Determinants |
|---|---|---|---|---|---|---|
| | (Sans & Quaglieri, 2016) | Impact on Hotel Revenue | Barcelona, Spain | AirBnB and Hotel Listings | Statistical Analysis | Policy Making |
| | (Yang & Mao, 2020) | Effects of Location on hotel and AirBnB | Houston, Texas | Monthly Revenue | Hausman-Taylor Model | Revenue |
| | (Choi, Jung, Ryu, Kim, & Yoon, 2015) | Impact on Hotel Revenue | Korea (Seoul, Busan, and Jeju) | Hotel Revenue and AirBnB Listings | Panel Regression | Hotel Revenue |
| | (Dogru, Mody, Line, et al., 2020) | Impact of AirBnB on Hotel Performance | USA | AirBnB and Hotel Listings | Regression | Hotel Revenues, Prices and Occupancy Rates |
| | (Dogru, Hanks, et al., 2020) | Impact of AirBnB on Hotel Performance | London, Paris, Sydney and Tokyo | AirBnB and Hotel Listings | Regression | Hotel Revenues, Prices and Occupancy Rates |
| | (Neeser, Peitz, & Stuhler, 2015) | Impact on Hotel Revenue | Norway, Finland, and Sweden | AirBnB and Hotel Listings | Logistic Regression | Hotel revenue per available room |
| | (Coyle & Yeung, 2016) | Impact on Hotel Revenue | 14 Cities in Europe | Number of Listings, Occupancy Rates and Average Revenue of AirBnB Hosts | Regression | Revenue and Occupancy Rate |
| | (Zervas, Proserpio, & Byers, 2017) | Impact on Hotel Revenue | Texas, USA | AirBnB and Hotel Listings from Smith Travel Research (STR), | Logistic Regression | Hotel Revenue |
| | (D. A. Guttentag & Smith, 2017) | Impact of AirBnB on Hotel | Canada | Online Survey | t-tests | Preference |
| | (Gunter et al., 2020) | Price Effects on AirBnB Demand | NYC, USA | AirBnB Listings | Spatial Durbin Model | AirBnB Listings Demand |

analytical approaches to study AirBnB listings include linear regression, logistic regression, ordinary least squares (OLS), t-test and text mining of reviews.

### 2.2 Earlier Research on Bikeshare Flows

The recent growth of bikeshare systems around the world has resulted in a number of research efforts examining bikeshare. Earlier research efforts can be broadly categorized into two groups. The first group of studies is focused on understanding user behavior, reasons for adopting bikeshare and user satisfaction from bikeshare systems using online surveys or questionnaires (see for example (Bachand-Marleau, Lee, & El-Geneidy, 2012; Buck et al., 2013; Fishman, Washington, & Haworth, 2014; Fuller et al., 2011; Schoner & Levinson, 2013)). The second group of studies examine bikeshare systems by conducting a quantitative analysis of ridership data. Given the focus of our current study, we restrict ourselves to a discussion of the second group of studies. Specifically, we provide a concise summary of the major research dimensions explored, urban regions considered for analysis, methodological approaches employed and major research findings from earlier research.

Several studies have examined bikeshare ridership data provided by bikeshare operator websites or downloaded through automated scripts from bikeshare websites. The most common dimensions of analysis in these research efforts include (a) system demand characterized as arrivals and departures from bike stations (Faghih-Imani & Eluru, 2016a, 2016b, 2017b; Faghih-Imani et al., 2014; Gebhart & Noland, 2014; Rixey, 2013; Rudloff & Lackner, 2014; Wang, Lindsey, Schoner, & Harrison, 2015; Yufei, Oukhellou, & Come, 2014), (b) factors affecting bikeshare operators to move bicycles to avoid excess bikes (or empty slots) at some stations (referred to as rebalancing demand) (Bouveyron, Côme, & Jacques, 2015; Faghih-Imani, Hampshire, Marla, & Eluru, 2017; Forma, Raviv, & Tzur, 2015; Fricker & Gast, 2016; Nair, Miller-Hooks, Hampshire, & Bušić, 2013; Pfrommer, Warrington, Schildbach, & Morari, 2014; Raviv, Tzur, & Forma, 2013; Vogel & Mattfeld, 2011), (c) destination station preferences for bikeshare users (El-Assi et al., 2017; Faghih-Imani & Eluru, 2015, 2017b) and (d) impact of bike share on the urban transportation system including reducing emissions, altering transportation mode share and competition across modes (see (Faghih-Imani, Anowar, Miller, & Eluru, 2017)). The various bikeshare systems analyzed in the literature include many urban regions such as New York (CitiBike), Montreal (BIXI), Paris (Velib), London (Santander), Chicago (Divvy), Hangzhou (Hangzhou Public Bicycle), Beijing (Beijing Public Bicyle), Melbourne (Melbourne Bike Share), and Brisbane (CityCycle).

The most commonly employed analytical approaches to study bikeshare systems include linear regression, linear mixed models, panel ordered logit models, negative binomial count models, multinomial logit (MNL), mixed multinomial logit, finite mixture multinomial logit model, and time series models and their variants (Buck et al., 2013; El-Assi et al., 2017; Faghih-Imani & Eluru, 2015; Faghih-Imani et al., 2014; Gebhart & Noland, 2014; Rixey, 2013; Rudloff & Lackner, 2014; Wang et al., 2015; Zhao et al., 2014). Major findings from these research efforts can be broadly summarized as follows. Bikeshare system usage at a station level is influenced by bikeshare infrastructure (such as number of stations and station capacity), bicycling infrastructure (such as presence of bike lanes), land use and built environment (such as population density, job density and points of interest), public transportation infrastructure (presence of bus/metro stops), and temporal and meteorological attributes (such as precipitation and temperature) (El-Assi et al., 2017; Faghih-Imani & Eluru, 2015, 2016a, 2016b; Faghih-Imani et al., 2014; Gebhart & Noland, 2014; Rixey, 2013; Wang et al., 2015). Destination choice studies found that bikeshare users prefer shorter trips with all else same (El-Assi et al., 2017; Faghih-Imani & Eluru, 2015). Bikeshare users trade-off on station distance with other conveniences such as access to points of interest and stations with larger capacity.

## 2.3 Earlier Research on TNC Flows

Ride hailing in its traditional form has received attention from various researchers (for example see (Faghih-Imani, Anowar, Miller, & Eluru, 2017) for detailed literature review of traditional taxi services). The research on TNC services is an emerging topic of interest in several fields including computer science, transportation, economics, and social sciences. In our analysis, we restrict ourselves to literature on TNC systems that are directly relevant from a transportation perspective.

Earlier research efforts focused on TNC ride hailing can be grouped into two streams. The first stream of studies explored TNC evolution, factors that affected usage, licensing and policy formulation, pricing mechanisms, and comparison across ride hailing services (with taxis or between various smart phone based ride hailing companies). These studies typically rely on questionnaire interviews, and online surveys for data collection. TNC evolution studies focused on the definition of ride hailing systems, how ride hailing services have evolved over time (Chan & Shaheen, 2012; Furuhata et al., 2013; Sun & Edara, 2015), investigated the challenges and opportunities presented by real-time services and highlighted various opportunities for future (Agatz, Erera, Savelsbergh, & Wang, 2012; Amey, Attanucci, & Mishalani, 2011). A TCRP report (Feigon & Murphy, 2016) examining shared modes of travel (such as bikesharing, carsharing, and TNC systems) by conducting surveys and interviews across seven urban regions (Austin, Boston, Chicago, Los Angeles, San Francisco, Seattle, and Washington, DC). The study concluded that individuals who adopt shared modes for their travel needs are more open to public transit alternatives. Further, these shared modes can serve as complementary modes to public transit. A set of studies explored the influence of various factors affecting TNC usage. For example, Cramer and Krueger (Cramer & Krueger, 2016) analyzed passenger service times for Uber and taxi across five major cities in the US. The authors concluded that availability of driver-passenger reviews, Uber's flexible labor supply model coupled with inefficient taxi regulations for passenger safety contributed to higher Uber utilization rates. Rayle et al. (Rayle, Dai, Chan, Cervero, & Shaheen, 2016) conducted a trip intercept survey to understand the source of TNC demand and concluded that nearly 50% of the demand is transferred from public transit and driving. Multiple studies explored pricing strategies employed by various ride hailing companies (L. Chen, Mislove, & Wilson, 2015; M. K. Chen & Sheldon, 2015; Guo, Liu, Xu, & Chiu, 2017). Studies examining Uber surge pricing strategies, concluded that surge pricing has a negative impact on demand. Smart et al. (Smart, Rowe, & Hawken, 2015) compared the performance of Uber and taxi services in terms of waiting time and cost using survey of riders in low income neighborhoods in Los Angeles. The data analysis found that Uber offered lower waiting times and provided service at a lower cost (even under surge pricing).

A second stream of studies conducted quantitative analysis using TNC usage data exploring trip patterns (a) to identify factors influencing TNC demand, (b) to understand TNC demand and its relationship with existing transportation modes. Earlier research has found that Uber demand is affected by temporal and weather patterns (Brodeur & Nield, 2016; Gerte, Konduri, & Eluru, 2018). Other factors that were found to affect ride hailing demand include land use attributes such as lower transit access time (TAT), higher length of roadways, lower vehicle ownership, higher income and more job opportunities (Alemi, Circella, Handy, & Mokhtarian, 2018; Correa, Xie, & Ozbay, 2017; Davidson, Peters, & Brakewood, 2017). Studies comparing the emerging ride hailing services with existing services such as public transit and bicycle sharing offer interesting results. Gerte et al. (Gerte, Konduri, Ravishanker, Mondal, & Eluru, 2019) found evidence for shifting taxi demand to smart phone based ride hailing services in New York City. Further, the study also found evidence of substitution relationship between ride hailing and bicycle share systems. Komaduri et al. (Komanduri, Wafa, Proussaloglou, & Jacobs, 2018) analyzed data from RideAustin, to examine the trip length and temporal distribution of the trips. A comparison of the adoption of RideAustin relative to public transit alternatives illustrated that individuals were choosing RideAustin to minimize travel time (highlighting the higher value of time for these travelers). Poulsen et al. (Poulsen et al., 2016) examined how the two systems that were introduced in the same time performed - Uber and Green taxis - in Manhattan area and found that the growth rate for Uber was substantially higher. Babar and Burtch (Babar & Burtch, 2017) compared the utilization rate of transit service in the US after the introduction of TNC services and found that utilization rate of bus service dropped while long-haul transit services (such as subway and commuter rail) experienced increasing utilization. The spectrum of quantitative methodologies employed in earlier studies include descriptive analysis, linear regression, logistic regression, difference in difference model and panel based random effects multinomial logit model.

**2.4 Earlier Research on Ride hailing Transformation**

Literature related to ride hailing vehicles can be categorized into three main streams: a) studies investigating various operational and quantitative aspects of taxis, b) studies investigating the evolution and various qualitative aspects of TNC based ride hailing and c) studies examining the relationship between various ride hailing systems and their interaction with public transportation.

The *first group of studies* focused on taxi services from different perspectives, including entry regulation (see Schaller (Schaller, 2007) for US and Canada regulation and Çetin and Eryigit (Çetin & Eryigit, 2011) for Istanbul regulation), demand and pricing (Chang & Chu, 2009; Milioti, Karlaftis, & Spyropoulou, 2015; Zhang & Ukkusuri, 2016), and impact of emerging technologies such as electric and autonomous vehicles (Burghout, Rigole, & Andreasson, 2015; Chrysostomou, Georgakis, Morfoulaki, Kotoula, & Myrovali, 2016; Jung, Chow, Jayakrishnan, & Park, 2014). Several studies analyzed different aspects of taxi operations including taxi passenger search schemes and routing of vacant taxis to improve the efficiency of taxi services (K. Wong, Wong, Yang, & Wu, 2008; R. Wong, Szeto, & Wong, 2014, 2015; Yang & Wong, 1998; Zhan & Ukkusuri, 2015; Zhang, Ukkusuri, & Lu, 2017). Crash injury severity and safety issues related to taxi services are also examined by several researchers (Dalziel & Job, 1997; Lam, 2004; Peltzer & Renner, 2003; Tay & Choi, 2016; Tseng, 2013).

The *second group of studies* explored TNC evolution, factors that affected usage, licensing and policy formulation, pricing mechanisms, and comparison across ride hailing services (with taxis or between various smart phone based ride hailing companies). These studies typically rely on questionnaire interviews, and online surveys for data collection. TNC evolution studies focused on the definition of ride hailing systems, how ride hailing services have evolved over time (Chan & Shaheen, 2012; Furuhata et al., 2013; Sun & Edara, 2015), investigated the challenges and opportunities presented by real-time services and highlighted various opportunities for the future (Agatz, Erera, Savelsbergh, & Wang, 2012; Amey, Attanucci, & Mishalani, 2011). A set of studies explored the influence of various factors affecting TNC usage. For example, Cramer and Krueger (Cramer & Krueger, 2016) analyzed passenger service times for Uber and taxi across five major cities in the US. The authors concluded that availability of driver-passenger reviews, Uber's flexible labor supply model coupled with inefficient taxi regulations for passenger safety contributed to higher Uber utilization rates. Multiple studies explored pricing strategies employed by various ride hailing companies (L. Chen, Mislove, & Wilson, 2015; M. K. Chen & Sheldon, 2015; Guo, Liu, Xu, & Chiu, 2017). Studies examining Uber surge pricing strategies, concluded that surge pricing has a negative impact on demand. Smart et al. (Smart, Rowe, & Hawken, 2015) compared the performance of Uber and taxi services in terms of waiting time and cost using survey of riders in low income neighborhoods in Los Angeles. The data analysis found that Uber offered lower waiting times and provided service at a lower cost. Another subset of studies conducted quantitative analysis using TNC usage data exploring trip patterns (a) to identify factors influencing TNC demand, (b) to understand TNC demand and its relationship with existing transportation modes. Factors that were found to affect ride hailing demand include temporal and weather patterns, land use attributes such as lower transit access time, higher length of roadways, lower vehicle ownership, higher income and more job opportunities (Alemi, Circella, Handy, & Mokhtarian, 2018; Correa, Xie, & Ozbay, 2017; Davidson, Peters, & Brakewood, 2017).

The *third group of studies* is comprised of research conducting comparative analysis using ride hailing usage data. The research conducted in this paper falls into this third category. A group of studies investigate the new age ride hailing demand considering relationship between ride hailing services with public transit system (Gerte et al., 2019; Komanduri, Wafa, Proussaloglou, & Jacobs, 2018; Murphy, 2016; Rayle, Dai, Chan, Cervero, & Shaheen, 2016). Rayle et al. (Rayle et al., 2016) conducted a trip intercept survey to understand the source of TNC demand and concluded that nearly 50% of the demand is transferred from public transit and driving. Studies comparing the emerging ride hailing services with existing services such as public transit and bicycle sharing offer interesting results. Gerte et al. (Gerte et al., 2019) found evidence for shifting taxi demand to smart phone based ride hailing services in New York City. Further, the study also found evidence of substitution relationship between ride hailing and bicycle share systems. Komaduri et al. (Komanduri et al., 2018) analyzed data from RideAustin, to examine the trip length and temporal distribution of the trips. A comparison of the adoption of RideAustin relative to public transit alternatives illustrated that riders were choosing RideAustin to minimize travel time (highlighting the higher value of time for these travelers). Poulsen et al. (Poulsen et al., 2016) examined how the two systems that were introduced in the same time performed - Uber and Green taxis - in Manhattan area and found that the growth rate for Uber was substantially higher. Babar and Burtch (Babar & Burtch, 2017) compared the utilization rate of transit service in the US after the introduction of TNC services and found that utilization rate of bus service dropped while long-haul transit services (such as subway and commuter rail) experienced increasing utilization.

## 2.5 Summary

This chapter presented a detailed summary of methodologies employed in earlier studies for predicting flows at different spatial unit for different attribute level. The data source along with the dependent and exogenous attributes used for analysis is described in detail in the subsequent chapter.

# CHAPTER 3: ANALYSIS OF HOSPITALITY DEMAND IN NEW YORK CITY USING AIRBNB DATA: A COPULA BASED COUNT MODEING APPROACH

## 3.1 Introduction

Travel and tourism industry is undergoing a transformation with the flourishing of online sharing economy marketplaces such as Uber (for taxi services), Eatwith (for community restaurants), and AirBnB (for accommodation). The shared housing market place AirBnB with its large inventory and wide reach across the globe is redefining the hospitality sector. AirBnB is unique in its design as it does not own any properties but provides a platform for ordinary people (sellers) to rent their residences (entire house/apartment or a room) to tourists (consumers) (Botsman & Rogers, 2011). AirBnB accommodation system is quite easy to use: a consumer searches for an entire home or private (or shared) room based on their travel dates and destination on the AirBnB website (www.AirBnB.com). The user is provided with a list of housing alternatives based on the user preferences. The success and wide adoption of the system is based on available review information and background check procedures for renters and tourists. AirBnB charges a service fee for each transaction. Initiated in 2008, popularity of this sharing hospitality platform has rapidly grown with over 200 million guests having stayed in about 3 million listings in more than 65,000 cities and 191 countries (AirBnB, 2017). In fact, since 2016, over 100 million people have enjoyed the accommodation through AirBnB while over 1 million new listings worldwide have been added to the market place.

The growth of AirBnB impacts transportation and urban systems along two major directions. *First*, AirBnB provides a unique snapshot of the hospitality industry and can serve as a surrogate for the health of tourism industry in the region. The number of available listings on AirBnB can serve as a proxy for tourist interest in the region. AirBnB provides renters with an opportunity to immediately respond to tourist demand by allowing for a simple listing process (without any substantial capital costs). In the event of a drop in tourist demand, renters on the website remove their listing. On the other hand, traditional hospitality industry with hotels respond to tourist demand slowly due to the large capital costs involved in increasing capacity. In addition, the traditional hospitality sector cannot dismantle their infrastructure as easily in response to the reduced tourist demand. Thus, with its ease of adding a listing, AirBnB listings provide a unique snapshot of the health of tourism industry. *Second*, an analysis of AirBnB listings will allow transportation and urban regional professionals examine the demand arising from these tourists on transportation and urban infrastructure. Cities such as New York that receive significant expenditures from tourists can provide improved services by enhancing infrastructure in response to emerging tourist locations.

The proposed research develops a framework to understand factors affecting AirBnB inventory. Drawing on NYC AirBnB listings data from a fine spatial and temporal resolution, the proposed study examines the ongoing transformation of sharing accommodation market from January 2015 to September 2017. For our analysis, monthly AirBnB inventory is represented at a disaggregate spatial resolution as the number of listing at a census tract level by listing type defined as (a) entire home or (b) private/shared room. The study develops an advanced econometric model framework relying on copula based model system. Specifically, our proposed approach accommodates for the presence of common unobserved factors affecting (a) the two dependent variables at the census tract (inventory by entire home and private/shared room) and (b) multiple repeated observations from 31 months of data. The framework takes the form a bivariate random parameter copula based negative binomial model. The proposed model framework is estimated using a host of independent variables including socio-demographic variables, transportation infrastructure variables and land use and built environment variables. The empirical analysis is augmented with a policy analysis conducted to illustrate how listings count is influenced by various exogenous attributes.

The rest of the chapter is organized as follows: The next section presents the methodological framework adopted in the analysis while section 3 provides a detailed description of the dataset with sample formation technique. Model results are presented in the fifth section followed by the policy analysis. Final section comprises with the concluding statements.

## 3.2 Econometric Methodology

The econometric framework for the joint model is presented in this section.

### 3.2.1 <u>NB Model</u>

Let $i$ be the index for CT $(i = 1,2,3, \dots, N)$ and $y_{mit}$ be the index for types of accommodation in time period $t$ $(t = 1,2,3, \dots, T)$ for a CT $i$; where $m$ takes the value of 1 for whole apartment/home and 2 for private or shared room. The NB probability expression for random variable $y_{mit}$ can be written as (Cameron, Li, Trivedi, & Zimmer, 2004):

$$P_{mit}(y_{mit}) = \frac{\Gamma(y_{mit}+\alpha_m^{-1})}{\Gamma(y_{mit}+1)\Gamma(\alpha_m^{-1})}\left(\frac{1}{1+\alpha_m\mu_{mit}}\right)^{\frac{1}{\alpha_m}}\left(1-\frac{1}{1+\alpha_m\mu_{mit}}\right)^{y_{mit}} \tag{3.1}$$

where, $\Gamma(\cdot)$ is the Gamma function, $\alpha_m$ is the NB dispersion parameter specific to room type group $m$ and $\mu_{mit}$ is the expected number of accommodations listed in CT $i$ for time period $t$. We can express $(\mu_{mit})$ as a function of explanatory variable $(x_{mi})$ by using a log-link function as: $\mu_{mit} = E(y_{mit}|\boldsymbol{x}_{mit}) = exp\big((\boldsymbol{\beta}_m + \gamma_{mi})\boldsymbol{x}_{mit} + \varepsilon_{mit}\big)$, where $\beta_m$ is a vector of mean effects to be estimated specific to room type group m and $\gamma_{mi}$ represents a vector of unobserved factors affecting count propensity associated with room type $m$ for CT $i$ and its associated zonal characteristics, assumed to be a realization from standard normal distribution: $\gamma_{mi} \sim N(0, \boldsymbol{\pi}_m{}^2)$. $\varepsilon_{mit}$ is a gamma distributed error term with mean 1 and variance $\alpha_m$.

3.2.2 Multivariate NB Model

The purpose of multivariate NB model is to examine counts of different types of AirBnB listings. We consider two types of AirBnB listings for our study approach: (a) whole apartment/home and (b) private/shared room at census tract (CT) level. For the multivariate approach, the equation system for modeling listings count across different listings types can be written by replacing the subscript $m$ with $j$ in equation 3.1. Thus, the probability for listings count for two different listings type $m$ can be represented as $P(c_{ijt})$, for which we can express $\mu_{mit}$ as a function of explanatory variables by using a log-link function as follows:

$$\mu_{mit} = E(c_{ijt}|\boldsymbol{z}_{ijt}) = exp((\boldsymbol{\delta_j} + \boldsymbol{\zeta}_{ij})\boldsymbol{z}_{ijt} + \varepsilon_{ijt} + \eta_{ijt}) \qquad (3.2)$$

where, $\boldsymbol{z}_{mi}$ is a vector of explanatory variables associated with CT $i$ and listings type $m$ . $\boldsymbol{\delta}_m$ is a vector of coefficients to be estimated. $\boldsymbol{\zeta}_{mi}$ is a vector of unobserved factors on listings count propensity associated with listings type $m$ for CT $i$ and its associated zonal characteristics, assumed to be a realization from standard normal distribution: $\boldsymbol{\zeta}_{mi} \sim N(0, \boldsymbol{\pi}_m{}^2)$. $\varepsilon_{mi}$ is a gamma distributed error term with mean 1 and variance $\alpha_m$. $\eta_{mi}$ captures unobserved factors that simultaneously impact number of AirBnB listings across two listings types for CT $i$. Here, it is important to note that the unobserved heterogeneity between total number of crashes across different collision types can vary across CT's. Therefore, in the current study, the correlation parameter $\eta_i$ is parameterized as a function of observed attributes as follows:

$$\eta_{mi} = \boldsymbol{\gamma_m} \boldsymbol{s}_{mi} \qquad (3.3)$$

where, $\boldsymbol{s}_{mi}$ is a vector of exogenous variables, $\boldsymbol{\gamma_m}$ is a vector of unknown parameters to be estimated (including a constant). In the current analysis, the multivariate NB model only allows for a positive correlation for total number of crashes across different collision types.

In examining the model structure of crash count across different collision types, it is necessary to specify the structure for the unobserved vectors $\boldsymbol{\zeta}$ and $\boldsymbol{\gamma}$ represented by $\boldsymbol{\Omega}$. In this paper, it is assumed that these elements are drawn from independent normal distributions: $\Omega \sim N(0, (\boldsymbol{\pi_m}^{\mathbf{2}}, \boldsymbol{\sigma}_m^2))$. Thus, conditional on $\boldsymbol{\Omega}$, the likelihood function for the joint probability can be expressed as:

$$L_i = \int_{\boldsymbol{\Omega}} \prod_{m=1}^{m} \left(P(c_{mi})\right) f(\boldsymbol{\Omega}) d\boldsymbol{\Omega} \tag{3.4}$$

Finally, the log-likelihood function is:

$$LL = \sum_{i} Ln(L_i) \tag{3.5}$$

All the parameters in the model are estimated by maximizing the logarithmic function $LL$ presented in equation 5. The parameters to be estimated in the multivariate NB model are: $\boldsymbol{\delta_m}$, $\alpha_m$, $\boldsymbol{\pi}_m$, and $\boldsymbol{\sigma_m}$.

### 3.2.3 Copula Multivariate NB Model

The focus of our study is to jointly model counts of AirBnB listings for: (a) whole apartment/home and (b) private/shared room at census tract (CT) level by employing a random parameters copula based bivariate NB modeling framework.

Let's assume $v_{il}$ is the expected number of listings in CT $i$ over a given time period for listings type $l$. We can express $v_{il}$ as a function of explanatory variable $(\boldsymbol{x}_{il})$ by using a log-link function as: $v_{il} = E(c_{il}|\boldsymbol{x}_{il}) = exp(\boldsymbol{\beta}_l \boldsymbol{x}_{il})$, where $\boldsymbol{\beta}_l$ is a vector of parameters to be estimated specific to listings type $l$.

By using copula based approach, correlation between random variables $y_{1i}$ and $y_{2i}$ can be explored. In constructing the copula dependency, let us assume that $\Lambda_1(y_{1it})$ and $\Lambda_2(y_{2it})$ are the marginal distribution functions of the random variables $y_{1i}$ and $y_{2i}$, respectively; and $\Lambda_{12}(y_{1it}y_{2it})$ is the joint distribution which can be generated as a joint cumulative probability distribution of uniform [0, 1] marginal variables $U_1$ *and* $U_2$ as below (Bhat & Eluru, 2009):

$$\begin{aligned} \Lambda_{12}(y_{1it}, y_{2it}) &= Pr( U_1 \leq y_{1it}, U_2 \leq y_{2it}) \\ &= Pr[\Lambda_1^{-1}(U_1) \leq y_{1it}, \Lambda_2^{-1}(U_2) \leq y_{2it} ] \\ &= Pr[U_1 < \Lambda_1(y_{1it}), U_2 < \Lambda_2(y_{2it}) ] \end{aligned} \tag{3.6}$$

The joint distribution (of uniform marginal variable) in equation 6 can be generated by a function $C_{\theta i}$ $(.,.)$ (Sklar, 1973), such that:

$$\Lambda_{12}(y_{1it}, y_{2it}) = C_{\theta i}(U_1 = \Lambda_1(y_{1it}), U_2 = \Lambda_2(y_{2it})) \tag{3.7}$$

where, $C_{\theta i}$ $(.,.)$ is a copula function and $\theta_i$ is the dependence parameter defining the link between $y_{1it}$ and $y_{2it}$. However, in our study, $y_{1it}$ and $y_{2it}$ are nonnegative integer valued events. For such count data, the probability mass function ($\zeta_{\theta i}$) is presented by using finite differences of the copula representation as follows (Cameron et al., 2004):

$$\begin{aligned} &\zeta_{\theta i}\big(\Lambda_1(y_{1it}), \Lambda_2(y_{2it})\big) \\ &= C_{\theta i}(\Lambda_1(y_{1it}), \Lambda_2(y_{2it});\ \theta_i) - C_{\theta i}(\Lambda_1(y_{1it} - 1), \Lambda_2(y_{2it});\ \theta_i) \\ &\quad - C_{\theta i}(\Lambda_1(y_{1it}), \Lambda_2(y_{2it} - 1);\ \theta_i) \\ &\quad + C_{\theta i}(\Lambda_1(y_{1it} - 1), \Lambda_2(y_{2it} - 1);\ \theta_i) \end{aligned} \tag{3.8}$$

where, $\Lambda_1(y_{1it})$ and $\Lambda_2(y_{2it})$ as the cumulative distribution function (CDF) of the NB distribution. The CDF of NB probability expression (as presented in equation 1) for $y_{mit}$ can be written as for a particular realization of $\gamma_{mi}$:

$$\Lambda_m(y_{mit}|\gamma_{mi}) = \sum_{k=0}^{y_{mi}} P_{mit}(y_{mit}|\gamma_{mi}) \tag{3.9}$$

The unconditional log-likelihood function (LL) with the joint probability expression in equation 4 by integrating over $\gamma_{mi}$ for all time periods can be written as:

$$LL = \sum_{i=1}^{N} \prod_{t} \int \zeta_{\theta i}\big(\Lambda_1(y_{1it}), \Lambda_2(y_{2it})\big)\, d\gamma \tag{3.10}$$

In our empirical analysis we select six different copula structures: 1) Gaussian, 2) Farlie-Gumbel-Morgenstern (FGM), 3) Clayton, 4) Gumbel, 5) Frank and 6) Joe (a detailed discussion of these copulas is available in (Bhat & Eluru, 2009)).

The level of dependence between the random variables can vary across CTs. Therefore, in the current study, the dependence parameter $\theta_i$ is parameterized as a function of observed attributes as follows:

$$\theta_i = fn(\boldsymbol{\delta}_m \boldsymbol{s}_{mi}) \tag{3.11}$$

where, $\boldsymbol{s}_{mi}$ is a column vector of exogenous variable, $\boldsymbol{\delta}_m$ is a row vector of unknown parameters (including a constant) specific to room type group $m$ and $\boldsymbol{f}_n$ represents the functional form of parameterization. Based on the dependency parameter permissible ranges, alternate parameterization forms for the six copulas are considered in our analysis (Nashad et al., 2016). We will employ the Bayesian Information Criterion to determine the preferred copula model.

## 3.3 Data

### 3.3.1 Data Source

The New York city, our study area is associated with 2166 Census tract and 5 boroughs with a population of about 8.5 million (Figure 3.1). New York City receives over 60 million foreign and American tourists each year. NYC have over 41000 AirBnB listings while around 92% situated in Manhattan and Brooklyn borough. Given these afore-mentioned implications, the proposed research conducts a comprehensive analysis of AirBnB listing in New York City region drawing on data from January 2015 to September 2017 (http://insideAirBnB.com/get-the-data.html). The listings dataset provides information on zip code, longitude and latitude, city and street name, accommodation information such as residence type (full apartment or private/shared room), number of bedrooms and bathrooms, price, amenities information and review of customers. The listings data is aggregated at a census tract level (2166 census tracts) in the New York City region.

In addition to the listing database, the explanatory attributes considered in the empirical study will also be generated at the CT level. The selected explanatory variables can be grouped into three broad categories: (1) built environment attributes such as number of restaurants and park area derived from New York City open data (https://nycopendata.socrata.com); (2) socio-demographic characteristics at the census tract/zip code level gathered from US 2010 census; (3) transportation infrastructure attributes.

3.3.2 Sample Formation

The first step in data assembly for analysis is sample formation to generate the dependent variables for the analysis (count of availability of home/room) from disaggregate listing data. The average density distribution of full apartment/home and private or shared room for 31 months for each census tract level of NYC was defined into 6 six categories start with no AirBnB and then from very low to very high that is shown in Figure 3.2. Of the 2166 census tracts, 120 tracts ending up with no AirBnB listings. In terms of the two dependent variables, around 17% of the census tracts have zero full apartment/home listings while the corresponding number for private/shared room is about 10%. Further, the figures indicate that major portion of the AirBnB listings are observed in Manhattan and Brooklyn boroughs. Given that the NYC tourism industry is concentrated in these two boroughs the trend is expected.

For the given study period, we aggregated monthly total number of available listings data for each month (total 31 months) for each census tract of NYC. To obtain a reasonable sample size for model estimation, 5 months listings data for each census tract were randomly selected. As a result of the random month selection, we ended up having 10230 samples observation finally. A summary of the dependent variable and independent variable data compilation procedure is presented in Figure 3.3.

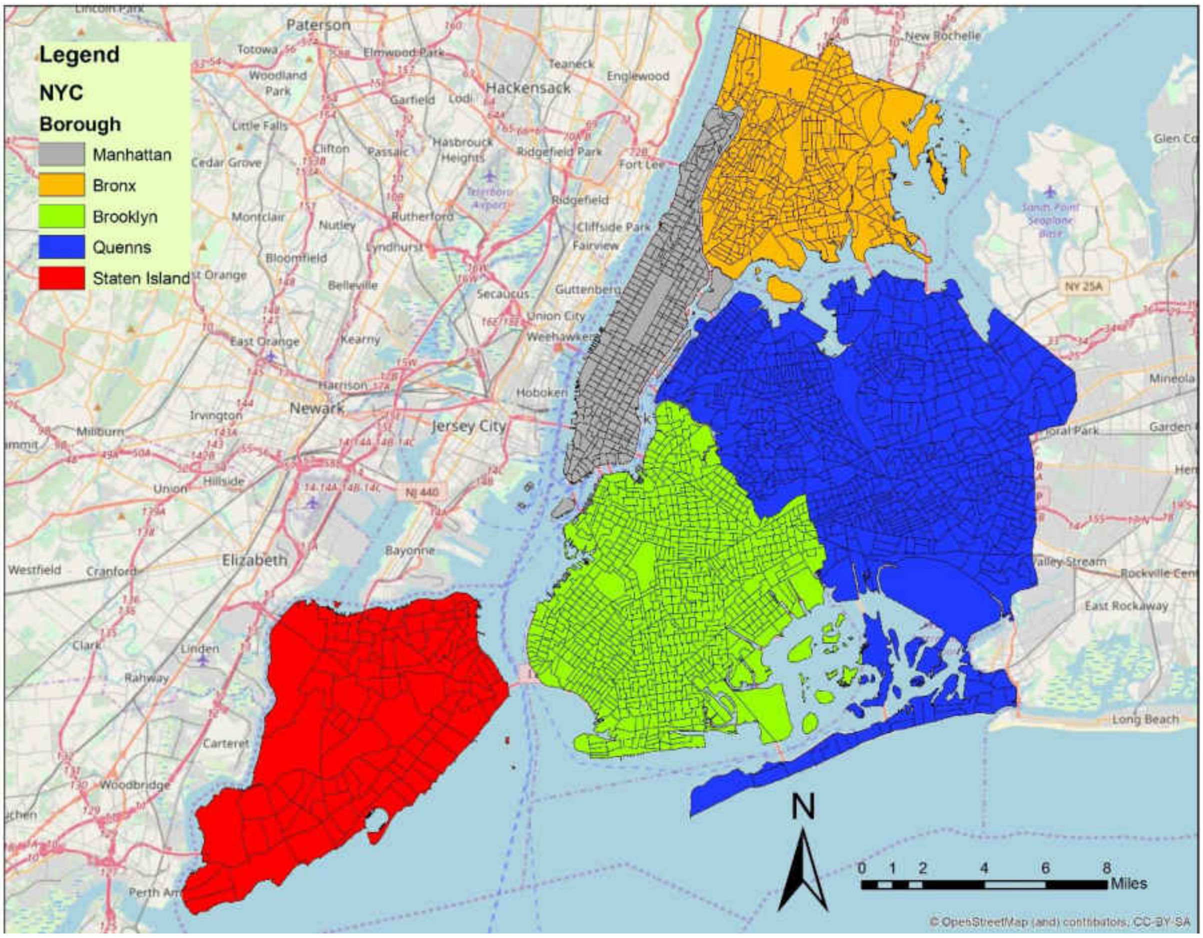
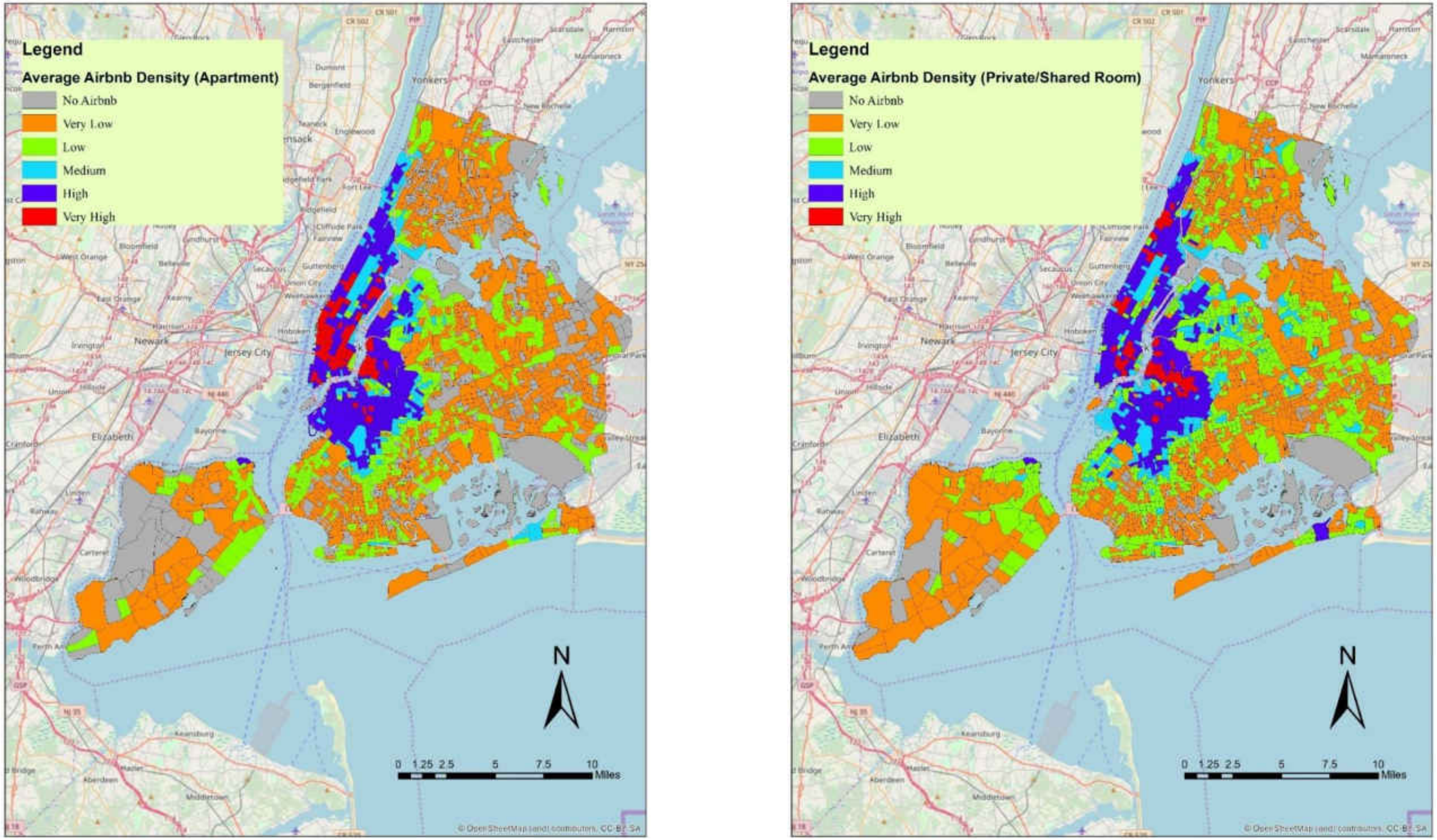



**Figure 3.1: Census Tract Zone of NYC**




**(a) Density Distribution of Whole Apartment/Home** **(b) Density Distribution Private or Shared Room**

**Figure 3.2: Density Distribution of Average Count of AirBnB (Apartment/Room)**

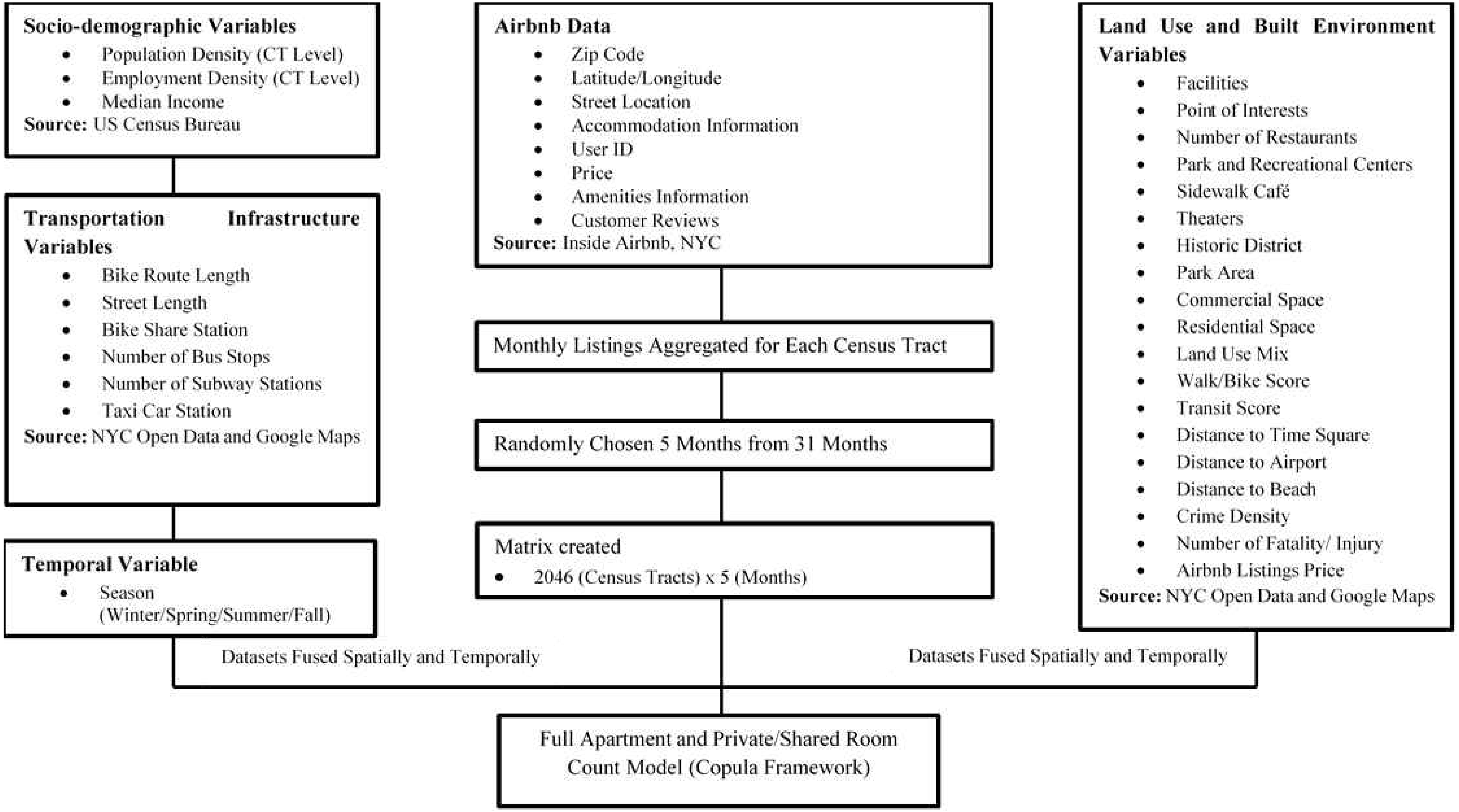



**Figure 3.3: Data Formation Flow Chart**

3.3.3 Independent Variable Generation

Several independent variables were generated in our study (see Figure 3.3). These can be grouped into four categories: 1) Socio-demographic variables, 2) Transportation infrastructure variables and 3) Land use and built environment variables. The socio-demographic attributes considered are population density, job density and median income. Population information was collected from US census 2010 and projected for corresponding year (2015-2017) at the census tract level. Job density data was estimated at the census tract level while median income was calculated at the census tract level for corresponding year.

Transportation infrastructure variables include number of bikeshare station, public transit stations in a census tract level. The variables created at the census tract level include length of bike routes, length of roads (minor and major roads). Number of subway stations and bus stops in the CT were generated to examine the influence of public transit on individual's preference of AirBnB location.

Several land use and built environment variables were considered including the number of facilities (schools, colleges, hospitals), the number of point of interests (museums, shopping malls), and the number of restaurants (including coffee shops and bars), total area of parks and commercial space (office, industry, retail) within each census tract. Few trip distance was also considered including distance of Times Square, nearest airport and beach from centroid of each census tract. While the actual trip might involve a different route, the shortest network distance would be an appropriate indicator of the distance traveled. Non-motorized vehicle score (average of walk score and bike score) and transit score associated with each AirBnB was considered at the census tract level. Total area of various land use profile together with mixed land use attribute was also considered to capture the preference land use for AirBnB. Average listings price (full apartment and private/shared room) for one night was estimated for each census tract level to capture effect of variation of listings price on AirBnB supply. Few safety related attributes such as total number of crimes, number of crashes considering number of fatality and injury was also created in a census tract level to get a clear view of AirBnB preferences. Finally, Seasonality is the only temporal variable considered. We consider winter (December-February), Spring (March-May), Summer (June-August) and Fall (September-November) as dummy variables. A descriptive summary of the analysis sample is presented in Table 3.2.

### 3.4 Empirical Analysis

#### 3.4.1 Model Specification and Overall Measures of Fit

Several models were estimated as part of our empirical exercise. These include: (1) Independent NB, (2) Mixed Independent NB, (3) Multivariate mixed NB, (4) Copula structures. Five copula structures were used in the empirical analysis; they are: 1) FGM, 2) Clayton, 3) Gumbel, 4) Frank and 5) Joe. The copula model estimation involved four considerations. *First,* five different models were estimated by considering the dependency parameter in the copula model to be the same across all CTs. *Second,* best three model estimated from first step were also estimated by considering the parameterization for copula dependency profile. *Third*, best copula model from first and second consideration were estimated to capture unobserved heterogeneity without considering dependency profile. *Finally*, dependency profile was added with unobserved heterogeneity in the same model from third step for analyze.

The performance of the estimated models was compared based on two goodness of fit measures best suited for comparing non-nested models: (1) Akaike information criterion (AIC) and (2) Bayesian Information Criterion (BIC). The AIC for a given empirical model is equal to:

## Table 3.1: Descriptive Summary of Sample Characteristics

| Variable Names | | Definition | CT Level | | |
|---|---|---|---|---|---|
| | | | Minimum | Maximum | Mean |
| *Dependent Variables* | | | | | |
| | Count of Entire Apartment | Total number of entire/whole apartment in CT | 0.000 | 225.000 | 9.15 |
| | Count of Private or Shared Room | Total number of Private or Shared Room in CT | 0.000 | 165.000 | 8.41 |
| *Sociodemographic Characteristics* | | | | | |
| | Total Population | Total number of populations in CT | 0.000 | 30260.000 | 4121.660 |
| | Population Density | Ln (Number of population in CT/Total area of CT in square miles) | 0.0000 | 12.450 | 10.472 |
| *Socioeconomic Characteristics* | | | | | |
| | Total Employment | Total number of jobs in CT | 0.000 | 15675.000 | 2394.190 |
| | Employment Density | Total number of jobs in CT/Total number of populations in CT | 0.000 | 1.000 | 0.573 |
| *Built Environment and Land Use Attributes* | | | | | |
| | CT area | Total area of CT in square miles | .0161 | 3.8177 | .128583 |
| | Facilities | Total number of facilities in CT | 0.000 | 135.000 | 16.931 |
| | Point of Interests | Number of point of interests in CT | 0.000 | 177.000 | 8.445 |
| | Park and Recreational Centers | Total number of park and recreational centers in CT | 0.000 | 3.000 | 0.036 |
| | Restaurants | Total number of restaurants in CT | 0.000 | 544.000 | 11.869 |
| | Sidewalk Cafe | Total number of sidewalk café in CT | 0.000 | 136.000 | 9.685 |
| | Theaters | Total number of theaters in CT | 0.000 | 23.000 | 0.057 |
| | Distance to Airport | Distance to the nearest airport from each CT | 5.186 | 10.637 | 9.144 |
| | Distance to Beach | Distance to the nearest beach from each CT | 15.299 | 15.310 | 15.304 |
| | Building Area | Ln (Total building footprint area of CT in square meters) | 0.000 | 5.560 | 2.733 |
| | Commercial Area | Ln (Total commercial area of CT in square meters) | 0.000 | 16.973 | 12.555 |
| | Residential Area | Ln (Total residential area of CT in square meters) | 0.000 | 16.730 | 13.975 |

| Variable Names | Definition | CT Level | | |
|---|---|---|---|---|
| | | Minimum | Maximum | Mean |
| Office Area | Ln (Total office area of CT in square meters) | 0.000 | 16.800 | 10.023 |
| Retail Area | Ln (Total retail area of CT in square meters) | 0.000 | 15.030 | 10.429 |
| Industrial Area | Ln (Total industrial area of CT in square meters) | 0.000 | 15.740 | 3.870 |
| Institutional Area | Ln (Total institutional area of CT in square meters) | 0.000 | 16.460 | 9.907 |
| Entertainment Area | Ln (Total entertainment area of CT in square meters) | 0.000 | 16.320 | 2.226 |
| Land use mix | Land use mix = $\left[\frac{-\sum_k (p_k(ln p_k))}{lnN}\right]$, where $k$ is the category of land-use, $p$ is the proportion of the developed land area devoted to a specific land-use, $N$ is the number of land-use categories in a CT | 0.000 | 0.92 | 0.325 |
| Buildings | Ln (Total Number of buildings in CT) | 0.000 | 8.085 | 5.821 |
| Floors | Ln (Total number of floors in CT) | 0.000 | 8.559 | 6.526 |
| Apartment Listings Price | Average whole apartment listings price per night (USD) | 0.000 | 3500.000 | 99.841 |
| Private or Shared Room Listings Price | Average private or shared room listings price per night (USD) | 0.000 | 3408.330 | 54.821 |
| Crime | Total number of crimes in CT | 0.000 | 2363.000 | 220.765 |
| Total Fatality | Total number of fatalities in CT | 0.000 | 4.000 | 0.171 |
| Pedestrian Fatality | Total number of pedestrian fatalities in CT | 0.000 | 4.000 | 0.111 |
| Bike Fatality | Total number of bike fatality in CT | 0.000 | 2.000 | 0.015 |
| Motor Vehicle Fatality | Total number of motor vehicle fatality in CT | 0.000 | 3.000 | 0.046 |
| Total Injury | Total number of injuries in CT | 0.000 | 197.000 | 20.060 |
| Pedestrian Injury | Total number of pedestrian injuries in CT | 0.000 | 45.000 | 4.697 |
| Bike Injury | Total number of bike injury in CT | 0.000 | 26.000 | 1.895 |
| Motor Vehicle Injury | Total number of motor vehicle injury in CT | 0.000 | 169.000 | 13.483 |
| Street Length | Ln (Street length of all type in mile per CT) | -0.090 | 4.64 | 1.451 |
| Bike Length | Ratio of bike length to street length | 0.000 | 0.58 | 0.086 |
| Walk Score | Walk Score in CT | 0.000 | 100.00 | 87.680 |

| Variable Names | | Definition | CT Level | | |
|---|---|---|---|---|---|
| | | | Minimum | Maximum | Mean |
| | Bike Score | Bike Score in CT | 0.000 | 95.000 | 66.500 |
| | Transit Score | Transit Score in CT | 0.000 | 100.000 | 83.240 |
| | Distance to Time Square (m) | Ln (Distance to Times Square in mile from CT) | -3.290 | 4.24 | 3.022 |
| *Transportation Infrastructure* | | | | | |
| | Bike Share Station | Total number of bikeshare stations in CT | 0.000 | 7.000 | 0.162 |
| | Bus Stops | Total number of bus stops in CT | 0.000 | 21.000 | 1.614 |
| | Subway Stations | Total number of subway stations in CT | 0.000 | 6.000 | 0.228 |
| | Taxi Car Station | Total number of taxi car stations in CT | 0.000 | 11.000 | 0.149 |
| | **Variable Names** | **Definition** | **Frequency (%)** | | |
| | Median Income | Low Median Income (Median income < 50,000) | 43.9 | | |
| | | Moderate Median Income (50,000 <= Median income <= 80,000) | 37.2 | | |
| | | High Median Income (Median income > 80,000) | 18.9 | | |
| | Historic District | Presence of listings on historic district or not | 29.8 | | |
| | Season | Spring (March-May) | 26.1 | | |
| | | Summer (June-August) | 25.0 | | |
| | | Fall (September-November) | 25.6 | | |
| | | Winter (December-February) | 23.2 | | |

$$AIC = 2k - 2ln(L) \tag{3.12}$$

where $k$ is the estimated number of parameters and $L$ denotes the maximized value of likelihood function for a given empirical model.

The empirical equation of BIC is:

$$BIC = -2ln(L) + K\, ln(Q) \tag{3.13}$$

where $ln(L)$ denotes the log likelihood value at convergence, $K$ denotes the number of parameters, and $Q$ represents the number of observations. Many of the earlier studies suggested that the BIC is the most consistent information criterion (IC) among all other traditionally used ICs (AIC, AICc, adjusted BIC) for number of segments selection in latent class models (Anowar, Yasmin, Eluru, & Miranda-Moreno, 2014; Bhat, 1997; Collins, Fidler, Wugalter, & Long, 1993; Dey, Anowar, Eluru, & Hatzopoulou, 2018a; Eluru, Bagheri, Miranda-Moreno, & Fu, 2012; Nashad et al., 2016; Nylund, Asparouhov, & Muthén, 2007; Yasmin, Eluru, & Ukkusuri, 2014). The advantage of using BIC is that it imposes substantially higher penalty than other ICs on over-fitting. The model with the lowest AIC and BIC value is the preferred model. The BIC and AIC values for the final specifications of all the models are presented in Table 3.2. Based on these values, copula models outperformed independent NB and multivariate NB model while mixed Gumbel copula with dependency profile parameterization model outperformed other copula models. The copula model BIC comparisons confirm the importance of accommodating dependence between full apartment and private/shared room count events in the macro-level analysis.

### 3.4.2 Estimation Results

We provide a discussion of results for the Mixed Gumbel copula model to present the effect of exogenous variables that presented in Table 3.3. To discuss the results briefly, 2nd and 3rd column of Table 3.2 represents the full apartment and private/shared room listings counts estimates respectively. Reader must note that a positive (negative) sign indicates that potential listings count increases (decreases) for the considered variable groups.

**Table 3.2: Model Fit Measures**

| Model | lnL | K | Q | BIC | AIC |
|---|---|---|---|---|---|
| ***Negative Binomial Count Models*** | | | | | |
| Negative Binomial | -45488.207 | 31 | 10230 | 91262.639 | 91038.414 |
| Multivariate Negative Binomial | -44924.200 | 32 | 10230 | 90143.859 | 89912.400 |
| ***Copula Model without Parameterization*** | | | | | |
| Copula (FGM) | -44725.867 | 31 | 10230 | 89737.959 | 89513.734 |
| Copula (Frank) | -44284.750 | 31 | 10230 | 88855.725 | 88631.500 |
| Copula (Gumbel) | -44278.714 | 31 | 10230 | 88843.653 | 88619.428 |
| Copula (Clayton) | -44498.249 | 31 | 10230 | 89282.723 | 89058.498 |
| Copula (Joe) | -44451.294 | 31 | 10230 | 89188.813 | 88964.588 |
| ***Copula Model with Parameterization*** | | | | | |
| Copula Parameterization (Frank) | -44185.310 | 34 | 10230 | 88684.545 | 88438.620 |
| Copula Parameterization (Gumbel) | -43832.890 | 36 | 10230 | 87998.171 | 87737.780 |
| Copula Parameterization (Joe) | -45104.380 | 33 | 10230 | 90513.452 | 90274.760 |
| ***Copula Mixed Model*** | | | | | |
| Copula Mixing (Gumbel) | -44289.455 | 32 | 10230 | 88874.369 | 88642.910 |
| **Mixed Copula Parameterization (Gumbel)** | **-43567.729** | **38** | **10230** | **87513.956** | **87217.458** |

3.4.2.1 Sociodemographic Characteristics

In terms of sociodemographic characteristics, the estimates indicate that both full apartment and private/shared room listings count are positively associated with higher population density at the census tract level. Employment density estimates indicate that CT with high job density are likely to experience more listings in both kinds e.g. apartment and private or shared room. It is expected that census tract with more job opportunity will attract individuals from other city or state to attend a job interview or presentation (see similar results (Deboosere et al., 2019; Sperling, 2015)). Also, an increase proportion of moderate income group in a CT increases the likelihoods of using their home as an AirBnB listing in count model components for both listings type. Moderate income family try to overcome their economic issues by giving permission to AirBnB platform to use their home as accommodation for tourists (Sperling, 2015).

**Table 3.3: Copula Count Mixed Model Results (Gumbel)**

| Variable Names | Apartment | | Private or Shared Room | |
|---|---|---|---|---|
| | **Estimate** | **t-stat** | **Estimate** | **t-stat** |
| Constant | -8.886 | -58.764 | -4.541 | -33.862 |
| ***Sociodemographic Characteristics*** | | | | |
| Population Density | 0.438 | 11.485 | 0.595 | 15.097 |
| Employment Density | 1.747 | 20.052 | 0.878 | 8.372 |
| Moderate Income (Base: Low and High Income) | 0.145 | 6.475 | 0.081 | 3.418 |
| ***Built Environment and Land Use Attributes*** | | | | |
| Average Listings Price | 1.628 | 68.571 | 0.984 | 57.448 |
| *Standard Deviation* | -- | -- | 0.068 | 11.147 |
| Point of Interests, Park and Recreational Centers | 0.267 | 4.481 | 0.285 | 5.382 |
| Restaurants and Sidewalk Café | 0.4678 | 1.848* | - | - |
| Historic District | 0.364 | 16.087 | 0.179 | 7.281 |
| Residential Density | 0.352 | 5.104 | 0.402 | 5.609 |
| Entertainment Area | 0.9416 | 3.575 | -- | -- |
| Land Use Mix | 0.528 | 6.390 | 0.785 | 8.585 |
| NonMV Score (Average of Walk Score and Bike Score) | 1.786 | 25.956 | 1.041 | 12.704 |
| ***Transportation Infrastructure*** | | | | |
| Bus Stops and Subway Stations | 0.251 | 4.699 | 0.232 | 4.120 |
| ***Road Network Characteristics*** | | | | |
| Bike Length Density | 1.195 | 11.547 | 0.826 | 7.531 |
| Distance to Time Square | -0.3652 | -22.673 | -0.345 | -22.026 |
| *Standard Deviation* | 0.132 | 13.464 | -- | -- |
| ***Seasonal Effect*** | | | | |
| Summer and Fall | 0.136 | 6.565 | 0.095 | 4.235 |
| *Standard Deviation* | -- | -- | 0.122 | 1.796* |
| *Dispersion parameter* | 0.558 | 34.411 | 0.806 | 33.200 |
| **Copula Parameter** | **Estimate** | | **t-stat** | |
| Constant | 0.967 | | 16.054 | |
| Average Listings Price | 0.221 | | 36.048 | |
| Historic District | 0.904 | | 11.931 | |
| Population Density | 0.266 | | 4.326 | |
| Point of Interests Park and recreational centers | 0.202 | | 13.227 | |
| Distance to Time Square | -0.245 | | -21.743 | |
| Log-likelihood at convergence | -43567.729 | | | |

* = attribute insignificant at 90% significance level

### 3.4.2.2 Built Environment and Land Use Attributes

With respect to built environment characteristics, average listings price in a CT is found to be a significant determinant with a positive impact. It is expected that people are more encouraged to be a host of AirBnB with higher listings price and to do so it will affect the rent of that neighborhood (D. Wang & Nicolau, 2017). As expected with increasing the number of point of interests together with various amusement park and recreational centers within a CT will increase the likelihood of listings count of that particular CT (Yang & Mao, 2020). Since NYC is one of the most tourist visited city, it is expected that people who visit Times Square are likely to find accommodation in the vicinity.

The variables considering built environment characteristics reveals that higher number of restaurants and sidewalk cafe are likely to result in increased number of apartment listings only. From Table 4, land use attributes play an important role in listings count for NYC. AirBnB listings situated in historic district increases the likelihood of apartment listings count. With respect to land use attributes, there are several attributes that found to be significant determinants for both the listings type. The noteworthy determinants regarding land use attributes that positively impact AirBnB listings count in a particular CT are region that used as a residential and entertainment zone. Also, mixed land use is positively associated with apartment and private/shared room listings count. As expected, tourist's with personal or professional purpose will have more interest on staying region with land use with various dimension. Considering transportation effect on land use attributes, there is a clear scenario captures that increasing proportion of non-motorize vehicle score has positive association with both listings counts at the CT level.

#### 3.4.2.3 Transportation Infrastructure

In terms of transportation facility, public transport system either in bus or subway format will increase the likelihood of both AirBnB listings type count. This result can be easily comparable with practical scenario that more public transit system will facilitate guest's stay in AirBnB much easier since major portion of tourist come from different city or state to stay in AirBnB. This transit facilitation criteria would be one of the major criteria for tourist to choose one particular AirBnB listings that may encourage host to establish their home as AirBnB listings.

#### 3.4.2.4 Road Network Characteristics

An increase in the length of bicycle route within the census tract (CT) results in an increased likelihood of the increasement of the AirBnB listings for both types (apartment and private/shared room). Visitors choose AirBnB listings of both apartment and private/shared room located in a particular CT that bring them closer to Time Square as highlighted by negative coefficient of CT centroid distance to Time Square.

#### 3.4.2.5 Temporal Effect

We tried different seasons along with interaction of seasons in the model. As expected, during warm and dry weather of summer and fall season have attracted tourist to travelling on NYC.

#### 3.4.2.6 Random Parameter Effect

The unobserved heterogeneity of the impact of average listings price is significant for private/shared room listings highlighting that the count associated with the private/shared room listings varies substantially across average listings price. Similar effect is also found for variable distance between Times Square to each CT for full apartment.

3.4.2.7 Dependency Effect

As specified in the result section, the estimated Gumbel copula based mixed bivariate NB model provides the best model fit in incorporating the correlation between the full apartment and private/shared room listings count events. In the last row panel of Table-4, dependency effects across various determinants by two listings types is presented in Copula parameter section. The various exogenous variables that contribute to the dependency include Average Listings Price, Historic District, Population Density, Point of Interests Park and recreational centers and Distance to Time Square. For the Gumbel copula, the first four attributes show the positive dependency while Distance to Times Square to listings attribute shows a negative dependency across CTs that supporting our hypothesis that the dependency profile varies across CTs. This provides support to our hypothesis that the dependency structures are not constant across all CTs and also the coefficient sign and magnitude reflects whether a variable increase or reduces the dependency and by how much. The proposed framework by permitting for such parameterizations allows us to improve model estimation results.

## 3.5 Policy Analysis

3.5.1 Elasticity Effects

From the sample of data not used for estimation, data for 5 months was randomly selected for each census tract for policy analysis. The elasticity effects are computed by evaluating the percentage change in counts in response to increasing the value of significant exogenous variables from best fit model by 10% (Used in safety studies ((J. Lee, Yasmin, Eluru, Abdel-Aty, & Cai, 2018)). The computed elasticities are presented in Table 3.4 (see (Eluru & Bhat, 2007)) for details methodology of elasticity calculations). Results presented in the Table-5 represent the percentage change of AirBnB counts due to 10% change in the independent variable. For example, the elasticity estimate for average AirBnB price variable indicates that a 10% increase in price will result in a 127.299% increase in Apartment count and a 56.914% increase in private or shared room count. All the other results can be interpreted similarly.

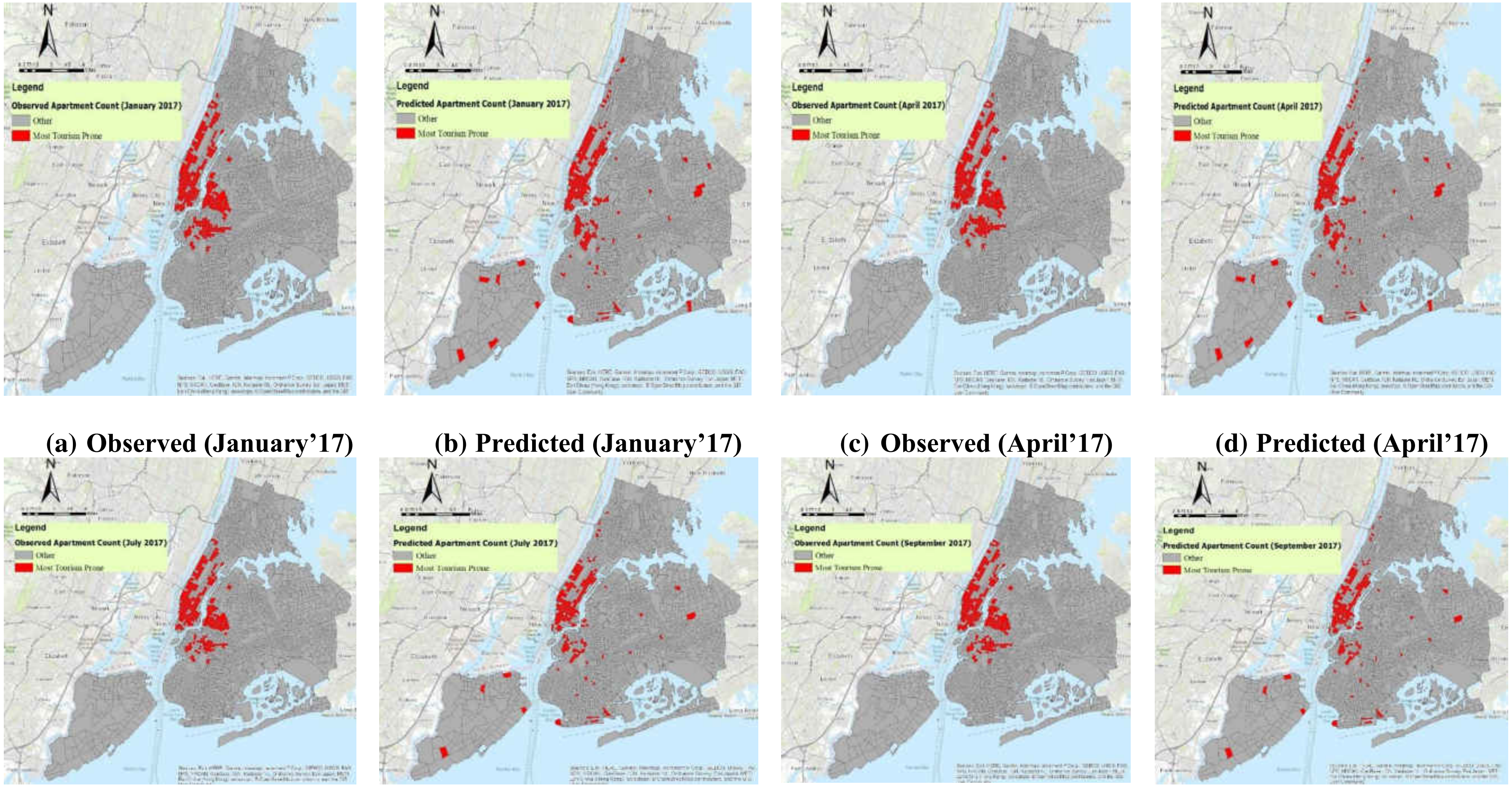
Based on elasticity effects results in Table 3.4, following observations can be made. First, elasticity effect on two dependent variables are different for various exogenous variables. Second, rank order of the top five important variable in terms of increasement for the expected number of both apartment and private or shared room counts include: average AirBnB price in CT, CT lies on historic district or not, median income per CT, effect of season and employment density. Third, increasing distance to Times Square from CT is the only variable which have negative impact on AirBnB counts for both types as expected. Fourth, private or shared room have higher elasticities relative to apartment counts for Point of Interests, park and recreational centers, land use mix and distance to Times Square from each CT. Fifth, an interesting finding from variation in elasticity effects for various exogenous variables is that with the increasing distance from Times Square to each CT have almost nine times more variation in elasticity for private or shared room than apartment count. Overall, the elasticity analysis results provide an illustration on how the proposed model can be applied to determine the critical factors contributing to increase in apartment and private or shared room AirBnB counts.

**Table 3.4: Elasticity Effects**

| Variable Names | Apartment | Private or Shared Room |
|---|---|---|
| ***Sociodemographic Characteristics*** | | |
| Population Density | 3.876 | 3.143 |
| Employment Density | 12.084 | 5.259 |
| Moderate Income | 15.123 | 8.202 |
| ***Built Environment and Land Use Attributes*** | | |
| Average Listings Price | 127.299 | 56.914 |
| Point of Interests, Park and Recreational Centers | 5.378 | 10.788 |
| Restaurants and Sidewalk Café | 0.990 | -- |
| Historic District | 25.214 | 29.645 |
| Residential Density | 8.641 | 2.545 |
| Entertainment Area | 0.082 | -- |
| Land Use Mix | 1.284 | 1.341 |
| NonMV Score (Average of Walk Score and Bike Score) | 8.714 | 6.092 |
| ***Transportation Infrastructure*** | | |
| Bus Stops and Subway Stations | 1.331 | 0.359 |
| ***Road Network Characteristics*** | | |
| Bike Length Density | 0.629 | 0.206 |
| Distance to Time Square (m) | -3.220 | -27.677 |
| ***Temporal Attributes*** | | |
| Season: Summer and Fall | 13.598 | 9.499 |

### 3.5.2 Spatial Distribution of Hotspots

To illustrate how our model can be used to identify zones with high tourist demand, we conduct a hot zone identification exercise. Hot zones are defined as the census tracts within the top 10 percentile of demand. With this definition, we compare the observed hot zones with the predicted hot zones from our model. We present the results for four months of data from the four seasons - January from Winter, April from Spring, July from Summer and September from Fall. The results are presented in Figure 3.4 for apartment listing type.




**(a) Observed (January'17)** **(b) Predicted (January'17)** **(c) Observed (April'17)** **(d) Predicted (April'17)**

**(e) Observed (July'17)** **(f) Predicted (July'17)** **(g) Observed (September'17)** **(h) Predicted (September '17)**

**Figure 3.4: Spatial Distribution of Most Tourist Zone as Apartment AirBnB Counts of NYC**

From the spatial distribution for observed AirBnB count, it is clearly seen that top tourist spot are dispersed throughout the Manhattan and few parts of Brooklyn borough for apartment. Further, the model predictions are reasonably close to the observed patterns.

### 3.6 Summary

In the first part of the dissertation, considering AirBnB as sharing accommodation system, we aim to analyze these three dimensions. First, by developing a model framework to count AirBnB listings at census tract level to capture the snapshot of accommodation supply for tourist in NYC. Second, capture the unobserved heterogeneity in the model together with correlation between those matrices. For this study purpose, a copula based negative binomial count model system is developed that implicitly recognizes shared common observed and unobserved factors for two types of AirBnB listings e.g. Apartment and Private or shared room in a census tract level. Given these afore-mentioned implications, the proposed research conducts a comprehensive analysis of AirBnB listings in New York City region drawing on data from January 2015 to September 2017. We found that mixed Gumbel copula model with dependency profile parameterization outperformed other copula models along with independent and negative binomial model. Finally, we validate the model by predicting AirBnB counts by it's two types and found that the predicted results are closely aligned for high demand destinations. This analysis will allow City planners and operators to better evaluate and improve tourism systems. We also conducted elasticity effects based on the best fit model results on validation dataset and found the top five important variable in terms of influencing the expected number of both apartment and private or shared room as: average AirBnB price in CT, CT lies on historic district or not, median income per CT, effect of season and employment density.

# CHAPTER 4: FRAMEWORK FOR ESTIMATING BIKESHARE ORIGIN DESTINATION FLOWS USING A MULTIPLE DISCRETE CONTINUOUS SYSTEM

## 4.1 Introduction

Transportation field is undergoing a transformative change in response to several technological innovations in recent years. A product of these technological transformations is the adoption of shared mobility systems such as bikesharing (such as CitiBike in New York City), car sharing (such as Zipcar or Car2Go), ridesourcing (such as Uber and Lyft) and ride-splitting (such as dynamic carpooling in urban regions). As highlighted in a recent Transit Cooperative Research Program report (1), understanding shared mobility adoption and usage provides an unprecedented opportunity to address existing mobility shortcomings in urban regions. In fact, public transit agencies and transportation planning agencies can enhance mobility and accessibility by incorporating these shared mobility alternatives within their planning frameworks. Among the shared mobility alternatives, bike sharing offers a sustainable transportation alternative in urban core regions and could be an effective solution to the last mile problem (2). In our research, we focus our attention on developing a research framework to contribute to our understanding of bikeshare origin destination flows.

About 1000 cities around the world have a bikeshare system in operation or in consideration for development (3). As reported by Richter, 2018 (4), the number of public use bicycles in the world have nearly quadrupled between 2013 and 2016. Further, a recent national association of city transportation officials (NACTO) report highlighted that of the 88 million trips made by bike share users in US between 2010-2016, 28 million were trips from 2016 only (5). Given the burgeoning growth in bikeshare system installations and their growing adoption for trip making, it is important to develop modeling frameworks to understand bike share demand flows in the system. An important mechanism for enhancing system adoption and usage is the development of current performance metrics (see (6)). As bikesharing is an emerging transportation mode, the current approaches being employed for analyzing system usage and performance measure are still in their infancy. In this study, we propose an enhanced framework to estimate usage dimensions of bikesharing at a system level.

To be sure, several earlier research efforts have explored approaches to model system level usage (7-10). These research studies examine the impact of bicycling infrastructure, land use and built environment, public transportation infrastructure, temporal and meteorological attributes on bikeshare system usage (defined as station level arrivals and departures). These models can be viewed as analogous to the trip generation and trip attraction models in the traditional trip based modeling approach. While these models provide important insights on variables affecting bikeshare usage, they do not provide any information on the system level flows between the stations. To elaborate, the approaches provide trip end information without the trip distribution relationship. To address this shortcoming, recent research has developed destination choice models at an individual trip level (7; 11; 12). In these studies, for every individual trip the choice of destination given the origin station is analyzed using a random utility based approach. The models developed at an individual trip level can be employed to obtain aggregate estimates of trip distribution (analogous to the gravity model). However, such an aggregation approach is purely a statistical construct and lacks behavioral support.

In this study, we remedy this drawback, by developing a model framework for bikeshare system usage as well as origin destination flows. Towards this end, we characterize system demand as origin level demand (number of trips) and allocate these trips to various destination stations (number of trips from an origin to destination) in the system. For the first variable, a linear mixed model is developed while the second variable is analyzed using a multiple discrete continuous model system that implicitly recognizes that the total arrivals across stations should add up to the total number of trips leaving the origin. The proposed framework is implemented for the New York City bikeshare system (CitiBike). The data drawn for the exercise includes bikeshare trips from January 2017 through June 2017 for the CitiBike system.

The rest of the chapter is organized as follows: The following section presents the methodological framework adopted in the analysis while section 3 provides a detailed description of the dataset with sample formation technique. Model results are presented in the fifth section followed by the policy analysis. Final section comprises with the concluding statements.

## 4.2 Econometric Modeling Framework

### 4.2.1 Linear Mixed Model for Station Level Weekly Origin Demand

The station level weekly origin demand variable is a continuous value and can be analyzed using linear regression models. However, the traditional linear regression model is not appropriate to study data with multiple repeated observations. In our empirical analysis, we observe the weekly demand at the same station for five weeks. Hence to recognize this, we employ a linear mixed modeling approach that builds on the linear regression model while incorporating the influence of repeated observations from the same station. The linear mixed model collapses to a simple linear regression model in the absence of any station specific effects.

Let $q = 1, 2, \ldots, Q$ be an index to represent each station ($Q$=574), $W = 1, 2, \ldots, 5$ be an index to represent the various weeks of data compiled for each station. The dependent variable (weekly demand) is modeled using a linear regression equation which, in its most general form, has the following structure:

$$y_{qw} = \beta X + \varepsilon$$

where $y_{qw}$ is the natural logarithm of weekly demand, $X$ is an L×1 column vector of attributes and the model coefficients, $\beta$, is an L×1 column vector. The random error term, $\varepsilon$, is assumed to be normally distributed across the dataset. In our analysis, the repetitions over weeks can result in common unobserved factors affecting the dependent variable. While a full covariance matrix can be estimated for the unobserved correlations, as we are selecting 5 random weeks from a sample of 26 weeks for each station, we decided to employ a simpler covariance structure. The exact functional form of the covariance structure assumed is shown below:

$$\boldsymbol{\Omega} = \begin{pmatrix} \boldsymbol{\sigma^2 + \sigma_1^2} & \boldsymbol{\sigma_1} & \cdots & \boldsymbol{\sigma_1} \\ \boldsymbol{\sigma_1} & \boldsymbol{\sigma^2 + \sigma_1^2} & \cdots & \boldsymbol{\sigma_1} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\sigma_1} & \boldsymbol{\sigma_1} & \cdots & \boldsymbol{\sigma^2 + \sigma_1^2} \end{pmatrix}$$

The covariance structure restricts the covariance across all five records to be the same. The parameters estimated in this correlation structure are $\sigma$ and $\sigma_1$ . The parameter σ represents the error variance of ε, $\sigma_1$ represents the common correlation factor across weekly records. The models are estimated in SPSS.

4.2.2 The MDCEV Model Structure for Destination Choice

According to Bhat and Eluru (Bhat & Eluru, 2010), we consider the following functional form for utility in this paper, based on a generalized variant of the translated CES utility function:

$$V(x) = \sum_{i=1}^{I} \frac{\gamma}{\alpha}\, \psi_i \left\{ \left( \frac{x_i}{\gamma} + 1 \right)^{\alpha} - 1 \right\} \qquad (4.1)$$

where $V(x)$ is a quasi-concave, increasing, and continuously differentiable function with respect to the consumption quantity ($I$x1)-vector $x$ ($xi \geq 0$ for all $i$), and $\psi_i$ associated with destination station $i$. $\psi_i$ represents the baseline marginal utility ($\psi_i$> 0 for all $i$), $\gamma$ is a translation parameter ($\gamma$ should be greater than zero) which enable corner solutions while simultaneously influencing satiation and $\alpha$ influences satiation ($\alpha \leq 1$).

The KT approach employs a direct stochastic specification by assuming the utility function $V(\boldsymbol{x})$ to be random over the population. A multiplicative random element is introduced to the baseline marginal utility of each good as follows:

$$\psi\ (z_i, \epsilon_i) = \exp\ (\beta' z_i + \epsilon_i) \tag{4.2}$$

where $z_i$ is a set of attributes characterizing destination station $i$, $\beta$ corresponds to a column vector of coefficients, and $\epsilon_i$ captures idiosyncratic (unobserved) characteristics that impact the baseline utility for good.

The overall random utility function of Equation (1) then takes the following form:

$$V(x) = \sum_{i=1}^{I} \frac{\gamma}{\alpha} \exp\ (\beta' z_i + \epsilon_i) \left\{ \left( \frac{x_i}{\gamma} + 1 \right)^{\alpha} - 1 \right\} \tag{4.3}$$

Following Bhat (Bhat, 2005, 2008), consider an extreme value distribution for $\epsilon_i$ and assume that $\epsilon_i$ is independent of $z_i$ ($i$ = 1, 2, …, $I$). The $\epsilon_i$'s are also assumed to be independently distributed across alternatives with a scale parameter normalized to 1. Due to the common role of $\gamma$ and $\alpha$, it is very challenging to identify both $\gamma$ and $\alpha$ in empirical application (see Bhat, (Bhat, 2008)). Hence, only $\gamma$ or $\alpha$ are estimated.

When the α- profile is used the utility simplifies to:

$$V(x) = \sum_{i=1}^{I} \frac{1}{\alpha} \exp\ (\beta' z_i + \epsilon_i) \{ (x_i + 1)^{\alpha} - 1 \} \tag{4.4}$$

When the $\gamma$ - profile is used the utility simplifies to:

$$V(x) = \sum_{i=1}^{I} \gamma \exp(\beta' z_i + \epsilon_i) \, ln\left(\frac{x_i}{\gamma} + 1\right) \tag{4.5}$$

In this study, $\gamma$ - profile is used.

The probability that an origin station has flows to the first M destination stations ($M \geq 1$) is:

$$P(e_1^*, e_2^*, e_3^*, \dots, e_M^*, 0,0, \dots,0) = \left[\sum_{n=1}^{M} d_n\right]\left[\sum_{n=1}^{M} \frac{1}{d_n}\right]\left[\frac{\prod_{n=1}^{M} e^{V_n}}{(\prod_{m=1}^{K} e^{V_i})^M}\right](M-1)! \tag{4.6}$$

where $(\sum_{n=1}^{M} d_n)\left(\sum_{n=1}^{M} \frac{1}{d_n}\right)$ is defined as Jacobian form for the case of equal unit prices across goods (Bhat, (Bhat, 2008)).

Where, $\boldsymbol{d_n} = \left(\frac{\boldsymbol{1-\alpha}}{\boldsymbol{e_n^*+\gamma}}\right)$

Unlike the traditional MDCEV model, in our context, the number of alternatives are substantially larger. Hence, we resort to estimating a single utility across alternatives (analogous to how multinomial logit based location choice models are estimated with a single utility equation).

## 4.3 DATA

### 4.3.1 Data Source

New York's CitiBike system is one of the major public bikeshare systems around the world and the largest in the United States. The CitiBike system was launched in May 2013 with 330 stations and 6,000 bicycles in the lower half of Manhattan and some part of northwest Brooklyn. In 2017, the system size expanded to 750 stations with 12,000 bicycles. According to CitiBike report, the number of annual subscribers were nearly 130,000 on July 2017. The trip itinerary dataset (from January 2017 to June 2017) of the CitiBike system is the primary data source employed (https://www.citibikenyc.com/system-data). The ridership dataset provides information on start and end time of trips, their origin and destination, geographic coordinates of stations (latitude and longitude), travel time or trip duration, user types, and age and gender for members' trips. The trip data was augmented with other sources including: (1) built environment attributes such as number of restaurants and park area derived from New York City open data (https://nycopendata.socrata.com); (2) socio-demographic characteristics at the census tract/zip code level gathered from US 2010 census; (3) the weather information corresponding to the Central Park station retrieved from the National Climatic Data Center (http://www.ncdc.noaa.gov/data-access).

#### 4.3.2 Sample Formation

For the given study period, the total number of available stations in CitiBike system was 644. Initially, we aggregated weekly trip data for each week (total 26 weeks) from each origin station to every possible destination station (643). The processing of large sample of trip data with other station level variables is substantially time-consuming and significantly increases the model run times (Faghih-Imani & Eluru, 2017a). To obtain a reasonable sample size for model estimation, 5 weeks trip data for each origin were randomly selected. As a result of the random week selection, we ended up having 70 stations with no trips. So, we eliminated those 70 stations (about 10% trips) from both origin and destination choice set. Finally, we had 574 stations for analysis. The location of CitiBike stations (574 stations) considered in this study is presented in Figure 4.1. We organized the dataset into two dimensions for our analysis; 1) For station level demand (aggregating total weekly trip at origin level) and 2) Trip distribution from origin to destination (aggregating weekly trip at O-D pair level). A summary of the dependent variable and independent variable data compilation procedure is presented in Figure 4.2.

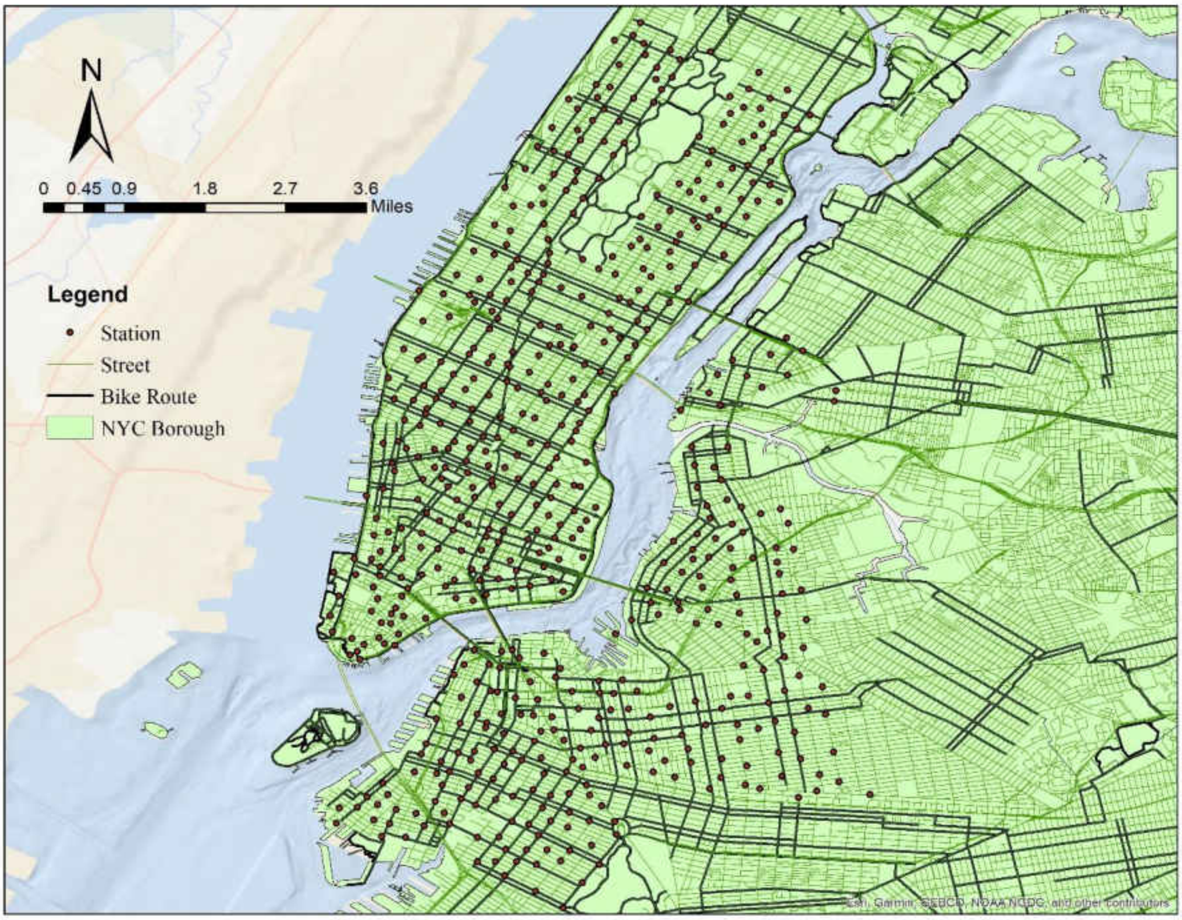



**Figure 4.1: NYC's Bicycle-Sharing System (CitiBike)**

4.3.3 <u>Independent Variable Generation</u>

Several independent variables were generated in our study (see Figure 4.2). These can be grouped into four categories: 1) Trip attribute, 2) Socio-demographic variables, 3) Bicycle and transportation infrastructure variables, 4) Weather attributes, 5) Temporal attributes and 6) Land use and built environment variables. <u>Trip attribute</u> includes the network distance between each origin-destination station pair estimated using the shortest path algorithm. While the actual trip might involve a different route, the shortest network distance would be an appropriate indicator of the distance traveled. The socio-demographic attributes considered are population density, job density and establishment density. Population information was collected from US census 2010 and projected for 2017 at the census tract level. Job density data was estimated at the census tract level while establishment density was calculated at the zip code level for 2016.

Bicycle and transportation infrastructure variables include CitiBike station attributes, bike route length, and public transit stations. For these attributes a 250-meter buffer around each station was created. The 250-meter buffer seems a reasonable walking distance based on the distances between CitiBike stations and the dense urban form of New York City (Kaufman, Gordon-Koven, Levenson, & Moss, 2015). The variables created at the buffer level include length of bike routes, length of roads (minor and major roads). The number of CitiBike stations and total dock's capacity within 250 meter buffer (excluding the station considered and its capacity) were estimated to capture the impact of neighboring stations on cycling trips. Number of subway stations and bus stops in the 250 meter buffer were generated to examine the influence of public transit on cyclist's preference of destination station. Weather variables include average temperature, relative humidity and precipitation over the week. Several interaction variables were also created. Seasonality is the only temporal variable considered. We consider winter (January-March) and Spring (April-June) as dummy variables.

Finally, several land use and built environment variables were considered including the number of facilities (schools, colleges, hospitals), the number of point of interests (museums, shopping malls), and the number of restaurants (including coffee shops and bars), total area of parks and commercial space (office, industry, retail) within 250 meter buffer, station elevation, and distance of destination from Times Square. Non-motorized vehicle score (average of walk score and bike score) and transit score associated with each CitiBike station was considered at the census tract level.

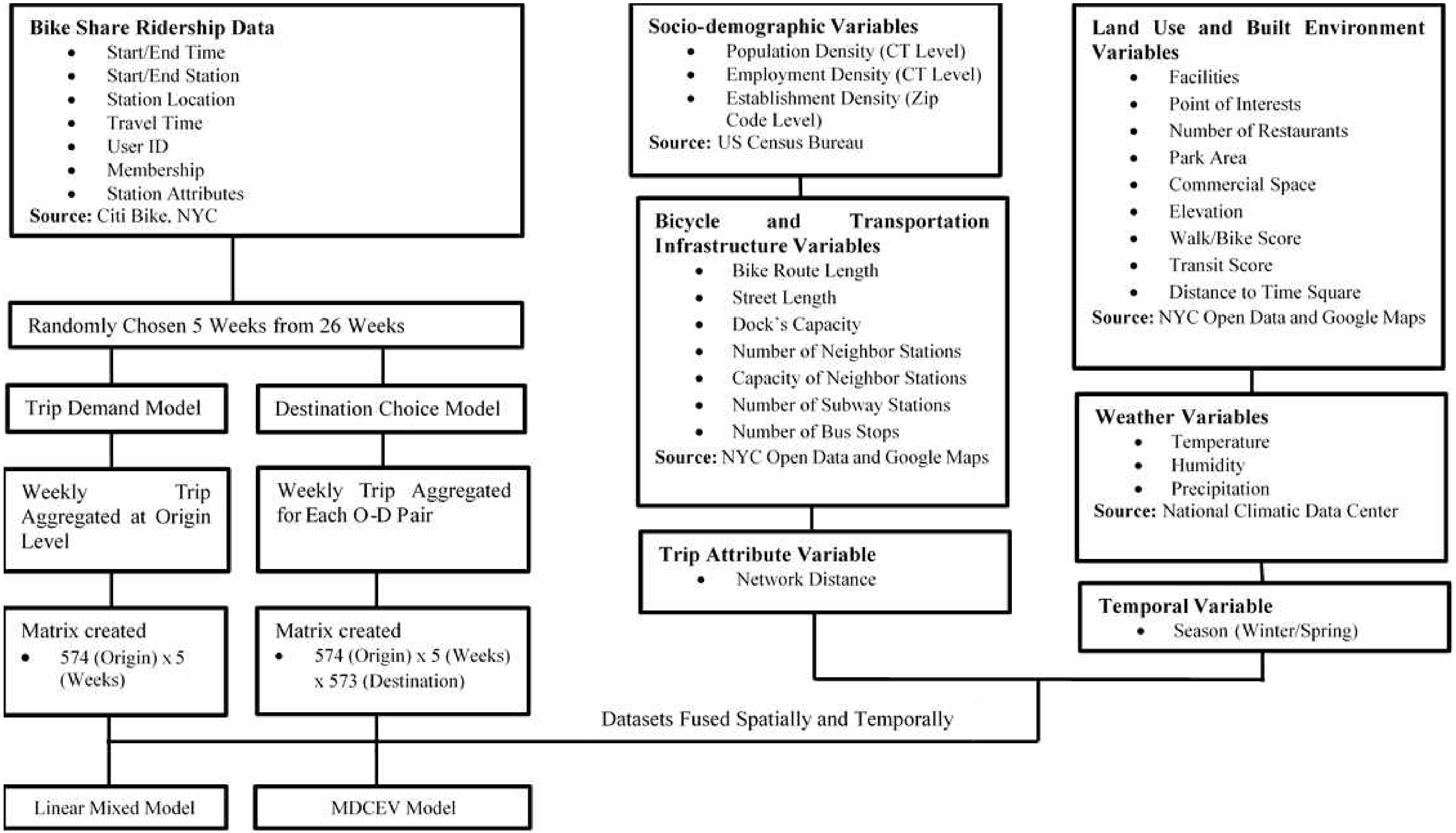
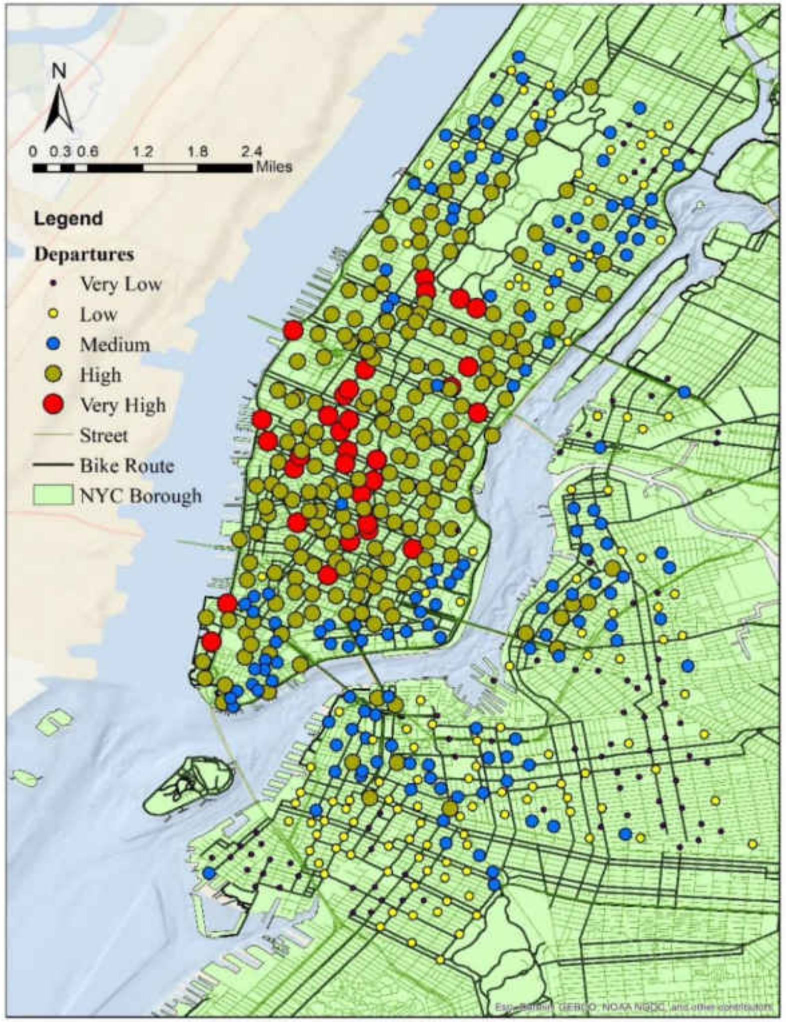
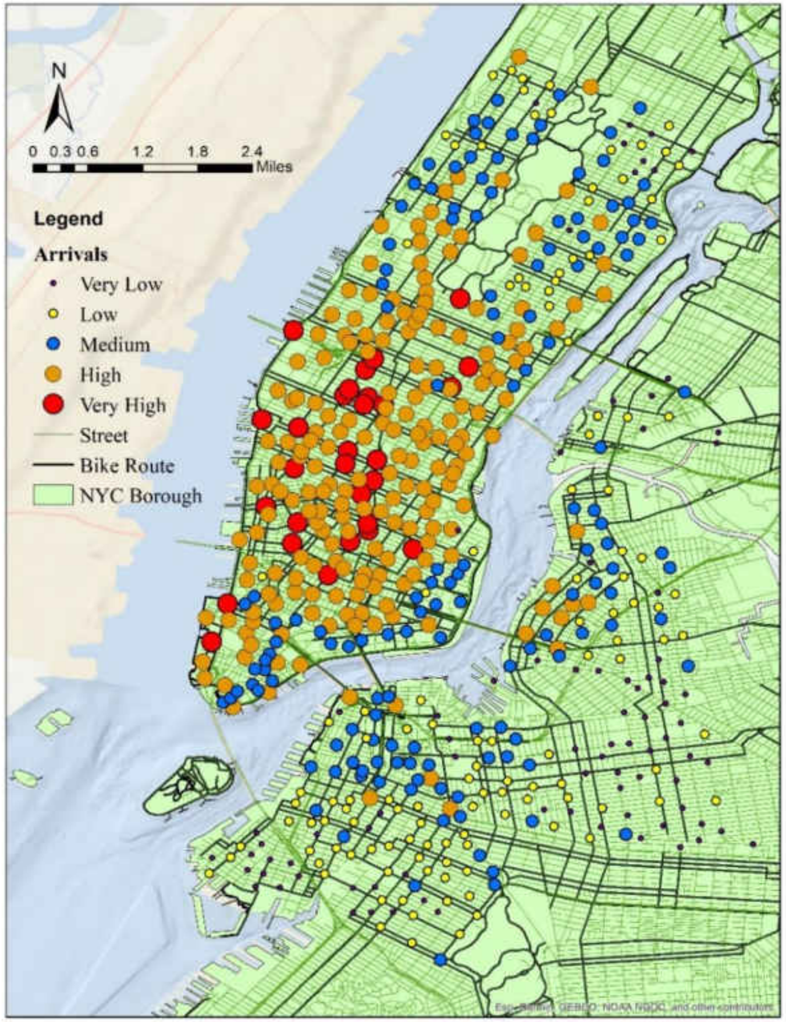



**Figure 4.2**: **Data Formation Flow Chart**




(a) Trip generation at origin stations.




(b) Trip attraction at destination stations.

**Figure 4.3: Bike Sharing Trips in NYC's CitiBike System**

4.3.4 Descriptive Analysis

A descriptive summary of the analysis sample is presented in Table 4.1. The number of weekly trips generated and attracted at each station is presented in Figure 4.3. In Figure 4.3, the number of trips generated (3a) and attracted (3b) to each station is categorized in five classes: Very Low (number of trips less than 500), Low (500-1000), Medium (1000-2000), High (2000-5000) and Very High (more than 5000). Overall, the visualization provides a brief overview of bicycle flows in NYC using the CitiBike system.

**Table 4.1: Descriptive Summary of Sample Characteristics**

| Continuous Variables | Min | Max | Mean | Std. Deviation |
|---|---|---|---|---|
| **Dependent Variable** | | | | |
| **Trip Demand** | | | | |
| Total Trip (Weekly per Origin) | 1.00 | 3726.00 | 402.17 | 390.06 |
| **Destination Choice** | | | | |
| Alternative Destination Chosen | 1.00 | 354.00 | 111.69 | 65.79 |
| Total Trip (Weekly O-D Pair) | 1.00 | 175.00 | 3.60 | 5.15 |
| **Independent Variables** | | | | |
| **Trip Characteristics** | | | | |
| Network Distance (km) (x 10^$^{-2}$) | 0.05 | 0.41 | 0.14 | 0.08 |
| **Socio-demographic** | | | | |
| Population Density (People per $m^2$ x 10^$^{-4}$) | 0.00 | 0.87 | 0.26 | 0.17 |
| Job Density (Jobs per Person) | 0.00 | 0.90 | 0.66 | 0.17 |
| Number of Establishment (per m2x 10^$^{-4}$) | 0.00 | 1.20 | 0.09 | 0.14 |
| **Bicycle and Transportation Infrastructure** | | | | |
| Length of Bicycle Facility in 250m Buffer (m x 10^$^{-4}$) | 0.00 | 0.91 | 0.24 | 0.17 |
| Length of Street in 250m Buffer (m x 10^$^{-4}$) | 0.14 | 0.84 | 0.38 | 0.10 |
| Station Capacity (x 10^$^{-2}$) | 0.07 | 0.67 | 0.32 | 0.10 |
| Number of Neighboring Station in 250m Buffer (x10^$^{-1}$) | 0.00 | 0.50 | 0.11 | 0.10 |
| Capacity of Neighboring Station in 250m Buffer (x10^$^{-3}$) | 0.00 | 0.27 | 0.04 | 0.04 |
| Number of Subway Stations in 250m Buffer (x10^$^{-1}$) | 0.00 | 0.70 | 0.06 | 0.09 |
| Number of Bus Stops in 250m Buffer (x10^$^{-1}$) | 0.00 | 1.10 | 0.22 | 0.22 |
| **Weather** | | | | |
| Temperature (°F) | 19 | 84 | 50.06 | 13.56 |
| Precipitation (in) | 0 | 3.02 | 0.16 | 0.44 |

| Continuous Variables | Min | Max | Mean | Std. Deviation |
|---|---|---|---|---|
| Humidity (%) | 26 | 98 | 61.44 | 17.5 |
| **Land Use and Built Environment** | | | | |
| Walk Score (x10^-2) | 0.69 | 1.00 | 0.97 | 0.05 |
| Transit Score (x10^-2) | 0.61 | 1.00 | 0.96 | 0.07 |
| Bike Score (x10^-2) | 0.45 | 0.95 | 0.85 | 0.09 |
| Number of Facilities in 250m Buffer (x10^-3) | 0.00 | 0.16 | 0.03 | 0.02 |
| Recreational Facilities in 250m Buffer | 0.00 | 2.00 | 0.08 | 0.30 |
| Number of Restaurants in 250m Buffer (x 10^-3) | 0.00 | 0.55 | 0.04 | 0.08 |
| Number of Sidewalk café in 250m Buffer (x10^-3) | 0.00 | 0.14 | 0.02 | 0.02 |
| Area of Parks in 250m Buffer (m2 x 10^-6) | 0.00 | 0.18 | 0.09 | 0.05 |
| Commercial Area in 250m Buffer (m2 x 10^-6) | 0.00 | 0.55 | 0.26 | 0.14 |
| Elevation (m x10^-3) | 0.00 | 0.16 | 0.04 | 0.03 |
| Distance to Time Square (m x 10^-5) | 0.58 | 1.32 | 0.52 | 0.28 |
| **Categorical Variables** | | | | |
| **Temporal** | **Percentage** | | | |
| Winter | 48.90 | | | |
| Spring | 51.10 | | | |

### 4.4 Estimation Results

In this section, estimation results from the two models are discussed. First, the results of the bikeshare demand model is discussed. Second, the trip distribution model results at destination level are discussed. The reader must note that we used same scaled parameter as presented in Table 1.

4.4.1 Trip Demand Model

4.4.1.1 Model Fit Measures

To evaluate weekly bikeshare demand at the origin station, a linear mixed model was estimated. The mixed model data fit was compared to the simple linear regression model data fit. The *Log-likelihood ratio* (LR) test statistic comparing these models was found to be 2015.0 which was higher than any corresponding chi-square value for 2 degrees of freedom ($\sigma$ and $\sigma_1$). Based on the LR test statistic, we can conclude that the linear mixed model offers the satisfactory fit for station level demand.

4.4.1.2 Results

The linear mixed model estimation results are presented in Table 3.2.

**Table 4.2: Linear Mixed Model Results**

| **Parameter** | **Estimates** | **t-stats** |
|---|---|---|
| Intercept | 1.253 | 3.273 |
| **Socio-demographic Attributes** | | |
| Job Density | 0.683 | 4.372 |
| **Bicycle Infrastructure and Transportation Attributes** | | |
| Station's Capacity | 2.468 | 8.407 |
| Number of Subway Stations in 250m Buffer | 0.383 | 2.491 |
| Bike Length in 250m Buffer | 0.871 | 5.524 |
| **Temporal Attributes** | | |
| Season: Winter | -0.784 | -53.378 |
| **Land Use and Built Environment Attributes** | | |
| Non-motorized vehicle score | 4.466 | 11.139 |
| Number of Facilities and Recreational Point in 250m Buffer | 3.256 | 4.158 |
| Distance to Time Square (m) | -18.116 | -16.599 |
| **Correlation Parameters** | | |
| $\sigma$ | 0.128 | 33.875 |
| $\sigma_1$ | 0.305 | 15.507 |
| **Restricted Log-Likelihood** | -1863.186 | |

*4.4.1.2.1 Socio-demographic Attributes*

Individuals are likely to make more trips using bikeshare in a location clustered with more job opportunities (see (Rixey, 2013; Wang et al., 2015) for similar results).

*4.4.1.2.2 Bicycle Infrastructure Variables*

People are more inclined to make trips from higher capacity (total number of bicycles) stations than lower capacity stations. Riders are more willing to make more trips from stations well served by bicycle facilities such as bicycle lanes (see (Buck et al., 2013) for similar results). As expected, number of subway stations positively impacts origin bike demand. This is plausible since bikeshare potentially serves as a last mile connection for some public transit users (similar results in (Nair et al., 2013)).

#### *4.4.1.2.3 Temporal variables*

There is a negative relationship between winter season and total weekly bicycle departures from a particular station compared to spring season. The result is expected as New York winter bikeshare usage is expected to be lower than spring bikeshare usage.

#### *4.4.1.2.4 Land Use and Built Environment Attributes*

This section highlights results regarding land use and built environment variables. Stations located in neighborhoods with high walkable and bikable facilities also increase bikeshare demand. Citibike stations near different facilities (schools, colleges, hospitals, office) and recreational locations (point of interests such as Times Square, museums, amusement parks, shopping malls.) increase demand. As expected, increasing distance from Time Square reduces bikeshare flows.

#### *4.4.1.2.5 Correlation Parameters*

The correlation parameters are statistically significant highlighting the role of common unobserved factors influencing the origin stations.

4.4.2 Destination Choice Model

4.4.2.1 Model Fit Measures

The final log-likelihood values for destination choice MDCEV model and equal probability MDCEV model are -534386813.50 and -597736907.30 respectively. The *log-likelihood ratio* (LR) test-statistic of comparison between the final model and the equal probability model is 126700187.60. The LR test-statistic value is significantly higher than the corresponding chi-square value for 20 additional degrees of freedom. Based on these values, we can see that the MDCEV destination choice model offers a reasonable fit.

4.4.2.2 Results

The best fit model results of destination choice are presented in Table 4.3.

**Table 4.3: MDCEV Model Results**

| **Parameter** | **Estimates** | **t-stats** |
|---|---|---|
| **Trip Attributes** | | |
| Network Distance (m) | -13.204 | -16014.381 |
| Network Distance x Winter | -0.847 | -867.060 |
| **Socio-demographic Attributes** | | |
| Population Density | 2.165 | 307.835 |
| Job Density | 0.607 | 696.644 |
| Establishment Density | 0.188 | 265.628 |
| **Bicycle Infrastructure and Transportation Attributes** | | |
| Station's Capacity | 1.397 | 852.412 |
| Bike Length in 250m Buffer | 0.588 | 1205.945 |
| Street Length in 250m Buffer | 0.003 | 3.150 |
| Number of Neighboring Stations in 250m Buffer | -0.467 | -151.695 |
| Capacity of Neighboring Stations in 250m Buffer | -0.448 | -52.984 |
| Number of Subway Stations and Bus Stops in 250m Buffer | 0.042 | 88.215 |
| **Land Use and Built Environment Attributes** | | |
| Transit Score | 1.604 | 824.049 |
| Non-motorized vehicle score | 5.259 | 2769.230 |
| Number of Restaurants and sidewalk cafe in 250m Buffer | 0.260 | 228.857 |
| Park Area in 250m Buffer | 0.093 | 34.682 |
| Number of Facilities in 250m Buffer | 3.256 | 648.628 |

| Parameter | Estimates | t-stats |
|---|---|---|
| Number of Recreational Points in 250m Buffer | 2.016 | 419.675 |
| Distance to Time Square (m) | -16.493 | -4801.280 |
| Elevation | -4.503 | -1182.673 |
| Commercia Area | 0.223 | 216.311 |
| **Satiation Parameters** | | |
| γ | 7.875 | 2350.980 |
| **Log-Likelihood at Convergence** | -534386813.520 | |

*4.4.2.2.1 Trip Attributes*

In the current research context, a negative coefficient was obtained for network distance of O-D pair. Intuitively, destinations further away are less appealing for cyclists. We also tried interaction of winter season with distance in the model. As expected, during cold weather the influence of distance is more burdensome for bikeshare users.

*4.4.2.2.2 Socio-demographic Attributes*

Stations located in Census tract with higher population density or heterogeneous land use mix are more likely to be chosen as destination stations (see (Faghih-Imani & Eluru, 2015, 2017b; Rixey, 2013; X. Wang, Lindsey, Schoner, & Harrison, 2015) for similar results). Similarly, job and establishment density also impacts station choice positively. The result probably highlights that bicycle-sharing systems are likely to be used for daily commute trips (see (Faghih-Imani, Eluru, & Paleti, 2017) for similar result).

*4.4.2.2.3 Bicycle Infrastructure Attributes*

Stations with increased dock capacity are more likely to be chosen (similar results in (El-Assi et al., 2017; Faghih-Imani & Eluru, 2015, 2017b)). An increase in the length of bicycle route within the 250-meter buffer of a destination station results in an increased likelihood of the station being chosen as destination (similar to findings of (El-Assi et al., 2017; Faghih-Imani

& Eluru, 2015, 2016, 2017b)). A similar result (albeit with lower magnitude) is obtained for street length variable.

Literature suggests that in addition to their own attributes, neighboring station attributes also affect destination choice behavior. In our study, we tested the impact of total number of stations and total dock capacity of neighboring stations in a 250m buffer. The number of stations and capacity in the station buffer offer surprising results. The two coefficients are negative highlighting that there is competition between bikeshare infrastructure. The result is quite different to what has been reported in earlier single discrete model approaches and warrants more investigation (see 7, 12 for different results). As the number of subway stations in the buffer increases, we observe that preference for that destination also increases.

*4.4.2.2.4 Land Use and Built Environment Attributes*

Intuitively, increased transit accessibility within the station buffer also increases the station's likelihood of being chosen as destination. As expected, stations located in neighborhoods with high walk and bike accessibility are preferred by cyclists. Cyclists prefer amenities around stations as indicated by the positive impact of number of restaurants and cafes in the vicinity of destination station. The CitiBike stations in the vicinity of parks are also more likely to be chosen. Individuals are likely to choose destination stations in a location with more facilities (such as museums, schools, colleges, university, hospitals). Visitors choose stations that bring them closer to Time Square as highlighted by negative coefficient of destination station distance to Time Square. Another important land use attributes that plays a significant role in choosing destination station is elevation of that station. People are less inclined to choose stations with steep slope for their trip. The presence of commercial area in the vicinity of destination station also increases the proclivity for the destination.

*4.4.2.2.5 Satiation Parameter*

As discussed earlier in the methodology section, the translation parameters $\gamma$ capture the extent of decrease in marginal utility across different destination stations. The translation parameter $\gamma$ is statistically significant at 95% level of significance, thereby implying that there are clear satiation effects in destination choice as distance of destination from Time Square increases. To elaborate, as the destination moves further away from Times Square, the satiation impacts are higher indicating fewer trips will be made to the destination.

## 4.5 Validation

For validation purpose, a hold-out sample was prepared in a similar fashion by randomly choosing 5 weeks from the rest of 21 weeks (5 weeks of total 26 weeks was used for sample). The same approach of choice set generation for estimation sample is exercised for validation sample (574 origins x 5 weeks x 573 destinations). The difference in the log-likelihood for the predicted and equal probability model is 48118 units clearly highlighting the enhanced fit of proposed model.

To further highlight the applicability of estimated model for predicting destination choice conditional on the origin, we categorize destination choices into four quartiles based on number of trips destined for both observed and predicted model. These four quartiles are defined as 1$^{st}$ quartile stations (trips destined are less than 25% of total originating trips), 2$^{nd}$ quartile stations (trips destined are 25-50% of total originating trips), 3$^{rd}$ quartile (trips destined are 50-75% of total originating trips) and 4$^{th}$ quartile stations (trips destined are more than 75% of total originating trips). We compute percentage of correctly classified predicted stations in each. The results of the evaluation are presented in Figure 4.4. The reader would note that the probability of correct classification varies across the four quartiles ranging from 18.88% though 51.8%. The result indicates that predicted model performs better in case of destination stations with higher demand. The proposed framework presents an innovative approach for examining bikeshare system usage and will allow bike sharing system planners and operators to better plan and manage their system.

### 4.6 Summary

In this chapter, considering bike sharing as one of the transportation sharing systems, this current study identifies two choice dimensions for capturing the bike share system demand: (1) station level demand and (2) how bike flows from an origin station are distributed across the network. A linear mixed model is considered to estimate station level demand while a multiple discrete continuous extreme value (MDCEV) model to analyze flows distribution is employed. The data for our analysis is drawn from New York City bikeshare system (CitiBike) for six months from January through June, 2017. For our analysis, we examine demand and distribution patterns on a weekly basis. A host of exogenous variables including trip attributes, socio-demographic attributes, bicycle infrastructure attributes, land use and built environment, temporal and weather attributes are considered. The model estimation results offer very intuitive results for origin demand and multiple discrete destination choice models. We validated the model by predicting trips to destined stations and found that predicted model performs well for high demand destinations. This analysis will allow bike sharing system planners and operators to better evaluate and improve bikeshare systems.

# CHAPTER 5: TRANSPORT NETWORKING COMPANIES DEMAND AND FLOW ESTIMATION: A CASE STUDY OF NEW YORK CITY

## 5.1 Introduction

Ride hailing services have been available as a mode of transportation since the early 17th century in the form of horse-drawn hackney carriages in Europe. With the advent of the automobile, taxis for hire have been the most common ride hailing transportation alternative. However, ride hailing has undergone a rapid transformation in the recent few years in response to the transformative technological changes including smart mobile availability, ease of hailing a ride using mobile applications, integration of seamless payment systems and real-time driver and user reviews. In fact, the convenience offered by transport networking companies (TNC) such as Uber, Lyft, and Via has allowed for a tremendous growth in ride hailing demand. For example, in New York City, the average daily trips by taxi (yellow taxi) was varying between 400 thousand and 500 thousand for the years 2010 and 2014 (Metcalfe & Warburg, 2012). However, since 2014, with the advent TNC services in the city, the total number of trips have increased. Specifically in 2018, the daily trips have increased to more than a million trips with traditional taxi accounting for nearly 300 thousand trips, and TNC services accounting for 700 thousand trips. These trends are not specific to New York City. A recent report analyzing reimbursed travel in the US has found that the share of Uber and Lyft has increased from 8% to 72.5% within 2014-2018 at the cost of taxi and rental car business share (Silver & Fischer-Baum, 2016). The prevalence of TNC services is also not restricted to US. Uber operates in over 60 countries, while Didi Express in China, Ola in India currently capture a large share of the ride hailing market in these countries. The immense growth in market share and the spread of these services across the world illustrate how the ride hailing market has undergone a rapid transformation in a short time frame.

The rapid transformation of the ride hailing market coupled with emerging shared mobility service expansions (such as Carshare, Bikeshare, and Scooter share) offers an unprecedented opportunity to address the existing mobility shortcomings in urban regions (as highlighted in a recent TCRP report (Feigon & Murphy, 2016). In fact, public transit and transportation planning agencies can enhance mobility and accessibility in a region by incorporating these shared transportation alternatives within their planning frameworks to provide holistic mobility options in denser urban regions. Specifically, dense urban regions with well-connected public transit systems can strategically target reducing the reliance on private automobile ownership (and use) by incorporating ride-hailing alternatives in trip planning tools. Further, by examining the spatio-temporal ride hailing data, transit agencies and shared mobility platforms can identify urban pockets with service needs to provide last mile connectivity. Towards understanding these patterns it would be beneficial to understand TNC demand and its spatial distribution in the region.

The current research effort contributes to this goal by developing quantitative models of TNC demand and flow distribution patterns. Using data from the NYC Taxi and Limousine commission, we conduct a comprehensive analysis of morning peak hour ride hailing data from Uber, Lyft, Juno and Via from 2018. The study develops (1) a demand component that estimates origin level TNC demand at the taxi zone level and (2) a distribution component that analyzes how these trips from an origin are distributed across the region. The former component is analyzed using linear mixed models and the latter component is analyzed using a multiple discrete continuous model system. The model components are developed using a comprehensive set of independent variables including aggregate trip attributes, transportation infrastructure variables, land use and built environment variables, weather attributes, and temporal attributes. The model estimates are validated using a hold out sample. Further, a policy exercise is conducted to illustrate how the proposed model system can be utilized for evaluating the impact of changes to independent variables.

The rest of the chapter is organized as follows: The following section presents the detailed description of the dataset with sample formation technique adopted in the analysis while section 3 provides methodological framework. Model results are presented in the fifth section followed by the policy analysis. Final section comprises with the concluding statements.

## 5.2 Data

### 5.2.1 Data Source

New York City with high residential density and large tourist population is an ideal market for ride hailing systems. The NYC Taxi and Limousine Commission (TLC) provides spatially aggregated trip data from all ride hailing companies (taxi, Uber, Lyft, Juno and Via) for public use (https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page). The trip itinerary dataset for 2018 for Uber, Lyft, Juno and Via was processed to obtain daily morning peak hour TNC usage patterns.. The dataset provides information on start and end time of trips, origin and destination defined as taxi zone ID, trip distance and vehicle license number. The trip data was augmented with other sources including: (1) built environment attributes derived from New York City open data (https://nycopendata.socrata.com); (2) socio-demographic characteristics at the census tract/zip code level gathered from US 2010 census data; (3) the weather information corresponding to the Central Park station retrieved from the National Climatic Data Center (http://www.ncdc.noaa.gov/data-access).

### 5.2.2 Sample Formation

A series of data cleaning and compilation exercises were undertaken for generating the sample data for estimation purposes. First, trips with missing or inconsistent information were removed. Second, trips longer than 500 minutes in duration (around 0.5% of all trips) were deleted considering that these trips are not typical ride-sharing trips. These trips could also be a result of two possibilities; either destination of those trips could be outside NYC or due to technical issues the trip information was recorded incorrectly. Third, trips that had the origin and destination outside of NYC taxi zone were also eliminated. Therefore, we focus on trips that originated and were destined within NYC taxi zone region only.

For the given study period (January 2018 to December 2018), the total number of available taxi zones in NYC was 260. Initially, we aggregated morning peak (6.30 am-9.30am) trip data for each day for each week (total 52 weeks) from each origin taxi zone ID to every possible destination taxi zone ID (260). The average number of daily trips generated and attracted at each taxi zone is presented in Figure 5.1. In Figure 5.1, the number of trips generated (Figure 5.1a) and attracted (Figure 5.1b) to each taxi zone is categorized into multiple classes from very low to very high. The figures clearly highlight the high TNC usage in Manhattan and airport locations (LaGuardia, John F. Kennedy International Airport and Newark airport).

For our analysis, to ensure that holiday weekends that are likely to have a different user patterns do not influence our analysis, we selected morning peak period trip data for 43 weeks without any holidays. The processing of the large sample of trip data is substantially time-consuming and significantly increases the model run times. To obtain a reasonable sample size for model estimation, we sampled following two steps; 1) 150 taxi zones were selected randomly from the total 260 taxi zones and 2) for each taxi zone one weekday was randomly selected for each week.

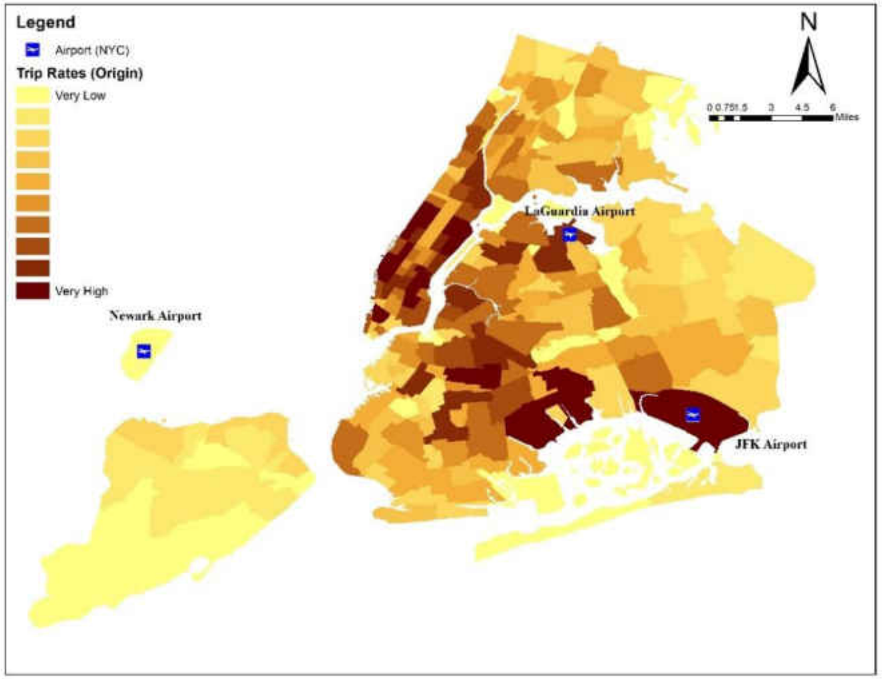
Thus, the data sampled had 150 taxi zone with 43 weekday morning peak trip data during 2018. We organized the dataset into two components for our analysis; 1) For zonal level origin demand (aggregating total daily morning peak trip at the origin level) and 2) Trip distribution from origin to destination (aggregating daily morning peak trip at the O-D pair level). Figure 5.2 provides a detailed flow chart of the independent and dependent variable data compilation procedure.

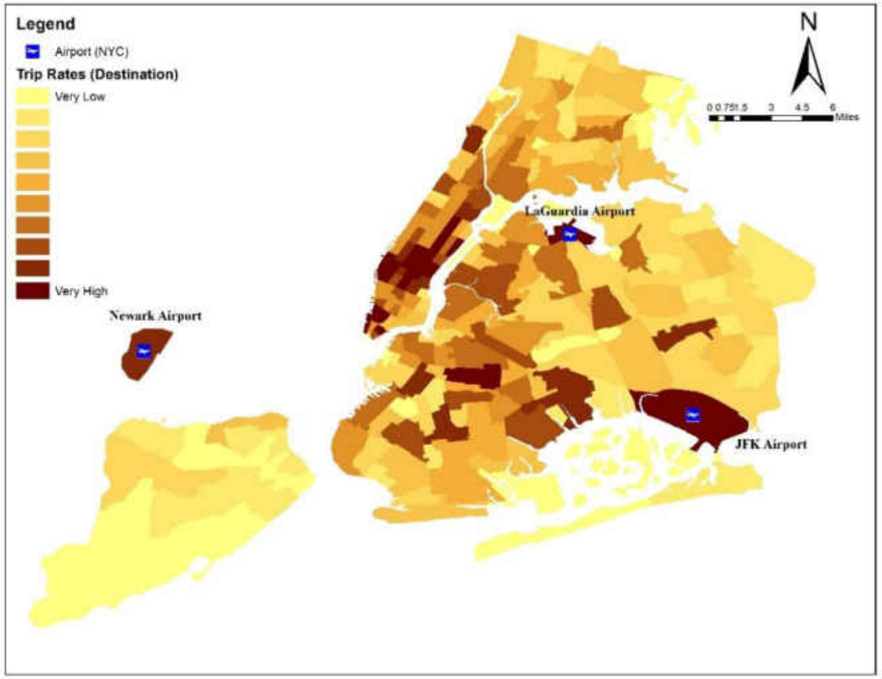
5.2.3 Independent Variable Generation

Several independent variables were generated in our study (see Figure 5.2). These can be grouped into five categories: 1) Trip attribute, 2) Transportation infrastructure variables, 3) Land use and built environment variables, 4) Weather attributes, and 5) Temporal attributes.

Trip attribute includes the network distance between each origin-destination taxi zone pair estimated using the shortest path algorithm tool of ArcGIS software. While the actual trip might involve a different route, the shortest network distance would be an appropriate indicator of the distance traveled. The variable will serve as a surrogate for travel time. As all the data is for morning peak, the impact of congestion is likely to be affecting all records similarly.

Transportation infrastructure attributes created at the taxi zone level include bike route length density (capturing the effect of availability of bicycle facilities on system usage), number of bikeshare stations, length of streets (minor and major streets). Number of subway stations and bus stops in the taxi zone were generated to examine the influence of public transit on rider's preference of destination station.




**(a) Trip generation at taxi zones** **(b) Trip attracted at destined taxi zones**

**Figure 5.1: Ride Hailing Trips in NYC's Taxi Zone Level**

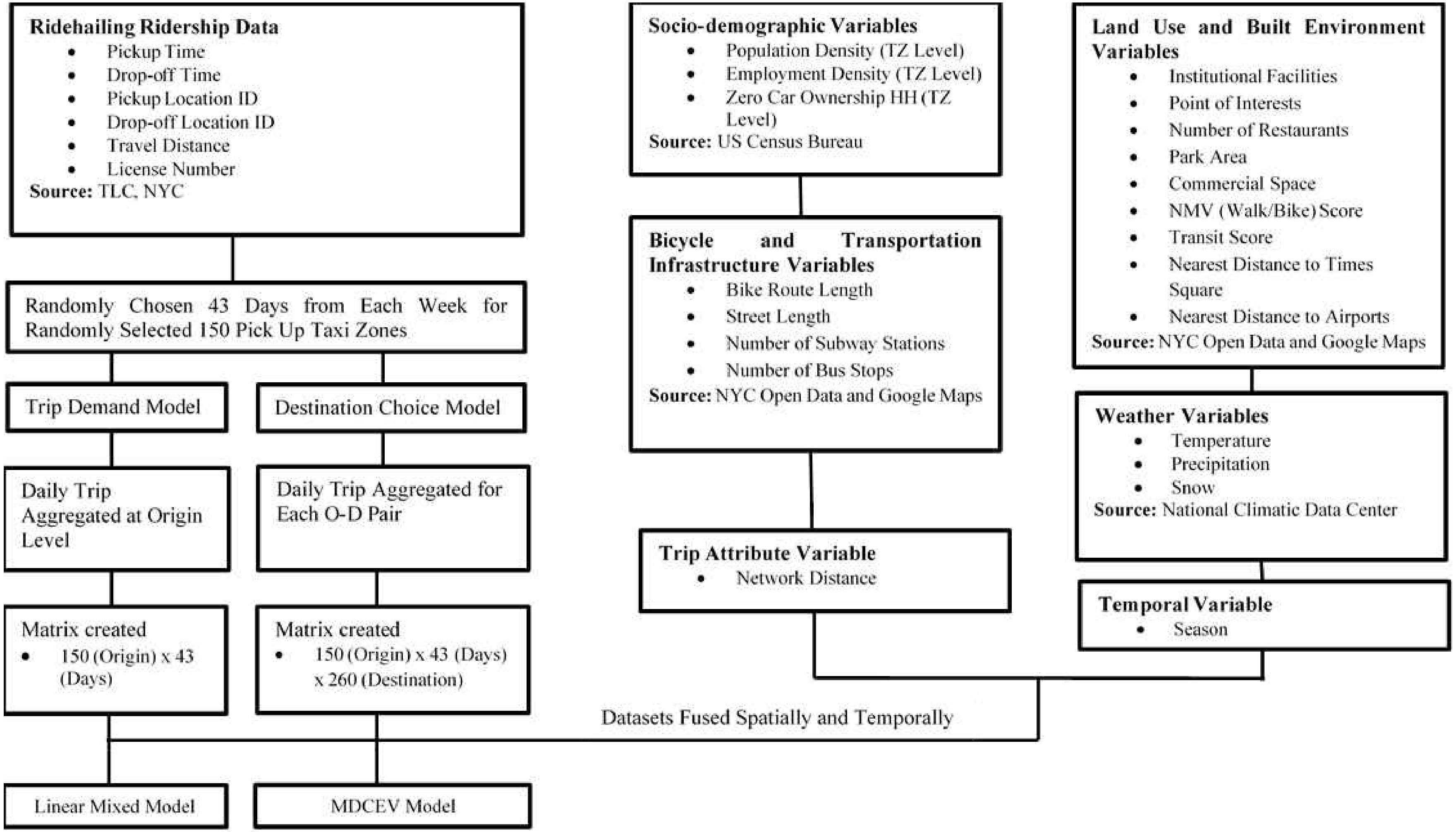



**Figure 5.2: Data Formation Flow Chart**

Several land use and built environment variables were considered including population density, job density and establishment density, the number of institutional facilities (schools, colleges, hospitals), the number of point of interests (museums, shopping malls), and the number of restaurants (including coffee shops and bars), total area of parks and commercial space (office, industry, retail) within each taxi zones. Distance of destination from Times Square and airport were estimated by using the shortest path algorithm tool of ArcGIS software. Airport indicator variable for the taxi zone was generated to examine the additional impact of airport destination. Population, job density and median income information was collected from US Census for 2014-2017 and extrapolated for 2018 at the census tract level considering average yearly population change from 2014-2017. Household car ownership information for 2018 was used to generate proportion of zero car ownership at taxi zone level to examine the impact of car ownership on riders' destination preferences. Non-motorized vehicle score (average of walk score and bike score) and transit score associated with each taxi zone was considered at the census tract level. Further, crime density and accident density were also generated at taxi zone level. Total number of crimes of all types for previous year (2017) was aggregated at census tract level and crime density was estimated by dividing with the corresponding year's population. In a similar manner, total number of accidents of all kind for each day of 2018 was considered to generate accident density.

Weather variables include average temperature, precipitation, and snow for that particular day. Several interaction variables were also created. Seasonality is the only temporal variable considered. We consider winter (December-February), Spring (March-May), Summer (June-August) and Fall (September-November) as dummy variables.

5.2.4 Descriptive Analysis

The data at an aggregate system level in the form of average number of trips by taxi zone for each week is presented in Figure 5.3. The various weeks with lower demand correspond to the weeks with holidays supporting our hypothesis that these weeks have a different demand pattern. The dependent variable distribution is generated to understand origin level demand and distribution of these flows across the study region. On average, 384 trips depart from each origin taxi zone in the morning peak hour and are destined to about 67 alternative taxi zones. The sample characteristics of the independent variables generated were suppressed due to space considerations.

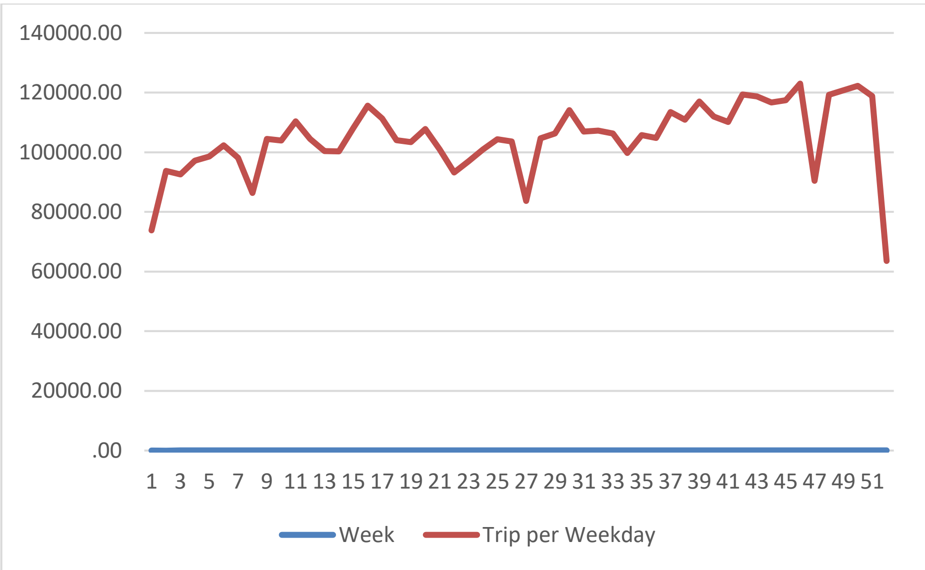



**Figure 5.3: Trip Rates of TNC demand by week**

## 5.3 Econometric Frameworks

5.3.1 Linear Mixed Model for Station Level Weekly Origin Demand

The taxi zonal level daily pick up demand variable is a continuous value and can be analyzed using linear regression models. However, the traditional linear regression model is not appropriate for data with multiple repeated observations. In our empirical analysis, we observe the daily peak hour demand at the same taxi zone for fourty-three weeks. Hence, we employ a linear mixed modeling approach that builds on the linear regression model while incorporating the influence of repeated observations for the same station. The linear mixed model collapses to a simple linear regression model in the absence of any station specific effects.

Let $w = 1, 2, \ldots, W$ be an index to represent each taxi zone ($W = 150$), $M = 1, 2, \ldots, 43$ be an index to represent the various day of weeks of data compiled for each pick up taxi zone. The dependent variable (daily peak hour demand) is modeled using a linear regression equation which, in its most general form, has the following structure:

$$y_{mw} = \beta X_{mw} + \varepsilon_{mw} \tag{5.1}$$

where $y_{mw}$ is the natural logarithm of weekly demand, $X$ is an $K \times 1$ column vector of attributes and the model coefficients, $\beta$, is an $K \times 1$ column vector. The random error term, $\varepsilon_{mw}$, is assumed to be normally distributed across the dataset. In our analysis, the repetitions over days can result in common unobserved factors affecting the dependent variable. While a full covariance matrix can be estimated for the unobserved correlations, as we are selecting 43 random days from a sample of 43 weeks for each tax zone, we decided to employ a simpler covariance structure. The exact functional form of the covariance structure assumed is shown below:

$$\Omega = \begin{pmatrix} \Omega^2 + \Omega_1^2 & \Omega_1 & \ldots & \Omega_1 \\ \Omega_1 & \Omega^2 + \Omega_1^2 & \ldots & \Omega_1 \\ \vdots & \vdots & \ddots & \vdots \\ \Omega_1 & \Omega_1 & \ldots & \Omega^2 + \Omega_1^2 \end{pmatrix} \tag{5.2}$$

The covariance structure restricts the covariance across all fourty-three records to be the same. The parameters estimated in this correlation structure are $\Omega$ and $\Omega_1$ . The parameter $\Omega$ represents the error variance of $\varepsilon$, $\Omega_1$ represents the common correlation factor across daily records. The models are estimated in SPSS using the Restricted Maximum Likelihood Approach (REML). The REML approach estimates the parameters by computing the likelihood function on a transformed dataset. The approach is commonly used for linear mixed models (Harville, 1977).

5.3.2 MDCEV Model for Destination Choice

According to Bhat and Eluru (Bhat, Sen, & Eluru, 2009), we consider the following functional form for utility in this paper, based on a generalized variant of the translated Constant Elasticity of Substitution (CES) utility function:

$$U(x) = \sum_{i=1}^{I} \frac{\gamma}{\alpha} \lambda_i \left\{ \left( \frac{x_i}{\gamma} + 1 \right)^{\alpha} - 1 \right\} \tag{5.3}$$

where $U(x)$ is a quasi-concave, increasing, and continuously differentiable function with respect to the consumption quantity ($I$x1)-vector ($x_i \geq 0$ for all $i$), and $\lambda_i$ associated with drop off taxi zone $i$. $\lambda$ represents the baseline marginal utility ($\lambda_i$> 0 for all $i$), $\gamma$ is a translation parameter ($\gamma$ should be greater than zero) which enables corner solutions while simultaneously influencing satiation and $\alpha$ influences satiation ($\alpha \leq 1$).

The KT approach employs a direct stochastic specification by assuming the utility function $U(x)$ to be random over the population. A multiplicative random element is introduced to the baseline marginal utility for each good (in our case destination) as follows:

$$\lambda\,(y_{iw}, \rho_{iw}) = \exp\,(\delta y_{iw} + \rho_{iw}) \tag{5.4}$$

where $y_{iwq}$ is a set of attributes characterizing drop off taxi zone $i$ during day $w$, $\delta$ corresponds to a column vector of coefficients, and $\rho_{iw}$ captures idiosyncratic (unobserved) characteristics that impact the baseline utility for destination stations. The overall random utility function of Equation (3) then takes the following form:

$$U(x) = \sum_{i=1}^{I} \frac{\gamma}{\alpha} \exp\,(\delta y_{iw} + \rho_{iw})\left\{\left(\frac{x_i}{\gamma} + 1\right)^{\alpha} - 1\right\} \tag{5.5}$$

Following (Bhat, 2005, 2008), consider a generalized extreme value distribution for $\rho_i$ and assume that $\rho_{iw}$ is independent of $y_{iw}$ $(i = 1,2,\ldots,I)$. The $\rho_{iw}$'s are also assumed to be independently distributed across alternatives with a scale parameter normalized to 1. Due to the common role of $\gamma$ and $\alpha$, it is very challenging to identify both $\gamma$ and $\alpha$ in empirical application (see (Bhat, 2008) for detailed discussion). Hence, either $\gamma$ or $\alpha$ parameter is estimated. When the $\alpha$ - profile is used, the utility simplifies to:

$$U(x) = \sum_{i=1}^{I} \frac{1}{\alpha} \exp\,(\delta y_i + \rho_i)\{(x_i + 1)^{\alpha} - 1\} \tag{5.6}$$

When the $\gamma$ - profile is used, the utility simplifies to:

$$U(x) = \sum_{i=1}^{I} \gamma \exp(\delta y_i + \rho_i)\, ln\left(\frac{x_i}{\gamma} + 1\right) \tag{5.7}$$

In this study, $\gamma$ - profile is used. Finally, the probability that an pick up taxi zone has flows to the first $D$ drop-off taxi zones $D \geq 1$ is:

$$P(e_1^*, e_2^*, e_3^*, \ldots, e_D^*, 0,0, \ldots, 0) = \left[\sum_{n=1}^{D} d_n\right]\left[\sum_{n=1}^{D} \frac{1}{d_n}\right]\left[\frac{\prod_{n=1}^{D} e^{U_n}}{(\prod_{d=1}^{K} e^{U_i})^D}\right](D-1)! \tag{5.8}$$

where $(\sum_{n=1}^{D} m_n)\left(\sum_{n=1}^{D} {}^{1}\!/_{m_n}\right)$ is defined as Jacobian form for the case of equal unit prices across goods (Bhat, 2008) where, $m_n = \left(\frac{1-\alpha}{e_n^* + \gamma}\right)$.

Unlike the traditional MDCEV model, in our context, the number of alternatives is substantially larger. Hence, we resort to estimating a generic parameter for each exogenous variable across alternatives (analogous to how multinomial logit based location choice models are estimated with a single utility equation).

## 5.4 Estimation Results

The mathematical details of the linear mixed model and multiple discrete continuous extreme value model are suppressed to save on space. The model estimation results from the two models are discussed – TNC demand model followed by the trip distribution model results.

### 5.4.1 Trip Demand Model

#### 5.4.1.1 Model Fit Measures

A linear regression model was estimated at first as benchmark for evaluating the linear mixed model. To compare these two models, a Log-likelihood ratio (LR) test was computed. The LR value was found to be 1915 which was higher than any corresponding chi-square value for 2 degrees of freedom. Based on the LR test statistic, we can conclude that the linear mixed model outperforms the simple linear regression model and offers satisfactory fit for the station level demand.

#### 5.4.1.2 Linear Mixed Model Results

The linear mixed model estimation results for morning peak hour TNC origin demand are presented in Table 5.1. The model estimation results offer intuitive findings. TNC demand, as expected is positively associated with population density. Increased median income of households within the taxi zones is found to increase demand for TNC trips (see (Correa et al., 2017; Smart et al., 2015) for similar results). The presence of airport in the taxi zone also contributes to increased TNC demand. Higher number of trips are likely to be generated from taxi zones with higher population than lower populated zones. The presence of different institutional facilities (such as schools, colleges, hospitals, and office) in the taxi zones increases the zonal demand. The presence of discretionary opportunities such as a higher presence of restaurants and sidewalk café also drives TNC demand. Taxi zones with higher proportion of residential area is positively associated with Peak hour morning TNC flows. The result illustrates the adoption of TNC service for morning commute activities from these zones. The results for precipitation variables highlight that in the presence of precipitation individuals are likely to make a trip via TNC services (see (Brodeur & Nield, 2016) for similar result). The results also indicate a positive influence of summer and fall season compared to winter and spring season. The finding is in line with earlier research (Brodeur & Nield, 2016). The result is also possibly reflecting the increased tourist activity during these seasons.

#### 5.4.1.3 Correlation Parameters

In the linear mixed model we estimate a parameter that recognizes the repeated measures of data for each taxi zone. The correlation parameter is statistically significant highlighting the role of common unobserved factors influencing the demand from taxi zones.

**Table 5.1: Linear Mixed Model Results for TNC Origin Demand**

| Parameter | Estimates | t-stats |
|---|---|---|
| Intercept | -1.679 | -3.903 |
| **Land Use and Built Environment Attributes** | | |
| Population Density | 1.261 | 8.869 |
| Median Income (x$10^{-3}$) | 8.035 | 4.079 |
| Airport as an Indicator | 0.804 | 4.079 |
| Number of Institutional Facilities in a Taxi Zone (x$10^{-3}$) | 0.195 | 1.655 |
| Number of Restaurants and Side cafe in a Taxi Zone (x$10^{-3}$) | 0.316 | 2.803 |
| Residential Area ($m^2$ x$10^{-6}$) | 0.316 | 2.803 |
| **Temporal Attributes** | | |
| Precipitation (cm) | 3.740 | 26.106 |
| Season: Summer and Fall (Base: Winter and Spring) | 1.548 | 8.574 |
| **Correlation Parameters** | | |
| $\Omega$ | 5.253 | 56.116 |
| $\Omega_1$ | 3.776 | 8.429 |
| **Restricted Log-Likelihood** | 37161.892 | |
| **Sample Size** | 6450 | |

5.4.2 TNC Distribution Model

5.4.2.1 Model Fit Measures

The final log-likelihood values for the estimated MDCEV model and equal probability MDCEV model are -1531122.801 and -1712633.216 respectively. The log-likelihood ratio (LR) test-statistic of comparison between the final model and the equal probability model is 363020.830. The LR test-statistic value is significantly higher than the corresponding chi-square value for 22 additional degrees of freedom highlighting that the MDCEV distribution model offers a reasonable fit.

5.4.2.2 MDCEV Model Results

The model results of TNC morning peak hour distribution model are presented in Table 5.2. The presentation of results is organized by the various variable categories. The reader would note that a single utility equation is estimated for all the destination zones (analogous to location choice model estimation for large number of alternatives). A positive (negative) coefficient indicates an increase (decrease) in the variable results in increasing the utility of the alternative destination.

**Table 5.2: MDCEV Model Results**

| Parameter | Estimates | t-stats |
|---|---|---|
| **Land Use and Built Environment Attributes** | | |
| Population Density | 0.462 | 22.824 |
| Job Density | 1.122 | 45.023 |
| Median Income ($x10^{-3}$) | 5.445 | 67.210 |
| Proportion of Zero Car HH | 1.376 | 78.465 |
| Transit Score ($x10^{-2}$) | 0.958 | 30.103 |
| Non-motorized vehicle score ($x10^{-2}$) | -1.807 | -51.698 |
| Number of Restaurants and sidewalk café in Taxi Zone ($x10^{-3}$) | 0.438 | 42.622 |
| Number of Institutional Facilities in Taxi Zone ($x10^{-3}$) | 0.194 | 8.528 |
| Number of Point of Interests and Recreational Points in Taxi Zone ($x10^{-3}$) | 1.401 | 41.801 |
| Commercial Area ($m^2$ $x10^{-6}$) | 1.641 | 87.265 |
| LU Mix | 0.723 | 35.999 |
| Airport Indicator | 3.702 | 335.179 |
| Times Square Distance (m x $10^{-3}$) | -0.378 | -66.091 |

| Parameter | Estimates | t-stats |
|---|---|---|
| **Trip Attributes** | | |
| Network Distance (m x $10^{-3}$) | -2.547 | -174.790 |
| **Transportation Infrastructure and Attributes** | | |
| Bike Lane Density in Taxi Zone | -0.730 | -22.787 |
| Number of Bikeshare Stations in Taxi Zone (x$10^{-2}$) | -0.108 | -26.258 |
| Street Length in Taxi Zone (m x $10^{-3}$) | 0.106 | 3.348 |
| Number of Bus Stops and subway stations in Taxi Zone (x$10^{-3}$) | 1.174 | 62.354 |
| **Temporal and Weather Attributes** | | |
| Network Distance (m x $10^{-3}$) x Winter | -0.577 | -5.659 |
| Network Distance (m x $10^{-3}$) x Temperature (°F x $10^{-2}$) | 2.460 | 10.983 |
| Times Square Distance (m x $10^{-3}$) x Precipitation (cm) | -0.031 | -7.267 |
| Network Distance (m x $10^{-3}$) x Precipitation (cm) | -0.721 | -13.517 |
| **Satiation Parameters** | | |
| Times Square Distance (m x $10^{-3}$) | 0.087 | 42.497 |
| **Log-Likelihood at Convergence** | -1531122.801 | |
| **Sample Size** | 1677000 | |

*5.4.2.2.1 Land Use and Built Environment Attributes*

Zones located in census tracts with higher population density are more likely to be chosen as destination locations. Similarly, job density also impacts destination preference positively. The results together point towards the adoption of TNC services for daily commute trips (see Correa et al., 2017 for similar result). Taxi zones with high income are preferred destination zones for TNC services. The model parameter for taxi zone level zero car household proportion highlights the increased adoption of TNC services among these zones Correa et al., 2017 found similar association with lower vehicle ownership households).

As expected, increased transit accessibility within a taxi zone increases the likelihood of the zone being chosen as a destination. On the other hand, the results indicate that zones with higher non-motorized score are less preferred destinations. While the result seems counterintuitive, it might be alluding to potential competition between TNC ride hailing and bicycle sharing systems in these zones. The presence of activity opportunities in the forms of restaurants and cafes, institutional facilities, and recreational centers and point of interests (POI) are positively associated with the destination zone preference. Taxi zone with higher commercial area serves as an attraction for TNC demand. The increase in land use mix value (range between 0 and 1) has a positive impact on destination zone preference.

The presence of airport in the destination taxi zone, as expected, increases the preference for the zone. The model also considers the influence of another major landmark in the region - Times Square. The parameter indicates that as the taxi zone is further from Times Square the preference of the zone as a destination reduces. The result illustrates how Times Square and its proximal zones serve as attraction centers for regular and tourist travel.

*5.4.2.2.2 Trip Attributes*

In the current research context, a negative coefficient was obtained for network distance of O-D pair. With the increasing distance to the destination, TNC demand distribution propensity reduces.

*5.4.2.2.3 Transportation Infrastructure and Attributes*

Several transportation infrastructure variables were considered in the demand distribution models. Of these variables, bike lane density, bikeshare stations, street length, bus stops and subway stations presented significant impacts on destination preferences. Taxi zones with higher bike length density (defined as ratio of bike length to overall roadway length) reduce the preference for the destination zone. The negative association with number of bikeshare stations within a taxi zone highlights that TNC demand is likely to be lower for a destination zone with more bikeshare stations. An increase in the street length within the destination zones results in an increased likelihood of the zone being chosen as destination (similar to findings of Correa et al., 2017). As the number of bus stops and subway stations in the taxi zone increases, we observe increased preference for that destination.

#### *5.4.2.2.4 Temporal and Weather Attributes*

The reader would note that temporal and weather attributes cannot be considered directly in destination distribution model. Hence, we interacted these variables with destination specific variables such as network distance and distance to Times Square. The results offer interesting results. In Winter, the negative influence of network distance increases further indicating that shorter trips are preferred (relative to other moths). The temperature variable interacted with network distance indicates that the influence of network distance is moderated by higher temperature i.e. as temperature increases the negative impact of network distance reduces. The precipitation variable interacted with network distance and distance to Times Square highlights the increase in sensitivity to travel time under precipitation conditions. The weather variables as a whole highlight how TNC distance impact is lower in good weather relative to poor weather.

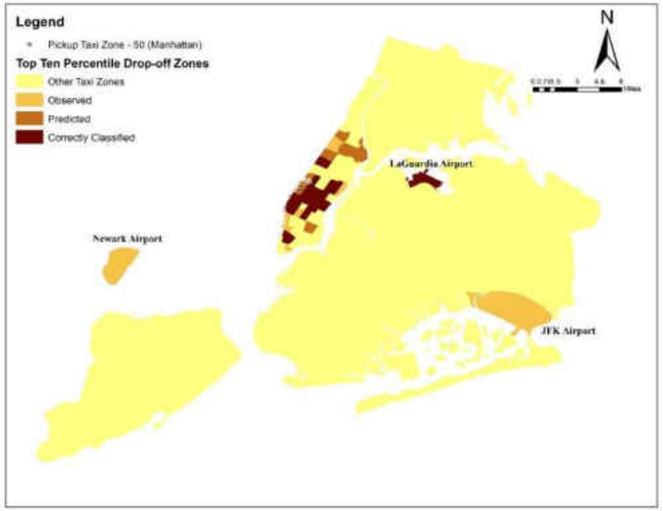
#### *5.4.2.2.5 Satiation Parameter*

We used distance to Times Square from taxi zones as a satiation parameter. In MDCEV model, the satiation parameter captures the extent of decrease in marginal utility across different destination zones. The satiation parameter is statistically significant at 95% level of significance, thereby implying that there are clear satiation effects in destination choice as distance of destination from Times Square increases. To elaborate, as the zone is further away from Times Square, the satiation impacts are higher indicating fewer trips will be made to the zone.

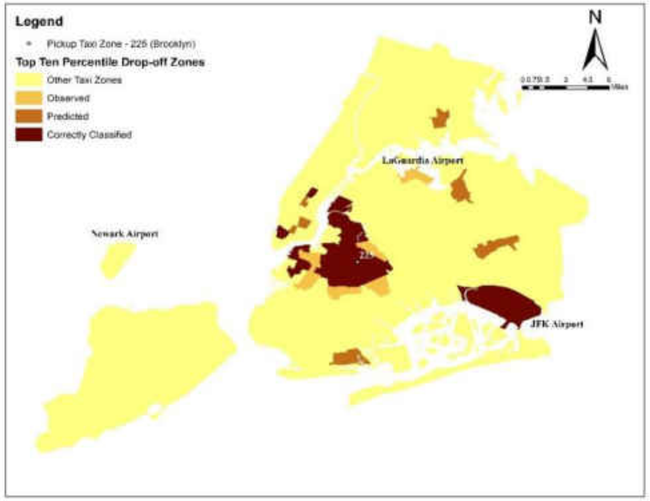
### 5.5 Validation Analysis Results

For validation purpose, a hold-out sample was prepared following the same procedure used to extract the estimation sample. After extracting 150 taxi zones for our base dataset, the remaining 110 taxi pick up zones were set aside for validation. Then we randomly chose 43 days from 43 corresponding weeks throughout the year for these 110 zones. The same approach of data preparation employed for estimation sample is exercised for validation sample (110 origins x 43 days x 260 destinations). Using the validation data, the model results from the estimation sample were used to generate a prediction measure in the form of predictive log-likelihood. The difference in the log-likelihood for the predicted and equal probability model is 3626720.830 units clearly highlighting the enhanced fit of the proposed model.

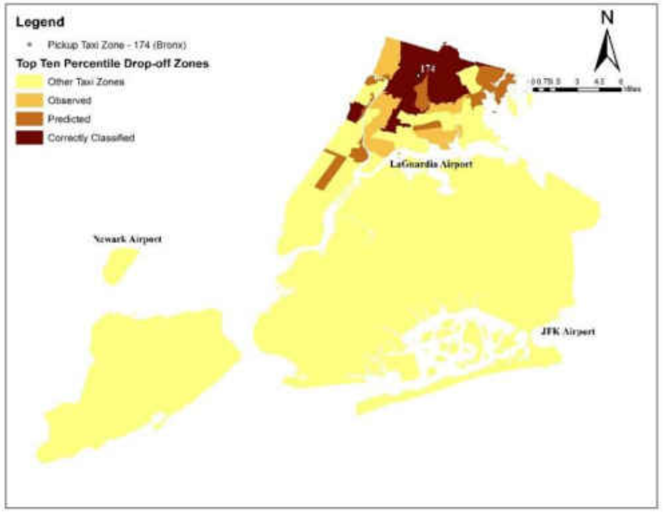
To further highlight the applicability of estimated model for predicting destination choice conditional on the origin, we estimated destined trips from each origin for each day at disaggregate level. Note that, zero trips to any destination for a week was also considered. To identify the preferred destination zones, top 10 percentile of preferred destination zones was captured for each pickup zone and validated with the top 10 percentile predicted destination zones. For the performance evaluation, we compute the correctly classified predicted trips for top 10 percentile destined zones for each taxi zone considering the total trips throughout the year. The reader would note that about 71% of the top destination zones were correctly classified. To provide a visual representation, we selected 5 random taxi zones from 5 NYC boroughs and predicted the top 10 percentile destination zones for them considering average daily morning peak hour trips throughout the year and compared them with observed top



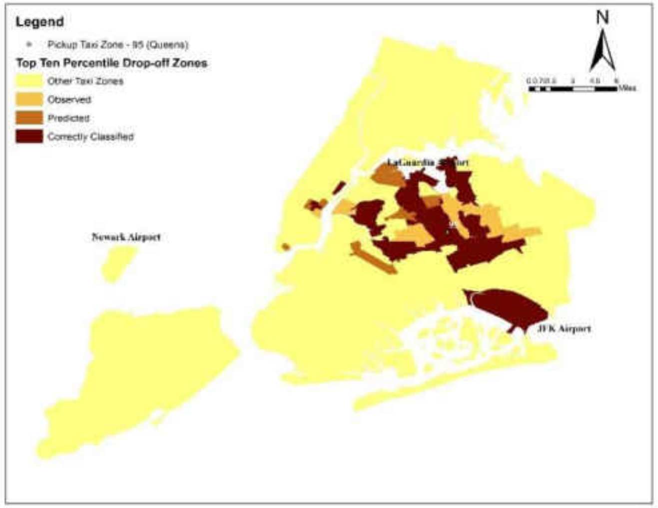

(a) Manhattan



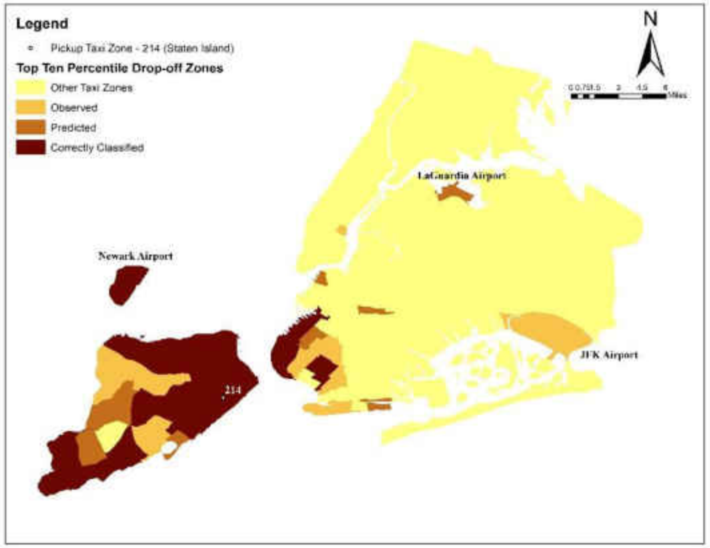

(b) Brooklyn




(c) Bronx




(d) Queens




**(e) Staten Island**

**Figure 5.4: Top 10 Percentile Destined Zones for Randomly Selected Pickup Zones from 5 NYC Borough**

destination zones for that particular zone (See Figure 5.4). Across the five boroughs, based on the observed and predicted measures from the Figure, taxi zones situated in Brooklyn offered the best prediction performance while taxi zone from Staten Island has inferior prediction performance. Overall, the two validation exercises, highlight the applicability of the proposed approach for TNC demand and distribution prediction.

### 5.6 Policy Illustration

The model results from Table 5.2 provide an indication of how the exogenous variables affect the network flows considering destination choice. However, they cannot provide the exact magnitude of the effect of these exogenous variables. Hence, elasticity effects computation considering changes of baseline marginal utility was used to evaluate the impact of exogenous variables on destination choice. The elasticity effects are computed by evaluating the percentage change in marginal utility of an alternative in response to increasing the value of exogenous variables from best fit model by 10%, 25% and 50% respectively. We selected five independent variables for presentation including job density, median income, network distance, institutional facilities and bus stops and subway stations. The computed elasticities are presented in Figure 5.5. Based on elasticity effects results in Figure 5.5, following observations can be made. *First,* the elasticity estimate for job density variable indicates that about 6.5, 17 and 37% increase in utility happens due to 10, 25 and 50% change in the independent variable. All the other results can be interpreted similarly. *Second*, rank order of the top three significant variable in terms of changes for the utility without considering positive or negative impact include network distance, job density and median income. *Third*, network distance between O-D can be considered as a proxy for travel time. The increasing value of this variable provides a snapshot of the impact of additional travel time due to traffic congestion or other safety incidents. Overall, the elasticity analysis results provide an illustration on how the proposed model can be applied to determine the critical factors contributing to increase in utility to choose a taxi zone as destination.

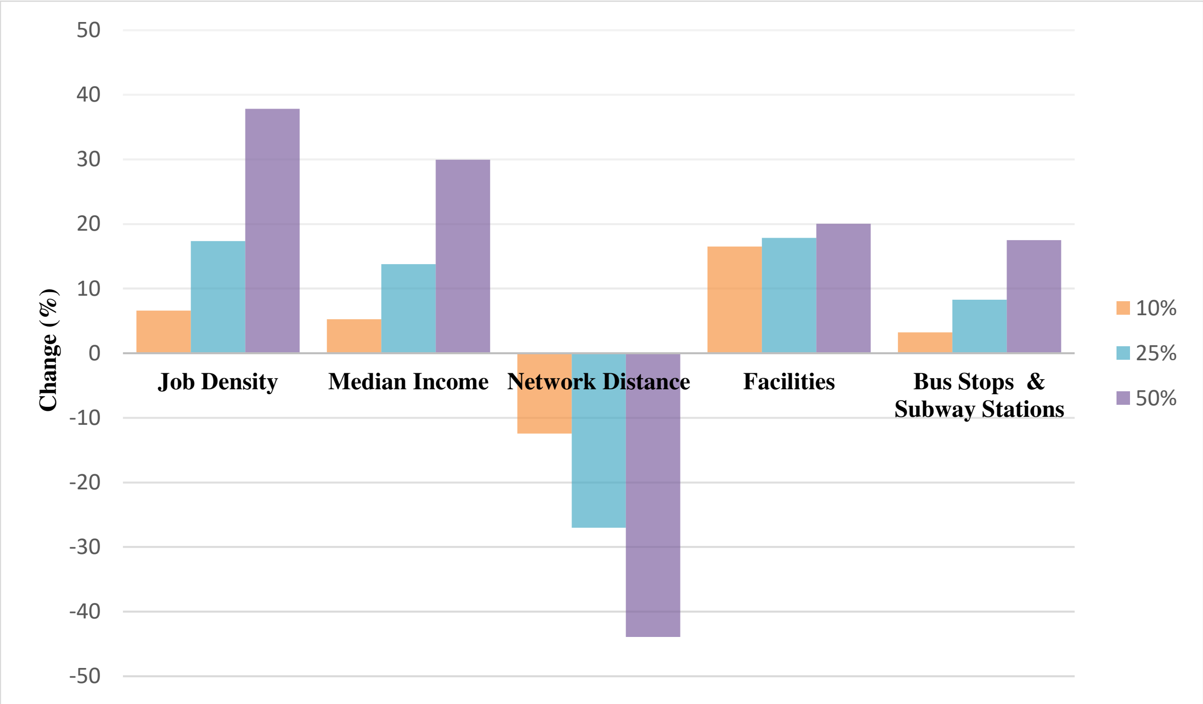



**Figure 5.5: Elasticity Effects Considering Utility Changes**

### 5.7 Summary

Given the burgeoning growth in transportation networking companies (TNC) based ride hailing systems and their growing adoption for trip making, it is important to develop modeling frameworks to understand TNC ride hailing demand flows at the system level. In the third part of the dissertation, we identify two choice dimensions: a demand component that estimates origin level TNC demand at the taxi zone level and (2) a distribution component that analyzes how these trips from an origin are distributed across the region. The origin level demand is analyzed using linear mixed models while flows from origin to multiple destinations is analyzed using a multiple discrete continuous model system (MDCEV). The data for our analysis is drawn from New York City Taxi & Limousine Commission (NYTLC) for twelve months from January through December 2018. For our analysis, we examine weekday morning peak hour demand and distribution patterns. The model components are developed using a comprehensive set of independent variables. The model estimation results offer very intuitive results for origin demand and distribution of flows across destinations. We validated the model by predicting trips to destination taxi zones and found that predicted model performs well in identifying high preference destination zones. In addition, elasticity effects are computed by evaluating the percentage change in baseline marginal utility in response to increasing the value of exogenous variables by 10%, 25% and 50% respectively.

# CHAPTER 6: TRANSFORMATION OF RIDE HAILING IN NEW YORK CITY: A QUANTITATIVE ASSESSMENT

## 6.1 Introduction

In most urban regions, individuals, who do not have access to or do not prefer to use personal vehicles, have the option of either using public transit, shared bicycling systems (for short distance trips) or a ride hailing service (such as taxi or Uber). While public transit systems are constrained by predefined routes and fixed schedules, bicycle sharing systems are limited by small distance range, ride hailing services at a cost provide individuals with convenient door-to-door car trips without the additional challenges associated with driving/bicycling (such as having to find a parking spot, concentrating on driving and physical effort of bicycling). In recent years, ride hailing has undergone a rapid transformation in response to the transformative technological changes including smart mobile availability, ease of hailing a ride using mobile applications, integration of seamless payment systems and real-time driver and user reviews. The convenience offered by transport networking companies (TNC) such as Uber, Lyft, and Via has allowed for tremendous growth in ride hailing demand. For example, in New York City, the average daily trips by taxi (yellow taxi) was varying between 400 thousand and 500 thousand for the years 2010 and 2014 (Silver & Fischer-Baum, 2016). However, since 2014, with the advent of TNC services in the city, the total number of trips have increased. Based on NYC TLC report (Silver & Fischer-Baum, 2016), from 2015 to 2018, TNC daily trips increased from 60,000 to 700,000 while traditional taxi (Yellow and Green together) daily trips declined from 450,000 to 285,000. The trend observed in NYC is not an exception. A recent report analyzing reimbursed travel in the US has found that the share of Uber and Lyft has increased from 8% to 72.5% from 2014-2018 at the cost of taxi and rental car business share (Rajagopalan & Srinivasan, 2008).

The TNC service induced transformation can be viewed as constituting two major components. The first component is the overall increase in ride-hailing demand possibly drawing from population of individuals driving, using public transit and even inducing newer travel. The second component of the transformation is the shift in the share of traditional taxi service demand toward TNC services (Gerte, Konduri, Ravishanker, Mondal, & Eluru, 2019). In a short time frame, in NYC, TNC services have increased their market share from 0 to nearly 70% by the end of 2018. While preliminary research has begun to explore the reasons for the transformation, it is safe to assume economists and social scientists will continue to examine the transformation for several years into the future.

The proposed study contributes to our understanding of this transformation by examining the NYC data from a fine spatial and temporal resolution by adopting an innovative joint econometric model system. The study examines two components of the transformation (a) the increase in ride hailing demand and (b) the shift from traditional taxi services to TNC services. The first component – taxi zone ride hailing demand - is analyzed adopting a negative binomial count model. The second component - share of traditional and TNC services demand - is analyzed using a multinomial fractional split model. The two model components are stitched together in a joint framework that allows for the influence of repeated observations as well as for the presence of common unobserved factors affecting the two components. The study employs trip level data from the NYC Taxi and Limousine Commission from January 2015 through December 2018 for the analysis. The data is aggregated by taxi zone for every month in the study period and analyzed by ride hailing alternatives: yellow taxi, green taxi and TNC services (including Uber, Lyft, Juno and Via).

The rest of the chapter is organized as follows: The following section presents the detailed description of the dataset with sample formation technique adopted in the analysis while section 3 provides methodological framework. Model results are presented in the fifth section followed by the policy analysis. Final section comprises with the concluding statements.

## 6.2 Data

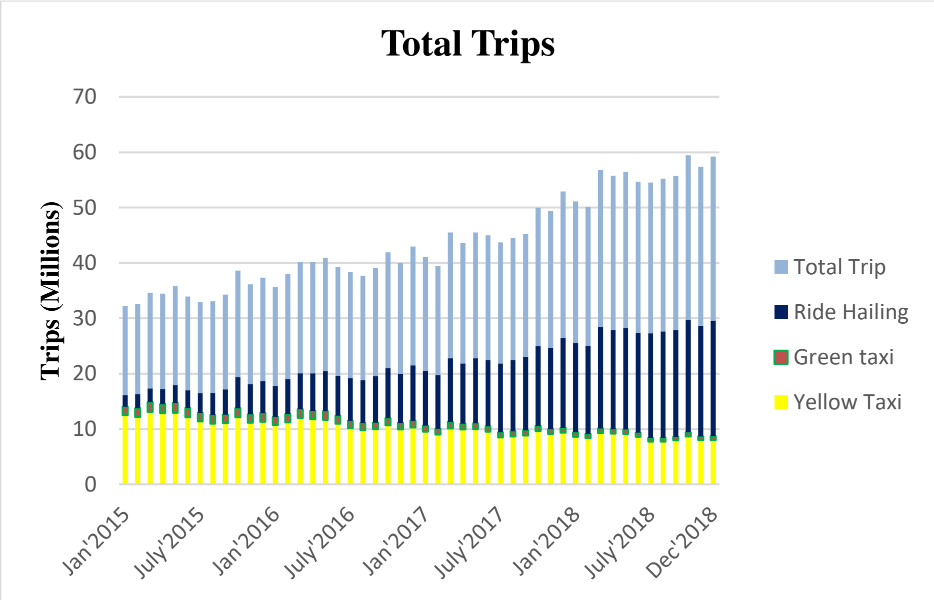
### 6.2.1 Data Source

The NYC Taxi and Limousine Commission (TLC) provides spatially aggregated trip data from all transportation networking companies (taxi, Uber, Lyft, Juno and Via) for public use (https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page). Yellow taxis are traditional and iconic ride hailing service in NYC while green taxis known as boro taxis and street-hail liveries started operation in August 2013. TNCs became operation at around a similar time frame. Thus, it is informative to examine how the share of green taxi and TNCs has evolved with time. The trip itinerary dataset was collected from 2015-2018 for yellow taxi, green taxi and TNC (Uber, Lyft, Juno and Via) for our analysis. The dataset provides information on start and end time of trips, origin and destination defined as taxi zone ID, trip distance and vehicle license number. The trip data was augmented with other sources including: (1) built environment attributes derived from New York City open data (https://nycopendata.socrata.com); (2) socio-demographic characteristics at the census tract/zip code level gathered from US 2010 census data; (3) the weather information corresponding to the Central Park station retrieved from the National Climatic Data Center (http://www.ncdc.noaa.gov/data-access).

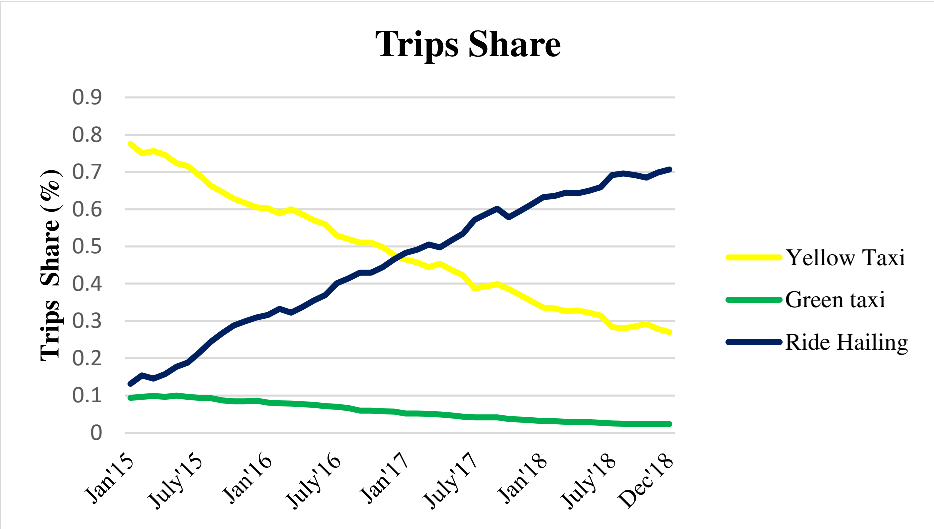
### 6.2.2 Sample Formation and Dependent Variable

A series of data cleaning and compilation exercises were undertaken for generating the sample data for estimation purposes. *First*, trips with missing or inconsistent information were removed. *Second*, trips longer than 500 minutes in duration (around 0.5% of all trips) were deleted considering that these trips are not typical ride-sharing trips. These trips could also be a result of two possibilities; either destination of those trips could be outside NYC or due to technical issues the trip information was recorded incorrectly. *Third*, trips that had the origin and destination outside of NYC taxi zone were also eliminated. Therefore, we focus on trips that originated and were destined within NYC taxi zone region only.

For the given study period (January 2015 to December 2018), the total number of available taxi zones in NYC was 259. Initially, we aggregated pickup data for each month from January 2015 to December 2018 for each origin taxi zone ID. Figure 6.1(a) represents the total trips generated in each month from January 2015 to December 2018 by each ride hailing alternatives while Figure 6.1(b) represents the proportion of total trips shared by yellow taxi, green taxi and TNC services. The evolving number of trips by ride hailing type offers clear depiction of how demand has increased as well as how TNC demand has surpassed traditional taxi demand. TNC service share crossed the share of yellow taxi in February 2017. Figure 6.1(b) represents the trips proportion shared by the three ride hailing alternatives from 2015 to 2018. The Figure highlights TNC's trip share increased from 13% to 70% from 2015-2018 while yellow taxis share declined from 77% to 27%. It is important to note that the share of green taxi dropped consistently to become almost negligible in 2018. The main reason we still retained green taxi as a separate alternative is to contrast two services (green taxi and TNCs) that started operation in the same time frame. For our analysis, we aggregated trip data for 48 months from January 2015 to December 2018. To obtain a reasonable sample size for model estimation, 24 months were randomly selected for each taxi zone for analysis.




**(a) Total Monthly Trips of All Ride Hailing Alternatives.**




**(b) Monthly Trips Share Between Three Ride Hailing Alternatives.**

**Figure 6.1: Dependent Variable Distribution**

### 6.2.3 Exogenous Variables

Several independent variables generated in our study are described below:

Transportation infrastructure attributes created at the taxi zone level include bike route length density (capturing the effect of availability of bicycle facilities on system usage), number of bikeshare stations, length of streets (minor and major streets). Number of subway stations and bus stops in the taxi zone were generated to examine the influence of public transit on rider's preference of mode choice.

Several land use and built environment variables were considered including population density, job density and establishment density, the number of institutional facilities (schools, colleges, hospitals), the number of point of interests (museums, shopping malls), and the number of restaurants (including coffee shops and bars), total area of parks and commercial space (office, industry, retail) within each taxi zones. Distance of destination from Times Square and airport were estimated by using the shortest path algorithm tool of ArcGIS software. Airport indicator variable for the taxi zone was generated to examine the additional impact of airport destination. Population, job density and median income information was collected from US Census for 2015-2017 and extrapolated for 2018. Household car ownership information for 2018 was used to generate proportion of zero car ownership at taxi zone level to examine the impact of car ownership on riders' trip count and mode choice preferences. Non-motorized vehicle score (average of walk score and bike score) and transit score associated with each taxi zone was considered at the census tract level. Further, crime density and accident density were also generated at taxi zone level. Total number of crimes of all types for previous year was aggregated at census tract level and crime density was estimated by dividing corresponding year's population. In a similar manner, total number of accidents for each month was considered to generate accident density.

Weather variables include average temperature, precipitation, and snow for that particular month of the year. Several interaction variables were also created. Seasonality is the one of the temporal variables considered. We consider winter (December-February), Spring

(March-May), Summer (June-August) and Fall (September-November) as dummy variables. Finally, we recognize that technology adoption cannot be explained by simply considering the variables described. To quantify the impact of time, we explicitly consider time elapsed since the beginning of TNC data collection (and other functional forms of the variable) as a temporal variable.

### 6.3 Methodology

The proposed joint econometric system jointly models "total number of trips" and "proportion of trips by type of ride hailing". The first variable is modeled using a Negative Binomial (NB) model and the second variable is analyzed using the multinomial logit fractional split (MNLFS) model. The mathematical details of the Joint NB-MNLFS model follows.

#### 6.3.1 NB Component

Let $i$ be the index for taxi zone $(i = 1,2,3,\dots,N)$ and $y_{it}$ be the ride hailing demand for a taxi zone $i$ in time period $(t = 1,2,3,\dots,T)$. The NB probability expression for random variable $y_{it}$ can be written as (Cameron, Li, Trivedi, & Zimmer, 2004):

$$P_{it}(y_{it}) = \frac{\Gamma(y_{it}+\alpha^{-1})}{\Gamma(y_{it}+1)\Gamma(\alpha^{-1})}\left(\frac{1}{1+\alpha\mu_{it}}\right)^{\frac{1}{\alpha}}\left(1-\frac{1}{1+\alpha\mu_{it}}\right)^{y_{it}} \tag{6.1}$$

where, $P_{it}$ is the probability that taxi zone $i$ has $y_{it}$ number of trips over time period of $t$. $\Gamma(\cdot)$ is the Gamma function, $\alpha$ is the NB dispersion parameter and $\mu_{it}$ is the expected number of trips listed in taxi zone $i$ for time period $t$ and can be expressed using a log-link function as:

$$\mu_{it} = E(y_{it}|\boldsymbol{x}_{it}) = exp\left((\boldsymbol{\partial} + \aleph_i)\boldsymbol{x}_{it} + \delta_{itj} + \varphi_{it}\right) \tag{6.2}$$

where, $\boldsymbol{x}_{it}$ is a vector of explanatory variables associated with taxi zone $i$ for time period $t$. $\boldsymbol{\partial}$ is a vector of coefficients to be estimated. $\aleph_i$ is a vector of unobserved factors on ride hailing demand propensity and its associated zonal characteristics assumed to be a realization from standard normal distribution: $\aleph_i \sim N(0, \varsigma^2)$. $\delta_{itj}$ captures unobserved factors that simultaneously impact total number of trips and proportion of trips by ride hailing type $j$ ($j = 1, 2,3; \mathrm{J} = 3$) for taxi zone $i$ and time period $t$. $\varphi_{it}$ is a gamma distributed error term with mean 1 and variance $\alpha$.

### 6.3.2 MNLFS Component

Let $z_{itj}$ be the fraction of trips by ride hailing type $j$ in taxi zone $i$ and time period $t$.

$$0 \leq z_{itj} \leq 1, \qquad \sum_{j=1}^{J} z_{itj} = 1 \tag{6.3}$$

Let the fraction $z_{itj}$ be a function of a vector $w_{itj}$ of relevant explanatory variables associated with attributes of taxi zone $i$ and time period $j$.

$$E[z_{itj}|w_{itj}] = Q_{itj}(\cdot)$$
$$0 < Q_{itj}(\cdot) < 1, \;\; \sum_{j=1}^{J} Q_{itj}(\cdot) = 1 \tag{6.4}$$

where $Q_{itj}(\cdot)$ is a predetermined function. The properties specified in equation (4) for $Q_{itj}(\cdot)$ warrant that the predicted fractional ride hailing types will range between 0 and 1 and will add up to 1 for each zone. In this study, a MNL functional form for $Q_{itj}$ in the fractional split model of equation (4). Then equation (4) is rewritten as:

$$E\left(z_{itj}\middle|w_{itj}\right) = Q_{itj}(\cdot) = \frac{\exp(\left(\boldsymbol{\beta}'_{\boldsymbol{j}} + \boldsymbol{\sigma_{ij}}\right) w_{itj} \pm \delta_{itj} + \xi_{itj})}{\sum_{j=1}^{J} \exp(\left(\boldsymbol{\beta}'_{\boldsymbol{j}} + \boldsymbol{\sigma_{ij}}\right) w_{ij} \pm \delta_{itj} + \xi_{itj})}, j \tag{6.5}$$

$$= 1,2,3, \dots,$$

where, $\boldsymbol{w}_{itj}$ is a vector of attributes, $\boldsymbol{\beta}'_j$ is the corresponding vector of coefficients to be estimated for ride hailing type $j$. $\boldsymbol{\sigma}_{ij}$ is a vector of unobserved factors assumed to be a realization from standard normal distribution: $\boldsymbol{\sigma} \sim N(0, \boldsymbol{\nu}_j{}^2)$. $\xi_{itj}$ is the random component assumed to follow a Gumbel type 1 distribution. $\delta_{itj}$ term generates the correlation between equations for total number of trips and trip proportions by ride hailing types. The $\pm$ sign in front of $\delta_{itj}$ in equation (5) indicates that the correlation in unobserved zonal factors between total trips and trip proportions by ride hailing type may be positive or negative. A positive sign implies that taxi zones with higher number of trips are intrinsically more likely to incur higher proportions for the corresponding ride hailing types. On the other hand, negative sign implies that taxi zones with higher number of trips intrinsically incur lower proportions for different ride hailing types. To determine the appropriate sign, we empirically test the models with both $'+'$ and $'-'$ signs independently. The model structure that offers the superior data fit is considered as the final model.

It is important to note here that the unobserved heterogeneity between total number of trips and trip proportions by ride hailing types can vary across taxi zones. Therefore, in the current study, the correlation parameter $\theta_{ij}$ is parameterized as a function of observed attributes as follows:

$$\delta_{itj} = \boldsymbol{\pi_j} \boldsymbol{\tau}_{itj} \tag{6.6}$$

where, $\boldsymbol{\tau}_{itj}$ is a vector of exogenous variables, $\boldsymbol{\pi_j}$ is a vector of unknown parameters to be estimated (including a constant).

In examining the model structure of total trip count and proportion of trips by ride hailing types, it is necessary to specify the structure for the unobserved vectors $\boldsymbol{\varsigma}, \boldsymbol{\sigma}$ and $\boldsymbol{\pi}$ represented by Ω. In this paper, it is assumed that these elements are drawn from independent realization from normal population:$\Omega \sim N(0, (\boldsymbol{\varsigma}^2, \boldsymbol{\nu}_j{}^2, \beth_j^2))$. Thus, conditional on Ω, the likelihood function for the joint probability can be expressed as:

$$\mathcal{L}_i = \int_{\Omega} P(y_{it}) \times \prod_{t=1}^{T} \prod_{j=1}^{J} \left(E\left(z_{itj} \middle| w_{itj}\right)\right)^{z_{itj}} f(\Omega) d\Omega \tag{6.7}$$

$z_{itj}$ is the proportion of trips in ride hailing category $j$. Finally, the log-likelihood function is:

$$\mathcal{LL} = \sum_{i} Ln(L_i) \tag{6.8}$$

All the parameters in the model are estimated by maximizing the logarithmic function $\mathcal{LL}$ presented in equation (8). The parameters to be estimated in the joint model are: $\boldsymbol{\partial}, \boldsymbol{\alpha}, \boldsymbol{\beta'}_j, \boldsymbol{\nu}_j$ and $\beth_j$**.** To estimate the proposed joint model, we apply Quasi-Monte Carlo simulation techniques based on the scrambled Halton sequence to approximate this integral in the likelihood function and maximize the logarithm of the resulting simulated likelihood function across individuals (see (Bhat, 2001; Eluru, Bhat, & Hensher, 2008; Yasmin & Eluru, 2013) for examples of Quasi-Monte Carlo approaches in literature).

## 6.4 Estimation Results

6.4.1 NB-MNL Fractional Split Joint Model

Table 6.1 presents the model estimation results of the joint NB-MNL fractional split model. The second column provides the results of the NB component while columns 3 through 5 present the results of the MNL fractional split model. The model results are discussed separately for total ridership demand and proportion by ride hailing alternatives.

6.4.1.1 Total Ridership Demand (NB Component)

A positive (negative) sign for a variable in the ride hailing demand component of Table 6.1 indicates that an increase in the variable is likely to result in more (less) ride hailing trips.

*6.4.1.1.1 Land Use and Built Environment Attributes*

As expected, zones located in census tracts with higher population density are more likely to be associated with higher number of trips. Similarly, increased job density and median income of in taxi zones is found to increase demand for ride hailing trips (see Correa et al. (Correa et al., 2017), Smart et al. (Smart et al., 2015) for similar results). The increased proportion of zero car households in urban areas increases demand for ride hailing (Correa et al. (Correa et al., 2017) found similar association with lower vehicle ownership households). As expected, increased transit accessibility within a taxi zone increases the propensity for higher ride hailing demand while taxi zones with higher non-motorized score reduce the appeal towards use ride hailing. It is possible that the presence of bicycle sharing serves as a competitive alternative for shorter trips.

The presence of activity opportunities in the form of restaurants and cafes, recreational centers and point of interests (POI) is positively associated with demand. Taxi zones with

## Table 6.1: Joint NB-MNLFS Model Estimation Results

| Joint Component | NB Model (Counts) | | MNLFS Model (Proportions) | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Ride hailing Type** | **Total Trips** | | **Yellow Taxi** | | **Green Taxi** | | **TNC** | |
| **Variable Name** | **Estimate** | **t-stat** | **Estimate** | **t-stat** | **Estimate** | **t-stat** | **Estimate** | **t-stat** |
| Constant | -1.426 | -10.40 | 2.688 | 9.44 | 0.639 | 1.42 | --- | --- |
| **Land Use and Built Environment Attributes** | | | | | | | | |
| Population Density | 0.245 | 2.12 | 2.069 | 4.35 | -3.813 | -3.55 | --- | --- |
| Job Density | 2.553 | 19.02 | | | --- | --- | 1.968 | 4.07 |
| Median Income ($x10^{-3}$) | 0.651 | 17.08 | 1.366 | 7.33 | --- | --- | --- | --- |
| Proportion of Zero Car HH | 1.003 | 9.70 | --- | --- | 3.508 | 5.28 | 0.830 | 1.85 |
| Transit Score ($x10^{-2}$) | 1.478 | 8.51 | --- | --- | --- | --- | --- | --- |
| Non-motorized vehicle score ($x10^{-2}$) | -1.189 | -6.34 | --- | --- | --- | --- | --- | --- |
| Number of Restaurants and sidewalk café in Taxi Zone ($x10^{-3}$) | 0.655 | 10.66 | --- | --- | --- | --- | -2.975 | -4.84 |
| Number of Point of Interests and Recreational Points in Taxi Zone ($x10^{-3}$) | 0.194 | 8.52 | 4.459 | 5.04 | --- | --- | --- | --- |
| Residential Area ($m^2$ x $10^{-6}$) | 1.5698 | 8.94 | --- | --- | --- | --- | --- | --- |
| Park Area ($m^2$ x $10^{-6}$) | 1.484 | 10.22 | 16.665 | 4.89 | -5.302 | -2.43 | --- | --- |

| **Joint Component** | **NB Model (Counts)** | | **MNLFS Model (Proportions)** | | | | | |
|---|---|---|---|---|---|---|---|---|
| Airport Indicator | 0.723 | 35.99 | 3.511 | 9.47 | --- | --- | --- | --- |
| Airport Distance (m x $10^{-3}$) | 4.089 | 60.66 | --- | --- | --- | --- | 0.313 | 2.63 |
| Times Square Distance (m x $10^{-3}$) | -1.047 | -35.77 | -2.384 | -14.33 | -0.511 | -2.65 | --- | --- |
| Accident Density (x$10^{-3}$) | --- | --- | --- | --- | -1.684 | -2.53 | --- | --- |
| **Transportation Infrastructure and Attributes** | | | | | | | | |
| Bike Lane Density in Taxi Zone | -1.522 | -8.97 | -2.111 | -2.22 | --- | --- | --- | --- |
| Number of Bikeshare Stations in Taxi Zone (x$10^{-2}$) | -0.059 | -2.65 | --- | --- | -0.322 | -1.97 | --- | --- |
| Street Length in Taxi Zone (m x $10^{-3}$) | 0.401 | 2.30 | -10.183 | -4.15 | --- | --- | --- | --- |
| Number of Bus Stops and Subway Stations in Taxi Zone (x$10^{-3}$) | 1.174 | 62.35 | -3.815 | -4.84 | --- | --- | --- | --- |
| **Temporal and Weather Attributes** | | | | | | | | |
| Times Square Distance (m x $10^{-3}$) x Summer (Season) | -0.577 | -5.65 | --- | --- | --- | --- | --- | --- |
| Time Elapsed as Month Sequel | 2.194 | 33.96 | -0.054 | -14.35 | -0.083 | -18.84 | --- | --- |
| Snow Depth (cm) | -0.031 | -7.26 | 0.281 | 2.97 | --- | --- | --- | --- |
| **Dispersion Parameters** | 0.160 | 27.45 | --- | --- | --- | --- | --- | --- |
| *Correlation* | --- | --- | 0.785 | 10.20 | --- | --- | 0.785 | 10.20 |

1

higher residential area are positively associated with ride hailing demand. The result potentially alludes to the adoption of ride hail service for commute activities from residential zones. As expected, availability of airport in taxi zones increases demand for ride hailing. The presence of park area in the taxi zone has a positive influence on ride hailing demand.

The study also considered the impact of landmarks such as Airports and Times Square on ride hailing demand. The presence of an airport in the taxi zone, as expected, contributes to higher ride hailing demand. Interestingly, as the distance of taxi zone from airports increases, the model indicates an increase in ride hailing demand. On the other hand, as the distance from Times Square increases, ride hailing demand is expected to reduce. The result is intuitive as Times Square and the proximal zones serve as attraction centers for regular and tourist travel.

*6.4.1.1.2 Transportation Infrastructure and Attributes*

Several transportation infrastructure variables such as bike lane density, bikeshare stations, street length, bus stops and subway stations were considered in the demand model. The parameter estimates for bike length indicate that probability of ride hailing trips decreases with increasing bike length density in the taxi zone. The negative association with number of bikeshare stations within a taxi zone highlights that ride hail trip demand is likely in competition with bikeshare demand (for shorter distance share). An increase in the street length within a taxi zone has a positive impact on demand. (similar to findings of Correa et al. (Correa et al., 2017)). The number of bus stops and subway stations in the taxi zone has a positive coefficient indicating an increment in ride hail demand. This result highlights the complementarity between ride hail and public transit alternatives.

#### *6.4.1.1.3 Temporal and Weather Attributes*

An interaction variable of summer season with Times Square distance from each taxi zone was used and the results highlight an interesting result. The results indicate that the ride hail demand in summer reduces faster than rest of the year as we move away from Times Square. The result clearly highlights the attraction of Times Square during summer months for visitors and their plausible adoption of ride hailing. Time elapsed variable that counts the month from January 2015 to December 2018 was used to find the impact of temporal trend attribute on ride hailing trip count. The result highlights the positive association with ride hailing representing how with time overall demand has increased. Finally, as the depth of snow in the taxi zone increases, ride hailing demand reduces. This is expected as trip generation across all modes is likely to reduce under snowy conditions.

### 6.4.1.2 Trip Proportion (MNL Fractional Split Component Model)

In the MNL fractional split model, a positive (negative) sign for a variable indicates that an increase in the variable is likely to result in higher proportion of trips for the corresponding alternative relative to the base alternative for that variable.

#### *6.4.1.2.1 Constant parameters*

The constant parameters have no substantive interpretation after introducing independent variables.

#### *6.4.1.2.2 Land Use and Built Environment Attributes*

In the context of land use and built environment attributes, population density in a census tract has significant impact on trip proportions. Increasing population has a positive impact on yellow taxi proportion and negative impact on green taxi proportion. The result seems reasonable since green taxi has regulations restricting on-street pickup. In a similar vein, with higher job density, the proportion of TNC proportion increases. The result potentially indicates preference among employed individuals for TNC. Taxi zones with high median income have positive association with yellow taxis proportion. The result probably reflects the indifference to typically higher fares of yellow taxi relative to TNCs. With increasing zero car ownership households, the likelihood of green taxi and TNC services trips proportion increases. Zero car household are inclined to adopting TNC services that are usually less expensive compared to taxis.

A negative association is observed for the presence of restaurants and cafes with TNC trip proportions while recreational centers and point of interests (POI) have an increased likelihood for the yellow taxi proportions. In terms of land use type, only proportion of park area variable has significant impact on trip proportions. The likelihood of yellow taxi trips increases for a high percentage of park area in a taxi zone while green taxi trip proportion reduces. As expected, availability of airport in taxi zones increases the inclination of choosing yellow taxis(See similar results for yellow taxi share for airport originated trips (Metcalfe & Warburg, 2012)). As the distance between taxi zone and airport increases, the share of TNC alternative increases. It is possible that TNC services are more readily available in these locations. As the taxi zones are further from Times Square, trip proportions for both taxis reduce reflecting their low accessibility as we move further away from Times Square. The results for accident density from the previous year reveal that taxi zones with higher accident density is likely to reduce green taxi proportion.

#### *6.4.1.2.3 Transportation Infrastructure and Attributes*

Several transportation infrastructure characteristics considered are found to be significant determinants of trip proportions by various ride hailing alternatives. Yellow taxi trip proportions are negatively associated with higher bike length density. Among transportation attributes, trip proportion of green taxi trips is found to be lower for taxi zones with higher bike sharing stations in vicinity while yellow taxi trip proportions are negatively associated with higher number of bus stops in taxi zones. An increase in the street length within a taxi zone results in a decreased of yellow taxi proportions.

#### *6.4.1.2.4 Temporal and Weather Attributes*

Elapsed time considering month is negatively associated with Yellow and green taxi trips proportions. The result suggests that yellow and green taxi trips number reduces with the time elapsed from January 2015 (as expected). The estimated snow depth variable implies a positive effect on yellow taxi trip proportions. It is possible that, under snowy conditions, the inventory of yellow taxi fleet is unchanged while the number of TNC services reduce.

#### *6.4.1.2.5Common Unobserved Parameters*

Several unobserved parameters were tested including: (1) correlation between demand component and ride hailing proportion components, (2) correlation across ride hailing proportion components and (3) random parameters in demand and proportion components. Of these tested parameters only common correlation between trip proportions of yellow taxi and TNC services was significant. The correlation between the two components could be either positive or negative. In our analysis, we found the positive sign to offer better fit. The results indicate that unobserved factors that increase the proportion of yellow taxi also increase the proportion of TNC services.

## 6.5 Performance Evaluation

The estimated models were used to predict the expected ridership at the taxi zone level and the proportion of the three ride hailing alternatives. These generated values were used to estimate the predicted number of trips by each ride hailing alternative. These estimated values are compared to the observed values to evaluate model performance. Three different measures: mean percentage error (MPE), mean absolute percentage error (MAPE) and root mean square error (RMSE) were computed based on the estimates from the joint model. A description of the measures follows:

MPE measures the prediction accuracy and is defined as:

$$\mathrm{MPE} = mean(\frac{\hat{y}_i - y_i}{y_i}) \tag{6.9}$$

The smaller the MPE, the better the model predicts observed data.

MAPE measure the error in terms of percentage and is defined as:

$$\mathrm{MAPE} = mean\left|\frac{\hat{y}_i - y_i}{y_i}\right| \tag{6.10}$$

The smaller the MAPE, the better the model predicts observed data. These measures of fit are generated at disaggregate level: across all crash types and across all observations.

Root Mean Square Error (RMSE) is basically the standard deviation of the residuals (prediction errors). It highlights how much data is concentrated around the best fit line.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \tag{6.11}$$

The measures were generated for the estimation sample as well as for the hold out sample. The hold-out sample was prepared following the same procedure used to extract the estimation sample. We used a sample of 20 months per taxi zone for validation. Fig. 2 presents the values of these measures for joint NB-MNLFS model for estimation and validation datasets. The results highlight that the joint NB-MNLFS model gives quite intuitive result across the various measures computed. The results also highlight the relatively small range of errors for estimation and validation datasets. The model performance does not worsen for validation dataset highlighting the appropriateness of the developed model for analyzing the data.

### 6.6 Policy Analysis

To illustrate how the proposed model can be adopted for future demand prediction, we conduct a hypothetical policy analysis. We consider the independent variables from 2018 to remain constant for the first 6 months of 2019 and examine the number of trips by ride hailing alternative. The model prediction values, thus generated are compared with the observed trips by ride alternative for the corresponding time period. The comparison of the observed and predicted trips by ride alternative are presented in Figure 6.3. The predicted TNC trips increased from 20 million to 25 million from December 2018 through June 2019 while yellow taxi trip reduced from 7.4 million to 6.4 million. Overall, the results clearly indicate a good match between observed and predicted trips by ride alternative. For Yellow taxi, the results compare favorably with slightly larger error in March 2019. From the figures, the reader would note that trips by green taxi have the largest deviation. However, this is an artifact of the small share of green taxi magnifying any shifts in number of trips. For TNC, the observed and predicted trips follow closely except for March 2019. To evaluate the exact mis-match in trip number by ride hailing alternative, we computed percentage error in prediction normalized to total number of trips. The estimated average percentage error for the three ride hailing

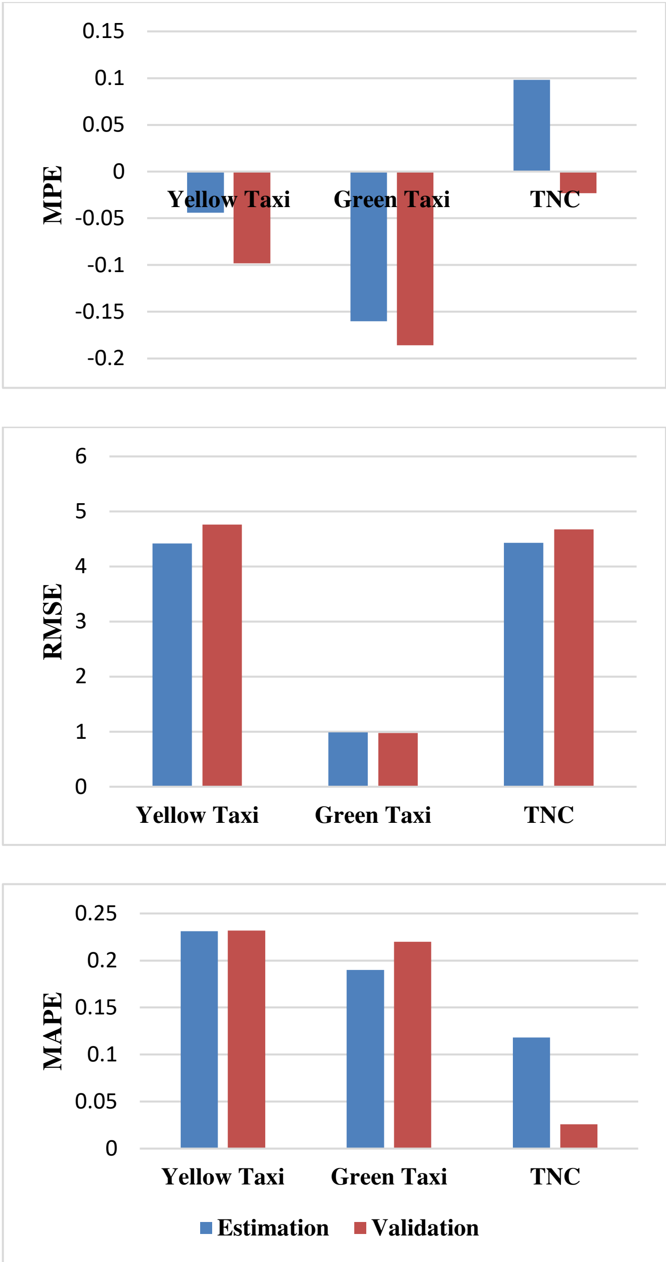



**Figure 6.2: Sample Predictive Performance Measure**

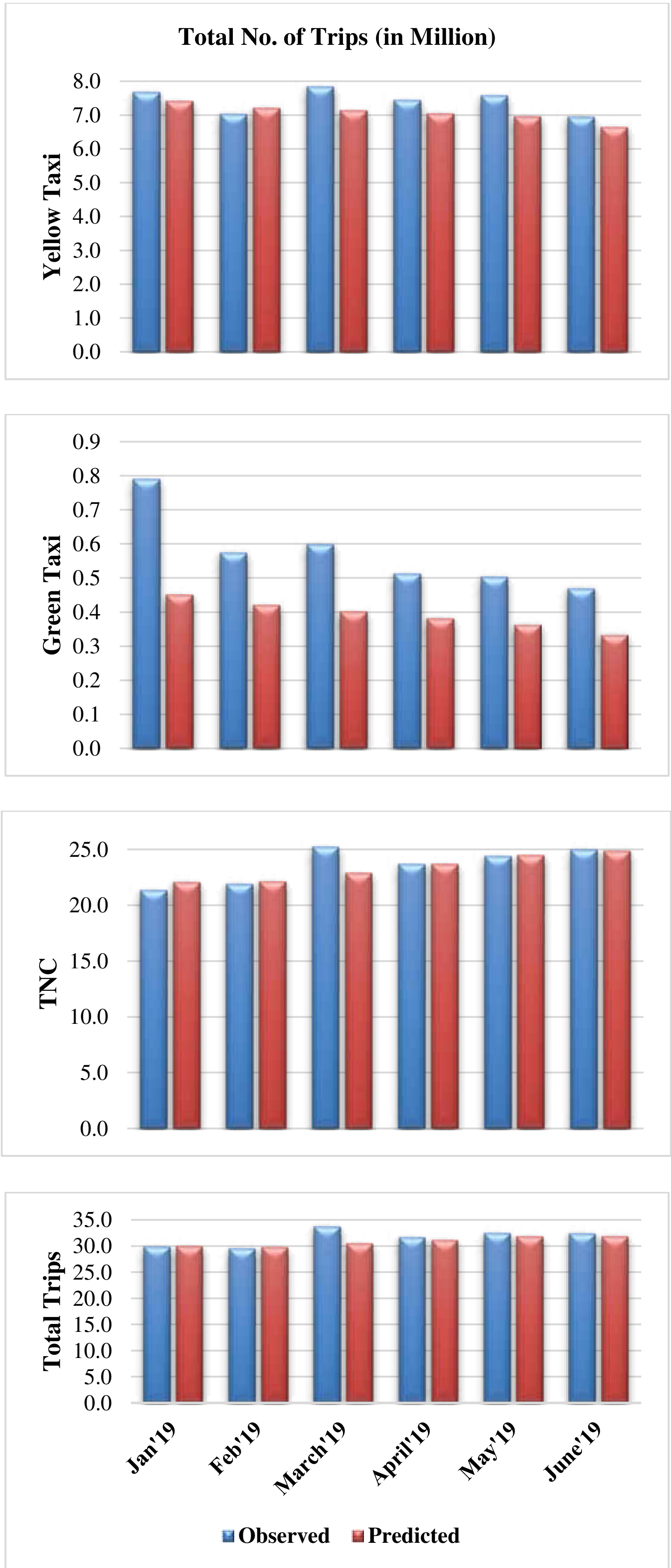



**Figure 6.3: Predicted Trip Comparison**

alternatives (yellow taxi, green taxi and TNC) is 1.29, 0.59 and 1.80% respectively with the range of these errors varying from a minimum of 0.53% through a maximum of 2.11% for yellow taxi, 0.42 through 1.13% for green taxi and 0.02 through 6.90% for TNC. These results also indicate that the maximum error for yellow taxi and TNC was for the month of March. We observed an anomaly in the data for the total number of ride hailing trips in March and this could be the reason for the slightly larger error. In spite of this discrepancy, the proposed model performs adequately. The comparison presented only documents the overall system level performance. The model outputs are provided at a fine spatial resolution that can be employed by city planners and ride hailing operators to effectively plan and manage for changing ride hailing patterns.

### 6.7 Summary

In this chapter, we develop an innovative joint econometric model system to examine two components of the transformation; (a) the increase in ride hailing demand and (b) the shift from traditional taxi services to TNC services. The first component is analyzed adopting a negative binomial (NB) count model while the second component is analyzed using a multinomial fractional split (MNLFS) model. The two model components are stitched together in a joint framework that allows for the influence of repeated observations as well as for the presence of common unobserved factors affecting the two components. The data for our analysis is drawn from New York City Taxi & Limousine Commission (NYTLC) for four years from January 2015 through December 2018. The data is aggregated by taxi zone for every month in the study period and analyzed by ride hailing alternatives: yellow taxi, green taxi and TNC. The model estimation considered a comprehensive set of independent variables including transportation infrastructure variables, land use and built environment variables, weather attributes, and temporal attributes. Several performance measures were generated using the joint NB-MNLFS model for estimation and validation datasets. The results illustrate the excellent performance of the proposed model. Further, to quantify the impact of time, we explicitly consider time elapsed since the beginning of TNC data collection in NYC as a surrogate variable and predicted trips by ride hailing alternative for future time periods.

# CHAPTER 7: CONCLUSIONS AND FUTURE RESEARCH SCOPE

## 7.1 Introduction

The objective of the dissertation is to develop advanced econometric frameworks to address methodological gaps in flow analysis of shared economy literature. Specifically, the primary focus of the current research is on advancing the state of the art in modeling flow or frequency variables for shared economy systems. In this study, we selected accommodation service (AirBnB), bikeshare service (Citi bike, NYC) and rideshare service (UBER/LYFT/Taxi). The proposed research endeavours to identify the various factors that affect the demand to assist policy makers in developing comprehensive planning solutions.

The current dissertation contributes substantially towards empirical and methodological perspectives for shared economy system demand analysis along six directions: (1) appropriate model framework, (2) investigate AirBnB supply as snapshot of AirBnB demand, (3) unobserved heterogeneity within count approach, (4) origin level shared mobility demand, (5) allocate shared mobility demand to the infinite number of alternatives and (6) shift from traditional taxi services to TNC services. In this chapter major conclusions from the earlier chapters are summarized. The rest of the chapter is organized as follows. Sections 7.2 through 7.5 discuss the findings of each chapter briefly alongside the methodological and empirical contributions of the dissertation. Section 7.6 concludes the dissertation by presenting the directions for future research scope.

## 7.2 Analysis of Hospitality Demand

In Chapter three, the current study proposes a copula based model framework together with simulation based multivariate frameworks to address correlation across various exogenous variables in sharing accommodation demand literature. To the best of the authors' knowledge, this is the first attempt to employ such copula based bivariate count models for AirBnB count literature to capture the unobserved heterogeneity with dependency profile. The data for our analysis is drawn from AirBnB listings (Inside AirBnB) for New York City for 31 months from January 2015 through June 2017. A host of exogenous variables including socio-demographic attributes, bicycle infrastructure attributes, land use and built environment, traffic attributes and roadway network attributes are considered. For our analysis, we examine five copula structures: (1) FGM, (2) Frank, (3) Gumbel, (4) Clayton and (5) Joe. Among all negative binomial model and copula framework, mixed Gumbel with parametrization for dependency fit the most suitable model. The model estimation results provide intuitive findings for significance of dependence profile on both listings count in the macro-level analysis. Several attributes like average listings price, number of point of interests and recreational points, transit accessibility, bike length in vicinity, and census tract level variables (such as population density, job density, and income) increase the likelihood of listings count while distance to Times Square decrease the likelihood of the likelihood of listings count.

The model estimates were also augmented by conducting policy analysis including elasticity analysis for both apartment and private or shared room separately and a spatial representation of hotspots for Apartment listings type only. Elasticity effects on two dependent variables are different for various exogenous variables. Rank order of the top five important variables in terms of increasement for the expected number for both apartment and private or shared room counts include: average AirBnB price in CT, historic district, median income per CT, effect of season and employment density. In addition to elasticity effects, a spatial distribution for observed and predicted count of top 10 percent was conducted. The spatial distribution of most tourism prone zone indicated that higher apartment prone zones were clustered around Manhattan borough of NYC. Overall, the policy analysis conducted provided an illustration on how the proposed model can be applied to determine the critical factors contributing to increase in tourism demand as AirBnB counts.

### 7.3 Bikeshare Demand and Origin Destination Flows

In Chapter four, the current study proposes a model framework for bikeshare system usage as well as origin destination flows. We identify two choice dimensions: (1) station level demand and (2) how bike flows from an origin station are distributed across the network. A linear mixed model is considered for modeling weekly origin station demand while a multiple discrete continuous extreme value model (MDCEV) is employed to analyze flows from origin to multiple destinations.

The data for our analysis is drawn from New York City bikeshare system (CitiBike) for six months from January through June, 2017. For our analysis, we examine demand and distribution patterns on a weekly basis. A host of exogenous variables including trip attributes, socio-demographic attributes, bicycle infrastructure attributes, land use and built environment, temporal and weather attributes are considered. The model estimation results provide intuitive findings for both station level demand and destination choice behavior. Several attributes like job density, number of facilities and recreational points, transit and bike accessibility, dock capacity, bike length in vicinity, and census tract level variables (such as population density, job density, and establishment density) increase the preferences for a destination while distance to Time Square, and winter season decrease the likelihood of choosing a destination. In addition to model estimation, a model validation effort was conducted using a hold out sample. The data fit relative to the equal probability MDCEV model highlighted the significant improvement in data fit for the estimated model. Finally, we employed our MDCEV model for prediction to compute the demand for destination stations across the system. We categorized the stations into four quartiles based on observed number of trips and computed the number of correctly classified stations based on our predictions. The result indicates that predicted model performs better in case of high demand destined stations.

### 7.4 Transport Networking Companies (TNC) Demand and Flow

Given the burgeoning growth in ride hailing systems and their growing adoption for trip making, it is important to develop modeling frameworks to understand ride hailing demand flows at the zonal level. Dense urban regions like NYC with well-connected public transit systems can strategically target reducing the reliance on private automobile ownership (and use) by incorporating ride-hailing alternatives in trip planning tools. However, current state-of-practice and travel demand models are not equipped to accurately examine the effects of these services. The research effort of Chapter five contributes to this goal by developing quantitative models of TNC demand and flow distribution patterns. We identify two choice dimensions: (1) a demand component that estimates origin level TNC demand at the taxi zone level and (2) a distribution component that analyzes how these trips from an origin are distributed across the region. The origin level demand is analyzed using linear mixed models while flows from origin to multiple destinations is analyzed using a multiple discrete continuous model system (MDCEV).

The data for our analysis is drawn from New York City Taxi & Limousine Commission (NYTLC) for twelve months from January through December 2018. For our analysis, we examine weekday morning peak hour demand and distribution patterns. The model components are developed using comprehensive set of independent variables including aggregate trip attributes, transportation infrastructure variables, land use and built environment variables, weather attributes, and temporal attributes. The model estimation results provide intuitive findings for both zonal level demand and flow distribution behavior. The model estimates are validated using a hold out sample set aside. The data fit relative to the equal probability MDCEV model highlighted the significant improvement in data fit for the estimated model. Several prediction exercises were also conducted to illustrate the value of the proposed model framework including identifying the top 10 percentile destinations and elasticity effect of changes to independent variables. The policy analysis results offer intuitive results and provide a mechanism for transportation planners to evaluate the impact of various changes on TNC demand and distribution.

## 7.5 Transformation of Ride Hailing

In Chapter six, the current study examines two components of the transportation networking companies induced transformation of ride hailing demand (a) the increase in ride hailing demand and (b) the shift from traditional taxi services to TNC services. The first component is analyzed adopting a negative binomial count model while the second component is analyzed using a multinomial fractional split model. The two model components are stitched together in a joint framework that allows for the influence of repeated observations as well as for the presence of common unobserved factors affecting the two components.

The data for our analysis is drawn from New York City Taxi & Limousine Commission (NYTLC) for four years from January 2015 through December 2018. The model estimation considered a comprehensive set of independent variables including transportation infrastructure variables, land use and built environment variables, weather attributes, and temporal attributes. Several performance measures were generated for the joint NB-MNLFS model for the estimation and validation datasets. The results clearly illustrate how the proposed model provides excellent match with estimation and validation datasets. Finally, a policy illustration is undertaken using independent variables from 2018 to estimate the trips by ride hailing alternatives and their proportions for the first 6 months of 2019. The results indicate that the predicted model tracks the evolving trends by ride hailing alternatives very closely.

### 7.6 Limitations and Future Research Scope

The summary of findings and the contributions of the dissertation in examining shared economy flow analysis are discussed in the preceding sections of this chapter. In this section, the limitations of the research efforts are discussed while offering potential research extensions for the future.

In Chapter three, we employ copula based bivariate count models for AirBnB count literature to capture the unobserved heterogeneity with dependency profile. While the study considers the effect of spatial unobserved heterogeneity in between exogenous variables, it would be more effective to incorporate temporal panel effect on this copula framework to enhance the model in our future work.

In Chapters four and five, we identified two choice dimensions for capturing the shared mobility system origin level demand and investigated how these trips flows from an origin level are distributed across the network. Unlike the traditional MDCEV model, in our context, the number of alternatives are substantially larger. Hence, we resort to estimating a single utility across alternatives (analogous to how multinomial logit based location choice models are estimated with a single utility equation). In our research context, bikeshare and TNC trips need to be allocated within 573 destination stations and 261 destination taxi zone respectively. Given the large number of alternatives, the model run times were substantially long affecting number of specifications we can test. In our analysis, unobserved effects arising from repetitions in the MDCEV model were not captured. Another potential avenue for future research is the consideration of sampling for MDCEV models (similar to sampling in MNL models).

Finally, in Chapter six, we examine shared mobility system demand transformation over the time period and the shift from traditional taxi services to TNC services by developing an innovative joint econometric model system. It might be interesting to enhance the study methodology by accounting for unobserved temporal effects (heteroscedasticity) across the multiple years of data. In future efforts, it might also be useful to include monthly economic indicators (such as employment and wages) in the model to control for macroeconomic condition.

## REFERENCES


Adamiak, C. (2018). Mapping AirBnB supply in European cities. *Annals of Tourism Research, 71*(C), 67-71.

Agatz, N., Erera, A., Savelsbergh, M., & Wang, X. (2012). Optimization for dynamic ride-sharing: A review. *European Journal of Operational Research, 223*(2), 295-303.

Ahn, J., Jeong, G., & Kim, Y. (2008). A forecast of household ownership and use of alternative fuel vehicles: A multiple discrete-continuous choice approach. *Energy Economics, 30*(5), 2091-2104.

Alemi, F., Circella, G., Handy, S., & Mokhtarian, P. (2018). What influences travelers to use Uber? Exploring the factors affecting the adoption of on-demand ride services in California. *Travel Behaviour and Society, 13*, 88-104.

Amey, A., Attanucci, J., & Mishalani, R. (2011). Real-time ridesharing: opportunities and challenges in using mobile phone technology to improve rideshare services. *Transportation Research Record, 2217*(1), 103-110.

Anowar, S., Eluru, N., & Hatzopoulou, M. (2017). Quantifying the value of a clean ride: How far would you bicycle to avoid exposure to traffic-related air pollution? *Transportation Research Part A: Policy and Practice, 105*, 66-78.

Anowar, S., Yasmin, S., Eluru, N., & Miranda-Moreno, L. F. (2014). Analyzing car ownership in Quebec City: a comparison of traditional and latent class ordered and unordered models. *Transportation, 41*(5), 1013-1039.

Babar, Y., & Burtch, G. (2017). Examining the impact of ridehailing services on public transit use. *Available at SSRN 3042805*.

Barron, K., Kung, E., & Proserpio, D. (2018). *The Sharing Economy and Housing Affordability*. Retrieved from

Bhat, C. R. (1997). An endogenous segmentation mode choice model with an application to intercity travel. *Transportation science, 31*(1), 34-48.

Bhat, C. R. (2001). Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model. *Transportation Research Part B: Methodological, 35*(7), 677-693.

Bhat, C. R. (2005). A multiple discrete–continuous extreme value model: formulation and application to discretionary time-use decisions. *Transportation Research Part B: Methodological, 39*(8), 679-707.

Bhat, C. R. (2008). The multiple discrete-continuous extreme value (MDCEV) model: role of utility function parameters, identification considerations, and model extensions. *Transportation Research Part B: Methodological, 42*(3), 274-303.

Bhat, C. R., & Eluru, N. (2009). A copula-based approach to accommodate residential self-selection effects in travel behavior modeling. *Transportation Research Part B: Methodological, 43*(7), 749-765.

Bhat, C. R., & Eluru, N. (2010). *The multiple discrete-continuous extreme value (MDCEV) model: formulation and applications.* Paper presented at the Choice Modelling: The State-of-the-art and The State-of-practice: Proceedings from the Inaugural International Choice Modelling Conference.

Bhat, C. R., Sen, S., & Eluru, N. (2009). The impact of demographics, built environment attributes, vehicle characteristics, and gasoline prices on household vehicle holdings and use. *Transportation Research Part B: Methodological, 43*(1), 1-18.

Bhowmik, T., Yasmin, S., & Eluru, N. (2018). A joint econometric approach for modeling crash counts by collision type. *Analytic methods in accident research, 19*, 16-32.

Botsman, R., & Rogers, R. (2011). What's mine is yours: how collaborative consumption is changing the way we live.

Brochado, A., Troilo, M., & Aditya, S. (2017). AirBnB customer experience: evidence of convergence across three countries. *Annals of Tourism Research, 63*, 210-212.

Brodeur, A., & Nield, K. (2016). Has uber made it easier to get a ride in the rain?

Burghout, W., Rigole, P. J., & Andreasson, I. (2015). *Impacts of shared autonomous taxis in a metropolitan area.* Paper presented at the Proceedings of the 94th annual meeting of the Transportation Research Board, Washington DC, 2015.

Cameron, A. C., Li, T., Trivedi, P. K., & Zimmer, D. M. (2004). Modelling the differences in counted outcomes using bivariate copula models with application to mismeasured counts. *The Econometrics Journal, 7*(2), 566-584.

Çetin, T., & Eryigit, K. Y. (2011). Estimating the effects of entry regulation in the Istanbul taxicab market. *Transportation Research Part A: Policy and Practice, 45*(6), 476-484.

Chan, N. D., & Shaheen, S. A. (2012). Ridesharing in North America: Past, present, and future. *Transport Reviews, 32*(1), 93-112.

Chang, S. J., & Chu, C.-H. (2009). Taxi vacancy rate, fare, and subsidy with maximum social willingness-to-pay under log-linear demand function. *Transportation Research Record, 2111*(1), 90-99.

Chen, L., Mislove, A., & Wilson, C. (2015). *Peeking beneath the hood of uber.* Paper presented at the Proceedings of the 2015 Internet Measurement Conference.

Chen, M. K., & Sheldon, M. (2015). Dynamic pricing in a labor market: Surge pricing and the supply of Uber driver-partners. *University of California (Los Angeles) Working Paper URL [http://citeseerx](http://citeseerx). ist. psu. edu/viewdoc/download.*

Chrysostomou, K., Georgakis, A., Morfoulaki, M., Kotoula, K., & Myrovali, G. (2016). Using Big Taxi GPS Data to Investigate Feasibility of Electric Taxis in Thessaloniki, Greece.

Choi, K. H., Jung, J. H., Ryu, S. Y., Kim, S.-D., & Yoon, S. M. (2015). The relationship between AirBnB and the hotel revenue: in the case of Korea. *Indian Journal of Science and Technology, 8*(26), 1-8.

Collins, L. M., Fidler, P. L., Wugalter, S. E., & Long, J. D. (1993). Goodness-of-fit testing for latent class models. *Multivariate Behavioral Research, 28*(3), 375-389.

Correa, D., Xie, K., & Ozbay, K. (2017). *Exploring the taxi and Uber demand in New York City: An empirical analysis and spatial modeling*. Retrieved from

Coyle, D., & Yeung, T. (2016). Understanding AirBnB in fourteen European cities. *The Jean-Jacques Laffont Digital Chair Working Papers*.

Cramer, J., & Krueger, A. B. (2016). Disruptive change in the taxi business: The case of Uber. *American Economic Review, 106*(5), 177-182.

Dalziel, J. R., & Job, R. S. (1997). Motor vehicle accidents, fatigue and optimism bias in taxi drivers. *Accident Analysis & Prevention, 29*(4), 489-494.

Davidson, A., Peters, J., & Brakewood, C. (2017). Interactive travel modes: Uber, transit, and mobility in New York City.

Dey, B. K., Anowar, S., Eluru, N., & Hatzopoulou, M. (2018). Accommodating exogenous variable and decision rule heterogeneity in discrete choice models: Application to bicyclist route choice. *PloS one, 13*(11), e0208309.

Dey, B. K., Anowar, S., & Eluru, N. (2019). A Framework for Estimating Bikeshare Origin Destination Flows Using a Multiple Discrete Continuous System.

Edelman, B., Luca, M., & Svirsky, D. (2017). Racial discrimination in the sharing economy: Evidence from a field experiment. *American Economic Journal: Applied Economics, 9*(2), 1-22.

Edelman, B. G., & Geradin, D. (2015). Efficiencies and regulatory shortcuts: How should we regulate companies like AirBnB and Uber. *Stan. Tech. L. Rev., 19*, 293.

El-Assi, W., Mahmoud, M. S., & Habib, K. N. (2017). Effects of built environment and weather on bike sharing demand: a station level analysis of commercial bike sharing in Toronto. *Transportation, 44*(3), 589-613.

Eluru, N., Bhat, C. R., & Hensher, D. A. (2008). A mixed generalized ordered response model for examining pedestrian and bicyclist injury severity level in traffic crashes. *Accident Analysis & Prevention, 40*(3), 1033-1054.

Eluru, N., Bagheri, M., Miranda-Moreno, L. F., & Fu, L. (2012). A latent class modeling approach for identifying vehicle driver injury severity factors at highway-railway crossings. *Accident Analysis & Prevention, 47*, 119-127.

Eluru, N., & Bhat, C. R. (2007). A joint econometric analysis of seat belt use and crash-related injury severity. *Accident Analysis & Prevention, 39*(5), 1037-1049.

Ert, E., Fleischer, A., & Magen, N. (2016). Trust and reputation in the sharing economy: The role of personal photos in AirBnB. *Tourism Management, 55*, 62-73.

Faghih-Imani, A., & Eluru, N. (2015). Analysing bicycle-sharing system user destination choice preferences: Chicago's Divvy system. *Journal of Transport Geography, 44*, 53-64.

Faghih-Imani, A., & Eluru, N. (2016). Incorporating the impact of spatio-temporal interactions on bicycle sharing system demand: A case study of New York CitiBike system. *Journal of transport geography, 54*, 218-227.

Faghih-Imani, A., & Eluru, N. (2017a). Examining the impact of sample size in the analysis of bicycle-sharing systems. *Transportmetrica A: transport science, 13*(2), 139-161.

Faghih-Imani, A., & Eluru, N. (2017b). A finite mixture modeling approach to examine New York City bicycle sharing system (CitiBike) users' destination preferences. *Transportation*, 1-25.

Faghih-Imani, A., Anowar, S., Miller, E. J., & Eluru, N. (2017). Hail a cab or ride a bike? A travel time comparison of taxi and bicycle-sharing systems in New York City. *Transportation Research Part A: Policy and Practice, 101*, 11-21.

Faghih-Imani, A., Eluru, N., El-Geneidy, A. M., Rabbat, M., & Haq, U. (2014). How land-use and urban form impact bicycle flows: evidence from the bicycle-sharing system (BIXI) in Montreal. *Journal of Transport Geography, 41*, 306-314.

Faghih-Imani, A., Eluru, N., & Paleti, R. (2017). How bicycling sharing system usage is affected by land use and urban form: analysis from system and user perspectives. *European Journal of Transport and Infrastructure Research, 17*(3).

Fang, H. A. (2008). A discrete–continuous model of households' vehicle choice and usage, with an application to the effects of residential density. *Transportation Research Part B: Methodological, 42*(9), 736-758.

Feigon, S., & Murphy, C. (2016). *Shared mobility and the transformation of public transit*.

Ferdous, N., Pinjari, A. R., Bhat, C. R., & Pendyala, R. M. (2010). A comprehensive analysis of household transportation expenditures relative to other goods and services: an application to United States consumer expenditure data. *Transportation, 37*(3), 363-390.

Fishman, E., Washington, S., & Haworth, N. (2013). Bike share: a synthesis of the literature. *Transport reviews, 33*(2), 148-165.

Fradkin, A., Grewal, E., & Holtz, D. (2018). The determinants of online review informativeness: Evidence from field experiments on AirBnB.

Furuhata, M., Dessouky, M., Ordóñez, F., Brunet, M.-E., Wang, X., & Koenig, S. (2013). Ridesharing: The state-of-the-art and future directions. *Transportation Research Part B: Methodological, 57*, 28-46.

Gerte, R., Konduri, K. C., & Eluru, N. (2018). Is there a limit to adoption of dynamic ridesharing systems? Evidence from analysis of Uber demand data from New York City. *Transportation Research Record, 2672*(42), 127-136.

Gerte, R., Konduri, K. C., Ravishanker, N., Mondal, A., & Eluru, N. (2019). Understanding the Relationships between Demand for Shared Ride Modes: Case Study using Open Data from New York City. *Transportation Research Record*, 0361198119849584.

Guttentag, D. (2015). AirBnB: disruptive innovation and the rise of an informal tourism accommodation sector. *Current issues in Tourism, 18*(12), 1192-1217.

Guttentag, D. A., & Smith, S. L. (2017). Assessing AirBnB as a disruptive innovation relative to hotels: Substitution and comparative performance expectations. *International Journal of Hospitality Management, 64*, 1-10.

Guo, S., Liu, Y., Xu, K., & Chiu, D. M. (2017). *Understanding ride-on-demand service: Demand and dynamic pricing.* Paper presented at the 2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops).

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association, 72*(358), 320-338.

Jäppinen, S., Toivonen, T., & Salonen, M. (2013). Modelling the potential effect of shared bicycles on public transport travel times in Greater Helsinki: An open data approach. *Applied Geography, 43*, 13-24.

Juul, M. (2015). The sharing economy and tourism: Tourist accommodation. *European Parliamentary Research Service. Disponível em: http://www. europarl. europa. eu/RegData/etudes/BRIE/2015/568345/EPRS_BRI (2015) 568345_EN. pdf. Acesso em, 26*(05), 2016.

Jung, J., Chow, J. Y., Jayakrishnan, R., & Park, J. Y. (2014). Stochastic dynamic itinerary interception refueling location problem with queue delay for electric taxi charging stations. *Transportation Research Part C: Emerging Technologies, 40*, 123-142.

Kaufman, S. M., Gordon-Koven, L., Levenson, N., & Moss, M. L. (2015). Citi Bike: The First Two Years.

Komanduri, A., Wafa, Z., Proussaloglou, K., & Jacobs, S. (2018). Assessing the impact of app-based ride share systems in an urban context: Findings from Austin. *Transportation Research Record, 2672*(7), 34-46.

Lam, L. T. (2004). Environmental factors associated with crash-related mortality and injury among taxi drivers in New South Wales, Australia. *Accident Analysis & Prevention, 36*(5), 905-908.

Lee, D. (2016). How AirBnB short-term rentals exacerbate Los Angeles's affordable housing crisis: Analysis and policy recommendations. *Harv. L. & Pol'y Rev., 10*, 229.

Lee, J., Yasmin, S., Eluru, N., Abdel-Aty, M., & Cai, Q. (2018). Analysis of crash proportion by vehicle type at traffic analysis zone level: A mixed fractional split multinomial logit modeling approach with spatial effects. *Accident Analysis & Prevention, 111*, 12-22.

Meddin, R., & DeMaio, P. (2016). The bike-sharing world map.

Metcalfe, G., & Warburg, J. (2012). A Policy agenda for the sharing economy. *The Urbanist*.

Milioti, C. P., Karlaftis, M. G., & Spyropoulou, D. (2015). Impact of the Financial Recession on the Taxi Market: Perceptions of Travelers and Taxi Drivers. *Transportation Research Record, 2536*(1), 67-75.

Murphy, C. (2016). Shared mobility and the transformation of public transit.

Nashad, T., Yasmin, S., Eluru, N., Lee, J., & Abdel-Aty, M. A. (2016). Joint modeling of pedestrian and bicycle crashes: copula-based approach. *Transportation Research Record: Journal of the Transportation Research Board*(2601), 119-127.

Neeser, D., Peitz, M., & Stuhler, J. (2015). Does AirBnB hurt hotel business: Evidence from the Nordic countries. *Universidad Carlos III de Madrid*, 1-26.

Nurul Habib, K. M., & Miller, E. J. (2009). Modelling activity generation: a utility-based model for activity-agenda formation. *Transportmetrica, 5*(1), 3-23.

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural equation modeling: A multidisciplinary Journal, 14*(4), 535-569.

Oskam, J., & Boswijk, A. (2016). AirBnB: the future of networked hospitality businesses. *Journal of Tourism Futures, 2*(1), 22-42.

Peltzer, K., & Renner, W. (2003). Superstition, risk-taking and risk perception of accidents among South African taxi drivers. *Accident Analysis & Prevention, 35*(4), 619-623.

Phaneuf, D. J., Kling, C. L., & Herriges, J. A. (2000). Estimation and welfare calculations in a generalized corner solution model with an application to recreation demand. *Review of Economics and Statistics, 82*(1), 83-92.

Pinjari, A. R., & Bhat, C. (2010). A Multiple Discrete–Continuous Nested Extreme Value (MDCNEV) model: formulation and application to non-worker activity time-use and timing behavior on weekdays. *Transportation Research Part B: Methodological, 44*(4), 562-583.

Pinjari, A. R., Bhat, C. R., & Hensher, D. A. (2009). Residential self-selection effects in an activity time-use behavior model. *Transportation Research Part B: Methodological, 43*(7), 729-748.

Poulsen, L. K., Dekkers, D., Wagenaar, N., Snijders, W., Lewinsky, B., Mukkamala, R. R., & Vatrapu, R. (2016). *Green Cabs vs. Uber in New York City.* Paper presented at the 2016 IEEE International Congress on Big Data (BigData Congress).

Proserpio, D., & Tellis, G. J. (2017). Baring the Sharing Economy: Concepts, Classification, Findings, and Future Directions. *Classification, Findings, and Future Directions (December 28, 2017).*

Quattrone, G., Proserpio, D., Quercia, D., Capra, L., & Musolesi, M. (2016). *Who benefits from the sharing economy of AirBnB?* Paper presented at the Proceedings of the 25th international conference on world wide web.

Rajagopalan, B., Pinjari, A., & Bhat, C. (2009). Comprehensive model of worker nonwork-activity time use and timing behavior. *Transportation Research Record: Journal of the Transportation Research Board*(2134), 51-62.

Rajagopalan, B., & Srinivasan, K. (2008). Integrating household-level mode choice and modal expenditure decisions in a developing country: Multiple discrete-continuous extreme value model. *Transportation Research Record: Journal of the Transportation Research Board*(2076), 41-51.

Rayle, L., Dai, D., Chan, N., Cervero, R., & Shaheen, S. (2016). Just a better taxi? A survey-based comparison of taxis, transit, and ridesourcing services in San Francisco. *Transport policy, 45*, 168-178.

Rixey, R. (2013). Station-level forecasting of bikesharing ridership: Station Network Effects in Three US Systems. *Transportation Research Record: Journal of the Transportation Research Board*(2387), 46-55.

Sans, A. A., & Quaglieri, A. (2016). Unravelling AirBnB: urban perspectives from Barcelona. *Reinventing the local in tourism: Producing, consuming and negotiating place, 73*, 209.

Schaller, B. (2007). Entry controls in taxi regulation: Implications of US and Canadian experience for taxi regulation and deregulation. *Transport policy, 14*(6), 490-506.

Silver, N., & Fischer-Baum, R. (2016). Public transit should be Uber's new best friend (2015).

Sklar, A. (1973). Random variables, joint distribution functions, and copulas. *Kybernetika, 9*(6), (449)-460.

Smart, R., Rowe, B., & Hawken, A. (2015). Faster and cheaper: How ride-sourcing fills a gap in low-income Los Angeles neighborhoods.

Sperling, G. (2015). How AirBnB combats middle class income stagnation. *Retrieved March, 2*, 2016.

Sun, C., & Edara, P. (2015). Is getting an Uber-Lyft from a sidecar different from hailing a taxi? Current dynamic ridesharing controversy. *Transportation Research Record, 2536*(1), 60- 66.

Tay, R., & Choi, J. (2016). Factors associated with crashes involving taxi owners and non-owners: A case of moral hazard and adverse selection? *Accident Analysis & Prevention, 87*, 78-82.

Tseng, C.-M. (2013). Operating styles, working time and daily driving distance in relation to a taxi driver's speeding offenses in Taiwan. *Accident Analysis & Prevention, 52*, 1-8.

von Haefen, R. H. (2004). Empirical strategies for incorporating weak complementarity into continuous demand system models. *Unpublished, Department of Agricultural and Resource Economics, University of Arizona*.

von Haefen, R. H., & Phaneuf, D. J. (2005). Kuhn-Tucker demand system approaches to non-market valuation *Applications of Simulation Methods in Environmental and Resource Economics* (pp. 135-157): Springer.

Wang, C., Komanduri, A., Viswanathan, K., Rossi, T., & West, R. (2018). *A Framework for Quantifying AirBnB Supply, Occupancy Rates and Travel Purpose to Support Visitor Modeling*. Retrieved from

Wang, D., & Nicolau, J. L. (2017). Price determinants of sharing economy based accommodation rental: A study of listings from 33 cities on AirBnB. com. *International Journal of Hospitality Management, 62*, 120-131.

Wang, X., Lindsey, G., Schoner, J. E., & Harrison, A. (2015). Modeling bike share station activity: Effects of nearby businesses and jobs on trips to and from stations. *Journal of Urban Planning and Development, 142*(1), 04015001.

Wong, K., Wong, S., Yang, H., & Wu, J. (2008). Modeling urban taxi services with multiple user classes and vehicle modes. *Transportation Research Part B: Methodological, 42*(10), 985-1007.

Wong, R., Szeto, W., & Wong, S. (2014). Bi-level decisions of vacant taxi drivers traveling towards taxi stands in customer-search: Modeling methodology and policy implications. *Transport policy, 33*, 73-81.

Wong, R., Szeto, W., & Wong, S. (2015). A two-stage approach to modeling vacant taxi movements. *Transportation Research Procedia, 7*, 254-275.

Yang, H., & Wong, S. (1998). A network model of urban taxi services. *Transportation Research Part B: Methodological, 32*(4), 235-246.

Yasmin, S., & Eluru, N. (2013). Evaluating alternate discrete outcome frameworks for modeling crash injury severity. *Accident Analysis & Prevention, 59*, 506-521.

Yasmin, S., & Eluru, N. (2018). A joint econometric framework for modeling crash counts by severity. *Transportmetrica A: transport science, 14*(3), 230-255.

Yasmin, S., Eluru, N., & Ukkusuri, S. V. (2014). Alternative ordered response frameworks for examining pedestrian injury severity in New York City. *Journal of Transportation Safety & Security, 6*(4), 275-300.

Zervas, G., Proserpio, D., & Byers, J. (2015). A first look at online reputation on AirBnB, where every stay is above average. *Where Every Stay is Above Average (January 28, 2015).*

Zervas, G., Proserpio, D., & Byers, J. W. (2017). The rise of the sharing economy: Estimating the impact of AirBnB on the hotel industry. *Journal of marketing research, 54*(5), 687-705.

Zhan, X., & Ukkusuri, S. V. (2015). Probabilistic urban link travel time estimation model using large-scale taxi trip data.

Zhang, W., & Ukkusuri, S. V. (2016). Optimal fleet size and fare setting in emerging taxi markets with stochastic demand. *Computer-Aided Civil and Infrastructure Engineering, 31*(9), 647-660.

Zhang, W., Ukkusuri, S. V., & Lu, J. J. (2017). Impacts of urban built environment on empty taxi trips using limited geolocation data. *Transportation, 44*(6), 1445-1473.

Zhao, J., Deng, W., & Song, Y. (2014). Ridership and effectiveness of bikesharing: The effects of urban features and system characteristics on daily use and turnover rate of public bikes in China. *Transport Policy, 35*, 253-264.