UC Irvine

UC Irvine Electronic Theses and Dissertations

Title

Potential for taxi ridesharing in New York City

Permalink

https://escholarship.org/uc/item/1bt1n7pm

Author

Cheng, Jiaqi

Publication Date

2018

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA,
IRVINE

Potential for taxi ridesharing in New York City

THESIS

submitted in partial satisfaction of the requirements
for the degree of

MASTER OF SCIENCE

in Engineering

by

Jiaqi Cheng

Thesis Committee:
Professor Jean-Daniel Saphores, Chair
Associate Professor Wenlong Jin
Associate Professor Jae Hong Kim

2018

© 2018 Jiaqi Cheng

DEDICATION

To

my parents and friends

in recognition of their worth

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to express the deepest appreciation to my committee chair, Professor Jean-Daniel Saphores. The door to Professor Saphores office was always open whenever I ran into a trouble or had a question about my research. He consistently allowed this thesis to be my own work, and always steered me in the right direction whenever I needed it. Without his guidance and persistent help this thesis would not have been possible.

I would also like to thank my committee members, Professor Wenlong Jin and Professor Jae Hong Kim, whose work and research always inspired me in the transportation engineering and urban planning field. They are truly dedicated mentors, and I have very fond memories of my time in their classes and talking to them.

Finally, I must express my very profound gratitude to my parents, my grandparents, and to my friends Jingyu Liu and Lu Xu for providing me with unfailing support and continuous encouragement, throughout my three years of study at UCI. They are the ultimate role models and the best gifts. I am so thankful that I have their unconditional support.

ABSTRACT OF THE THESIS

Potential for taxi ridesharing in New York City

By

Jiaqi Cheng

Master of Science in School of Engineering
University of California, Irvine, 2018
Professor Jean-Daniel Saphores, Chair

Taxi ridesharing (TRS) is the urban transport alternative that matches separate individual rides to a shared ride, with similar spatial and temporal features. TRS provides a variety of benefits, including saving money for customers, reducing operating costs for taxi operators, cutting the emissions of greenhouse gases and of various air pollutants, and decreasing vehicle miles traveled (VMT). After developing a static one-to-one matching algorithm to evaluate the potential for this approach, I analyzed the year 2014 records of a "big data" public dataset of over 160 million records of canary yellow taxi trips in New York City. I found that approximately 48% to 52% of taxi trips could be shared, with relatively small monthly variation. For the whole year 2014, VMT could be reduced by 98.5 million miles, customers could save \$400 million, emissions of CO_2 could be reduced by 87.8 million lb., and gasoline consumption could be decreased by 4.5 million gallons. To understand the characteristics of the areas where taxi ride sharing is most promising, I also estimated a Tobit model with census tract socio-economic and land use variable. Results suggest that education, age (generation), car ownership, and employment density have a significant impact on taxi ridesharing.

INTRODUCTION

As cities continue to grow, there is increasing pressure on urban environmental quality and urban infrastructure in general, and on transportation services in particular. Fortunately, recent developments in information and communication technology (Cohen, 2015, p. 87) have spawned the emergence of a robust sharing economy, which could relieve some of the urban growing pains. According to the National League of Cities (2015, p. 4 & 7), from creative technologies and business models to redefined concepts of justice and happiness, the sharing economy is creating major shifts in cities around the world. In particular, sharing rides (or vehicles) and shifting to active modes has the potential to decrease congestion in increasingly crowded urban areas. In this context, the goal of this thesis is to explore the potential for taxi ridesharing in New York City based on "one-to-one" matching and to explain this potential in Manhattan based on land use and socio-economic characteristics.

Conventional taxis offer a nice example of how an established industry can be disrupted by information technology and now by sharing. Building on their success in promoting new forms of mobility, companies like Uber and Lyft are now promoting ridesharing services (via Uberpool and Lyft Line) that match individual who want to travel at approximately the same time with drivers. The emergence of these services corresponds to the increasing emphasis on more sustainable consumption, especially in mega cities like New York City (e.g., see Chen, Liu & Chen, 2010, p. 1; Santi *et al.*, 2014, p. 13290).

The idea of sharing taxi rides is not new. In fact, it has been used in some developing countries for several years (Hosni, Naoum-Sawaya, & Artail, 2014, p. 304). In recent years, several approaches have been proposed to foster taxi ride sharing in developed countries. A number of studies have focused on dynamic ‘many-to-many’ approach, where trips with different origins and destinations can be shared (Martinez, Correia, & Viegas, 2015, p. 476; Santos & Xavier, 2015, p. 6729; Santi *et al.*, 2014, p. 13294). Although this approach appears to have an enormous potential, it has seen few field tests apart from Uber Pool (Barann, Beverungen, & Muller, 2017, p. 85), and it suffers from a number of problems, including its complexity and people’s willingness to share rides with complete strangers without any say.

An alternative proposed by Barann, Beverungen, and Muller (2017, p. 86-88) is the ‘one-to-one’ approach, which only matches trips that have close origins and destinations. Although the ‘one-to-one’ alternative has a lower potential than the ‘many-to-many’ approach, it is much easier to implement and it is likely to receive a higher degree of customer acceptance.

In this thesis, I explore the potential of the ‘one-to-one’ by analyzing for a whole year yellow taxi rides for New York City (NYC) from the open dataset managed by the NYC Taxi & Limousine Commission. The volume of data makes this research a big data endeavor. Understanding the potential for taxi ridesharing entails estimating the percentage of trips that can be merged or shared, and quantifying some of the resulting benefits (e.g., on avoided VMT, gasoline saved, and pollutants reduction).

To further understand the potential for taxi ridesharing in NYC, I estimated a Tobit model at the census tract level based on land use and socio-economic variables from the census. To my knowledge, this is the first study to explain the potential for taxi ridesharing.

My methodology and empirical results should be useful to taxi operators but also to planners and decision-makers concerned with implementing more sustainable transportation systems in large urban areas.

CHAPTER 1: LITERATURE REVIEW

In this chapter, I first briefly introduce some concepts related to taxi ridesharing, before reviewing some research on the benefits of taxi ridesharing with an emphasis on studies that also analyzed NYC taxi data. I then provide a brief overview of taxi demand modeling.

Ridesharing

According to Cohen and Munoz (2015, p. 2), Posesn (2015, p. 406), and Furuhata, *et al.* (2013, p. 29), ridesharing refers to the joint and collaborated trip of at least one driver and one rider by matching riders with similar time schedules and itineraries to a shared vehicle. Organized ridesharing has become very popular in recent years. Approaches range from simple online bulletin boards to more elaborate decision support systems that provide automatic matching (Hosni, Naoum-Sawaya, & Artail, 2014, p. 304; Chen, Liu, & Chen, 2010, p. 2). Organized matching can either be provided by matching agencies (such as Uber) or by ridesharing operators. When matching is performed by a matching agency, it involves either static carpooling with pre-arranged trips, or dynamic ridesharing. Through information and communication technology, dynamic ridesharing allows automatic matching of single, non-recurring, short-notice, and on-demand requests. Conversely, when matching is performed by ridesharing operators, vehicles are assigned to pick up and drop off passengers (Furuhata *et al.*, 2013, p. 29).

Present work related to ride matching grew out of models of the dial-a-ride problem, which consists of defining a set of minimum cost routes to satisfy a set of transportation requests (Bellman, 1962, p. 61). Hosni et al. applied an incremental cost heuristic approach to estimate the maximum total profits of the shared taxi model (Hosni, Naoum-Sawaya & Artail, 2014, p. 303).

Factors that can deter carsharing have also received some attention. Psychological barriers were well studied by Horowitz and Sheth (1977, p. 1-2), who found that the loss of privacy and extra time requirements are deterrents to most potential carpooling opportunities. Another problem found by Chan and Shaheen (2012, p. 96) is that people have a need for personal space and time, and that some people prefer to avoid uncomfortable social situations. Personal security is also a concern when sharing a ride with strangers, although this is a perceived risk. Bhardwaj *et al.* (2016, p. 108) pointed out that carpooling could be unsuccessful because of psychological barriers associated with riding with strangers and inconvenient scheduling.

Taxi ridesharing (TRS)

Many-to-many approach

Over the last ten years, researches have studied a number of different TRS approaches. Proposed approaches for TRS vary depending on routing patterns, matching constraints, matching dynamics, and matching goals. Most of the proposed approaches assume that riders can embark and disembark anytime during a trip. This possibility was introduced to solve a serious TRS assignment problem (Ma, Zheng, & Wolfson, 2013, p. 411).

From a theoretical perspective, trip sharing is traditionally seen as an instance of "dynamic pickup and delivery" (Yang, Jaillet, & Mahmassani, 2004, p. 135), in which several goods or customers must be picked up and delivered efficiently at specific locations within well-defined time windows. Such problems are typically solved using optimization methods, in which a function of system variables is optimized subject to a set of constraints. Since these optimization problems often involve binary or discrete decision variables, their computational feasibility heavily depends on the number of variables and the size of the problem, e.g., the number of customers and locations considered. Most previous taxi studies have therefore focused on small-scale routing problems, such as within airport perimeters (Marin, 2006, p. 195). Large urban taxi systems, in contrast, involve thousands of vehicles performing hundreds of thousands of trips per day, which makes them intractable by conventional methods.

More recently, data-driven methods have been developed and popularized. One notable example is a simulation model proposed by Santi *et al.* (2014, p. 13290-13294). In their study, they propose a graph-based approach where nodes represent taxi trips, and two nodes are connected if trips can be shared. Their model aims to maximize the number of sharable trips or minimize the total time needed to accommodate all trips. Their results show that cumulative trip length can be reduced by 40% or more, and taxi sharing system can also be effective in cities with a lower taxi density than New York City. The structure of the shareability network is heavily dependent on the maximum number of shared trips k per sharing and the maximum waiting time Δ that a customer can tolerate, which, however, have some limitations. According to Masayo *et al.* (2015, p. 896), their solution is only feasible for certain scenarios such as when k=2. For larger values of k or Δ, this problem becomes NP-hard and requires much longer computational time.

Ma *et al.* (2013, p. 410-415) designed a large-scale taxi ridesharing system to efficiently serve in real-time requests sent by taxi users to reduce total distance traveled. According to them, they were the first to study dynamic ridesharing for a large number of taxis. Their methodology splits a region into grid cells so that distance between any two spots can be calculated "heuristically", allowing them to make minimum computations of shortest path, but at the cost of decreased accuracy (Masayo *et al.*, 2015, p. 894). Their empirical results suggest that the proposed service could provide 40% additional taxi rides for users while saving 15% of travel distance compared with no taxi sharing (provided the ratio of the number of requests to that of taxi is 5). Moreover, with their approach the current taxi fleet could accommodate a 25% increase in demand. Masayo et al. (2015, p. 888), however, argue that the approach proposed by Ma *et al.* (2013) cannot be scaled up to very large systems.

To improve on the work by Santi *et al.* (2014, p. 290-294) and Ma *et al.* (2013, p. 410-415), Masayo *et al.* (2015, p. 888-594) proposed a data-driven simulation framework for developing citywide taxi sharing solutions under a wide range of ride-sharing scenarios. Unlike in Santi *et al.* (2014, p. 290-294), trips do not need to be known in advance, as their framework is designed to handle real-time ridesharing. Also, their simulation model can take into account customer preferences (e.g., maximum waiting time that customers can tolerate), and taxi constraints (e.g., passenger capacity and maximum number of trips that can be shared). Their algorithm is linear in the number of trips and it relies on an efficient indexing scheme, which combined with parallelization, makes their method scalable. To evaluate the effectiveness of their model, they tested it on over 360 million trips in 2011 and 2012 in New York City. In one simulation, they processed over 150 million trips (approximately one-year of data) in under 10 minutes. Interestingly, they report very high sharing rates: when only two trips can be shared and the maximum waiting time is 3 minutes, 94% of trips can be shared; when up to three trips can be shared and the waiting time is up to 5 minutes, almost all trips can be shared. The scalability of their model makes it possible to perform large-scale studies that explore a wide range of what-if scenarios. One limitation of their approach is the difficulty to scale the storage of the full distance matrix for large road networks.

One-to-one Approach

Despite of its many potential benefits, the 'many-to-many' approach has a number of shortcomings. The main problem is that it may not be well accepted. Indeed, many passengers are reluctant to accept picking up or dropping off other passenger during a trip as they may be concerned about their safety or their comfort. Moreover, the many-to-many approach brings about operational difficulties and requires complex algorithms. To simplify the sharing problem, Barann, Beverungen and Muller (2017, p. 87-88) proposed a 'one-to-one' approach based on static taxi routing. Their approach focuses on merging rides with close origins and close distinations, which makes it easier for people to decide if they want to join with other passengers before embarking in a taxi. Since there is distance constraint between rides' origins and destinations, the start and end points are unchangeable if a matching is established. In contrast with the high sharing possibility of dynamic 'many-to-many' approach, this static approach has a lower sharing percentage: after analyzing 5 million taxi trips data in New York City, their results indicate that 48% of them can be matched for shared rides.

Advantages and disadvantages of taxi ridesharing

Taxi ridesharing has a number of potential benefits. They include reducing travel costs for customers by splitting fares (Santi et al., 2014, p. 13290) and decreasing VMT, which in turn can reduce congestion, gasoline consumption, and cut the emissions of local air pollutants and of greenhouse gases (Ma, Zheng, & Wolfson, 2013, p. 410). Taxi sharing can be particularly valuable in high demand situations by reducing the overall waiting time (Santos & Xavier, 2015, p. 6729). For taxi operators, operating costs (e.g., fuel consumption and car depreciation) are reduced (Hosni, Naoum-Sawaya, & Artail, 2014, p. 304), which should improve their profitability as they face increase competition from transportation network companies (such as Uber and Lyft).

However, one-to-one taxi sharing also has a few disadvantages. Overall, it slightly increases service and waiting time (Santi *et al.*, 2014, p. 13290-13291; D'Orey, Fernandes and Ferreira, 2012, p. 140-141). Moreover, passengers may have privacy and safety concerns about sharing a vehicle with people they do not know. Furthermore, reducing the cost of taxi services may induce a rebound effect, leading to an increase in demand accompanied by negative environmental effects and more congestion (Santi *et al.*, 2014, p. 13290).

Taxi and Limousine Commission (TLC) Trip record dataset

The New York City dataset has already received some attention. Ferreira, Poco, Vo, Freire, and Silva (2013, p. 2149) proposed a model to visualize taxi trips and provide information about origins and destinations to study mobility inside the city. Their model supports a wide range of spatio-temporal queries, and it is scalable system to support interactive responses.

Wallsten (2015) examined how ridesharing companies such as Uber and Lyft affect demand for traditional taxi services, using taxi data from the TLC Trip record dataset with private government datasets. He argued that more options for transportation users mean better service quality for them, partly based on a reduction in consumer complaints about NYC taxis following the rise of Uber in New York City.

Gonzales, Yang, Morgul, and Ozbay (2014) analyzed TLC Trip record data to identify the factors that drive demand for taxi services, taking into account variations by location and time of day. Using OLS, they developed demand models for taxi trip generation and mode choice that accounts for key features of transit in the communities where trips are generated. In their trip generation model, the following six variables have the most explanatory power: transit accessibility, population size, median age, percent of population educated beyond bachelor's degree, median income per capita, and number of job opportunities. Their study also illustrates how big data from taxis and transit systems can be combined with demographic and socioeconomic information to develop demand models.

Qian and Ukkusuri (2015, p. 31-35) drew data from TLC trip record dataset to implement a geographically weighted regression (GWR) model to explain the spatial heterogeneity of taxi rides. Their results show that the GWR model performs better than a global regression model partly because taxi ridership is highly sensitive to local variations in urban land use.

Zhan *et al.* (2013, p. 37) developed a model to estimate the link travel times of taxi trips. Their goal was to efficiently estimate hourly average link travel times, infer possible paths for each trip and then estimates the link travel times by minimizing the differences between expected and actual path travel times.

Other papers of interest based on the TLC trip record dataset include Donovan and Work (2015), who analyzed the resilience of taxi systems to Hurricane Sandy, and Yazici, Kamga, and Singhal (2013, p. 37-38), who developed an improved algorithm to pick-up travelers at John F. Kennedy (JFK) Airport.

Taxi demand modeling

Several recent studies have focused on understanding the demand for taxis in order to better manage them and design better regulations. For example, Toner (2010) and John and David 2014) investigated the elasticity of taxi fares and service attributes. More specifically, John and David (2014) developed segment-specific mode choice models to obtain direct elasticities of interest for costs and service level features. Schaller (2005) conducted multiple regression models to estimate taxi demand for 118 cities in the United States. An interesting finding is that the number of households with no vehicles and the number of people commuting by subway are significantly correlated with the number of taxi cabs.

Gonzales *et al.* (2014) modeled taxi pick-ups and drop-offs in New York City using multivariate regression. The authors then estimated the number of hourly pick-ups and drop-offs within each census tract. Their results indicate that population size, age, education, income, and total jobs matter for estimating taxi demand.

Since the number of taxi trips made is a non-negative continuous variable, some researches have used Tobit models to quantify the primary factors influencing taxi demand. For example, Kattan, Barros and Wirasinghe (2010, p. 16) used estimated Tobit models to explain work trip demands by taxi in Canadian cities. They reported that the number of work trips made by public transport and the number of low-income households both have positive effects on the number of taxi work trips.

Some models have also been developed to predict mode choices that include both public transportation and taxis (e.g., Gebeyehu & Takano, 2008), but these papers are not reviewed here.

CHAPTER 2: DATA

Three sets of data were collected for this study: 1) NYC taxi trips data provided by the NYC Taxi & Limousine Commission; 2) Census tract socio-economic data from the United States Census Bureau; and 3) Land use data from the NYC Department of City Planning. They are described in turn.

2.1 Taxi Trips Data

To analyze the potential of implementing a one-to-one TRS system in NYC, I extracted data from the 'TLC Trip Record Data', which was created by the NYC Taxi & Limousine Commission (NYC Taxi & Limousine Commission, n.d.). This publicly available dataset covers over 1.3 billion individual taxi trips in New York City starting from January 2009.

This dataset records both canary yellow and apple green taxi trips information. Canary yellow taxis are allowed to pick up passengers anywhere in the five NYC boroughs. Taxis painted apple green, which appeared in August 2013, can pick up passengers in the Bronx, Brooklyn, Queens (excluding LaGuardia Airport and John F. Kennedy International Airport), Staten Island, and Upper Manhattan. Both canary yellow and apple green taxis have the same fare structure. They are operated by private companies and licensed by the New York City Taxi and Limousine Commission (TLC).

Data in the TLC Trip Record dataset include the following fields:

1) Pick-up and drop-off dates/times;
2) Pick-up and drop-off longitude and latitude (GPS tracking technology is used in all New York City taxis, regulated by the TLC; Hochmair, 2016, p. 47);

3) Trip distances (in miles);

4) Trip fares (fees paid by customers);

5) Trip rate (standard, JFK, Newark, Nassau, Westchester, negotiated fare, or group ride);

6) Payment types (credit card, cash, or other types), and

7) Passenger counts reported by driver, on a per-trip basis.

To keep this study manageable, I extracted year 2014 trip data for the canary yellow, which represents 165,104,266 records. Only yellow cabs data were used for a couple of reasons. First, in 2014 green taxis were still fairly uncommon (they appeared in August 2013). Second, a key motivation for introducing apple green taxis was to better serve people with disabilities. However, it is more challenging to engage in ridesharing with a disability because the type of ridesharing considered in this study involves some walking. According to Gates *et al.* (2006, p. 38), physically disabled people walk the slowest among all age groups. Indeed, the average walking speed for disabled people using motorized wheelchairs is commonly assumed to be 3.8 ft/s, which is substantially slower than 4.54 ft/s (U.S. ROADS, 1997), the average walking speed of non-disabled people.

2.1.1 Data Cleaning

To detect and remove inaccurate or incomplete records, I thoroughly cleaned up the data as follows.

1) Like Donovan and Work (2015, p. 6), and Barann, Beverungen and Muller (2017, p. 88), I removed from my dataset trips that took place totally or in part outside of the

5 NYC boroughs, which is the area served by yellow cabs. Figure 1 shows my study area, which is outlined by the red dotted line.

2) For all trips, the reported travel distances should be longer than the straight-line distance between origin and destination.

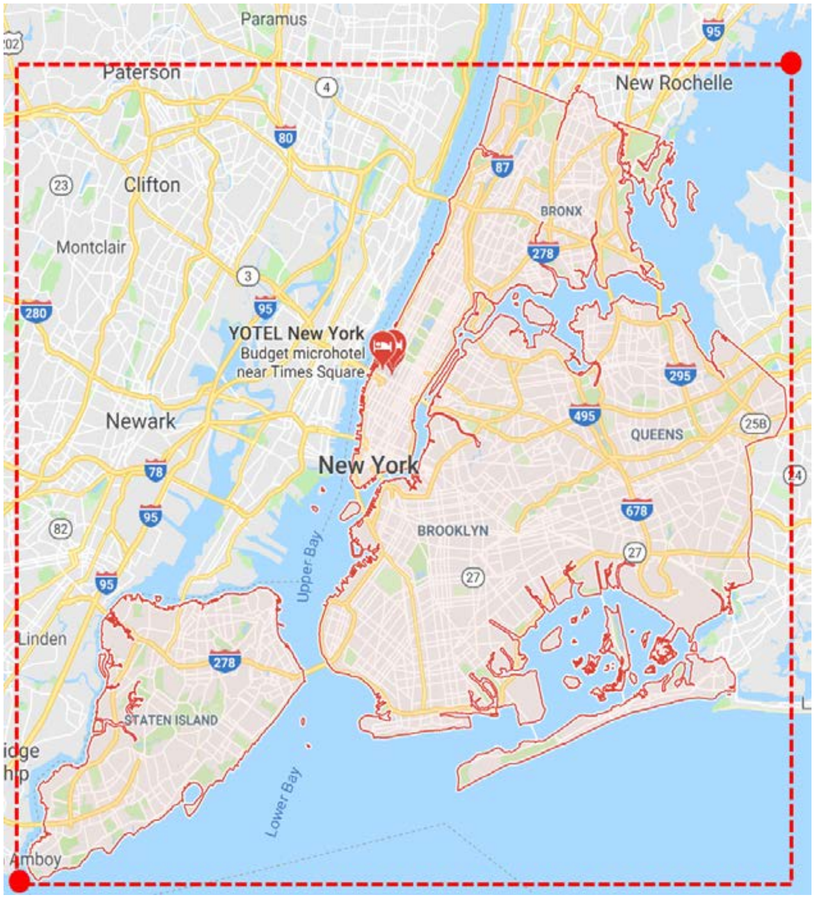

Figure 1: Study Area

3) Trips with unrealistic travel time were discarded. More specifically, trips with a duration under one minute or longer than 2 hours were dropped as the former do not justify taking a taxi and the latter likely involve the taxi to be idle or to drive outside

of NYC. A look at the distribution of trip travel times shows that ~99.8% of trips in my dataset last between 1 minute and 2 hours (see Figure 2 for a histogram of trip durations in December 2014).

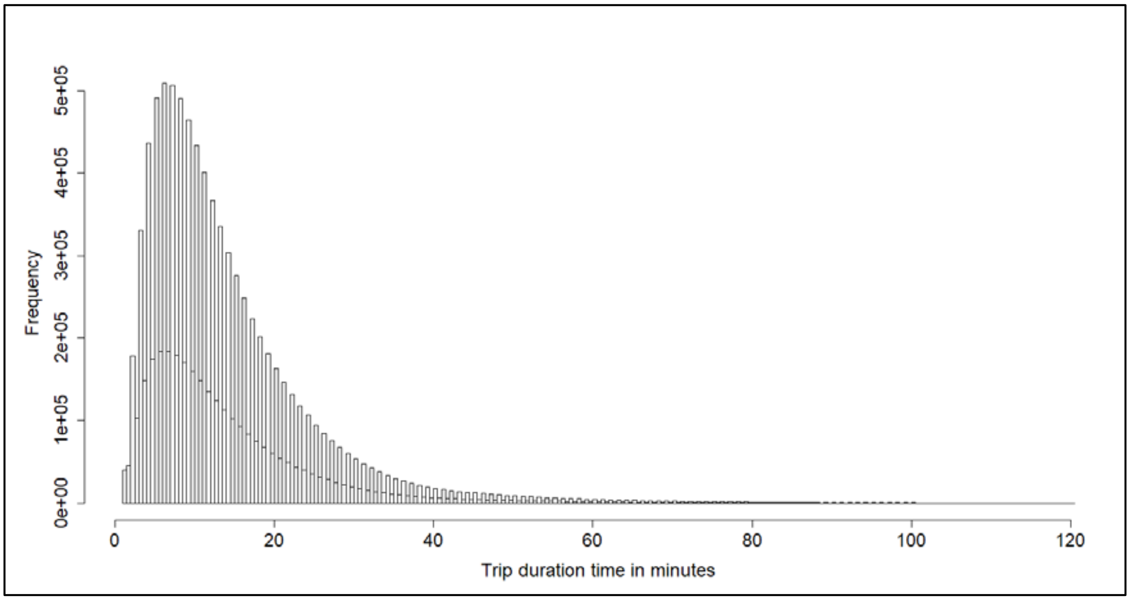

Figure 2: Histogram of trip durations for December 2014

4) Trips with unrealistic trips distances were discarded. More specifically, trips shorter than 0.124 miles (200 meters) and longer than 50 miles were dropped. Trips shorter than 0.124 miles likely correspond to data entry errors. The 50-mile cutoff distance is motivated by the 25 mph speed limit in New York City in 2014 (Meyer, 2016), and our 2 hours travel time maximum from Step 3, as the maximum distance that can legally be travelled in 2 hours at 25 mph is 50 miles. In 2014, ~99.2% of NYC yellow taxi rides were between 0.124 mile and 50 miles.

5) All trips with no passenger or with 5 or more passengers (including 5) were dropped. Indeed, the maximum number of passengers allowed in a canary yellow taxicab is 5 (NYC Taxi & Limousine Commission, n.d.), so only trips with 1 to 4 passengers have the potential to be shared.

6) Trips with a rate code of 5 (negotiated fare) or 6 (group fare on a fixed route) were dropped because it was either impossible to calculate how much would be saved by sharing a ride (rate code 5) or no savings were possible (rate code 6).

7) Trips with an average speed under 0.062 mph or over 25 mph were dropped. As mentioned above, the minimum trip distance is 0.124 mile, and the maximum trip travel time is 2 hours, so the minimum average speed is 0.124/2, or 0.062 mph. Moreover, since the NYC speed limit is 25 mph, I excluded trips with an average speed above that number.

8) Only trips with fares between $2.5 and $250 were analyzed. According to the NYC Taxi & Limousine Commission (n.d.), the initial charge is $2.5, plus 50 cents per 0.2 mile. Please note that only 19 rides in my entire dataset have a fare of $2.50; all others are at least $3.00. Here, I surmised that expensive trips were unlikely to be shared because they correspond either to unusual circuits or to a let of waiting.

Initially there were 165,104,266 records for year 2014 in dataset. After cleaning the data, 137,446,493 records were left, which corresponds to removing ~17% of the records from the original dataset. Table 1 shows the impact of data cleaning for each month. It shows that the percentage of dropped records ranges between 16% to 18% with an average of 17% of all records.

Table 1: Data cleaning results

	Raw number of taxi trips	% of rides removed	Number of taxi trips after data cleaning
January	13,780,652	16.9%	11,451,023
February	13,062,610	16.4%	10,915,993
March	15,427,495	17.0%	12,799,114
April	14,618,256	17.1%	12,121,212
May	14,773,634	17.3%	12,217,676
June	13,811,776	17.5%	11,390,968
July	13,105,903	16.8%	10,897,796
August	12,688,442	17.9%	10,421,550
September	13,373,468	16.3%	11,192,873
October	14,231,705	15.9%	11,974,728
November	13,217,276	15.9%	11,120,043
December	13,013,049	15.9%	10,943,517
Year 2014	165,104,266	16.8%	137,446,493

2.1.2 Summary statistics after data cleaning

Tables 2 to 5 show summary statistics after cleaning the data based on trip distance, duration, average speed, and fare. Some key features are visualized in Figures 3 to Figure 6.

As can be observed, there is no substantial variations from one month to the next. A higher median value is often associated with a higher standard deviation. For example, average trip distance and trip duration are highest in June, and so is the June standard deviation for these variables. Furthermore, as expected, trip distance, trip duration, and trip fare are highly positively correlated since longer trip distances mean longer trip times and higher fares.

From Table 2 and Figure 3, the maximum monthly trip distance varies the most compared to other summary statistics. The median value is very stable ranging from 1.60 to 1.76, which shows that taxi trips in NYC are typically short.

Table 2: Summary statistics for average trip distance by month in 2014 (miles)

Month	Min	1st Qu.	Median	Mean	3rd Qu.	Max.	Std.Dev.
January	0.13	1.00	1.60	2.33	2.78	45.50	2.35
February	0.13	1.00	1.65	2.39	2.80	48.65	2.42
March	0.13	1.00	1.70	2.43	2.90	45.70	2.47
April	0.13	1.00	1.70	2.48	2.90	44.87	2.56
May	0.13	1.07	1.74	2.60	3.00	46.34	2.74
June	0.13	1.07	1.74	2.61	3.00	45.61	2.76
July	0.13	1.06	1.70	2.53	2.96	44.89	2.62
August	0.13	1.09	1.76	2.61	3.00	46.64	2.75
September	0.13	1.05	1.70	2.60	3.00	45.50	2.80
October	0.13	1.03	1.70	2.57	2.94	43.80	2.74
November	0.13	1.00	1.70	2.52	2.90	49.65	2.67
December	0.13	1.00	1.66	2.53	2.90	47.55	2.75

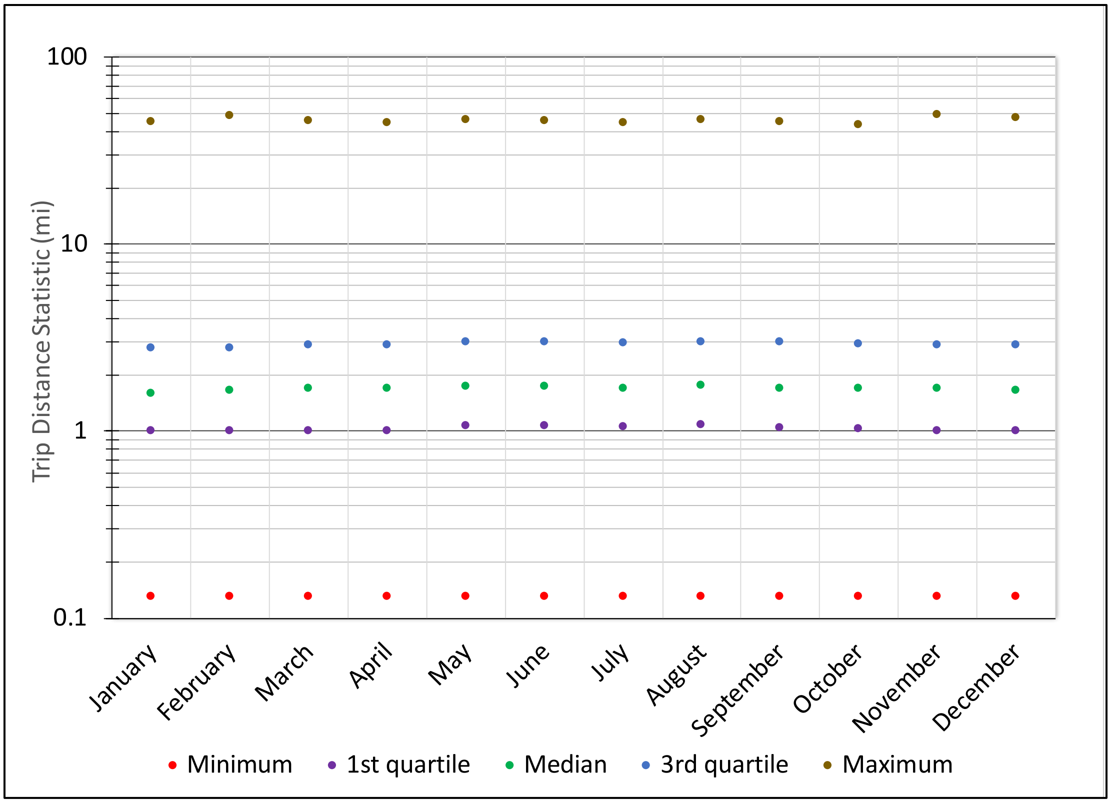

Figure 3: Taxi average trip distance statistics by month for 2014 (mi)

From Table 3 and Figure 4, we first note that the minimum and maximum trip time are the same for each month (1 minute and 2 hours), as a result of the data cleaning process. The median trip time for February, March, April, July and August is 10 minutes, but for the other months it is 11 minutes.

Table 3: Summary statistics for average trip duration by month in 2014 (min)

Month	Min	1st Qu.	Median	Mean	3rd Qu.	Max	SD
January	1.0	6.0	9.8	11.9	15.0	120.0	8.8
February	1.0	6.1	10.0	12.3	15.9	120.0	9.1
March	1.0	6.0	10.0	12.1	15.3	120.0	8.8
April	1.0	6.3	10.2	12.7	16.0	120.0	9.4
May	1.0	6.7	11.0	13.4	17.0	120.0	10.3
June	1.0	6.7	11.0	13.4	17	120.0	10.1
July	1.0	6.3	10.1	12.7	16.0	120.0	4.4
August	1.0	6.4	10.2	12.8	16	120.0	9.6
September	1.0	6.9	11.0	13.7	17.2	120.0	10.6
October	1.0	7.0	11.0	13.6	17.1	120.0	10.2
November	1.0	6.9	11.0	13.6	17.3	120.0	10.4
December	1.0	7.8	11.0	13.8	17.7	120.0	10.7

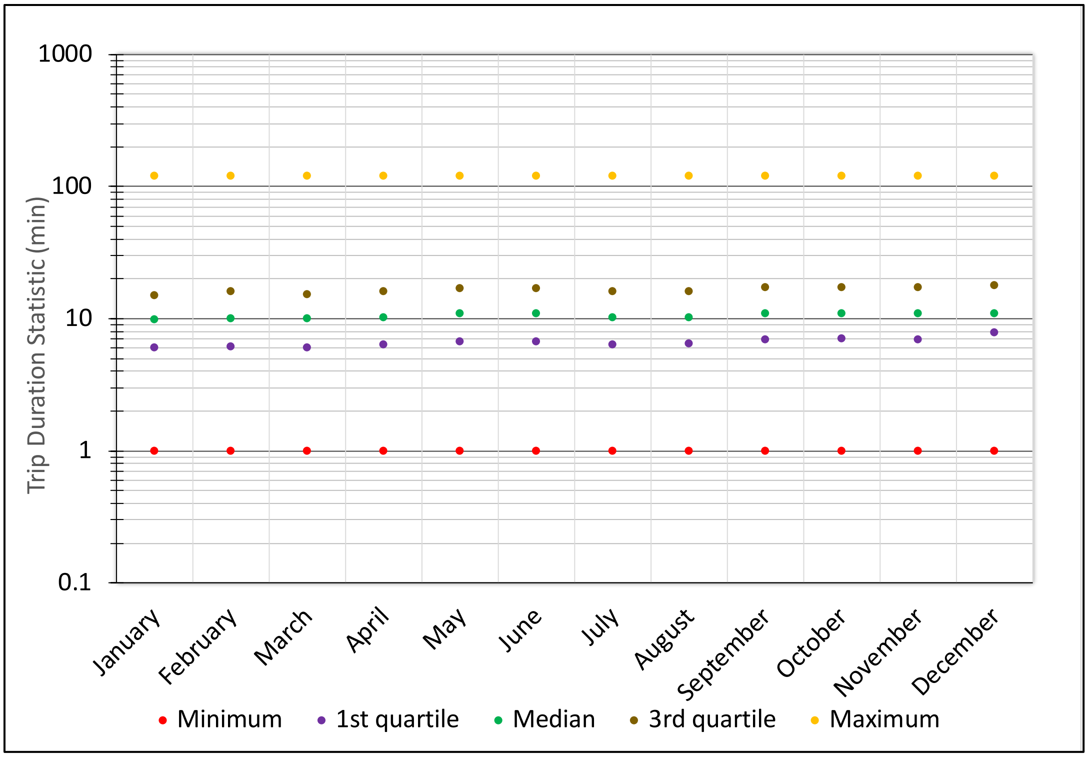

Figure 4: Taxi average trip duration statistics by month for 2014 (min)

As shown on Table 4 and Figure 5, the maximum average trip speed is 25 mph (the speed limit), and average speed (11 mph) reflects congestion in a busy urban area.

Table 4: Summary statistics of taxis average speed by month for 2014(mph)

Month	Min	1st Qu.	Median	Mean	3rd Qu.	Max	SD
January	0.12	8.29	11.00	11.65	14.40	25.00	4.64
February	0.12	8.10	10.75	11.42	14.10	25.00	4.56
March	0.12	8.44	11.17	11.80	14.55	25.00	4.61
April	0.10	8.04	10.82	11.51	14.31	25.00	4.70
May	0.12	7.95	10.75	11.43	14.25	25.00	4.71
June	0.12	7.94	10.83	11.49	14.40	25.00	4.78
July	0.11	8.27	11.20	11.80	14.74	25.00	4.78
August	0.11	8.45	11.25	11.90	14.76	25.00	4.72
September	0.10	7.75	10.57	11.24	14.06	25.00	4.73
October	0.10	7.66	10.37	11.11	13.83	25.00	4.67
November	0.13	7.54	10.28	10.96	13.68	25.00	4.66
December	0.11	7.41	10.20	10.89	13.68	25.00	4.73

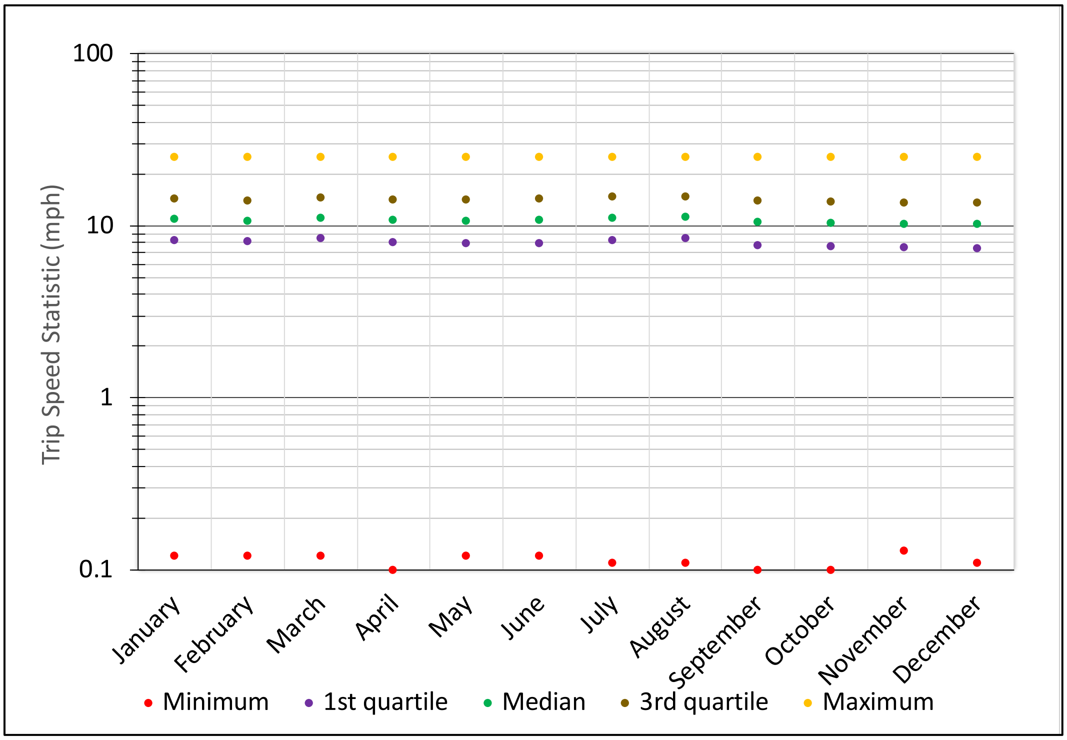

Figure 5: Taxi average ride speed statistics by month for 2014(mph)

From Table 5 and Figure 6, we can observe that the median taxi fare ranges from $10.5 to $11.4. Although the maximum monthly fare exceeds $200, it is below $250.

Table 5: Summary statistics of trip taxi fares by month for 2014(US dollars)

Month	Min	1st Qu.	Median	Mean	3rd Qu.	Max	SD
January	3.00	7.80	10.50	12.91	15.00	230.08	8.64
February	3.00	8.00	10.80	13.24	15.50	220.83	8.90
March	3.00	8.00	10.75	13.23	15.50	222.50	8.96
April	3.00	8.00	11.00	13.59	15.62	239.66	9.29
May	3.00	8.00	11.40	14.20	16.30	227.50	10.00
June	3.00	8.30	11.40	14.20	16.50	222.00	10.10
July	3.00	8.00	11.00	13.70	16.00	232.00	9.50
August	3.00	8.00	11.00	13.90	16.00	202.50	9.80
September	3.00	8.30	11.40	14.30	16.50	238.50	10.20
October	3.00	8.30	11.40	14.20	16.50	215.50	10.00
November	3.00	8.00	11.30	14.10	16.30	226.80	9.90
December	3.00	8.00	11.30	14.17	16.50	242.00	10.10

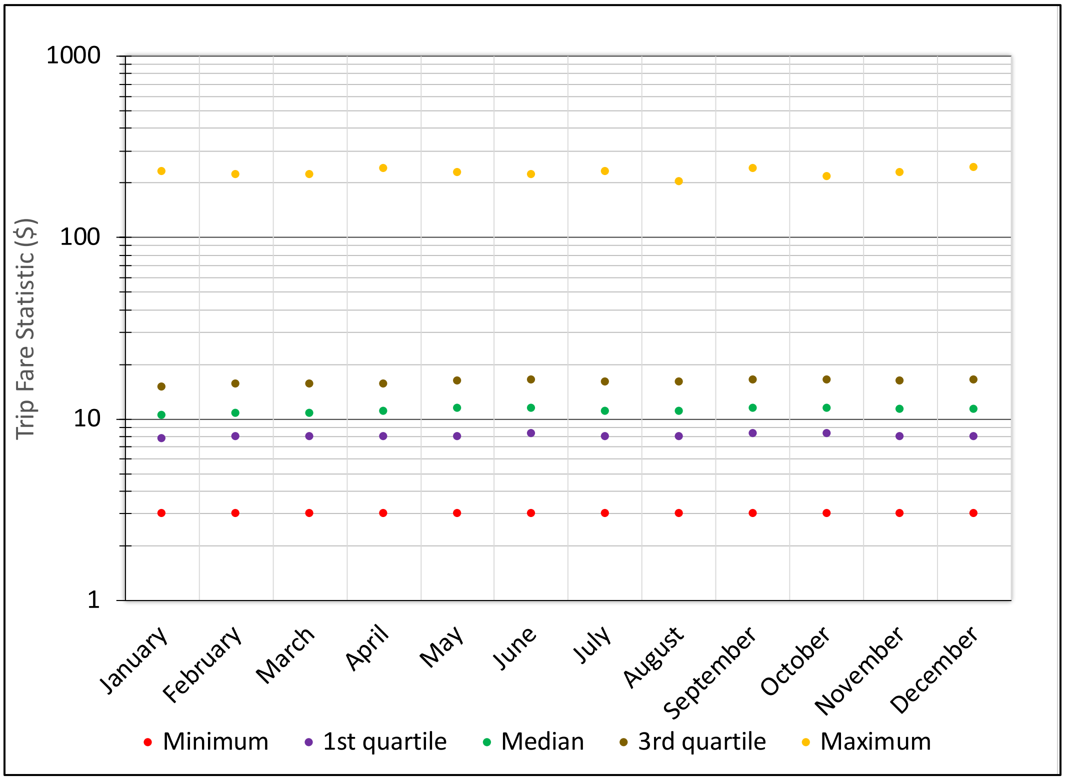

Figure 6: Taxi trip fare statistics by month for 2014($)

2.1.3 Temporal distribution

Figure 7 shows the frequency of trips during each hour during a typical day in December 2014. Trip numbers are decreasing after midnight until 6 AM, after which they pick up. The smallest number of trips is between 4 AM and 6 AM. From 6 AM to 3 PM, the number of trips increases gradually, especially during the morning peak hour (6 AM to 9 AM). It then decreases around 4 PM to 5 PM and increases again after 5 PM. The demand for trips reaches a peak at 7 PM and is highest between 7 PM to 11 PM, which reflects that New York City is "The City that Never Sleeps".

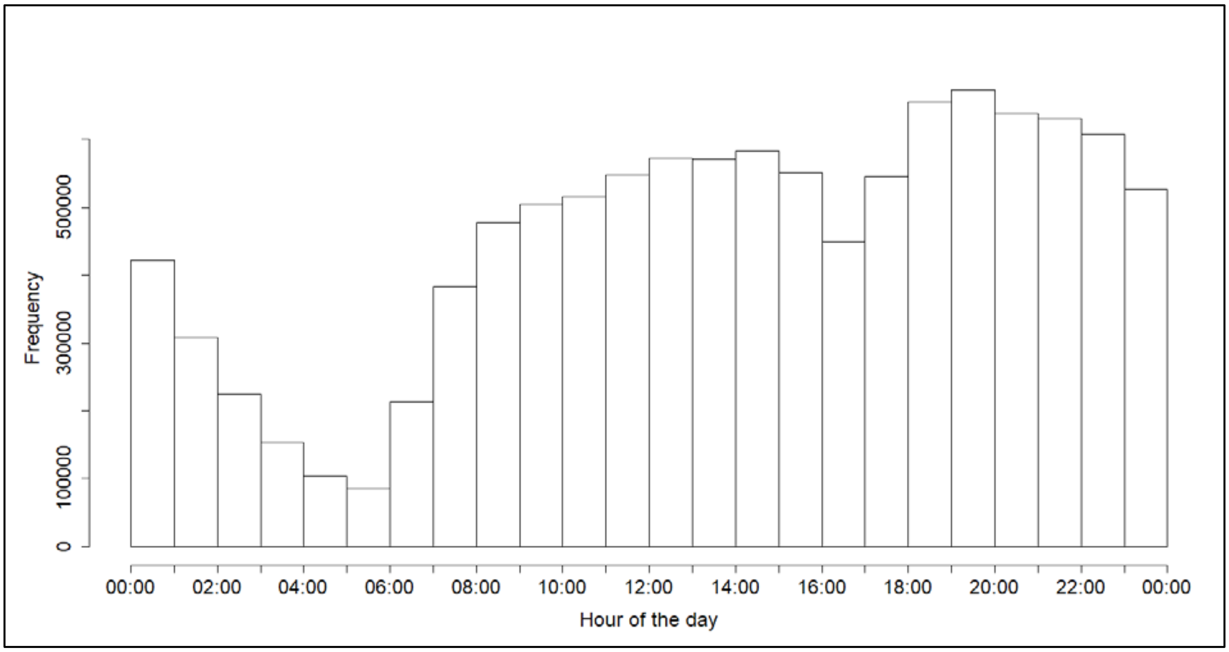

Figure 7: Histogram of number of trips by hour of the day, in December 2014

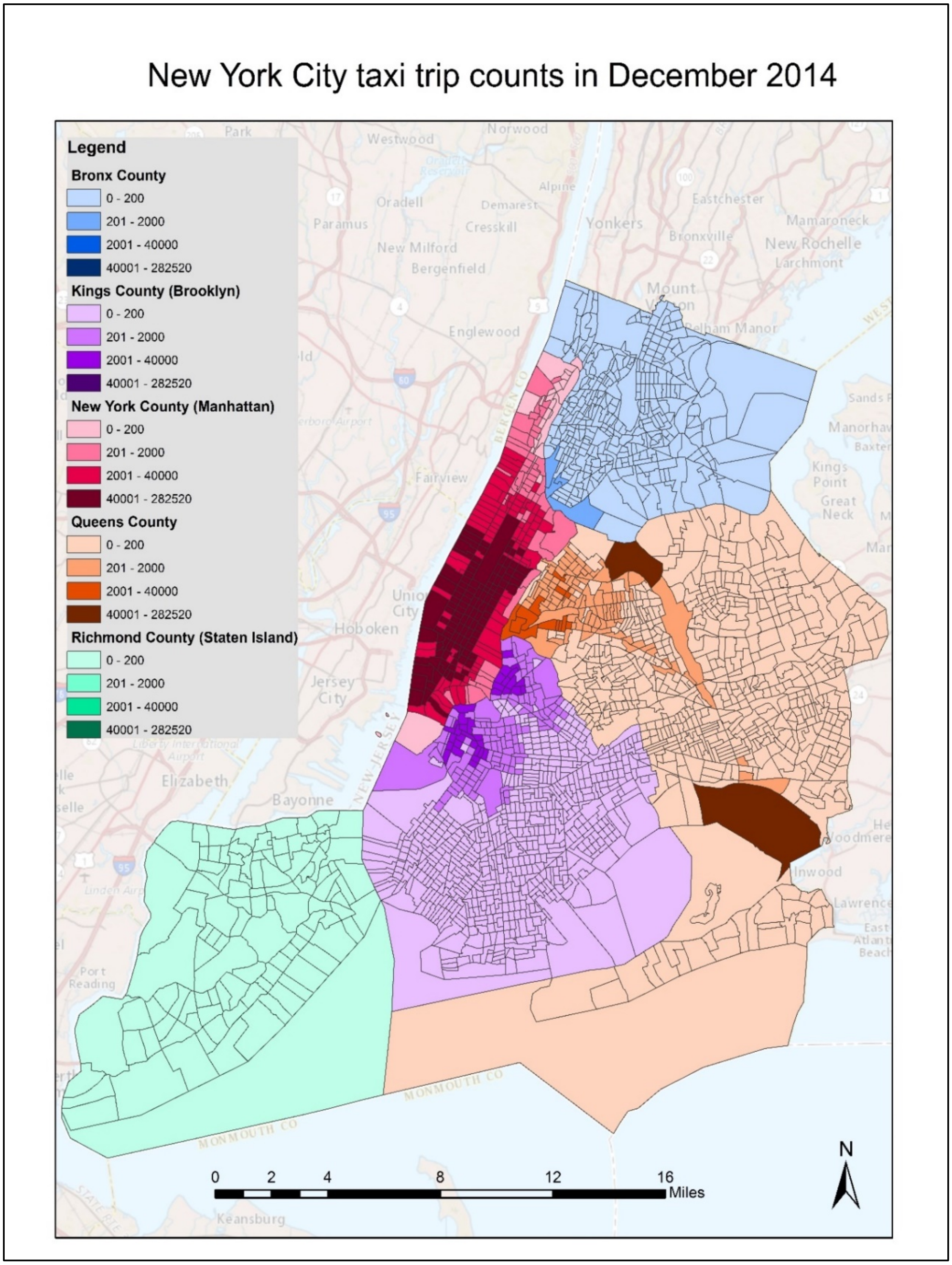

Figure 8: Taxi pick-up counts in NYC in December 2014

2.1.4. Spatial distribution

Using GIS to allocate trip origins to census tracts, Figure 8 shows taxi pick-up counts in December 2014. A cursory exploration suggests that other months have similar spatial distributions. A Manhattan-centric pattern can be observed here, with most trips originating in the south of Central Park in Manhattan. Outside of Manhattan, LaGuardia and John F. Kennedy, the two airports in Queens, also show a high demand for taxi trips. Conversely, there is very little activity in Queens, the Bronx, Staten Island, and areas of Brooklyn that are farther from Manhattan.

Note, however, that Figure 8 reflects only the activity of canary yellow taxis. As mentioned above, some areas in outer boroughs are served by apple green taxis.

2.2 Social-economic data by census tracts

2.2.1 Data cleaning

Since yellow cabs mostly serve Manhattan, I will focus on Manhattan for my statistical model that explains the potential for sharing taxis. My dependent variable will be the number of shareable rides per acre. These data result from my matching algorithm (see below). My explanatory variables will be socio-economic and land use variables for 2014. All social-economic data were retrieved from United States Census Bureau.

In 2014, Manhattan had 288 census tracts. To estimate the model, I removed 10 census tracts because their population was under 100, which made it impossible to get a good estimate for socio-economic variables such as income or education. Although census tracts with a small population typically offer very few opportunities for sharing taxi rides, there are exceptions such as the census tract that contains Central Park. Its population is 4, but it is the source of approximately 17000 taxi trips annually.

2.2.2 Explanatory Variables

Table 6 describes my explanatory variables. There are seven categories in total: population, age, ethnicity, income, education, employment and car ownership. Population density is defined as total population per square mile (in 1000s). According to Kumar and Lim (2006, p. 570) and Beutell and Wittig-Berman (2008, p. 509), we can currently distinguish between five generations: generation Z (born after 1994), generation Y (born between1980 and 1994), generation X (born between 1963 and 1981), baby boomers (born between 1946 and 1964) and matures (born before 1946). For ethnicity, I distinguished between Whites (the baseline), Blacks or African Americans, Asians, and others. I also include a variable for Hispanic or Latino status. Income is mean annual income per capita (in $1,000s). For education, I consider 4 categories: less than high school, high school, some college or associate degree, and bachelor's degree or higher. For employment variable, I consider employment density (thousands of jobs per square mile, as I expect that it will impact the demand for taxis) and the unemployment rate, obtained by dividing the number of unemployed by the labor force over 16 years old or above.

A number of explanatory variables are percentages. For example, for age and ethnicity, I calculated the percentage of each generation and each race of total population. For education, first I summed up the four categories of people with known education, and then calculated the percentage of each sub-category. I multiplied all percentages by 100 for model coefficients to be of a similar order of magnitude. It helped for discussing results since one unit in this case is 1%.

Table 6: Explanatory variables

Explanatory Variable	Description
Population	Population density: Total population per square mile (in 1000s)
Age	Generation Z: Population with age less than 19 years till 2014 Generation Y: Population with age 20 to 34 years old till 2014 Generation X: Population with age 35 to 49 years old till 2014 Baby boomers: Population with age 50 to 69 years old till 2014 Matures: Population with age 70 till 89 years old till 2014
Ethnicities	White, Black or African American, Asian, and other races; Hispanic or Latino
Income	Mean annual income per capita
Education	Less than high school High school Some college or associate degree Bachelor's degree or higher
Employment	Unemployment rate: number of unemployed persons / labor force over 16 years old or above Employment density: job counts per square mile (in 1000s)
Car ownership	Percentage of household with no vehicle available

2.3 Land Use Data

A GIS shapefile containing land use information by census tract was retrieved from the NYC Department of City Planning website (http://www1.nyc.gov/site/planning/data-maps/open-data/dwn-pluto-mappluto.page). I organized the data by calculating the percentage of land in each census tract devoted to each of 12 types of land use.

Finally, Table 7 shows summary statistics for my independent variables after removing census tracts with a low or no population (census data are incomplete for census tracts).

Table 7: Summary statistics of explanatory variables (N=278)

Explanatory variables	Min	1st Qu.	Median	Mean	3rd Qu.	Max	SD
Population							
Population density (in 1000s)	2.173	45.640	82.233	99.602	126.750	474.120	77.399
Age (% of)							
Generation Z	0.00	11.00	16.00	16.48	21.00	44.00	7.16
Generation Y	5.00	24.00	29.00	30.80	36.00	64.01	11.19
Generation X	5.00	19.00	21.00	21.48	24.00	47.00	5.21
Baby boomers	5.00	18.00	22.00	22.00	25.00	52.01	6.28
Matures	0.00	5.00	8.00	9.45	12.00	30.00	5.53
Ethnicity (% of)							
White	5.03	29.37	66.26	57.22	80.19	99.57	26.68
Black or African American	0.00	2.40	6.33	15.60	18.53	85.03	20.32
Asian	0.00	4.18	9.32	12.25	15.38	83.72	12.70
Other races	0.00	4.82	8.31	14.92	19.75	68.95	15.43
Hispanic or Latino	1.00	7.00	13.00	22.82	33.00	93.00	22.31
Income							
Mean annual income per capita (in $1000s)	10.31	26.01	62.57	68.32	101.76	247.85	46.18
Education (% of the CT population with)							
Less than a high school education	0.00	2.00	7.00	12.66	22.00	53.00	12.78
A high school education	1.00	6.00	10.00	13.63	20.00	37.00	8.51
Some college or associate degree	2.00	12.00	16.00	17.31	21.75	64.00	7.65
A bachelor's degree or higher	6.00	33.00	64.50	56.39	78.75	90.00	24.43
Employment							
Unemployment rate	0.00	5.00	7.00	8.39	11.00	68.00	6.43
Employment density (in 1000s)	0.36	10.53	32.51	152.43	142.64	3079.05	331.55
Car Ownership							
Percentage of household with no vehicle	42.00	73.00	79.00	77.59	84.00	96.00	8.57
Land Use (% of)							
One & Two Family Buildings	0.00	0.00	0.00	2.32	3.00	22.00	4.10
Multi-Family Walk-Up Buildings	0.00	1.00	7.00	11.03	17.75	48.00	11.38
Multi-Family Elevator Buildings	0.00	8.00	18.00	23.08	33.75	96.00	19.43
Mixed Residential & Commercial Buildings	0.00	12.00	19.00	21.92	31.00	92.00	14.89
Commercial & Office Buildings	0.00	1.00	4.00	13.45	15.00	87.00	19.98
Industrial & Manufacturing	0.00	0.00	0.00	1.22	1.00	18.00	2.62
Transportation & Utility	0.00	0.00	0.00	3.67	2.00	59.00	9.15
Public Facilities & Institutions	0.00	5.00	9.00	12.98	16.00	88.00	14.05
Open Space & Outdoor Recreation	0.00	0.00	1.00	6.01	5.00	81.00	13.70
Parking Facilities	0.00	0.00	1.00	1.52	2.00	21.00	2.69
Vacant Land	0.00	0.00	1.00	2.28	2.75	44.00	4.80
NA	0.00	0.00	0.00	0.48	0.00	68.00	4.28

CHAPTER 3 METHODOLOGY

3.1 Algorithm to match taxi rides

3.1.1 Matching constraints

The goal of this algorithm is to find all the trips that have the potential to be shared. There are two roles in my taxi ridesharing algorithm: main rider and participant. Main rider is the ride that initiates ridesharing, and participant is the ride that joins the ridesharing. Within a certain time window, to fulfill ridesharing, passengers on the participant ride needs to walk to the origin of the main rider from its origin. They also need to walk to their final destination after arriving at the main rider's destination. For simplicity, I set the maximum trip number to be shared is 3, so there can be two cases for taxi sharing: one main rider and one participant (Figure 10), and one main rider and two participants (Figure 11). As can be seen in Figure 9, O_p is the Origin of participant *p*, O_m is the origin of main rider *m*, D_p is the destination of participant *p*, D_m is the destination of main rider *m*. To initiate the shared ride, *p* needs to walk to *m*, and also needs to walk to its destination D_p after being dropped off at D_m. Figure 10 and 11 show the whole process of taxi ridesharing.

To find main riders and participants, I analyzed all trips in my dataset by pick-up time order, which obeys 'first-come-first-serve' rule. This method mimics a real-world system, in which trips are matched according to pick-up time. For each incoming main rider, the algorithm searched for potential candidates as participants within a certain time window. Since there are some constraints to match participant to main rider, if a candidate was found that satisfied all the constraints, then I will say that the rides can be shared, or rides can be merged, for both main rider and participants.

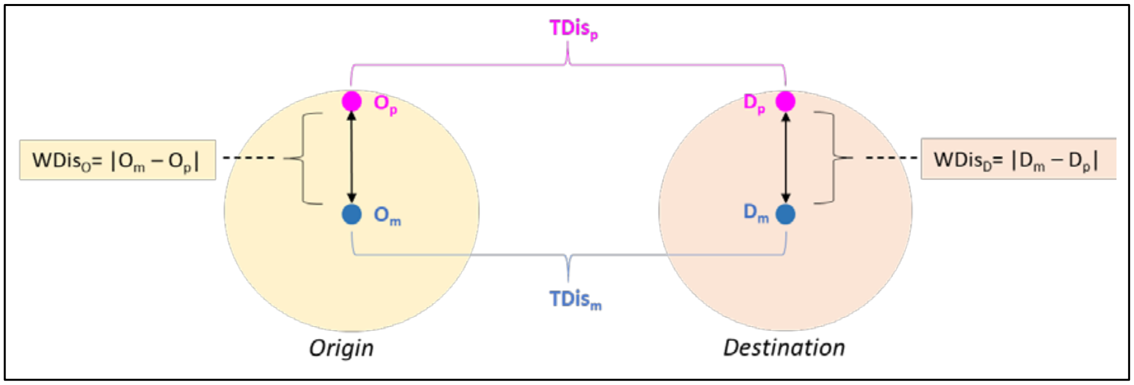

Figure 9: Walking between origins and destinations for people sharing a taxi

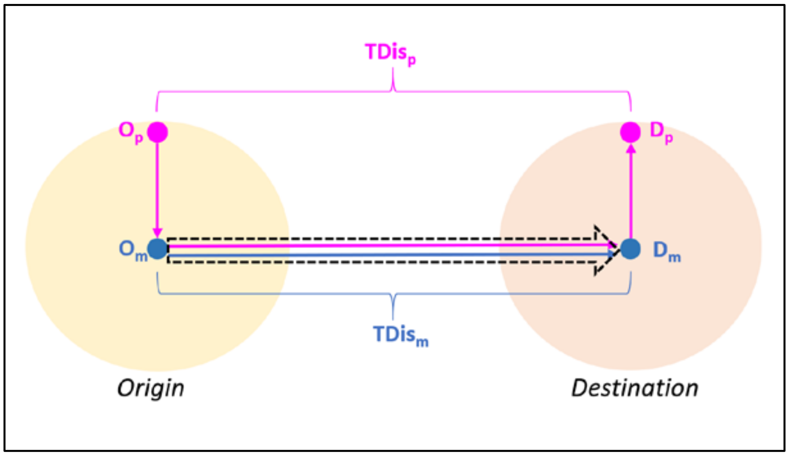

Figure 10: Case a: One Main Rider *m* and one Participant *p*

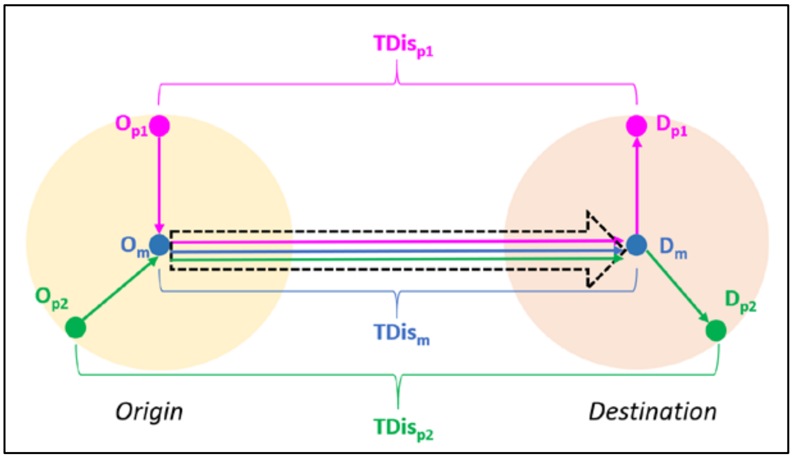

Figure 11: one main rider (*m)* and two Participants (*p1 & p2)*

Five constraints need to hold to match a participant with a main rider: number of people in a trip, time, capacity, distance, and fares (see Table 8 for a summary).

(1) Number of trip participants: in most cases, there is only 1 or 2 passengers per ride and the capacity of the NYC taxi is 5, so I set 3 to be the maximum number of trips that could be merged (designated by NT).

(2) Time constraint: this time window specifies the time gap between the pick-up time of two rides. Only trips within this time window may be shared. Here, I set this time window to 5 minutes. It also provides an upper bound on how far each participant can walk to catch the taxi and to his/her real destination. Since this time window applies to both origin and destination, the maximum total delay is 10 minutes.

(3) Capacity constraint: the taxi capacity in NYC is 5, which is the number that I used to restrict the number of participating passengers.

(4) Distance constraint: both the distance between main rider's origin and each participant's origin, *and* the distance between the main rider's destination and each participant's destination should be less than 0.3 miles. Moreover, the cumulated walking distance of each participant should not exceed his/her original trip distance.

(5) Fares constraint: for the shared ride, the fares will be evenly split between main rider and the participant. To simplify the pricing mechanism, main rider's original cost divided by the number of merging trips should be less than any single ride's original fares. Reducing costs of money is one of the benefits and reasons to encourage people join ridesharing, if the fares of taxi sharing do not decrease, no one is willing to participate. Furthermore, since consumers pay less due to taxi sharing, to prevent a negative influence on the operator's and drivers' incomes, a constant $2.5 surcharge

fee was added to each ride's payment in ridesharing (Barann, Beverungen, & Muller, 2017, p. 88).

Table 8: Constraints description

Constraints		Description	Notation and Criteria
(1)	Number of trip participants	Maximum number of trips to be merged	$NT \leq 3$
(2)	Time	Maximum departure time gap between two trips	$Time_j - Time_i \leq 5$ mins
(3)	Capacity	Maximum number of passengers in one single trip	$NP \leq 5$
(4)	Distance	Maximum walking distances between participant and main rider's origins and destinations	$WDis_O \leq 500$ meters; $WDis_D \leq 500$ meters
		Participant's aggregate walking distances of origin and destination should not exceed participant's original trip distance	$WDis_O + WDis_D < TDis_p$
(5)	Fares	The fare for each participant in a shared ride should be less than original fares	$Fares_m/NT + \$2.5 \leq Fares_m$ $Fares_m/NT + \$2.5 \leq Fares_p$

There are some assumptions in this algorithm:

(1) All taxi customers are willing to join ridesharing as long as the constraints are satisfied. Personal constraints were not considered, such as gender or smoking.

(2) Every participant will arrive at their main rider's pickup place on time, and they all will accept that they will need to walk up to 0.3 mile at their origin and at their destination based on a straight-line distance measurement.

3.1.2 Algorithm flowchart

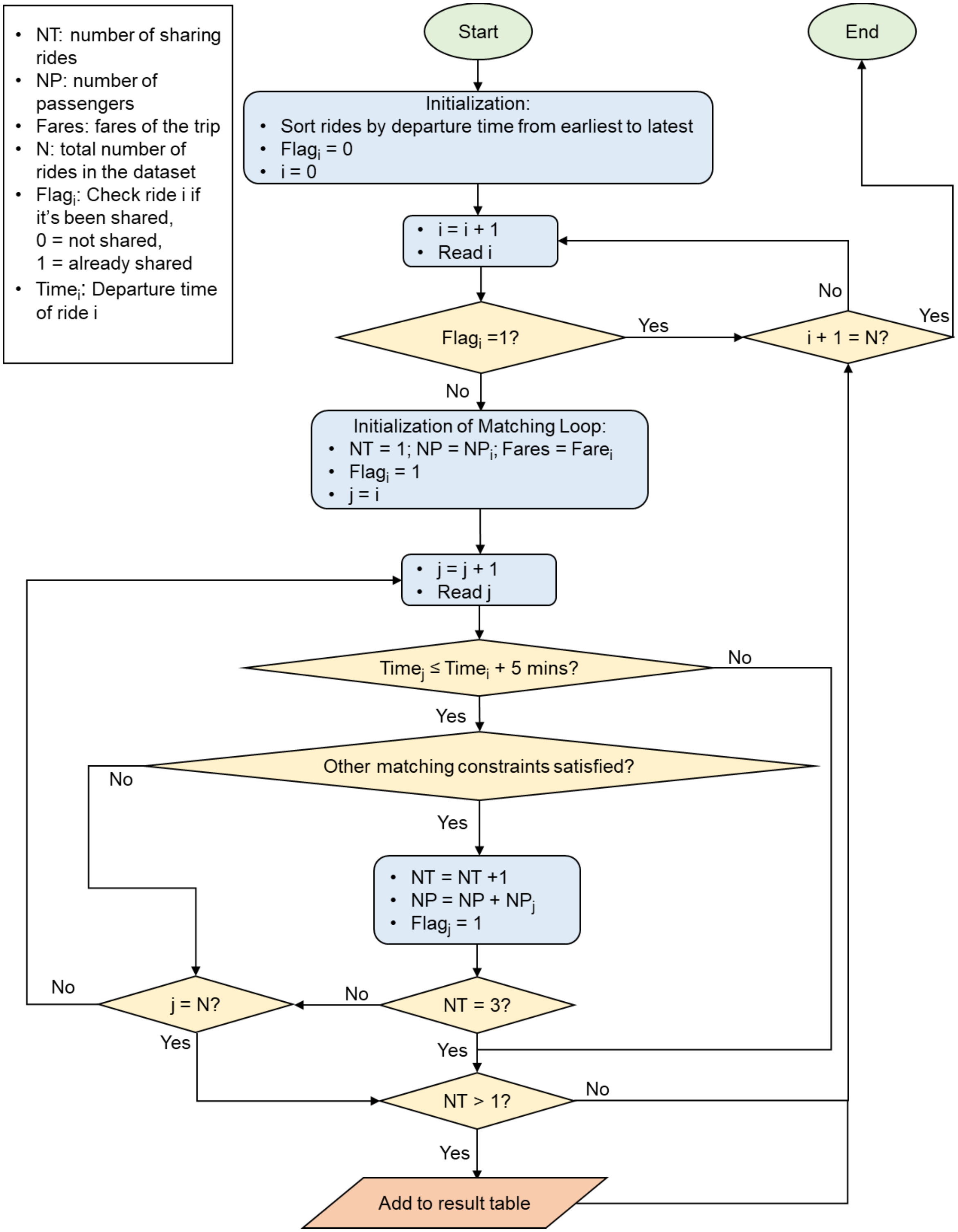
• NT: number of sharing rides
• NP: number of passengers
• Fares: fares of the trip
• N: total number of rides in the dataset
• Flag$_i$: Check ride i if it's been shared, 0 = not shared, 1 = already shared
• Time$_i$: Departure time of ride i
Start
End
Initialization:
• Sort rides by departure time from earliest to latest
• Flag$_i$ = 0
• i = 0
• i = i + 1
• Read i
Flag$_i$ =1?
Yes
i + 1 = N?
No
Yes
No
Initialization of Matching Loop:
• NT = 1; NP = NP$_i$; Fares = Fare$_i$
• Flag$_i$ = 1
• j = i
• j = j + 1
• Read j
Time$_j$ ≤ Time$_i$ + 5 mins?
No
Yes
Other matching constraints satisfied?
No
Yes
• NT = NT +1
• NP = NP + NP$_j$
• Flag$_j$ = 1
No
j = N?
No
NT = 3?
Yes
Yes
NT > 1?
No
Yes
Add to result table

Figure 12: Flowchart for the taxi matching algorithm

3.2 Benefits of taxi ridesharing

Thanks to taxi ridesharing, VMT can be reduced and taxi customers could save money. For each taxi ridesharing, to calculate saved VMT, the original trip distance of participating customers was aggregated, since participants do not need to travel separately, and the distance traveled by the shared taxi was subtracted.

Likewise, to calculate potential monetary savings to customers sharing a taxi, I summed the fares the main rider and each participant would have paid if they had traveled separately, and then subtracted the cost of the shared ride. Table 9 summarizes my methodology.

Table 9: Benefits calculation from each taxi ridesharing

	Before taxi ridesharing	After taxi ridesharing	Benefits
VMT (miles)	$TDis_m + TDis_p$	$TDis_m$	$TDis_p$
Taxi fares ($)	$Fares_m + Fares_p$	$Fares_m + 2.5*NT$	$Fares_p – 2.5*NT$
Saved CO_2 (lb)			$TDis_p * 0.8907$
Saved gas (gallons)			$TDis_p * 0.0454$

Moreover, sharing taxis can also reduce CO_2 emissions and gasoline consumption. To estimate these benefits, I followed a rough approach from the Green Vehicle Guide from the United States Environmental Protection Agency (EPA). According to this source, the average passenger vehicle emits ~404 grams (or 0.8907 pounds) of CO_2 per mile (United States Environmental Protection Agency, n.d.). From EPA's greenhouse gas equivalencies calculator, we also know that one pound of emitted CO_2 corresponds to 0.051 gallons of gasoline consumed (United States Environmental Protection Agency, n.d.). Therefore, total saved CO_2 (lb.) can be calculated by multiplying 0.8907 by saved VMT, and avoided gallons of fuel can be obtained by multiplying avoided CO_2 emitted by 0.051. The total benefits in terms of saved VMT, saved expenses for customers, avoided CO_2 emissions and reduced gasoline consumption are then obtained by summing over all shared taxi rides.

3.3 Tobit model

The Tobit model can describe the linear relationship between a non-negative dependent variable and a set of explanatory variables. In this thesis, I explain the number of sharable rides per acre in each census tract, which is a non-negative continuous variable, with census tract level socio economic and land use variables. I therefore estimate a left censored Tobit model, which can be described as follows (Greene, 2012):

$$y_i = \begin{cases} \alpha + \beta X_i + \varepsilon_i , & if\ \alpha + \beta X_i + \varepsilon_i > 0 \\ 0, & if\ \alpha + \beta X_i + \varepsilon_i \leq 0 \end{cases} \tag{1}$$

where ε_i is an error term assumed to be normally distributed ($\varepsilon_i \sim N(0,\sigma^2)$) and X_i contains explanatory variables for census tract "*i*". At the outset, I hypothesized that the potential demand for sharing taxi rides could depend on population density and various characteristics of residents of a census tract (generation, ethnicity, income, education, unemployment rate, car ownership) along with job density and the land use. Table 7 presents summary statistics for my explanatory variables.

3.3.1 Data cleaning

To explain the potential demand for shareable taxi rides, I focus on Manhattan to keep this study manageable. Of the 288 Manhattan census tracts, ten had a 2014 population under 100 (or just zero). Some of the socio-economic variables (such as income or education) were not available for these census tracts so I removed them from my dataset. My final dataset has 278 census tracts.

Table 10: Results of Multicollinearity Tests

	Variable	VIF	Keep or
Population	Population density (in 1000s)	1.364	✓
	Percentage of Generation Z	3.503	✓
Age	Percentage of Generation Y	6.070	✓
	Percentage of Generation X	2.579	✓
	Percentage of Matures	4.786	✓
Ethnicity	Percentage of Black or African American	4.405	×
	Percentage of Asian	3.669	×
	Percentage of other races	17.501	×
	Percentage of Hispanic or Latino	21.233	×
Income	Mean income per capita (in $1000s)	6.097	✓
	Percentage of population with less high school education	11.320	✓
Education	Percentage of population with high school education	5.107	✓
	Percentage of population with college or associate degree	2.580	✓
Employment	Employment density (in 1000s)	2.292	✓
	Unemployment rate	1.408	✓
Car	Percentage of household with no vehicle available	1.838	✓
Land use	% of one & two family buildings	1.718	✓
	% of multi-family walk-up buildings	2.426	✓
	% of mixed residential & commercial buildings	1.701	✓
	% of commercial & office buildings	3.610	✓
	% of industrial & manufacturing	1.276	✓
	% of transportation & utility	1.377	✓
	% of public facilities & institutions	1.747	✓
	% of open space & outdoor recreation	1.580	✓
	% of parking facilities	1.348	✓
	% of vacant land	1.206	✓
	% of other land uses	1.188	✓

3.3.2 Testing for Multicollinearity

Prior to estimating my model, I perform a multicollinearity check using variance inflation factors (VIF) (Greene, 2012). Multicollinearity exists when there is substantial correlation between the explanatory variables in a model, which can inflate standard errors and bias significance tests. Generally, a VIF above 10 indicates high correlation and is a cause for concern. Table 10 displays VIF results. To avoid multicollinearity, I removed ethnicity variables from my explanatory variables.

CHAPTER 4 RESULTS

4.1 Algorithm Results

Table 11and Figure 13 show the results from the taxi ride sharing algorithm. The sharing percentages were calculated by dividing raw trip numbers by the number of shareable taxi trips. These percentages oscillate around 50%, with a low of 48.74% in August and a high of 52.56% in March. As can be observed from Figure 13, the total (raw and clean) number of rides and the number of shareable rides are roughly parallel.

Table 11: Results from sharing algorithm by month for 2014

	Original trip number	Percentage of trips removed	Number of trips after data cleaning	Shared trip number	Percentage of sharing
January	13,780,652	16.9%	11,451,023	7,163,110	51.98%
February	13,062,610	16.4%	10,915,993	6,842,893	52.39%
March	15,427,495	17.0%	12,799,114	8,109,250	52.56%
April	14,618,256	17.1%	12,121,212	7,678,895	52.53%
May	14,773,634	17.3%	12,217,676	7,570,269	51.24%
June	13,811,776	17.5%	11,390,968	7,016,977	50.80%
July	13,105,903	16.8%	10,897,796	6,543,859	49.93%
August	12,688,442	17.9%	10,421,550	6,184,734	48.74%
September	13,373,468	16.3%	11,192,873	6,937,811	51.88%
October	14,231,705	15.9%	11,974,728	7,558,243	53.11%
November	13,217,276	15.9%	11,120,043	6,922,029	52.37%
December	13,013,049	15.9%	10,943,517	6,808,257	52.32%
Year 2014	165,104,266	16.8%	137,446,493	85,336,327	51.69%

Table 12 and Figure 14 present the detailed benefits from taxi ridesharing. Substantial benefits can be obtained: in year 2014, the estimated total VMT saved is 98.5 miles, resulting in avoided gasoline consumption of 4.5 million gallons and associated emissions of 87.7 million pounds of CO_2. Moreover, taxi customers could save $401.7 million.

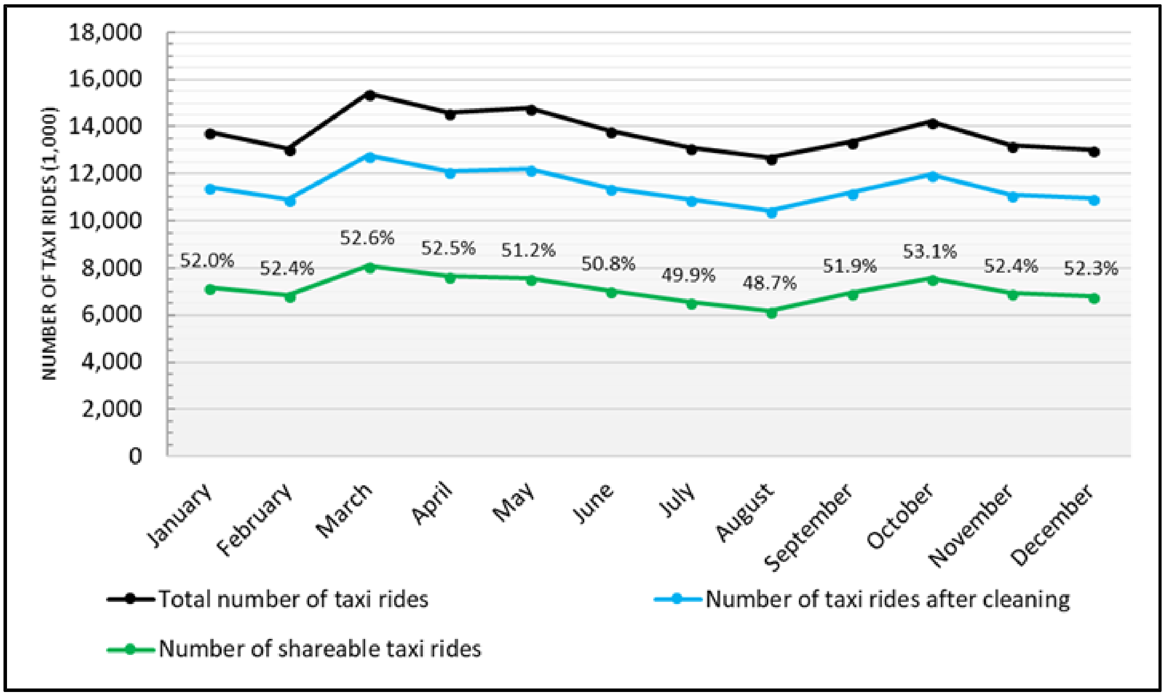

Figure 13: Summary of sharing results

Since CO_2 emission and gasoline consumption are tightly connected, an increase in saved VMT is accompanied with an increase in avoided gasoline consumption and of CO_2 emissions. According to NYC Taxi & Limousine Commission (2013, p. 15), New York City taxi fleets generate approximately 315,491 metric tons of CO_2 a year. Using different assumptions, Tseng *et al.* (2018, p. 1) estimated this number to be around 242,900 metric tons instead. Based on these two estimates, New York City taxis emit 535.5[1] million to 695.5 million pounds of CO_2 per year. As a result, if all sharing possibilities found by my algorithm were implemented CO_2 emissions from taxis would be cut by 12.6% to 16.4%.

[1] 1 metric tons = 2204.62 pounds

Table 12: Benefits from taxi ride sharing

	Total VMT saved (miles)	Total money saved for customers ($)	Money saved per ride sharing ($)	Total saved CO_2 (lb)	Total saved gas (gallons)
January	8,101,246	31,047,327	4.33	7,215,508	368,280
February	7,591,816	30,583,135	4.47	6,761,776	345,121
March	9,202,296	36,409,065	4.49	8,196,177	418,333
April	8,793,157	35,744,651	4.65	7,831,770	399,733
May	9,005,900	36,855,669	4.87	8,021,253	409,405
June	8,344,877	34,349,490	4.90	7,432,502	379,355
July	7,551,844	30,386,866	4.64	6,726,174	343,304
August	7,206,886	28,476,130	4.60	6,418,932	327,622
September	8,197,967	34,252,156	4.94	7,301,654	372,676
October	8,900,076	37,227,021	4.93	7,926,999	404,594
November	7,981,747	33,758,931	4.88	7,109,075	362,847
December	7,663,500	32,683,177	4.80	6,825,623	348,380
Year 2014	98,541,312	401,773,614	4.71	87,767,443	4,479,650

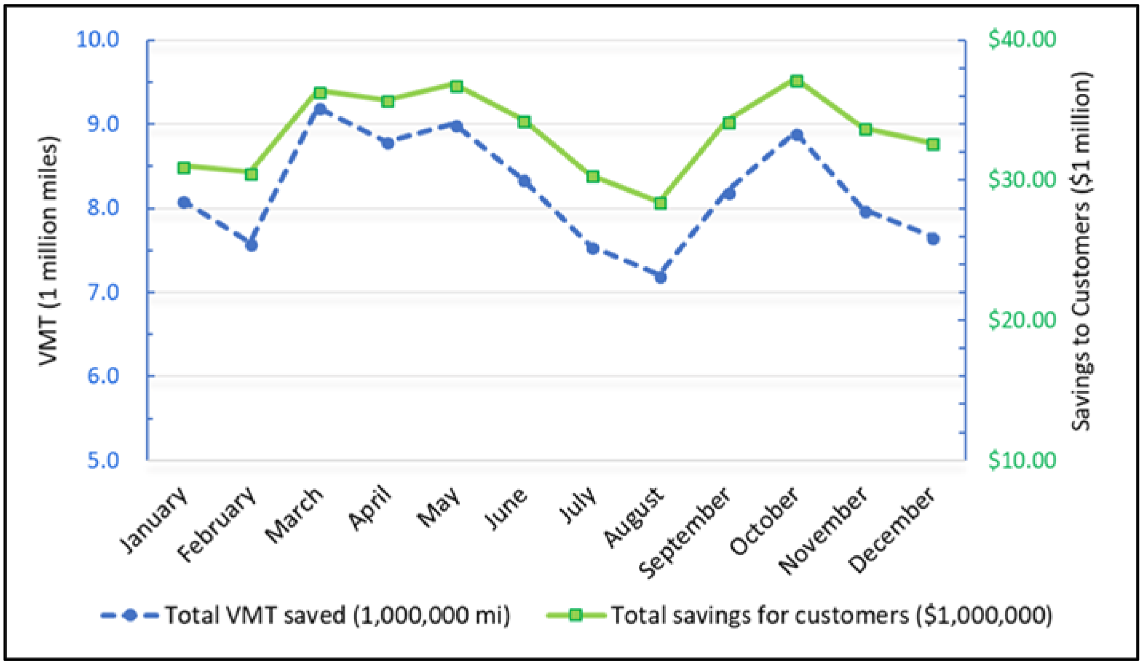

Figure 14: Benefits from taxi ridesharing

I also calculated how much taxi customers could save per ride if all sharing possibilities were realized (see Table 11). Overall, ride sharing would lower the average cost of a NYC taxi ride by $4.71.

4.2 Tobit model result

The parameters of the Tobit regression were estimated using maximum Likelihood Estimation in the statistical package R. To reduce heteroscedasticity, the dependent variable was transformed to be the logarithm of one + the density of shareable rides (one was added because some census tracts had no shareable rides). Results are reported in Table 13.

Table 13: Results of Tobit regression model for taxi rides sharing potential

	Coefficient	Elasticity at mean value
Population density (in 1,000s per square mile)	0.00784***	0.78088
% of CT members from Generation Z	-0.05803**	-0.95633
% of CT members from Generation Y	-0.03415	-1.05182
% of CT members from Generation X	-0.08605***	-1.84835
% of CT members from the GI Generation	-0.03172	-0.29975
Mean income per capita (in $1000s)	0.02440***	1.66701
% of CT population with less than a high school education	-0.09314***	-1.17915
% of CT population with high school education	0.02801	0.38178
% of CT population with some college or associate degree	-0.09231***	-1.59789
CT employment density (in 1000s, per square miles)	0.00060*	0.09146
CT unemployment rate	-0.10988***	-0.92189
% of households with no vehicle available	0.10912***	8.46662
% of one & two family buildings	-0.00556	-0.01290
% of multi-family walk-up buildings	-0.00674	-0.07434
% of mixed residential & commercial buildings	0.02026**	0.44410
% of commercial & office buildings	0.01845*	0.24815
% of industrial & manufacturing	0.10748**	0.13113
% of transportation & utility	0.01462	0.05366
% of public facilities & institutions	0.02828***	0.36707
% of open space & outdoor recreation	0.00599	0.03600
% of parking facilities	0.02572	0.03909
% of vacant land	-0.00702	-0.01601
% of other land uses	0.01814	0.00871
Intercept	1.90656	NA
Standard deviation of error	1.50592***	NA

Note: Statistical significance: * = 0.10, **=0.05, ***=0.01

As can be observed from Table 13, thirteen of my explanatory variable coefficients are statistically significant. This includes population density, the percentage of residents from generations Z and X, mean income per capita, two education variables (the percentage of people with less than a high school education, and with a college or an associate degree), employment density (in 1000s), the unemployment rate, the percentage of households with no vehicle available, and several land use variables.

Since the dependent variable is the logarithm of the shareable taxi density plus one (added since a couple of census tracts in Manhattan have no shareable ride in 2014), model coefficients can be interpreted as semi-elasticities. Hence, for explanatory variable X, a 1% increase in X results in a (β*100) percent change in shareable taxi density plus one, where β is the coefficient of X in the Tobit model. Elasticities calculated at the mean value of each variable are presented in the last column of Table 13 and briefly discussed below.

As expected, population density is statistically significant. A 1% increase in population density would increase the percentage of shareable taxi rides per acre by 0.78% at the mean population density.

The age composition of the population matters to some extent. Interestingly, the coefficients of all four generational variables are negative, which means that compared to Baby Boomers (my baseline) other generations are less likely to take taxis that could be shared. However, this difference is only significant for members of the X and Z generations: a 1% of increase in the percentage of Generation Z members results in a 0.95% decrease in shareable taxi rides for an area with a “mean age composition” (1.85% in the case of Generation X).

Unsurprisingly, income also matters. As mean income increases, so does the potential for shareable rides. More precisely, a 1% increase in per capita income results in a 1.66% increase (at the mean) in shareable taxi rides. Interestingly, the impact of education is not monotonic here as increasing the percentage of people with less than a high school education or with some college negatively impacts the potential for taxi shareable rides (the coefficients for both of these variables are significant), but not the percentage of people with just a high school education (the baseline is people with a college education or higher).

Employment density also plays a role here, but not quite as strong as I expected, although some of the impacts of employment may have been captured by land use variables (see below). Results show that a 1% increase in employment density results in a 0.09% increase in shareable taxi rides (again at the mean employment density value). As expected, an increase in the unemployment rate has the opposite effect. A 1% increase in the unemployment rate results in a 0.92% drop in the potential for shareable taxi rides.

The variable with the largest impact in this model is the percentage of households with no available vehicles. As this percentage increases by 1%, at the mean, the percentage of shareable taxi rides jumps by 8.47%, which is substantial.

Finally, as mentioned above, land use is important for explaining the potential for shareable taxi rides. Compared to multi-family buildings with elevators (my baseline here), increasing the percentage of mixed-residential & commercial buildings, of public facilities, of commercial & office buildings, and even of industrial and manufacturing facilities increases the potential for shareable taxi rides (with elasticities at the mean that range from 0.44% for mixed residential & commercial buildings to 0.13% for industrial and manufacturing facilities).

CHAPTER 5 CONCLUSIONS

In this thesis, I first presented a “one-to-one” algorithm to obtain an upper bound on the percentage of Canary Yellow taxi trips in 2014 in New York City that could be shared. My “one-to-one” algorithm matched rides with close origins and destinations, and accounted for requirements such as walking time, walking distance, taxi fees, and passenger counts. Results indicate that up to half of taxi trips could be shared. Sharing taxi rides has potentially substantial benefits: in 2014, VMT could be reduced by up to 98.5 million miles, which could save 4.5 million gallons and close to 88 million pounds of CO_2. Moreover, depending on the pricing scheme adopted, customers could save up to $400 million.

Second, based on the results from the matching algorithm, I estimated a Tobit model to explain the potential for sharing taxi rides at the census tract level in Manhattan based on land use and socio-economic characteristics. Results show that population and job densities, land use, as well as education, age by generation, and most importantly the lack of car availability explain the potential for taxi ride sharing.

These results are not without limitation. First, my model does not account for potential spatial effects (a spatial Tobit model may be more appropriate here). Second, given the available data, my results only provide an upper bound on taxi ride sharing since there is no information about the gender, ethnicity, and age of taxi riders and their willingness to share a ride. Third, the benefits of taxi ride sharing presented here are only estimates based on average values. Finally, if taxi sharing lowers substantially the price of taking taxis, it could induce new demand, which could not be accounted for here. Expanding on these ideas is left for future work.

REFERENCES

Barann, B., Beverungen, D., & Muller, O. (2017, July). An open-data approach for quantifying the potential of taxi ridesharing. *Decision Support Systems, 99*, 86-95.

Bellman, R. (1962). Dynamic Programming Treatment of the Travelling Salesman Problem. *Journal of the ACM*, 61-63.

Beutell, N. J., & Wittig-Berman, U. (2008). Work-family conflict and work-family synergy for generation X, baby boomers, and matures. *Journal of Managerial Psychology, 23*(5), 507-523.

Bhardwaj, D., Khan, A., Patil, S., & Dhoot, R. (2016). Real Time Taxi Rde Sharing. *International Journal of Computer Science and Information Technologies, 7*(1), 108-111.

Chan, N. D., & Shaheen, S. A. (2012). Ridesharing in North America: Past, Present, and Future. *Transportation Review, 32*(1), 93-112.

Chen, P., Liu, J., & Chen, W. (2010). A fuel-saving and pollution-reducing dynamic taxi-sharing protocol in VANETs. *2010 IEEEE 72nd Vehicular Technology Conference Fall*, 1-5.

Cohen, B., & Munoz, P. (2015). Sharing cities and suatainable consumption and production: towards an integrated framework. *Journal of Cleaner Production, 134*, 87-97.

Donovan, B., & Work, D. (2015). Using coarse GPS data to quantify city-scale transportation system resilience to extreme events. *2015 Transportation Research Board Annual Meeting.*

D'Orey, P., Fernandes, R., & Ferreira, M. (2012). Empirical evaluation of a dynamic and distributed taxi-sharing system. *2012 15th Int. IEEE Conf. Intell. Transp. Syst. IEEE*, 140-146.

Ferreira, N., Poco, J., Vo, H., Freire, J., & Silva, C. (2013). Visual exploration of big spatio-temporal ruban data: a study of New York City taxi trips. *IEEE Transactions on Visualization and Computer Graphics, 19*, 2149-2158.

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

Gates, T. J., Noyce, D. A., Bill, A. R., & Van Ee, N. (2006). Recommended Walking Speeds for Pedestrian Clearance Timing Based on. *Transportation Research Record: Journal of the Transportation Research Board*, *1982*, pp. 38-47. Washington, DC. Retrieved from https://pdfs.semanticscholar.org/fc81/aff9a47546a8034cf0d94438ed7466c97326.pdf

Gebeyehu, M., & Takano, S.-e. (2008). Modeling the Relationship between Travelers' Level of Satisfaction and Their Mode Choice Behavior. *Transportation Research Forum, 47*(2), 103-118.

Gonzales, E. J., Yang, C., Morgul, E. F., & Ozbay, K. (2014). *MODELING TAXI DEMAND WITH GPS DATA FROM TAXIS AND TRANSIT.* Mineta National Transit Research Consortium.

Greene, W. H. (2012). *Econometric Analysis, 7th Edition.* New York.

Hochmair, H. H. (2016). SPATIO-TEMPORAL PLATTERN ANALYSIS OF TAXI TRIPS IN NEW YORK CITY. *Transportation Research Record, 2542*, 45-56.

Horowitz, A., & Sheth, J. (1977). Ride Sharing to Work: An Attitudinal. *Transportation Research Record*, 1-8.

Hosni, H., Naoum-Sawaya, J., & Artail, H. (2014). The shared-taxi problem: formulation and solution methods. *Transportation Research Part B: Methodological, 70*, 303-318.

John, R. M., & David, H. A. (2014). Demand for taxi services: new elasticity evidence. *Transportation, 41*(4), 717-743.

Kattan, L., Barros, A., & Wirasinghe, S. (2010). Analysis of work trips made by taxi in Canadian cities. *Journal of Advanced Transportation, 44*, 11-18.

Kumar, A., & Lim, H. (2006). Age differences in mobile service perceptions: comparison of Generation Y and baby boomers. *Journal of Services Marketing, 22*, 568-577.

Ma, S., Zheng, Y., & Wolfson, O. (2013). T-share: a large-scale dynamic taxi ridesharing service., (pp. 410-421).

Marin, A. (2006). Airport management: Taxi planning. *Annals of Operations Research, 143*, 191-202.

Martinez, L., Correia, G., & Viegas, J. (2015). An agent-based simulation model to assess the impacts of introducing a shared-taxi system: an application to Lisbon (Portugal). *Journal of Advanced Transportation, 49*, 475-495.

Masayo, O., Huy, V., Claudio, S., & Juliana, F. (2015). A scalable approach for data- driven taxi ride-sharing simulation. *IEEE International Conference on Big Data*, (pp. 888-897). Washington, DC.

Meyer, D. (2016, April 1). *STREETSBLOG NYC.* Retrieved from DOT Releases Borough-by-Borough Speed Limit Maps: https://nyc.streetsblog.org/2016/04/01/city-releases-borough-by-borough-speed-limit-maps/

National League of Cities. (2015). *CITIES, THE SHAARING ECONOMY AND WHAT'S NEXT.* Retrieved from https://www.nlc.org/sites/default/files/2017-01/Report%20-%20%20Cities%20the%20Sharing%20Economy%20and%20Whats%20Next%20final.pdf

NYC Department of City Planning. (n.d.). Retrieved from New York City Borough Boundary: https://www1.nyc.gov/assets/planning/download/pdf/data-maps/open-data/nybb_metadata.pdf?ver=17d

NYC Taxi & Limousine Commission. (n.d.). Retrieved from TLC Trip Record Data: http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml

NYC Taxi & Limousine Commission. (n.d.). Retrieved from FAQs - Passenger: http://www.nyc.gov/html/tlc/html/faq/faq_pass.shtml

NYC Taxi & Limousine Commission. (n.d.). Retrieved from Taxicab Rate of Fare: http://www.nyc.gov/html/tlc/html/passenger/taxicab_rate.shtml

NYC Taxi & Limousine Commission. (2013). *Take Charge: A Roadmap to Electric New York City Taxis.* Retrieved from http://www.nyc.gov/html/tlc/downloads/pdf/electric_taxi_task_force_report_20131231.pdf

Posen, H. (2015). Ridesharing in the sharing economy: should regulators impose Uber regulations on Uber? *Iowa Law Review, 101*, 405-433.

Qian, X., & Ukkusuri, S. (2015). Spatial variation of the urban taxi ridership using GPS data. *Applied Geography, 59*, 31-42.

Santi, P., Resta, G., Szell, M., Sobolevsky, S., Strogatz, S., & Ratti, C. (2014). Quantifying the benefits of vehicle pooling with shareability networks. *Proceedings of the National*

Academy of Science of the United States of America, 111, 13290-13294. Retrieved from http://www.pnas.org/content/111/37/13290.full.pdf

Santos, D., & Xavier, E. (2015). Taxi and ride sharing: a dynamic dial-a-ride problem with money as an incentive. *Expert Systems with Applications, 42*, 6728-6737.

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

Toner, J. (2010). The welfare effects of taxicab regulation in English towns. *Economic Analysis and Policy, 40*(3), 299-312.

Tseng, C.-M., Chau, S. C.-K., & Liu, X. (2018). Improving Viability of Electric Taxis by Taxi Service Strategy Optimization: A Big Data Study of New York City. *IEEE Transactions on Intelligent Transportation Systems*, 1-13. doi:10.1109/TITS.2018.2839265

U.S. ROADS. (1997). *Study Compare Older and Younger Pedestrian Walking Speeds*. Retrieved from http://www.usroads.com/journals/p/rej/9710/re971001.htm

United States Environmental Protection Agency. (n.d.). *Green Vehicle Guide*. Retrieved 2018, from EPA: https://www.epa.gov/greenvehicles/greenhouse-gas-emissions-typical-passenger-vehicle

United States Environmental Protection Agency. (n.d.). *Greenhouse Gas Equivalencies Calculator*. Retrieved 2018, from https://www.epa.gov/energy/greenhouse-gas-equivalencies-calculator

Wallsten, S. (2015, June). The Competitive Effects of the Sharing Economy: How is Uber Changing Taxis? *TECHNOLOGY POLICY INSTITUTE*.

Yang, J., Jaillet, P., & Mahmassani, H. (2004). Real-time multivehicle truckload pickup and delivery problems. *Transport Science, 38*(2), 135-148.

Yazici, M., Kamga, C., & Singhal, A. (2013). A big data driven model for taxi driver's airport pick-up decisions in New York City. *2013 IEEE International Conference on Big Data*, (pp. 37-44).

Zhan, X., Hasan, S., Ukkusuri, S., & Kamga, C. (2013). Urban link travel time estimation using large-scale taxi data with partial information. *Transportation Research Part C: Emerging Technologies, 33*, 37-49.