Essays on the Market Impacts of Regulatory Regimes

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Matthew H. Shapiro

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
Doctor of Philosophy

Thomas J. Holmes, Advisor

May, 2018

Acknowledgements

This work and my development as an economist all stem from the patience, guidance, and support of my advisor, Tom Holmes. I would not have successfully navigated the academic and personal tribulations of the program without his mentorship. I also owe a debt of gratitude to my committee members, Joel Waldfogel and Naoki Aizawa, for their time discussing these projects. Joel's enthusiasm in particular was always a great asset in the most trying parts of this process.

This research has also benefitted from chats with Amil Petrin, Doireann Fitzgerald, Maria Ana Vitorino, Morris Kleiner, Dominic Smith, Mons Chan, Keaton Miller, who inadvertently provided the linchpin data idea for my first chapter, Thomas Quan, Thomas Youle, Ethan Singer, Imke Reimers, Malin Hu, as well as well as other participants in the Applied Micro workshop at Minnesota, seminar participants during the 2018 job market, and attendees of the NBER Digitization Tutorial. My greatest thanks go to Boyoung Seo, the coauthor of my final chapter and one of my closest friends.

Finally, I would like to thank the Department of Economics staff and the patient phone operators of the NYC Transportation and Limousine Commission. I like to think my persistent calls contributed to the decision to move their data online.

This dissertation was financially supported in part by Richard and Ellen Sandor and a data grant from the Indiana University Kelley School of Business.

Dedication

To my patient friends and less patient family, particularly Sheldon Shapiro, my grandfather and role model.

Abstract

This dissertation contains three essays, which focus on markets featuring heavy government intervention. The first two study the effects of Uber's entry into the taxi industry of New York City. The final essay, coauthored with Boyoung Seo, studies intervention in the growing market for electric vehicles in California.

In the first chapter I quantify the magnitude and distribution of the welfare offered by Uber's cab-to-customer matching technology. I combine publicly available transportation data with data scraped from Uber and traffic cameras in New York City to estimate a model of demand for transportation services and imbed it in a spatial equilibrium framework in which Uber and taxis compete. Uber's matching advantage depends on the density of the market. In consumer welfare terms, the introduction of Uber added only $0.10 per ride in the densest parts of New York but over $1.00 in the least dense. These results imply Uber's appeal in its densest market has depended on advantages independent from its matching technology, including its lower regulatory burden.

In the second chapter I document the potential of digitization to reduce statistical discrimination. First, I find that the search behavior of hail taxis, even controlling for profitability, highlights statistical discrimination against certain consumers. Second, Uber has mitigated the negative externalities in the cab markets among these consumers. A reasonable hypothesis is that Uber's matching technology permits contracts without the cost of undirected searching in previously avoided areas of the city.

In the final chapter, my coauthor and I assess the efficacy of vehicle subsidy programs and investment in a charging station network on demand for electric vehicles. In contrast to previous literature, we consider heterogeneity in tastes for electric vehicles and price elasticities across demographics, as well as the heterogenous marginal benefits of charging stations, and demonstrate the importance of both dimensions in correctly identifying the impact of subsidies and charging stations on demand. We use zip code-level data on vehicle purchases in California to estimate a random coefficient discrete choice model of automobile demand capable of proposing more efficient incentive structures.

Contents

Acknowledgements **i**

Dedication **ii**

Abstract **iii**

List of Tables **viii**

List of Figures **x**

1 Density of Demand and the Benefit of Uber **1**

1.1 Introduction 1

1.2 Evolution of Uber and Taxis in the New York City Market 7

1.2.1 Brief Description of the Market 8

1.2.2 Mapping the Growth of Uber Over Time 10

1.2.3 Constructing Wait and Price Data for Uber and Taxis 12

1.2.4 Contrasting Taxi and Uber 15

1.3 Structural Model 16

1.3.1 Demand 17

1.3.2 Supply 19

1.3.3 Equilibrium 27

1.4 Data and Estimation 28

1.4.1 Demand 29

1.4.2 Supply 36

1.5 Results and Discussion 40

1.5.1 Model Estimates . . . 40
1.5.2 Welfare Analysis and Counterfactuals . . . 42
1.5.3 Counterfactuals on Uber Restrictions and a Congestion Tax . . . 45
1.5.4 Concluding Comment . . . 46
1.6 Conclusion . . . 47
1.7 Figures . . . 50
1.8 Tables . . . 68

2 Mitigating Preference Externalities Through Digitization: An Application to Transit Markets 71
2.1 Introduction . . . 71
2.2 Data Sources and Construction . . . 73
2.2.1 Data Sources . . . 73
2.2.2 Data Generation . . . 74
2.2.3 Identifying Decision Nodes . . . 75
2.2.4 Taxi Choice Variable . . . 76
2.2.5 Constructing Choice Characteristics . . . 77
2.2.6 General Patterns . . . 78
2.3 Model and Estimation . . . 79
2.3.1 Choice Model . . . 79
2.3.2 Estimation . . . 80
2.3.3 Cleaning Data for Estimation and Endogeneity . . . 81
2.4 Results . . . 82
2.5 Discussion . . . 84
2.6 Conclusion . . . 85
2.7 Tables and Figures . . . 86

3 Heterogeneous Effects of Subsidy and Infrastructure Investment in Electric Vehicles Adoption 94
3.1 Introduction . . . 94
3.2 The Market for Electric Vehicles in CA . . . 97
3.2.1 Range Anxiety . . . 98
3.2.2 High Price . . . 100

3.3 Electric Vehicle Purchase Patterns . . . 101
3.4 Data . . . 102
3.5 Demand Specification . . . 105
3.6 Estimation . . . 107
3.6.1 Reducing the Parameter Space . . . 107
3.6.2 Maximum Likelihood Moments . . . 108
3.6.3 Matching Income Distribution Moments . . . 109
3.6.4 Subsidy Instrument Moments . . . 110
3.6.5 GMM Estimator . . . 111
3.7 Identification . . . 111
3.7.1 Identifying Heterogeneity . . . 111
3.7.2 Endogeneity . . . 112
3.8 Conclusion . . . 113
3.9 Tables and Figures . . . 115

References **125**

Appendix A. Appendix to Chapter 1 **132**
A.1 Data Sources and Construction . . . 132
A.1.1 Collecting Uber and Lyft Characteristic Data . . . 132
A.1.2 Constructing Wait Time Data . . . 133
A.1.3 Classifying For-Hire Vehicle Data . . . 137
A.1.4 Determining Morning Taxi Commutes . . . 138
A.1.5 Routing . . . 140
A.2 General Appendix . . . 141
A.2.1 Uber's Expansion Pattern Across Geographies . . . 141
A.2.2 Transportation Habits . . . 143

Appendix B. Appendix to Chapter 2 **146**
B.1 Route Planning and Timing . . . 146
B.2 Constructing Shifts . . . 149
B.3 Identifying Decisions and Intersection Passing . . . 150

Appendix C. Appendix to Chapter 3 **152**
C.1 Construction of Charging Station Variable 152

List of Tables

1.1 **Comparison of Taxi / Uber Platforms** 68
1.2 **Data Summary: Transit Choice** 68
1.3 **Data Summary: Transit Characteristics** 68
1.4 **Demand Estimation Results** 69
1.5 **Demand Estimation Results** 69
1.6 **Supply Estimation Results** 69
1.7 **Summary of Changes in Each Counterfactual, by Density Quartile** 70
2.1 **Percentage of Searching Trips with Intersection Crossing** 86
2.2 **Variables and Aggregation Methods** 87
2.3 **Descriptive Statistics for Regression Variables,** $t_r = 1$ 91
2.4 **Estimating Economic Versus Social Factors** 92
2.5 **Percent Change in Share from One Standard Deviation in Variable** 93
2.6 **Areas Benefitting from Ride Growth, June 2013 to June 2016** . 93
3.1 **Characteristics of Top 4 PEV and PHEV Models in California** 115
3.2 **Charging Speed by Charger Type** 115
3.3 **Desired Range Vs. Actual Range** 115
3.4 **Cumulative Charging Stations in California** 117
3.5 **Sponsored Charging Infrastructure Projects** 117
3.6 **Charging and Driving Behavior** 118
3.7 **Government Monetary Incentives for PEV** 120
3.8 **Income Distribution Conditional on a Vehicle or PEV Purchase** 121
3.9 **Poisson Regression of Electric Vehicles Sold, by Model** 122
3.10 **CHTS and CVRP Survey Summary** 123

3.11 **Heterogeneous Income and Political Distribution by County** . . 124
3.12 **LODES Residence and Workplace Summary** 124
A.1 **Summary Statistics for SAR** 142
A.2 **SAR Regressions** 143
A.3 **Count by Number of Modes** 145
B.1 **Sample Data for Determining Shifts** 149
B.2 **Searching Trips for Sample Data** 150

List of Figures

1.1 **Exclusion Zones for Taxi Service in NYC** 50
1.2 **Monthly Pickups by Cab Type** 51
1.3 **Monthly Pickups by Cab Type and Zone** 52
1.4 **Comparison of Uber and Taxi Price and Wait Times** 53
1.5 **Distribution of the Relative Price of Uber and Taxi, 2015** . . . 54
1.6 **Distribution of the Relative Price of Uber and Taxi, 2016** . . . 55
1.7 **Geographic Density Measure over NYC** 56
1.8 **Total Pickups by Taxi Zone, Q1 2016** 57
1.9 **Uber Share of Transit Market by Density, Q2 2016** 58
1.10 **Traffic Camera Locations** 59
1.11 **Sample of Traffic Camera Image Processing** 60
1.12 **Zones Included in the Estimation** 61
1.13 **Uber's Share of Pickups Around the Exclusion Border, Q3 2014** 62
1.14 **Average Estimated Taxi Wait Time Elasticity by Area** 63
1.15 **Estimated α_l Parameters over Density** 64
1.16 **Compensating Variation per Uber Ride, 2013 to 2016** 65
1.17 **Fitted Compensating Variation over Geographic Density** 66
1.18 **Compensating Variation per Uber Ride, After Banning Uber** . 67
2.1 **Percentage of Searching Trips with Intersection Crossing by Hour** 86
2.2 **Discretization of Cab Driver's Decision** 87
2.3 **Relative Choice of Taxis to Move North versus South, by Zone** 88
2.4 **Relative Median Income of Pickups After Moving North** 89
2.5 **Relative Race Index of Pickups After Moving North** 90
3.1 **Dissatisfaction with Charging Infrastructure** 116

3.2 **Public Charging Station Subsidy Availability, December 2016** . 118
3.3 **Charging Access at Work is Important** 119
3.4 **Distribution of Battery Charge at the Start of Charging Events** 119
3.5 **Full-Electric Vehicle Subsidy Available, December 2016** 120
3.6 **Heterogeneous Income and Political Distribution by County** . . 123
A.1 **Image Captured from Uber Application** 133
A.2 **Locations Sampled for Scraping in NYC** 134
A.3 **Stylized City with One Taxi** 134
A.4 **Deviation of Implied Uber Dataset from True Dataset** 138
A.5 **Recorded Trips Grouped by Number of Segments** 144
A.6 **Recorded Trips by Number of Segments, Excluding Walking** . 145
B.1 **Sample of Accurate Route Estimate** 147
B.2 **Percent Error in Estimated V. Actual Trip Distance** 148
B.3 **Determining Choice from Data** 151

Chapter 1

Density of Demand and the Benefit of Uber

1.1 Introduction

Since the company's founding in 2009, Uber and the ride-sharing business at large have transformed the once stagnant taxi cab market. Recent research by Cohen et al. (2016) estimates that Uber delivered a consumer surplus of nearly $6.8 billion dollars to the United States in 2015 alone. The magnitude and distribution of these consumer benefits, however, depend in large part on whether Uber can facilitate transactions that were otherwise cost prohibitive or impossible under existing services. In this paper I propose that Uber's technological advantage in matching consumers over these existing taxis highly depends on a market's density of potential demand. For New York City, I find that this advantage shrinks significantly with density. The technology difference translates to highly heterogenous consumer surplus gains from Uber; I estimate that they vary by over a factor of ten from the least dense to most dense areas in the study.

One of Uber's principal innovations in the transportation market is the way the platform matches consumers to drivers. Compared to a system in which people must physically hail a taxi, Uber's technology has effectively allowed potential customers to hail cabs blocks or miles from their location. While telephone-based dispatch services offered an analogous service, Uber and similar companies refined the system by using

geo-positioning to minimize the time a customer must wait for a driver. How advantageous this system is over hailing, however, depends on a market's geography. In very dense markets, like central Manhattan, where vacant cabs drive through most streets frequently, physically waving down a taxi can result in a match quickly. By comparison the same customer may wait longer for the contracted Uber. The Uber driver must navigate to her location and try to identify the correct person to pick up on a busy street. This simple intuition drives the central hypothesis of the paper. Uber's technology in matching consumers is advantageous in less dense markets but evaporates and can even be detrimental in highly dense areas.

This paper quantifies the technological matching advantage Uber has over taxis, the extent to which it depends on a market's density of potential demand, and the implications for the consumer value of Uber in these different areas. The New York City taxi market has an ideal setting to study this relationship. Besides being the largest taxi market in the United States, the city features wide variation in density from central Manhattan to less dense Manhattan and the outer boroughs.[1] This geography offers key variation across which to contrast the demand for taxis and Uber.

To study the development of the market I use publicly available trip-level data on the pickups of taxis and for-hire vehicles like Uber. These rich records permit a study of the New York over both space and time. I augment this dataset with two unique sources on consumer wait times for Uber and taxis. In the first I scraped the Uber app on a simulated Android phone to collect wait time and surge price data for that service in 47 locations across the city at different times of day. For taxis I follow Frechette et al. (2016) in using the pickup data to estimate a measure of the time consumers wait for taxis. I calculate these wait times at a granular level with respect to location and time, and discipline these estimates by using scraped traffic feeds to record the frequency of taxi traffic at key intersections throughout the city. I treat these wait times as the interface the consumer has to each of these platforms' technologies. Simple patterns in the trends of Uber and taxis over time in conjunction with these wait times delivers a hint at the major result of the paper. In the less dense parts of NYC, the taxi market has expanded with the growth of Uber. The wait times for an Uber in these areas are

[1] Approximately 20% of US taxi cab drivers are based out of New York and the city hosted 10% of Uber's 2 billion global rides in 2016.

much lower than for taxis. In Manhattan, however, Uber has cannibalized the share of taxis in the market without much overall expansion. In these same areas, the wait time for an Uber is often no better or even worse than for a taxi.

These stark findings motivate a model with a focus on controlling for geographic heterogeneity to isolate the effects of density. I break up this modeling problem into two parts. In the first, I develop a standard discrete choice demand model for transportation in granular submarkets of New York City. The demand model incorporates not only taxi and Uber but also alternative transportation choices, most importantly public transit options. Another important feature of the demand side is that I permit unobserved heterogeneity in the tastes for different choices across the city. This heterogeneity allows the model to capture consumer preferences for Uber over alternatives that I cannot directly model, such as better quality vehicles or the ability for consumers to screen drivers. All of these features are critical to ensure that I can separate out the demand for taxis as a function of geographic density from the quality of these outside options in the particular submarket. Because consumers do not care about technology differences across taxi and Uber per se but rather the prices and wait times they experience, both of which I measure directly, I estimate this portion of the model separately from supply.

The results from this demand model deliver immediate results on the change of consumer surplus from Uber across density. Following the methodology of the new product literature (see Petrin (2002)), I compare my time of study in 2016 to market data from 2013 as a baseline when Uber did not exist.[2] I estimate a \$0.10 per ride compensating variation to Uber riders giving up their Uber in the densest parts of the city but up to \$1.00 per ride to Uber riders in the least dense markets. As a percent of revenue per rides in the same area, values range from approximately 2% to 10%.

The second part of the model, the supply side, adds several features. The first is to estimate the relative efficiency of taxis versus Uber in areas of different density. This efficiency is a key set of parameters in the structural model. The second is to allow for equilibrium counterfactuals that test whether the consumer surplus brought by Uber is driven by technology or the strict regulatory regime capping the number of yellow taxis at levels below that which Uber operates. The supply model I employ is an extension of

[2] Uber technically did operate in the city already. Section 2 makes the argument that Uber's presence was small enough at that time to ignore its impact on the market.

the spatial equilibrium model introduced by Buchholz (2017) building off the oblivious equilibrium framework of Weintraub et al. (2008). The model itself, however, differs in two critical ways. First, I allow alternative platforms — in this case Uber — to exist in the same market as taxis. Second, I leverage the model and my collected data on wait times to estimate taxi matching efficiency in each of the submarkets I study rather than imposing it as the same for the entire city or large subregions of the city. This alteration alone is essential to allow the efficiency of taxi matching to change with submarket density. A key result from the estimation of this model is that matching efficiency for taxi cabs is indeed highly correlated with market density.

The counterfactuals fed to the supply model focus on tracking welfare and service quality changes from alterations to the taxi regulatory regime. In one key counterfactual I replace every Uber with a yellow taxi to determine the net impact on service quality by simply allowing yellow cabs to sidestep quota regulations. I find yellow taxis decidedly cannot replace Uber to match the same service quality measured in wait times, particularly in the less dense areas of the city. The result highlights that Uber offers genuine technological change in the less dense markets, where previous transactions with taxis were infeasible. In the densest markets of the city, matching technology is insufficient to explain why consumers have switched from taxi to Uber. Measuring service quality as wait times alone yields a result that dense markets would be better off with taxis replacing Uber.

Another principal counterfactual assesses a real policy in consideration in the debate around NYC transportation services. In this counterfactual I introduce a congestion tax on Uber for pickups in parts of Manhattan. Although I assume the incidence falls entirely on Uber, consumers in Manhattan substitute back to taxis as Uber service quality diminishes for a given volume of demand. Ultimately, the cost is born by the outer areas of New York since the tax reduces how much Manhattan can subsidize Uber drivers' operations in the outer parts of the city, and they exit the market.[3]

To my knowledge this paper is the first to discuss the magnitude and geography-based distribution of the welfare benefits offered by Uber's matching technology. It is not, however, the first to evaluate the consumer welfare impact of having a better

[3] As recently as March 2018 NYC policy makers have considered congestion taxes of the form proposed at the end of this paper. See https://www.nytimes.com/2018/03/31/nyregion/congestion-pricing-new-york.html.

matching technology for taxi-like services in New York. The closely related work of Frechette et al. (2016) and Buchholz (2017) both develop models of the taxi market in NYC. Frechette et al. (2016) explores this question through an aggregate model of the market to measure the cost of search frictions. In counterfactual simulations of the market, they introduce an Uber-like matching system by having a dispatcher link potential consumers to the closest cab within a mile. This paper builds off their findings in several dimensions. This work has the benefit of data on both Uber and taxis. The data allow me to explicitly model heterogeneity in the demand for these two services along more than just the wait time. On the supply side the data are crucial for estimating the relative efficiency of these two platforms across densities. Ultimately, a key finding from this paper mirrors a major result from their counterfactuals. They show that a dispatching system is a larger improvement over the existing technology when demand is thinner across the course of the day. This paper makes the same claim but across locations in the city.

The model in Buchholz (2017), in contrast, allows rich spatial heterogeneity in the market for taxis. In a counterfactual the research simulates the introduction of a more efficient matching technology by assuming that cabs can perfectly reach customers in each of several locations across the city. This paper shows, however, that the value of these services depends on several qualities of these different submarkets. One crucial advantage that Uber has over taxis, for example, is to match with consumers far from their current location. The incremental value of this difference over regular taxis depends on the density of the location, per the results in this paper. The real Uber matching system can also be disadvantageous in some locations, an aspect missed by modeling Uber as a hail cab with better efficiency. The value of Uber's matching technology also depends on the substitutability of taxis with alternative forms of transportation. How elastic customer response is to better matching depends on these outside options, which I explicitly control by estimating demand for a full model of a city's transportation services.

Recent research has also leveraged Uber data to directly measure the value of the company and its impact on the taxi industry. Cohen et al. (2016) estimates an aggregate measure of the consumer surplus generated by Uber. They suggest Uber has generated a potential $6.8 billion consumer surplus in the United States for 2015 alone. The

result cannot account for significant intra-city and inter-city variance in this surplus. Additionally, total surplus changes depend on Uber's impact on existing services over time. Depending on whether Uber expands the market for taxi services or simply displaces it, welfare gains estimated from short-term analysis may be mitigated by accounting for the quality of alternatives. Other research quantifies the benefits of Uber to the labor side of the market. Hall and Krueger (2015) and Chen et al. (2017) identify and quantity the labor-side welfare impact from Uber's flexible supply model, in contrast to traditional cab systems that work on fixed shift times.

This research additionally complements a recent explosion of research around both Uber and the taxicab industry, a few of which directly contrast Uber with traditional taxis. Cramer and Krueger (2016) look at the dimension of utilization rate, how often a cab is occupied, as a measure of Uber's relative efficiency. While not directly modeling these effects, they attribute utilization differences to the matching technology, Uber's scale, regulations, and the flexible supply model. Notably, they find Uber's utilization rate is significantly higher in all cities, save NYC. This paper models the first three of those forces though using expansion rather than utilization as a performative measure. Berger et al. (2017) examine the interaction of Uber and incumbents through the lens of employment and earnings. Their results contrast with a story of Uber rapidly destroying incumbents in the market, a result echoed in the NYC taxi market particularly. Bian (2017) focuses on the existence of network effects in the matching processes for these platforms. To some extent I capture this effect by allowing matching efficiency to vary with the scale of transactions; consumers may indirectly respond to the size of the "network" through the waiting time.

Naturally, Uber's signature surge pricing has itself been the subject of intense study. While I currently treat surge pricing as exogenously set by Uber in my estimation, other papers endogenize their problem. Castillo et al. (2017) find surge pricing solves a problem inherent in dispatch models of taxi operation — in contrast to street hail models. Flexible prices can deter demand from growing beyond supply capacity, thus preventing hypercongestion. Bimpikis et al. (2016) focus on the more familiar allocative role of surge pricing. They highlight the increasingly important role surge pricing plays in the profitability of *unbalanced* markets, in which some areas of the market feature much higher levels of demand than others at a given price. Both pieces of research

suggest that the "non-matching" innovations Uber brought to the market, dynamic pricing and the flexible supply side, could be critical to its sustainability.[4]

For transportation policy makers this literature and the new results developed in this paper might inform the vigor with which they try to regulate Uber and similar ride-sharing services. If the difference from existing taxis — granted few cities feature incumbents of the style in NYC — is largely in regulation avoidance, incumbents have little to fear from regulating ride-sharing services on equal footing. As of the time of writing, New York City has already begun moving in this direction. The city has lowered the barriers between incumbents and Uber on the supply side by introducing a universal taxi license. They have also attempted to crack the shift change rules with a pilot program allowing flexible driving hours. In contrast London has taken steps to ban Uber entirely.

The paper proceeds as follows. Section 2 offers a brief overview of the recent history of taxi services in NYC and a larger view of Uber's expansion and regulatory fights across the country. In this section I also describe my methodology for collecting price and wait time data for taxis and Uber; I then use it to develop descriptive evidence of the impact of Uber on the existing taxi market. Sections 3 and 4 present the demand and supply models to take it to the data. Section 5 presents the results from this estimation along with counterfactual analysis. Finally, Section 6 concludes with thoughts on future research.

1.2 Evolution of Uber and Taxis in the New York City Market

While the last five years have brought volatility to the New York City transportation market, two key platforms have emerged (or remained): traditional street-hail taxis, which I will refer to as taxis, and Uber. This section will briefly describe the broader changes in the market over these past several years. The focus then moves to taxis and Uber, describing their operation and the regulatory regime they face. Here, I explain my method of constructing prices and wait times to explore quality differences between

[4] Enormous subsidies from cheap financial capital are likely critical as well. I am not familiar with any research isolating this dimension.

the two platforms.

The section provides evidence of two key facts. The market for cab services has expanded but the magnitude of that expansion is highly correlated with an area's geographic density. Additionally, Uber has largely displaced traditional taxis in dense markets while contributing to an expanding market elsewhere. These two facts lay the groundwork for identifying the magnitude of Uber's technological advantage over traditional taxis in the structural model.

1.2.1 Brief Description of the Market

Prior to Uber's launch into the NYC market in May 2011, taxicab services were split primarily between street-hail yellow taxis and pre-arranged dispatched cabs, known as for-hire vehicles (FHVs). The Taxi and Limousine Commission (TLC) governed these two segments markedly differently. FHVs are restricted from picking up street hails (legally) but operated under comparatively free circumstances. Hard caps did and do not exist for FHVs nor are companies operating FHVs subject to strict price controls. In practice NYC has long had many more FHVs than hail cabs, but they were still far less ubiquitous on the road. While hail cabs completed hundreds of thousands of trips per day, FHVs completed on average about 20 or 30 thousand.

In contrast hail taxis are highly regulated. Two regulations through the period of interest are of particular importance. The first is the infamous medallion system limiting the number of yellow taxis on the road. From 2011 to the present day, the number of available medallions has fluctuated from 13,237 to 13,587. Because these medallions are required for the operation of yellow cabs and, until recently, in high demand, they became famous for their prices peaking in 2014 at over $1 million. For independent medallion owners, mandated to be around 40% of the market, the entrance costs are obvious; for cab drivers leasing cabs from corporate owners, the costs are passed down in the required weekly leasing payment, itself regulated by the TLC to prevent gouging drivers. The second important regulation are the strict price controls on ride fares.

The cap on the number of medallions and the stagnant FHV market left many parts of the city poorly serviced. Then Mayor Bloomberg perceived the problem as significant enough in early 2011, perhaps notably before the arrival of Uber, to warrant an initiative

to improve taxi service in the outer boroughs.[5] The solution, Boro (or "green") cabs, first hit the roads in August 2013. Many were simply converted livery cabs previously falling under the umbrella of FHV regulation.

For the purpose of this paper, green cabs also face two important regulations. These cabs can operate exactly like street hail yellow taxis, including the price regime, except at John F. Kennedy and LaGuardia Airports and in Manhattan below East 96th and West 110th Streets.[6] Figure 1.1 illustrates this exclusion zone over a map of the city. The second is that green cab vehicle licenses are capped, like yellow taxis, but they are available for a fixed fee. In the first year the TLC made 6,000 available for a price of $1,500. Every year thereafter licenses were available at a price of $3,000.[7] These regulations introduce a type of taxi with identical technology but a lower cost to enter the market, and, importantly, the geographic restrictions introduced a new supply of taxis in areas once scarcely visited by yellow taxis.

Uber, of course, significantly differs from both yellow and green taxis in terms of matching technology and regulation.[8] Uber drivers are prohibited from picking up street hails, though the dispatching model of ride-sharing companies circumvents that process by using the a phone app to make the match. While Uber drivers in New York City do not face the same legal operating costs as taxi drivers, they are still more regulated than drivers from other cities.[9] First, Uber is regulated just as any FHV is, and drivers must obtain a special license through the same process. Uber itself must also operate through base stations and satisfy the same administrative costs and burdens as other FHV companies. These similarities allay concerns that driver quality might

[5] This initiative actually came a year after the arrival of Uber. The State of the City address in January 2011 was the first acknowledgement of outer borough service as a priority: "Why shouldn't someone in the Bronx, Brooklyn, Queens, or Staten Island be able to hail a legal cab on the street? 97 percent of yellow cab pick-ups happen in Manhattan or at the airports — even though 80 percent of New Yorkers live outside of Manhattan."

[6] These cabs, unlike yellow taxis, can also continue to serve pre-dispatched rides from their base. My data only include the hail rides they complete.

[7] Although there is an overall cap of 18,000 green cabs, only about 8,000 have been purchased to the present day; hence, the supply cap has not been binding.

[8] App-based matching services for taxis first arrived for taxis in September 2015. CMT and Verifone, the two credit card operators covering approximately 8,000 covering approximately 14,000 yellow and green taxis, respectively, licensed different companies for their apps. Unfortunately, it is unclear what the uptake on these services has been. They were unsuccessful enough to require relaunching the programs in April 2016 toward the end of the period of interest.

[9] See Kleiner (2017).

be an unobserved, but important, difference between Uber and taxis.[10] Table 1.1 summarizes the key differences across yellow and green taxis and Uber.

The regulatory systems for these three platforms coalesce to create uneven relative costs for the different taxi technologies across New York City. In the green taxi exclusion zone of Manhattan, the hail technology requires a hefty fee, in the form of the medallion, to pick up potential consumers. The Uber model, along with other ride-sharing services, completely avoid this cost. Past the exclusion zone, the different technologies are on a more level playing field.

1.2.2 Mapping the Growth of Uber Over Time

Uber's ubiquity may be enough to convince one of the magnitude of Uber's operation in NYC. Figure 1.2 illustrates its expansion in terms of total monthly pickups over time. These data come from the TLC, which publishes detailed trip-level data for green and yellow taxis and, since 2015, pickup location by taxi zone and time data for FHVs, including Uber. The figure delineates two important cutoffs. The first is the introduction of green cabs in July 2013; it sees a steady rise in its pickups for about six months before largely flatlining, despite the availability of more permits after August 2014. The second is January 2015 when FHV data, including for Uber, is first consistently available. Lyft, which I can also separate from the rest of the FHV data, is part of the "Other FHV" category and a relatively small operation compared to Uber in this time period. The big takeaway, however, is Uber's performance compared to the yellow taxis over this time period. From when I first have data on Uber pickups in April 2014 — these early data are not available consistently through to 2015 — to July 2016, yellow taxis shrank from twenty-six times the operation of Uber to two times the size.

The story is incomplete, however, without looking at Uber's expansion across the city and the relationship of its growth with market density. Ideally, I would measure density as the number of potential consumers per unit of area and time. Ignoring the dimension of time, this ideal measure is particularly farfetched to obtain in New York because of the notorious difficulty in measuring the movement of population over the course of the day. These challenges motivate restrictions I introduce later in the paper

[10] Hall et al. (2017a) explicitly tests for quality differences in Uber drivers potentially arising by NYC regulations and find little evidence of a difference.

to limit the demand analysis to morning commuters. For now I use a different proxy for the potential density of demand, capturing relative geographic density across the city. In particular I use zoning data from New York's Property Land Use Tax Lot Output (PLUTO) to calculate the ratio of building space to the surface area of each taxi zone. Figure 1.7 depicts this measure overlaying the city.[11] As a cursory check for the usefulness of this measure, I note that it is positively correlated with the total taxi pickups normalized by the size of each taxi zone in Figure 1.8. In the model sections of this paper I will distinguish between this geographic density and transaction density, the latter referring to the volume of matches in a given area and time and can be considered analogous to market scale.

As seen in Figure 1.7 the exclusion zone marker, denoted by the thick black line, provides a natural break for a cursory analysis to compare the development of the taxi market in areas of different geographic density. The green exclusion zone, i.e. most of Manhattan, is, geographically, the densest part of the city while density drops off quickly in the other boroughs. Figures 1.3a and 1.3b map monthly pickup data by taxi or FHV type for pickups originating inside and outside the green cab exclusion zone, respectively. Outside of the exclusion zone, Uber is the largest platform in the market by July 2016 and the growth has, largely, not been at the expense of yellow and green taxis. Inside the exclusion zone, Uber remains the smaller competitor but has gained at the expense of the yellow taxis. Year-on-year monthly pickup growth inside the exclusion zone from 2015 to 2016 is, on average, slightly higher than 5% and outside the zone upward of 40%. Even in absolute terms growth is higher outside the exclusion zone by a factor of 2. By the end of the second quarter of 2016, Uber served approximately the same number of rides in and out of the exclusion zone but commanded a larger market share in most taxi zones outside (see Figure 1.9).

In Appendix A.2.1 I conduct a more granular analysis of the relationship between patterns in the market and density with observations at the level of taxi zone. The two key facts from the wider analysis hold even at finer levels. First, the market for taxi-like services has expanded relatively more in less dense zones of the city. Second, Uber and taxis are nearly perfect substitutes, in terms of 1:1 replacement of rides from

[11] I also carried out the analysis from this section using the ratio of building space to the length of road bordering and running through the area. The change results in no qualitative differences.

yellow taxis to Uber, over time in the most dense zones.

These trends paint a stark picture that motivates the theory in this paper. In the city's geographically densest markets, those in the exclusion zone, overall growth is minimal and hence Uber's gains come at the expense of yellow taxis. Enough consumers prefer Uber to taxi to switch, but the offering is not sufficiently valuable relative to alternatives to draw from other sources of transportation. Outside this area yellow and, importantly, green service plateaus early despite the latter operating below its supply cap. The ascendency of Uber and ride-sharing services in these sparse areas do not cannibalize existing business but is part of pure expansion. A natural hypothesis to draw is that potential green cab drivers found it insufficiently profitable to service areas that Uber can or was insufficiently attractive to generate the demand.[12]

1.2.3 Constructing Wait and Price Data for Uber and Taxis

The two key pieces of ride quality data by which consumers in this paper compare Uber and taxis are their prices and wait times. Unfortunately, only price data for taxis are readily available. Trip-level data from the TLC break down various components of each ride, like trip distance and time and, most relevantly, total trip fare. Additionally, because the taxi fare structure is fixed, I can approximate the cost of counterfactual rides with decent accuracy. The following sections detail how I collect the remaining price and wait data.

Uber Wait Times and Price

The trip records provided by the TLC for FHVs, which include Uber, do not provide the same information on prices. While Uber has a fixed fare structure for each of its products, its well-known *surge pricing* mechanism multiplies these prices in locations with especially heavy demand, relative to the available supply. Since these fees are quite substantial relative to baseline prices, controlling for them as best as possible is necessary for demand estimation.

To collect price information for Uber, I emulated an Android phone and automated

[12] One caveat about reading expansion from the available data is that new entrants may have cannibalized business from community vans, like "dollar vans." These services are more comparable to public transit than taxis.

the process of feeding the Uber app locations throughout New York City over the period of March to June 2016. I could then scrape the relevant data from the app itself. Every hour and a half, I scraped the expected wait time — then labeled "ETA" in the app — and surge price, which at the time showed up as a "[#]x" warning before agreeing to the contract. Appendix A.1.1 has more details on this process. To interpolate wait times at locations and times of day not sampled I used a simple inverse distance weighting method over time and space sampling only points not separated by bodies of water.[13] I use the same method of interpolation for surge prices with a more significant drop off to account for the sharper change in surge prices by area and over time.[14]

Besides interpolation another potential source of measurement error is the accuracy of the data reported to the consumer, i.e. the information I collect via scraping. Cohen et al. (2016) reveal that actual surge prices charged are slightly different than the prices shown to consumers. They report, however, these differences are marginal. Further, since consumers have no way to learn "true" price, this issue should not affect their transportation decision. On the other hand if wait times are consistently over or under-estimated, this error will be unaccounted for when I use these data in the demand estimation.

Taxi Wait Times

The final piece of critical information is the wait time for taxis. Frechette et al. (2016) is the first paper of which I am aware that attempts to convert the raw TLC data into information on how long consumers must wait for taxis. Two issues prevent me from adopting their methodology wholesale. First, this paper's research question revolves around spatial heterogeneity in the competition of taxis and Uber; hence, I need a metric for wait times at a fairly granular level. Second, the data available from the TLC has changed in the intervening years. Fitting their simulated model requires knowledge of the search time for each cab. For the data available in that time period, it was possible to track cabs over the course of the day. Since 2013 the TLC removed these identifiers

[13] Future versions of this paper will employ a more sophisticated distance metric using actual travel distance for the spatial dimension.

[14] Uber uses zones, unknown to me, to set surge prices so it less reasonable to assume a smooth transition of prices over locations. On the temporal dimension surge prices are highly volatile (see Diakopoulos (2015)).

because of privacy concerns. Without the ability to track a cab, the data do not indicate how much time each taxi spent searching for a new passenger.

In my adjusted procedure I generate a probabilistic count of taxis on each of New York's street segments throughout the day.[15] I use this count to estimate how long it takes for a vacant cab to pass through each segment without a pickup. Ideally, I would know precisely when customers arrive on a street segment to determine the time for a cab to pass them. This procedure does not assume a structure on customer arrival and measures wait time as how long a marginal consumer on each segment would have to wait for a cab to arrive. Hence, in reality some consumers in the data might be more or less "lucky" in catching a cab than what I measure. Since I cannot follow individual cabs, instead I keep track of how far the taxi could have traveled in an allotted time. Within this travel area is every possible route the cab could have taken based on a street map of NYC. For tracking purposes subsequent pickups are randomly assigned to any cab that could have made it to the location in time. This assignment generates a guess of the path of cabs between drop off and pickup. The full details of this algorithm and the additions that follow are described in Appendix A.1.2.

To improve the guess of which paths vacant taxis took, I scraped traffic camera data available publicly via the New York City Department of Transportation.[16] Figure 1.10 marks each camera location by a blue dot. I captured a roughly continuous feed — each camera updates a still image approximately every 4 or 5 seconds — for every camera 6 times a day with an additional measurement on Fridays and Saturdays for the months of September and December 2015.[17] On average I captured 600 images from each half hour block per camera. To follow a single camera for the entire two month period requires processing approximately 200,000 images. I spent the equivalent of 2 months working time selectively processing the images of roughly 150 cameras across the city.[18]

Figure 1.11 shows a screen capture of the program I put together to process the

[15] The area I track is limited to the areas covered in the estimation portion of the paper (see Figure 1.12).

[16] *Source:* http://nyctmc.org

[17] The specific times I started the camera captures were 3 AM, 9 AM, 12 AM, 2 PM, 5 PM, and 10 PM. I also captured 8 PM on Friday and Saturday. I also collected data in June 2016, closer to the period of estimation, but these images have not been processed.

[18] For the curious, or sympathetic, reader I first attempted to process the images using machine learning-based computer vision programs. Unfortunately, that these cameras produce static images rather than a video feed confounded my best attempts at that solution. Additionally, it proved difficult

images from the cameras. The key questions are those asking about "new and empty" taxis, that is, those marked vacant by a light atop the taxi in the image.

1.2.4 Contrasting Taxi and Uber

With taxi and Uber prices and wait times calculated, I close the section by contrasting these two services in select parts of the city. Figure 1.4 highlights a few popular origins and destinations to contrast the price and wait time for taxis and Uber over the course of the day using 2016 prices. The taxi price in the image is the median for all observed trips in March 2016. Uber prices are calculated using the fare structure at the time, the average weekday surge price for the time of day, and the median trip distance and time. Reported wait times are the average weekday wait for the time of day; note that they do not depend on the destination.

A few facts are obvious. Uber is typically cheaper, but it depends on the surge price at the time. They do not, however, always bear the advantage in wait time. In Williamsburg, a less geographically dense location outside of Manhattan, the wait time for an Uber is on average lower across the day, even though both green and yellow taxis can pick up there. In Manhattan, particularly the locations I have chosen, Uber's advantage depends on the time of day. The wait for taxis is higher during the shift changes around 5AM/PM but otherwise lower on average. While these wait times are functions of the market outcome and not the matching process alone, they may hint at the conditions driving the patterns from Figures 1.3a and 1.3b. From the wait perspective alone, Uber is hardly better and often worse than taxis in the most dense markets of the city. These are also the areas of the city where the competition between Uber and taxi has exhibited business stealing. Other factors may drive taxi customers to Uber, but they are not sufficient to grow the market. Meanwhile, in the outer boroughs, or at least in the Williamsburg neighborhood of Brooklyn, Uber offers substantial improvement in this quality dimension. The outer boroughs are also where there has been a significant expansion of the market.

---

to consistently determine whether a cab's vacancy light is on in areas of vastly different lighting. In future work the program I plan to outsource the processing of the cameras I missed and check the work I did to further refine my wait time measurements.

I conclude by noting a significant competitive advantage Uber, and other non regulated companies, have against taxi companies: the ability to change fare structures. Uber has had two major rate cuts in the past several years. The first in July 2014 made the prices of the cheap UberX option roughly comparable to traditional taxis. The second in February 2016 solidified Uber's price advantage. Figure 1.5 compares the actual price of all yellow taxi trips in March 2015 at the indicated locations against a fitted Uber trip cost.[19] At this point slightly more than half of Uber trips would have been cheaper than taxis. The difference between a more or less expensive ride largely depends on the time and distance of each trip since Uber charges by the minute and mile differently than taxis. Figure 1.6 illustrates the same distribution but in March 2016. After the price drop in February 2016 few Uber rides would have been more expensive than the same taxi trip. The "surge price tail" drives most of the cases where taxis are cheaper.

While I will not be modeling Uber's decision to change its fare structure, the difference in and out of the exclusion zone to this February 2016 highlight each market's capacity to grow. In Manhattan the effect of the price change is a bump in rides that follows an increase in rides for all taxis in February 2016. Meanwhile the period was followed by stagnant growth relative to the previous year, which did not feature a price change. One might worry that the price drop has a hollow effect because of the negative impact it has on supply. The expansion pattern outside of the exclusion zone, however, nullifies this argument. February 2016 marked an inflection point in the growth rate of Uber in these areas of the city. This differential response emphasizes the need to model the demand for these services with attention to heterogeneity across the city.

1.3 Structural Model

In this section I develop a structural spatial model of the NYC transportation market. The model features two sets of decision makers. On the demand side consumers arrive in separate zones across the city at specific times and make a discrete decision in their choice of transportation to a pre-determined destination. These choices are not limited

[19] The figure uses the surge price profile from 2016, but in practice affects a small percentage of rides.

to taxis and Uber, but also include the city's public transportation network. On the supply side I do not model the decision-making process of the principal governing taxi and Uber drivers — the medallion owners or the division of Uber governing decisions for NYC — rather I focus on the search decisions of the drivers themselves. This portion of the model is an extension of Buchholz (2017), allowing multiple, here two, taxi platforms to interact with each other as their respective drivers search for customers throughout the city. The immediate purpose of explicitly modeling the supply side of the market is to generate the distribution of cabs over the course of the day, figures which are not observable from the available data, and estimate the efficiency with which taxis match to consumers in areas of different geographic and transaction density.

1.3.1 Demand

Define a sub-market of New York as a zone l and time period t. The entire market for New York City is the collection of zones L and time periods throughout the day T. In a given sub-market I define the maximum potential demand $\tilde{Q}_l^t$, that is the total number of consumers who are looking for a ride in l at t. At time t a fraction of those consumers in l, $\gamma^t(l, l')$, seek to travel to location l'. Note also that $\sum_{l'} \gamma^t(l, l') = 1$.

A consumer i in zone l traveling to l' faces a choice set C of transportation options to complete the journey. These options may include public transit, taxis, Uber, walking, or an outside option, depending on availability. Both green and yellow taxis are considered part of the same choice.[20] Although standard for the literature, I justify the assumption that consumers typically only use one transportation option per trip in Appendix A.2.2. Besides the characteristics of the options, the choice might also depend on consumer i's income bracket g. The utility this particular consumer derives from choosing option $j \in C$ has the following form:

$$\begin{aligned} u_{ij}^t &= \alpha_{g(i)} p_j^t(l_i, l_i') + \beta_w w_j^t(l_i, l_i') + X_j^t(l_i, l_i')\beta + \xi_j^t(l_i) + \varepsilon_{ij}^t \\ &\equiv V_{ij}^t + \varepsilon_{ij}^t \end{aligned} \tag{1.1}$$

Going forward I will suppress these time superscripts. p_j is the cost in dollars of taking

[20] This simplification requires the consumer who wants a taxi would take either a yellow or green taxi, whichever responded to the hail first. It is possible consumers with a penchant for one color wait longer for the ride of their desired choice, but that seems out of character for New Yorkers.

the transit choice from l_i to l_i'. $\alpha_g(i)$, the marginal utility of money and hence value of time, depends on the commuter's income bracket, g. w_j is the associated waiting time, that is time not in transit, the commuter must expend with this transit choice. X_j is a vector of other trip characteristics, including the travel time and associated walking distance. The way consumers in this model can respond to congestion both explicit in counterfactuals and in reality is through the travel time term, as congestion will impact the travel time for Uber and taxis. ξ_j is transit choice-location-time unobserved demand. This term may reflect latent time-area preferences for the transit choice or unobserved qualities of the transit choice common regardless of location. For example, if consumers have a taste for a greater relative taste for Uber in the morning to avoid waiting in the cold, this preference would be picked up by ξ_j. Finally, ε_{ij} is an additional idiosyncratic taste shock. Consumer i then chooses option j if and only if

$$u_{ij} \geq u_{ik}, \forall k \in C$$

In total this utility model features parameters of interest $\theta_d = (\alpha', \beta_w, \beta')$.

Following the standard process from the discrete choice literature, this decision can be summarized as a choice probability. Let A_{ij} be the set of $\varepsilon_i = (\varepsilon_{0i}, \ldots, \varepsilon_{Ji})$ rationalizing choice j, that is $A_{ij} = \{\varepsilon_i | u_{ij} \geq u_{ik}, \ \ \forall k \in C\}$. The probability commuter i chooses transit option j is then $q_i(j; \theta_d)$, where

$$q_i(j; \theta_d) = \int_{A_{ij}} dP(\varepsilon_i) \tag{1.2}$$

Apart from the destination, the only distinguishing feature of each consumer is her income. Let $f_l(d_i)$ be the distribution of income classifications in location l. I assume this distribution is fixed for the location regardless of time of day *and the destination.* Then the share and total quantity of consumers traveling from l to location k who choose j is, respectively,

$$q_{lk}(j; \theta_d) = \int_{\{i | l_i' = k\}} q_i(j; \theta_d) f_l(d_i) d(i) \tag{1.3}$$

$$Q_{lk}(j; \theta_d) = \tilde{Q}_l \gamma(l, k) q_{lk}(j; \theta_d) \tag{1.4}$$

For the entire sub-market l (at time t), the share of consumers who choose j is then

$$q_l(j;\theta_d) = \sum_k \gamma(l,k) q_{lk}(j;\theta_d) \tag{1.5}$$

$$Q_l(j;\theta_d) = \tilde{Q}_l q_l(j;\theta_d) \tag{1.6}$$

Assumptions in the estimation section will permit an explicit formulation for Equation 1.2 and its dependents.

Demand in the Supply Model

In the supply model, taxi and Uber drivers take most features of each location in the city as exogenous. For example, subway and bus schedules and their impact on a consumer's choice decision is an exogenous feature of each time and location. To drivers the relevant demand information is $Q^t_{lk}(j;\theta_d)$ for $j \in \{Uber, Taxi\}$, where l and k are the origin and destination, respectively. To emphasize that price and wait time are relative qualities of Uber and taxi that can change over the course time[21] , one can rewrite this function as $q_j(p, w|l, k, t)$, where p and w are vectors of the price and wait time for the two taxi platforms, for each $j \in \{Uber, Taxi\}$. $q_j(p, w|l, t)$ is the same function integrated over destinations.

The supply model proceeds by describing the searching and matching procedure for taxis and Uber. The setup largely follows Buchholz (2017) with a few critical differences. First, I explicitly contrast the matching technology of an Uber and taxi. Additionally their matching efficiency depends on the area of the city they serve. Finally, the model predicts consumer wait times in the submarkets across the city thus allowing consumer demand to respond to changes in service quality in counterfactuals.

1.3.2 Supply

There is a fixed supply of yellow taxis, green taxis, and Uber cars operating throughout the city. These totals are denoted V_x, V_g, and V_u, respectively. Taxi and Uber drivers operate among the locations L of New York. I assume drivers, regardless of platform,

[21] Whereas Uber and taxis have no impact on the quality of other transit options in the area, for example.

attempt to maximize individual profits by picking up passengers over the course of a shift. Both Uber's system of loosely ride-based commission and taxi's system of weekly lease payments from drivers to owners are consistent with this goal. Throughout the day, composed of distinct five-minute time periods $t = \{1, \ldots, T\}$, vacant cabs search for consumers and occupied cabs travel to the destination location designated by their passenger. Any two locations $l, k \in L$ in the city are linked by a distance δ_{lk} and a time to travel τ_{lk}, the latter of which can change with the volume of traffic in the adjoining areas. τ_{ll} will denote an additional time to travel within a location for Uber only.

Arrival of Passengers

Having discussed demand in the previous section, I begin by fitting it into the supply model. To accommodate that the total number of consumers looking for rides in a particular submarket is likely not deterministic, I let $\tilde{Q}_l^t$ be the parameter of a Poisson distribution. In combination with the shares derived from the demand model, *on average* the demand function for taxis and Uber will be

$$Q_j(p, w|l, t) = \tilde{Q}_l^t \gamma^t(l, k) q_j(p, w|l, k, t) \tag{1.7}$$

for $j \in \{Taxi, Uber\}$ where $\gamma^t(l, k)$ is the location transition matrix introduced in the demand section. Holding fixed (p, w), the Poisson setup allows random variation in the scale of demand but not the relative demand between Uber and taxis. I assume that drivers for all platforms are aware of the distributions of $\tilde{Q}_l^t$ across the city over time and the demand parameters θ_d, i.e. they have enough information to form expectations over $Q_j(p, w|l, t)$. For now I again rewrite these functions for simplicity; denote $Q_{l,u}^t$ the demand for Ubers and $Q_{l,b}^t$ the demand for taxis in location l at time t.[22]

Searching

I start with the discussion of the matching process for taxis, both green and yellow. The critical difference in their operation compared to Uber is the process by which they match to consumers.

[22] Recall I assume consumers are indifferent between hailing a yellow or green taxi.

Consider a period t and location $l \in L$. Each location l is also associated with a dummy x_l, which takes a value of 1 when green taxis are not allowed to pick up in that area. The number of vacant taxis in location l at the beginning of the period is given by $v^t_{l,b} = (1-x_l)v^t_{l,g} + v^t_{l,x}$, where $v^t_{l,g}$ and $v^t_{l,x}$ are the number of vacant green and yellow taxis, respectively.

Taxis can only match with potential passengers within the same location. This process is

$$m^t_b(v^t_{l,b}, Q^t_{l,b}) = v^t_{l,b}\left(1 - \left(1 - \frac{\alpha^t_l}{v^t_{l,b}}\right)^{Q^t_{l,b}}\right) \tag{1.8}$$

Note that every location has a different efficiency parameter

$$\alpha^t_l = f_l(m^t_b) \tag{1.9}$$

which can be further specified as a function of the area and the scale, or transaction density, of the market. Section 1.4 will introduce assumptions on the form to make it tractable in estimation. Critically this function imposes no forced relationship between geographic density and efficiency.

Suppressing time the expected probability of finding a passenger from the perspective of a cab in a period is

$$p_{l,b} = \frac{E_{Q_{l,b}}[m_b(v_{l,b}, Q_{l,b})|\tilde{Q}_l]}{v_{l,b}}$$

Passengers in l seeking to travel to k, however, are not indifferent between traditional taxis and Uber. Therefore, I also specify the joint probability of finding a match and that match traveling to location k. Let $q_b(l,k) \equiv q_b(p, w|l, k, t)$, suppressing the t, that is the share of consumers who prefer taxis conditional on traveling from l to k.

$$p_{lk,b} = \frac{q_b(l,k)}{\sum_k q_b(l,k)} p_{l,b} \tag{1.10}$$

Hence $p_{lk,b}$ is the probability that a cab is matched to a passenger and that passenger requests to be taken to location k. Clearly $\sum_k p_{lk,b} = p_{l,b}$, the probability of matching at all.

From the perspective of a consumer, we can derive a similar expected probability of

being matched to a taxi.

$$p_{l,c} = E_{Q_{l,b}}\left[\frac{m_b(v_{l,b}, Q_{l,b})}{Q_{l,b}}|\tilde{Q}_l\right] \tag{1.11}$$

Note that the expectation is still over the total consumers who have arrived, and it is assumed the number of vehicles which will pass through the location at the time is known. This probability yields an expected wait time for the consumer to match a taxi. If $p_{l,c}$ were constant over time it would be given by the mean of the geometric distribution, i.e.

$$w_{l,b} = 1/p_{l,c}$$

measured in terms of the number of periods. Because $p_{l,c}$ adjusts with time, however, I can approximate the wait time using a survival function. Let $S_l^t(x)$ be the probability of matching in the xth period after initial arrival in period t:

$$S_l^t(x) = p_{l,c}^{x+t} * \prod_{k=1}^{x-1}(1 - p_{l,c}^{k+t}) \tag{1.12}$$

assuming that $x + t \leq T$ else it is 0. The expected wait time (in terms of number of periods) is then approximated by

$$w_{l,b}^t = \sum_{k=1}^{\bar{T}} k * S_l^t(k) \tag{1.13}$$

where necessarily $t+\bar{T} \leq T$. Note consumers waiting for taxis are assumed to disappear if they are not matched in their own period (and potentially born anew in the next period), but it is still possible to calculate the time it would have taken to catch a cab.

The first key difference between an Uber and taxi is an Uber driver need not be in the same location as a potential passenger to match to her. Indeed, their dispatching technology is the central focus of their potential advantage over taxis. I *assume that Uber permits guaranteed matching* at the expense of variation in the time needed for Uber to get to its passenger. With taxis there is uncertainty whether a match will be made but once the taxi is contracted the customer is in the cab. Although Uber assigns customers to the nearest vacant driver, at least at the time the paper covers, the model

will not specify the location of Uber vehicles with enough precision to replicate reality. Therefore, I impose a rule to approximate this process.

At a given time $v_u = \{v_{l,u}\}_l$ describes the distribution of vacant Uber cabs across any of the locations in the city. Likewise, $Q_u = \{Q_{l,u}\}_u$ is demand across the city.[23] The share of consumers in l matched to an Uber in location k is given by the logit form

$$p^k_{l,c} = \frac{\exp(v_k/(1+\tau_{lk})\mathbb{1}(\tau_{lk} \leq \bar{\tau}))}{1 + \sum_{k'} \exp(v_{k'}/(1+\tau_{lk'})\mathbb{1}(\tau_{lk'} \leq \bar{\tau}))} \tag{1.14}$$

where locations with a larger mass of vacant cabs and closer are likeliest to be the source of the match. $\bar{\tau}$ is the farthest distance an Uber would be dispatched for a pickup. Hence the probability an Uber in location k is assigned to a passenger in location l is determined by

$$p^k_{l,u} = (Q_{l,u}/v_{k,u})p^k_{l,c} \tag{1.15}$$

Again, let $q_u(l,k) \equiv q_u(p, w|l, k, t)$, the share of consumers who prefer Uber conditional on traveling from l to k. Then

$$p^{k'}_{lk,u} = \frac{q_u(l,k)}{\sum_k q_u(l,k)} p^{k'}_{l,u} \tag{1.16}$$

is the probability an Uber in location k' is matched to consumer in l requesting to be taken to location k. Note I do not allow Uber drivers to reject rides.[24]

It straightforward to check that every passenger will indeed be matched, that is

$$Q_{l,u} = \sum_l p^k_{l,u} v_{k,u}$$

[23] Since mid 2016 Uber has begun assigning passengers to non-vacant cabs. Presumably, it would be a straightforward extension to consider here, where all Uber cabs, occupied or otherwise would be considered for the match, with the wait time also accounting for the time to drop off the last customer. That is a needless complication for the period of consideration, though.

[24] Ge et al. (2016) show that drivers do indeed discriminate against passengers. This would be relevant to consumers in my model if consumers in certain areas have to wait longer in ways not captured by the estimate given by the Uber app. The data I collected on wait times prior to finalizing the transaction, however, do not allow me to determine if approximated (pre-finalization) and actual wait times differentially diverge depending on one's location.

The result of this assignment process is how long a passenger must wait for the contracted Uber and how long the Uber must drive (at its expense) to pick up that passenger. I assume when a consumer opens the Uber app to observe the wait time, the observer is marginal and sees the average result of the process above, that is

$$w_{l,u} = \sum_k \tau_{lk} p^k_{l,u} \tag{1.17}$$

Congestion

A final implication of the search behavior of taxis and Uber drivers is their impact on congestion throughout the city. I model this impact by allowing transit speeds across zones to differ with the level of taxis and Ubers searching in each of the zones. Consider two proximal taxi zones l and k,

$$\tau^t_{lk} = g_{lk}\left(\sum_j (v^t_{l,j} + e^t_{l,j}), \sum_j (v^t_{k,j} + e^t_{k,j})\right) \tag{1.18}$$

where e designates the count of employed cabs and g a function that can depend on the zone pair. τ^t_{lk} for locations that are not proximal remain the shortest route between l and k, considering the changes from traffic. In Section 1.4 I introduce assumptions on the function to estimate it independent of the rest of the model. In counterfactual simulations consumers will then respond to congestion through their preferences on travel time.

Static Profits

I assume drivers on all platforms are only paid for rides.[25] The fare structure, however, depends on the type of platform. For taxis the fare structure is set by regulations with a fixed price of ϕ_b and distance-based fare π_b. Hence profits for a ride taking a passenger from l to k are

$$\Pi^x_{lk} = \phi_b + \pi_b \delta_{lk} - c_{lk}$$

[25] In reality Uber has flexibility around this assumption I cannot capture. In the short-run Uber might offer driver incentives detached from the fare structure.

where c_{ij} are the costs, i.e. fuel, of travel on the trip. For green cabs I make the adjustment

$$\Pi^g_{lk} = (1 - x_l)\left[\phi_b + \pi_b \delta_{lk} - c_{lk}\right]$$

that is, I force profits for green cabs to be 0 in areas where they should not pick up.

Uber's fare structure is slightly different. In addition to a commission taken from revenue, Uber utilizes surge prices, which multiply revenues by some factor, and time-based fares. Hence profits for a ride taking a passenger from l to k are

$$\Pi^u_{lk}(s) = \kappa \times s\left[\phi_u + \pi_{u,1}\delta_{lk} + \pi_{u,2}\tau_{lk}\right] - c_{lk}$$

where $s \geq 1$ is the surge factor and κ is the commission.[26]

States and Payoffs

Ultimately, the interesting behavior of the taxis in the model is the decision of where to locate in their search for passengers. The object of interest from this model is the state of the world S encapsulating the location of taxis and Uber at any given time and driving this behavior. All cabs keep track of 7 sets of states. First is cab i's own location at t, l^t_i for cab i. The rest of the market is captured by the count of vacant green, yellow, and Uber taxis in each location, $v^t_{l,g}, v^t_{l,x}, v^t_{l,u}$, respectively, and the count of cabs in transit $e^t_{k,g}$, etc., where k indexes the number of periods until the cab arrives at its destination. Finally, the drivers keep track of the distribution of surge prices s^t_l. For the estimation I assume the distribution of surge prices are known in advance, not an unreasonable assumption in the medium run if surge prices follow general daily patterns.

Hence, the full state for any cab i is

$$S^t_i = \{l^t_i, \{v^t_{l,j}\}_{j,l}, \{e^t_{l,j}\}_{j,l}, \{s^t_l\}_l\}$$

I assume all vacant drivers have a belief on the complete state $S = \{S^t_i\}_{i,t}$. While it is a stretch to assume that taxis have knowledge of surge prices, it is more reasonable that

[26] Because I do not currently model the decision of Uber as a platform, this s is taken as exogenous and read from data by location and time of day as explained in Section 1.2.3.

they form an expectation of it over time. Nonetheless, I do not model this dimension of uncertainty for now. Given S, taxis can then evaluate the expected dynamic value of any location l.

$$V_{l,b}^{t}(S) = E_{p_l|\tilde{Q}_l,S^t}\left[\sum_k p_{lk,b}^{t}\left(\Pi_{lk}^{b} + V_{k,b}^{t+\tau_{lk}}\right) + (1-\sum_k p_{lk,b}^{t}) E_{\varepsilon_k^{t+1}}\left[\max_{k\in C(l)} V_{k,b}^{t+\tau_{lk}} - c_{lk} + \varepsilon_k\right]\right] \tag{1.19}$$

for $b \in \{x, g\}$. Note that $C(l)$ is the choice set of alternative locations given a cab starts in l. $C(l)$ includes all adjacent, i.e. proximal, locations and l itself. The first half of the expectation is the probability that in location l the taxi makes a match and is sent to location k with a passenger. The second half of the expectation is the probability that it makes no match within the time period and must make a decision about where to search next by maximizing the net present value of profits. I allow an additive idiosyncratic shock ε to that decision, which follows the extreme value distribution, useful both for modeling the search choice decision and capturing unobserved heterogeneity in searches.

Uber cabs have a different value function by nature of the matching assignment process.

$$V_{l,u}^{t}(S) = E_{p_l|\tilde{Q}_l,S^t}\left[\sum_k \sum_{k'} p_{kk',u}^{l,t}\left(\Pi_{kk'}^{u}(s_k^t) - c_{lk} + V_{k',u}^{t+\tau_{lk}+\tau_{kk'}}\right) + (1-\sum_k \sum_{k'} p_{kk',u}^{l,t}) E_{\varepsilon_k^{t+1}}\left[\max_{k\in C(l)} V_{k,b}^{t+\tau_{lk}} - c_{lk} + \varepsilon_k\right]\right] \tag{1.20}$$

The major difference is in the first term. The first probability after the summation is the probability that an Uber in location l at t is matched to a consumer in location k who requests a destination in k'. Two additional differences from taxis are that the payoff from the trip depends on the surge price of the customer's location and, even after match is made, the driver faces a cost in traveling from location l to the pickup point in k. The time superscripts on the continuation value are also slightly different. Taxis are idle only during the search process. Uber drivers are idle both while searching and while picking up passengers. These differences identify what could be one heuristic for

how cab drivers make their decisions. A first-order concern for taxis is the probability they will find a match in a given area. That concern is supplanted by the idle time for an Uber driver, as described in Castillo et al. (2017).

Choice Problem and Transition Beliefs

The critical choice problem facing drivers is where to search in the event they do not get matched in period t. Save the form of the continuation value, the decision problem faced in Equations 1.19 and 1.20 are identical. Both unmatched Uber drivers and taxis in location l solve the following problem

$$k^* = \arg\max_{k \in C(l)} \left(V_{k,j}^{t+\tau_{lk}} - c_{lk} + \varepsilon_k \right) \tag{1.21}$$

where $C(l)$ is the lists of location adjacent to l. Although the problem is the same for $j \in \{b, u\}$, the motivations for moving are not. While we expect taxis to search to maximize their probability of matching, the related incentive for an Uber driver is to minimize the distance they would need to travel upon being matched to a consumer.

Because ε is assumed to follow the Extreme Value I distribution with scale σ_ε, the probability of choosing a particular search location k is given by the logit formula

$$\sigma_{l,j}(k|S^t) = \frac{\exp(E[V_{k,j}^{t+\tau_{lk}} - c_{lk}|S^t])}{\sum_{k' \in C(l)} \exp(E[V_{k',j}^{t+\tau_{lk'}} - c_{lk'}|S^t])} \tag{1.22}$$

for a cab of type j starting in location l. For any type of cab in location l at t, then, this function determines the optimal movement to new locations. Designate that vector $\sigma_{l,j}^t$. There are then three matrices to determine the entire transition matrix for empty cabs σ_j^t.

1.3.3 Equilibrium

To ensure the tractability of the model, I utilize the oblivious equilibrium concept of Weintraub et al. (2008) adapted for the taxi industry in Buchholz (2017). I assume that drivers of all cab types hold beliefs over competitors of all other cab types regarding both their policy functions and spatial distribution. Along with complete knowledge of demand and the surge price schedule, every driver can project the evolution of the state

S^t over the course of the day. Let Σ^t denote this transition belief for the state at time t.

The **equilibrium** is the set of states $\{S^t\}$, transition beliefs Σ^t, policy functions σ_j^t for all j and t given an initial state S^0 and the total number of cabs V_j for $j \in \{x, g, u\}$ such that the following conditions hold

1. Taxis match in each location l at the start of period t according to Equation 1.8. They transition according to $p_{lk,b}^t$ once conditioned on matching. Uber drivers also match and are assigned customers according to Equation 1.15 with transitions conditional on contracting defined by Equation 1.16. Together these transitions determine the aggregate movement of occupied cabs. Let $\nu(e_j^{t+1}|e_j^t, m^t, p_u^t)$ be the transition kernel where e_j^t is the distribution of cabs of type j at time t, p_u^t is the assignment of Uber drivers to passengers across all locations, and m^t are taxi matches.

2. Vacant drivers at the end of every period move according to the solution of Equation 1.21, that is $\sigma_j^t(S^t, \tilde{\Sigma}^t)$ based on the state and beliefs for all cab types, where $\tilde{\cdot}$ denotes beliefs. Let $\mu(v_j^{t+1}|v_j^t, \tilde{\sigma}_j^t, S^t)$ be the transition kernel of vacant taxis.

3. The realized state transition is the combined movement for employed and vacant cabs, along with the exogenous change in surge prices, s^t. Hence $\Sigma(S^{t+1}/s^{t+1}|S^t/s^t) = \nu(e_j^{t+1}|e_j^t, m^t, p_u^t) + \mu(v_j^{t+1}|v_j^t, \tilde{\sigma}_j^t, S^t)$.

4. Rational expectations ensure that $\Sigma^t = \tilde{\Sigma}^t$ for all t.

Existence follows from the standard arguments.

1.4 Data and Estimation

Estimation of the model proceeds in three steps. Because I measure price and wait times directly, I estimate demand separately as the first step of this procedure. I then integrate Equations 1.4 and 1.6 into the estimation of the supply model. Finally, the congestion and efficiency term functions are estimated at the end.

This section proceeds by describing the data and estimation procedure for demand. It concludes by repeating the exercise for supply.

1.4.1 Demand

Rather than modeling the transit demand for the entire city throughout the day, I make two limitations on the scope of the estimation. The first is that I estimate the demand system based on Monday through Thursday morning commuting patterns. The reason for this limitation is multifold. Residency population distributions are more accurate representations of the actual distribution of population on weekday mornings. Second, other than what the MTA can track through card users, the census publishes the most persistent datasets on transit choice, but those choices are limited to morning commutes. Finally, changes to the transit system can influence a consumer's extra-marginal decision about making a trip. I lack the data to meaningfully estimate a model that captures this dimension. In modeling morning commutes, which are more likely to occur with regularity, I can mitigate this particular issue. Additionally, granular data exist linking home and work locations. I explain extending the demand estimates beyond these commuting times in the discussion of the supply model.

The second limitation reduces the geographic scope of the estimation. Its purpose is simply to avoid areas with near zero shares of taxi users. Figure 1.12 shows the extent of the coverage. In total 129 taxi zones, or 350 census tracts, are included. One concern is that these areas are some of the most dense in the city. Because the matching efficiency for taxis will be estimated for each zone in the supply section, it should not bias those estimates. Results reported over measures of density, however, will obviously be truncated to these higher density locations.

This model is not one of a daily choice problem for each consumer. Rather, as details of the data sources lay out, I model the *typical* choice a consumer would make on her way to work. The major assumption is that the demand parameters governing the consumer problem for morning commute trips will also apply to any another trips taken in the day.[27]

[27] This assumption is not entirely out of left field, but a particular concern is that, say, consumers are more wait time elastic in their choices for the morning commute.

Data Sources

Three sources of data constitute the consumer choice data for the demand estimation. The first is the 2008 New York Customer Travel Survey conducted by the Metropolitan Transportation Authority (MTA) over the period of May through November 2008. While the period covered is long before the key time period of interest in mid-2016, the dataset offers more extensive details on transit behavior and rider demographics than any other available.[28] The survey provides the typical work transit mode of over $10,000$ people living in the five boroughs of New York. In addition to their transit choice the survey takers indicate the census tract of their home and work, time of departure, and income bracket. These observations are designated $M = \{M_i\}_{i=1}^{N_1}$.

The next set of choice data, from the American Community Survey, allow for the construction of choice shares at the temporospatial divisions outlined in the model section. The 5-year surveys report estimates of transit choice by the time of departure in 30-minute to 1-hour blocks and by census tract. I require the construction of shares for two time periods, 2008 and 2015.[29] I can then aggregate the tract data up to the taxi-zone level. The long horizon in calculating the estimates of these shares is the principal disadvantage in using this dataset. In an environment I have speculated has changed significantly over the past several years, 5-year estimates dampen the extent of the transformation. Additionally, the ACS does report taxi service usage but does not distinguish between different types of taxis. These observations are designated $B = \{B_i\}_{i=1}^{N_2}$.

To augment the ACS data, the final set of choice data utilizes the yellow, green, and Uber pickup data from the TLC already exhibited elsewhere in the paper. As alluded to in the discussion of the other datasets, I favor this source over the ACS for its ability to differentiate between traditional taxis and Uber and the potential to take advantage of daily variations in the characteristics — and choice — of these two products. These variations are arguably more subject to change from day to day than those impacting the choice to take public transit or walk. The unfortunate shortcoming is that I am

[28] Future versions of this paper will certainly take advantage of updated transit surveys. The National Household Travel Survey for 2016 to '17 is slated for release in 2018, for example. The ACS Public Use Microdata Sample also includes rich demographics but obscures home and work locations in geographies too large to be useful in this context.

[29] The 2015 data can be updated to 2016 as soon as the Census Bureau releases it in late 2017.

unable to distinguish work commuters from other passengers. In the main body of the paper I attempt to tease out which rides are morning commutes, the details of which are in Appendix A.1.4.[30] I use observations from the end of March 2016 through June 2016, corresponding to the period during which I collected characteristic data for Uber rides, and designate these observations $K = \{K_i\}_{i=1}^{N_3}$. A summary of these data sources is in Table 1.2.

The observable trip characteristics in the demand model are travel times, walking distance, price, and wait times. Nearly every characteristic depends on the starting location, the destination, and the time of departure. Consider a taxi trip I observed at 9:00AM from location l to location k. Of the menu of choices, the only characteristics read off data are the price and traveling time for the taxi. Every other characteristic must be simulated. Table 1.3 summarizes which characteristics are observed or simulated when the choice is the observed choice.

To simulate travel times for any choice, I use a mix of OpenTripPlanner and a separately built graph of NYC's road network. OpenTripPlanner provides directions similar to Google Maps, but trips can be planned around an arbitrary public transit schedule. Since my data are now historic, this feature was important to accurately reflect the contemporaneous schedules. Appendix A.1.5 further details its usage and application to vehicle-based choices. I also use this program to gauge walking distances. I assume for vehicle-based choices that the walking distance is negligible.

Prices and wait times are the final characteristics and the only two observable characteristics differentiating Uber from taxis. Prices for public transit are calculated using contemporaneous prices along each route simulated. Section 1.2 described the process for gathering price and wait times for taxis and Uber. The route planner also generates wait times for public transit options.

An additional assumption I impose on the data is that all Uber rides I observe are through UberX, the most popular choice at the time. Although I collected price and wait time for all of Uber's choice offerings in New York City, The TLC dataset does not distinguish which type of Uber passengers used to complete a particular ride.

The final sources to address concern individual characteristics, destinations, and

[30] In future robustness tests I will check the impact on the demand parameters by dropping this filtering.

income. I simulate income at the census tract level — sub-geographies of taxi zones — using 5-year ACS data. I use the Longitudinal Employer-Household Dynamics (LEHD) Origin-Destination Employment Statistics (LODES) for 2014 to simulate home tract-to-work tract commute flows for later periods, and the Census Transportation Planning Product (CTPP) estimated based on ACS data from 2006 to 2010 for the 2008 period. These two datasets are used to empirically estimate the origin-destination matrix $\gamma(l, k)$.

Estimation Procedure

To estimate the model in Section 1.3, I require a few additional assumptions. First, I assume ε_{ij} are iid across individuals but multivariate normal across products, following the set up in Goolsbee and Petrin (2004). Hence ε_i follows the distribution below

$$\varepsilon_i = \begin{bmatrix} \varepsilon_{tran} - \varepsilon_0 \\ \varepsilon_{taxi} - \varepsilon_0 \\ \varepsilon_{uber} - \varepsilon_0 \\ \varepsilon_{walk} - \varepsilon_0 \end{bmatrix} \sim MVN \left(0, \begin{bmatrix} 1 & \sigma_{tr,tx} & \sigma_{tr,u} & \sigma_{tr,w} \\ \cdots & \sigma_{tx}^2 & \sigma_{tx,u} & \sigma_{tx,w} \\ & \cdots & \sigma_u^2 & \sigma_{u,w} \\ & & \cdots & \sigma_w^2 \end{bmatrix} \right) \tag{1.23}$$

or $\varepsilon_i \sim MVN(0, \Omega)$, introducing an additional six demand parameters. I depart from the standard logit setup because of the striking and close substitution from yellow taxis to Uber laid out Section 1.2. The second assumption, worth restating here, is that (θ_d, Ω) govern preferences in all time periods covered.

Given the first assumption I can calculate predicted shares for each market, here defined as 30-minute blocks (t) for each of the taxi zones (l) covered in Figure 1.12 for each separate month in March to June 2016. For each taxi zone l, ns individuals are simulated *for each census tract*, so each taxi zone is partitioned by census tracts $l = \{l^1, \ldots, l^n\}$. Income is drawn independently for each of the simulated consumers according to the distribution $f_{l^i}(d_i)$, where d_i is the vector of indicators for income bracket in the census tract. The desired destination tract k^i is also drawn independently based on $\gamma(l^i, k^i)$, pinned down empirically with data described in the previous section. Each of the ns individuals is then assigned a time-of-departure block by weights generated from ACS data. Let $\pi_l^t(l^i)$ be the fraction of departures in time block t originating from

tract l^i. Given (θ_d, Ω), predicted market shares for the zone are

$$q_l^t(j;\theta_d,\Omega) = \sum_k \pi_l^t(l^k) \left[\frac{1}{ns} \sum_{k_i=1}^{ns} \int_{A_{k_i j}} dP(\varepsilon_i) \right] \tag{1.24}$$

with the integral computation here over the MVN idiosyncratic errors.

I proceed with the standard technique introduced by Berry (1994) to "concentrate out" $\{\xi_j^t(l)\}_j$, the unobserved choice taste parameters for each market. Shares are constructed for these markets in 2008 and 2016 using the ACS data and denoted $q_l^{t,DATA}(j)$.

Before the inversion, I make a slight alteration to Equation 1.1 to account for a peculiarity with the wait times. For taxis and Uber the wait time is independent of the destination. I split up wait time into two additive components $w_j(l,l') = [d_j w_j(l) + (1-d_j) w_j(l,l')]$, where d_j is an indicator for a taxi-like choice. I can then define $\delta_j(l;\theta_d,\Omega) = \beta_w d_j w_j(l) + \xi_j(l)$. Typically, the Berry inversion carries the extra value of reducing the parameter space by the terms linear in δ. In this case β_w still remains in the form $\beta_w(1-d_j)w_j(l,l')$, but, the inversion still plays two important roles in this estimation. First, isolating w for taxis and Uber in a linear equation introduces an opportunity to instrument wait times to handle with endogeneity concerns. Unfortunately, the same technique cannot be used with price — price is always a function of the starting point and destination — but controlling for the component of the error, a location-time-specific taste for Uber, that is likely correlated with price can mitigate endogeneity concerns if not directly address it. For all other transit choices, prices and wait times should not be meaningfully responsive to these unobserved tastes.

The Berry technique in this case requires the restriction that shares in the data match predicted shares from the model at the half-hour, taxi-zone block level. That is,

$$q_l^{t,DATA}(j) - q_l^t(j;\theta_d,\Omega,\delta) = 0 \quad \forall j,l,t \tag{1.25}$$

Berry (1994) demonstrates that for each (θ_d, Ω), there exists a unique $\delta(\theta_d,\Omega)$ for Equation 1.25 to hold. Given this restriction I utilize four sets of moments to identify the demand parameters.

The first set of moments are the score of a maximum simulated likelihood estimator using the travel survey data M. Unlike the following sets of moments, these are less

sensitive to simulation errors. Let $j(i)$ denote the work commute choice of individual i. From equation 1.2

$$L(\theta_d, \Omega; M) = \frac{1}{N_1} \sum_{i=1}^{N_1} w_i * \log q_i \left(j(i); \theta_d, \Omega \right)$$

where w_i is the sample weight for observation i. The score of the log likelihood function yields $|\{\theta_d, \Omega\}|$ moment conditions

$$E[\psi_1(\theta_{d0}, \Omega_0, M_i)] \equiv E\left[\frac{\partial L(\theta_{d0}, \Omega_0; M_i)}{\partial \theta_{d0}} \right] = 0 \tag{1.26}$$

with corresponding sample moment at arbitrary (θ_d, Ω), $\frac{1}{N_1} \sum_{i=1}^{N_1} \psi_1(\theta_d, \Omega, M_i)$.

The next set of moments match aggregate income statistics in the ACS by commuting choice, similar to the matching moments used in Petrin (2002). In its 5-year data the ACS reports a breakdown of transit choices by income. The sample statistics to match are read straight off the data and denoted by

$$\hat{q}_{l^i kj} \equiv \hat{q}_{l^i} \left(income_r \in [\aleph_k, \aleph_{k+1}] | r \text{ uses option } j \text{ for transit} \right)$$

for location l^i. Let the population analog be $\mu_{l^i kj}$. The corresponding statistics generated by aggregate model predictions are derived by the following equation.

$$q_{l^i kj}(income_r \in [\aleph_k, \aleph_{k+1}] | r \text{ uses option } j \text{ for transit}; \theta_d, \Omega) = \frac{\sum_{r=1}^{ns} \left[\mathbb{1}(income_r \in [\aleph_k, \aleph_{k+1}]) \int_{A_{rj}} dP(\varepsilon_r) \right]}{\sum_{r=1}^{ns} \int_{A_{rj}} dP(\varepsilon_r)}$$

for the simulated individuals r corresponding to tract l^i. The moment restriction imposes that at the true parameter the model's aggregated statistic and the population should be equal.

$$E[\psi_2(\theta_{d0}, \Omega_0, B_i)] \equiv \mu_{l^i kj} - q_{l^i kj}(income_r \in [\aleph_k, \aleph_{k+1}] | r \text{ uses option } j \text{ for transit}; \theta_{d0}, \Omega_0) = 0 \tag{1.27}$$

for all j, k and tracts.

The third set takes advantage of the much richer data for taxi and Uber usage, relative to the other transit options. The set roughly attempts to pin down the elements of Ω. The first method tracks unexpected disturbances in the subway lines — historic records are available for most Manhattan lines — that create delays (and hence wait times) for commuters who would typically take the subway.[31] For those days I measure the substitution toward taxis and Uber. Denote $\hat{\Delta}^e_{lj}$ the change in the share of $j \in Taxi, Uber$ from the change in wait time on the subway, that is an empirical measure of the cross wait elasticity without holding fixed all the other factors likely to have changed from the disturbance day e. Let $\Delta^e_{lj}(\theta_d, \Omega)$ be the same measure from the model replicating the observed characteristics from the event day. On average over all events E, these measures should match.

The second method utilizes day-to-day variation in the wait times between taxi and Uber. The procedure follows the same set up as the subway method. Here, I define an event — these are more uncommon than subway delays — as deviations from long-term average wait times for taxi cabs. I estimate an empirical measure of the cross wait elasticity as with the subway and force the model to match that on average.

$$E[\psi_3(\theta_{d0}, \Omega_0, K_i)] \equiv \frac{1}{|E|} \sum_{e=0}^{|E|} \hat{\Delta}^e_{lj} - \Delta^e_{lj}(\theta_d, \Omega) = 0 \tag{1.28}$$

The final set of moments are simply instruments on the wait time in the linear equation for $\delta(\theta_d, \Omega)$. I adapt instruments from Frechette et al. (2016) and measure average traffic speeds in rings one and two zones away from each location l at the specified time block t, denoted Z^t_l.[32] These traffic speeds should be correlated with wait times of both Uber and taxis but not otherwise affecting the choice. Hence,

$$E[\psi_4(\theta_{d0}, \Omega_0, B_i)] \equiv E[\xi^t_{lj}(\theta_{d0}, \Omega_0) Z^t_{lj}] = 0 \tag{1.29}$$

[31] Note also this time period is before disturbances were the exception not the rule.

[32] Relevant comments I received in seminars and hope to include in future iterations of this paper are better instruments to put here. First, as I argue in this very paper, traffic speeds impact demand. A better instrument would be unexpected changes to traffic speeds, like through accidents. Alternatively it is well documented how weather acts like a supply shifter in this market.

for all l, t and $j \in \{Taxi, Uber\}$ where $\xi_{lj}^t(\theta)$ is the value of ξ implied from the share inversion for a given value of θ.

GMM Estimator

The sets of moments can be stacked into a single vector. Formally, it is assumed that θ_{d0}, Ω_0 uniquely satisfies

$$E\left[\psi(\theta_{d0}, \Omega_0, M, B, K)\right] = E\begin{bmatrix}\psi_1(\theta_{d0}, \Omega_0, M)\\ \psi_2(\theta_{d0}, \Omega_0, B)\\ \psi_3(\theta_{d0}, \Omega_0, K)\\ \psi_4(\theta_{d0}, \Omega_0, B)\end{bmatrix} = 0 \tag{1.30}$$

with sample analog $\hat{\psi}(\theta_d, \Omega, M, B, K)$. The GMM estimate $(\hat{\theta}_d, \hat{\Omega})$ is the solution the following criterion function

$$\hat{\theta}_d, \hat{\Omega} = \arg\min_{\theta_d, \Omega} \hat{\psi}(\theta_d, \Omega)' W \hat{\psi}(\theta_d, \Omega)$$

where W is the weighting matrix derived via Hansen (1982). To calculate standard errors, I follow the adjustment on the standard formula offered by Goolsbee and Petrin (2004) to manage sampling errors in the observed market shares.

1.4.2 Supply

Data and Setting

For the supply estimation I focus on the area covered by the demand estimation, illustrated in Figure 1.12. While I could extend the coverage area, limiting myself and the cabs to these locations allows me to most fully take advantage of the demand derived from the first stage of the estimation. The day is played out over the course of 5-minute periods from 6am to 4pm Monday through Friday every month of March through June 2016. I assume all taxis play the same strategy for a time period in a given month. To account for the fact that Uber is still expanding over the time period, particularly outside of Manhattan, I re-solve for the relative demand of taxis and Uber using the

available price and wait data each month. Location-time-level unobserved preferences are assumed fixed for the entire period.

Extending the results from the demand model to apply here requires one additional assumption. The unobserved demand tastes $\xi_j^t(l)$ for each transportation type are only estimated for times covered by the demand model, that is up to noon. Since the greatest volatility in these values is across l rather than t within a fixed month, I substitute in the average unobserved taste for each location l beyond noon. In future robustness work, I can limit the supply analysis to the first half of the day at the expense of forcing taxis out of their shifts artificially early.

The second complication to address is the matter of total supply on the road for the various services. Yellow taxis follow the standard 5 to 5 work shift, the timing of which motivates the daily 6am to 4pm simulation. Based on advertisements for green cab shift partners on the website *nycitycab.com*, I assume green cabs largely follow the same shift pattern. The TLC also publishes records for the average green and yellow cabs on the road for each month, a record that I use as the assumed number of cabs under operation that month.

Uber drivers, on the other hand, have flexible shifts. The number of cabs on the road weekly is bounded on the high end by TLC reports indicating the unique vehicle dispatches from each of the FHV bases in New York, including the Uber bases. Considering the result in Hall and Krueger (2015) that as of 2014 nearly 42% of NYC UberX drivers drive less than 15 hours a week — barely a full shift — anything close to the upper bound as the number of Uber vehicles on the road during the day would be any overestimate. Instead I look to the pickup data itself to form a first guess at a bound on supply. I look for the maximal simultaneous (within 5 minutes) pickups on a day and average it for the month. Approximating the "consistent" supply on the road V_u, however, is built into the estimation procedure.

There are several components of the model I can directly construct using data. First $\gamma^t(l,k)$, the transition probability for time t I treat differently in this section than the last. Since the period of interest covers more than morning commutes, I need an approximation of destination preferences outside the time period. Relying on the fact that Uber and taxis are close substitutes, I estimate $\gamma^t(l,k)$ using *old* yellow and green taxi data from 2013 before Uber took a chunk out of their businesses. In areas outside

of Manhattan the precision of these estimates decline with the observed trips taken. However, as detailed in the model section, riders are not indifferent between Uber and taxi conditional on a destination. I use the current demand model to pin down their relative preference.

The other model parameters read from data are profits. Travel distances are approximated by observed taxi rides between locations, as are travel times. While travel times are determined in the model, for the estimation these travel times are estimated using actual travel times read from the taxi data. The cost of operation scales with distances traveled and by mile is the average monthly fuel cost against the average fuel economy of the taxi fleet. The final component for profits are surge prices for Uber. The estimation period is precisely the time for which I collected surge prices from the Uber app, March to June 2016. Finally, I assume $\bar{\tau}$ in Equation 1.14 is 16, based on the observation in data that no wait time for Uber was ever over 16 minutes. Finally, the commission rate for Uber, $\kappa = 0.238$, is based on the estimates reported in Castillo et al. (2017), which uses data from Uber in Manhattan.

Estimation

For a given vector of parameters $\theta_s = (\sigma_\varepsilon, \{\alpha_l^t\}, \{\tilde{Q}_l^t\}, V_u)$ I need to solve the equilibrium. Here, I follow the oblivious equilibrium procedure developed under Weintraub et al. (2008) to significantly simplify this procedure.

I assume the city is serviced by enough of every type of taxi to treat them as a continuum. Under this assumption transition probabilities become deterministic; the predicted transition paths throughout the day are thus also deterministic. The value of this insight from the oblivious equilibrium literature is that every driver need not keep track of or have beliefs on the behavior of all drivers when making their own search decisions. Instead each driver makes decisions conditional on industry averages.

Let $S_0 = \{S_0^t\}_t$ be the initial guess of the state with S_r the state for each iteration of the algorithm. Denote $m_{l,j}^t$ the matches of type j in location l. The only adaptation to the algorithm from Buchholz (2017) is to check for convergence of all taxi types and Uber.

To derive $\tilde{Q}$ for a guess of the parameters $\{\theta_s / \tilde{Q}_l^t\}$, I feed a guess to the Equilibrium Algorithm and use the distribution of taxis (not Uber) to invert $\tilde{Q}$ from Equation 1.8.

Algorithm 1: Equilibrium Algorithm

Set $r = 0$ and $\tilde{Q} = \tilde{Q}_0$ generated by the fixed point algorithm described
Use the guess of S_r^T to calculate $V_j^T(S_r^T, \tilde{Q}^T)$ for all $j \in \{x, g, u\}$
for $t = T - 1$ *to* 1 **do**
    Guess S_r^t to calculate $V_j^t(S_r^t, \tilde{Q}^t)$ $\forall j$
    for *t=1 to T-1* **do**
        Derive $\sigma_r(k|S_r^t)$ for all t, j, l
        Realized transitions and σ yield the actual state $\tilde{S}^{t+1}$
        Update the state S_{r+1}^{t+1} for the next period using $\tilde{S}^{t+1}$
        Update continuation values $V_j^{t+1}(S_{r+1}^{t+1})$
        Set $r = r + 1$
    end
end
while $|V_{j,r}^t - V_{j,r-1}^t| > \varepsilon \quad \forall t, j$ **do**
    Repeat 4 to 11
end

I repeat the process to convergence.

For the other parameters $\{\theta_s / \tilde{Q}_l^t\}$ I use simulated method of moments. I match the average vacancy time for cabs, the utilization rate, and wait times. Uber wait times and the average utilization rate taken from Cramer and Krueger (2016) address the scale of Uber. The more vehicles on the road, the lower the utilization rate, for a given level of rides. But, clearly wait times also change as a function of the number of Uber drivers on the road. Since demand in this portion of the estimation is not responsive to changes in wait times approximated by the model (but rather the wait times in data), more Uber cabs should unambiguously decrease wait times.

In Section 1.3 I detailed how matching probabilities are linked to wait times. Because these wait times are estimated in terms of periods, I convert it to a continuous value by assuming that the probability of being matched within any five-minute period t is uniform conditional on being matched in that period. Matching taxi wait times pins down α_l^t. To estimate the function for α_l^t, I assume a simple form for Equation 1.9

$$\alpha_l^t = \alpha_l + \theta^{t'} \alpha_{tod} + \alpha_{dens} m_{l,b}^t + \varepsilon_l^t \tag{1.31}$$

where t indexes the time and θ^t is a vector of time-of-day dummies. Additionally, the number of matches are adjusted for the street access area of the location. β_{dens} should capture the relationship of transactional density with matching efficiency, while α_l is a location fixed effect that may pick up factors like street layout. The nature of the market helps identify this density term; near the green exclusion border there is an artificial glut of supply from the green taxis. Figure 1.13 demonstrates a large drop off in the share of Uber pickups around the green exclusion border. Assuming conditions are similar around the border, this hints at a larger available supply from competitors, in particular green and yellow taxis.

Finally, I estimate the congestion function from Equation 1.18 after assuming the following form

$$\tau_{lk}^t = \beta_{lk} + \theta^{t'}\beta_{tod} + \beta_l traf_l^t + \beta_k traf_k^t + \varepsilon_{lk}^t$$

where $traf_l^t = (v_{l,j}^t + e_{l,j}^t)$ for $j \in \{x, g, u\}$. Each route between proximal zones can have a separate intercept, which I estimate as a fixed effect.

1.5 Results and Discussion

1.5.1 Model Estimates

The parameter estimates from the model appear in several tables and figures in the Appendix. Tables 1.4 and 1.5 present the result of the demand-side estimation. Table 1.6 presents the result for the estimated volume of Uber cabs, the degree of uncertainty σ_ε governing the location decision of taxi and Uber drivers, and the estimate of transaction density on efficiency α_{dens}.

A few validating and interesting facts emerge from the demand estimates alone. First, price elasticities, as conventional wisdom would posit, diminish with income. This fairly general pattern is reflected in the estimates for α_g in Table 1.4. Additionally, consumers do not value all time equally. Consumers penalize waiting time much more significantly than travel time, a finding reflected in transportation literature. Finally, the off-diagonal components in Table 1.5 reflect inherent substitutability of different transit choices. Expectedly, Uber and taxis have the highest degree of substitutability even controlling for the fact that the quality of service they offer, in terms of price and

wait time, is already quite similar in most areas of the city.

Figure 1.14 is another attempt at validating the results from the demand estimation. The figure maps the average taxi wait time elasticity in each taxi zone covered by the estimation. Quite sensibly, the areas with better substitutes have the highest elasticities along the dimension of wait time . This pattern also appears to be reflected down the central axis of Manhattan where most of the subway lines run. This figure, in part, previews one part of the value of Uber in different parts of the city. In Manhattan consumers can easily substitute away from Uber were it to disappear. In other areas of the city, where Uber has a larger share of traffic, we would expect these consumers to be hit by both a much sharper drop in the market quality (along wait time) and also have a lower wait elasticity.

The final supply results are presented in Figure 1.15 and should be taken along with the result that $\alpha_{dens} = .0048$. These figure maps the residual parameters (α_l) from Equation 1.31 over building to ground density with values normalized by the maximum. Despite controlling for transactional density, I still find $corr(\alpha_l, density_l) > 0$.[33] The residual values range from 0.34 to 1.32. To get a sense of their meaning fix the transactional density at 0, and suppose an area had 100 potential and 100 potential cabs. In the location with $\alpha_l = 0.34$, we would expect 29 matches. In the location with $\alpha_l = 1.32$, we would predict only 73.5 matches.

To account for the role of transactional density in efficiency via α_{dens}, consider the zone with the most transactions per unit area per 5 minutes. Appropriately, this zone contains Penn Station, which has anywhere from 50 to over 250 pickups per 5 minute period per square mile over the course of the day. α_l for this zone is 0.92. Absent any transactional density, the matching function would predict 60 matches with 100 cabs and 100 consumers. Ignoring time of day effects, the predicted matches accounting for the transactional density ranges from 68 to 88. Hence the transactional density alone improves efficiency in a particular area by 20%. All other zones, by definition, do not receive close to the same boost in efficiency from transactional density.

[33] Earlier versions of the paper in which I set $\alpha_l^t = \alpha_l$ exhibited a much stronger correlation.

1.5.2 Welfare Analysis and Counterfactuals

The supply results alone in Figure 1.15 combined with the impact of transactional density present a stark estimate of the relative advantage Uber's technology might have over taxis. This section attempts to determine the significance of this relative advantage from the perspective of consumers across different parts of the market. I then tease out the importance of Uber's different matching regime in servicing the city through a series of counterfactuals that change the regulatory burden on yellow taxis. Finally, I consider a set of regulations geared toward leveling the playing field between taxis and Uber through geographic restrictions and congestion taxes. Table 1.7 summarizes the spatially heterogenous results of the counterfactuals separated out by density quantiles.

First, to avoid having all welfare results depend too heavily on the supply model, I carry out a standard welfare analysis dependent solely on the estimated demand model. Following the tactic of Petrin (2002), I treat Uber as a new product category introduced to the market by the period of analysis in March to June 2016. I compare the welfare of consumers choosing Uber to their alternative in a baseline year 2013, when I assume Uber is still an small platform in the market. Technically, Uber entered New York in May 2011, but, as I argued in Section 2, the scope of the operation even as late as early 2014 was negligible compared to yellow taxis. Finally, I break down the compensating variation per ride for each of the 143 zones in the analysis.

In extrema, these compensating variation values range from $1.00 per ride in eastern Queens, where it is approximately 10% of revenue from the ride, to $.10 per ride in central Manhattan, where it is approximately 2% of revenue from the ride. Figure 1.16 illustrates the range of results for all the zones in the study over a map of the city. Figure 1.17 depicts the same information smoothed out over the measure of geographic density. In total the change in consumer surplus per day for these riders works out to be $73,000 for an average total of 120,000 transactions per day, a factor of 10 difference from the estimates in Cohen et al. (2016).[34] In Table 1.7, the results from the exercise

[34] I cannot directly assess the contrast in these findings. One theoretical difference, however, is in the demand model used to generate price elasticities. My paper estimates demand on longer term transit choice decisions, which should yield consumers with higher price elasticity than in Cohen et al. (2016). Higher price elasticity would also translate to a lower consumer surplus. Second, this work demonstrates that where *in a city* consumers are sampled can impact one's estimate of the value of Uber (or any new transit option).

are recorded under "C1".

Taxi Entry Policy Counterfactuals

The counterfactual exercises proceed in a series of steps to understand the importance of Uber's technology itself, versus its relatively high level of supply, in providing service to different markets in the city. For every counterfactual I start from the previous baseline and "shock" the system with the particular regulatory adjustment for that counterfactual. Unlike in the supply estimation, demand and supply interact with each other in the determination of the new equilibrium. I handle this problem through the following algorithm.

Algorithm 2: Counterfactual Algorithm

1. Shock the system with a regulatory change, e.g. eliminating Uber
2. Holding fixed the demand system, allow cabs to re-optimize their location decisions
3. Re-calculate the wait times and congestion based on these decisions
4. Feed the wait times and travel times back into the demand system and allow consumers to re-optimize
5. Repeat 2-4 until convergence

Three remaining issues are unaccounted for in these counterfactuals. The first is the potential evolution of unobserved tastes for taxis once the regulatory change has been applied. I simply hold these unobserved tastes fixed at the average daily level per taxi zone as of June 2016. The second issue is that I have no proof guaranteeing convergence. While computationally that has not been issue at the current parameter values, I intend to address this potential problem in the future. Finally, a multiplicity of equilibria are a near certainty in this setting because of the feedback from demand to supply and vice versa. I am only presenting one potential new equilibrium. Future robustness work may require testing the sensitivity of the equilibria.

The first of these counterfactuals ("C2" in Table 1.7) is meant to replicate the results of the previous welfare analysis conducted, along with simulating the now unlikely possibility of banning Uber in New York as London has done. I ban Uber as of March 2016 and re-solve the model. The welfare changes here approximate those from the demand-based analysis and serve as a sanity check for using the supply model in further

counterfactuals. Figure 1.18 maps the relative difference in the calculations across the zones in the city. While the qualitative pattern of the welfare patterns is somewhat different than the in the baseline welfare calculation, the quartile averages across density groups are fairly similar. The results do suggest, however, that modeled congestion, as a countervailing force against taxis congregating in the densest areas of the city, may be *too* strong. Compared to models without congestion in earlier versions of the paper, the calculated compensating variation in Manhattan is higher and likewise lower outside Manhattan. Looking at the change in rides, the percentage declines (or in the best case, growth) in Manhattan are worse than before. Congestion pushes taxi cabs out of attractive dense locations, and, in this case, that effect could be over shooting the real impact.[35]

The final counterfactual ("C3" in Table 1.7) in this section asks how many yellow taxis would be needed to at least match the service quality (gauged by wait time) consumers have in the initial data. I again carry out this counterfactual by adding a stock of yellow taxis weighted by the initial distribution of cabs from the previous counterfactual. Because of network externalities there are infinite possible equilibria resulting from this regulatory change. I choose to handle it by adding an additional 500 taxis, allow the model to reach an equilibrium, and I then add the next 500. Ultimately, I find New York City would require 32,000 additional yellow taxis; note that this quantity is greater than the 21,356 Uber cabs previously servicing the market.[36] The result is born from two factors. First, enough yellow taxis must enter the densest markets to make the less dense markets relatively more appealing. Additionally, because of yellow taxis' technological disadvantage in the outer boroughs, more are required to achieve the same level of service quality from the baseline number of Uber cabs.[37] Finally, note that this analysis makes no claim about the profitability of these routes. Depending on the outside options for these drivers, the market may never achieve this level service even were the regulatory cap on yellow taxis to completely disappear.

[35] In earlier versions of the paper, not accounting for congestion, the difference was larger across quartiles. The convergence across the specifications is encouraging evidence of the supply framework. On the other hand the qualitative patterns of benefits diverge more than in previous models.

[36] Without congestion this estimate approached 36,500.

[37] A corollary counterfactual in the works is redoing this problem but adding only "green" taxis. Green taxis eliminate the first factor driving up this number; that is their service starts in less dense markets.

1.5.3 Counterfactuals on Uber Restrictions and a Congestion Tax

For the counterfactuals in this section, I consider a set of less draconian policies geared toward evening the playing field between Uber and taxis. As I discussed in Section 1.2 Uber's unfair regulatory advantage stems from being able to pick up passengers in Manhattan at no additional cost. The old technology requires a high fee to pick up in these areas through the medallion system. Hence, in the section I simulate two counterfactuals: one in which Uber is banned from picking up in the green exclusion zone. The other implements a tax on pickups for Uber for every pickup in the green exclusion zone.

The simulation of these counterfactuals is slightly different than for the previous set. Obviously, both Uber and taxis are searching in the city simultaneously. Because Uber continues to operate, I need an additional assumption on counterfactual surge prices. For now, I assume that the surge pricing scheme stays constant. Future iterations will estimate a surge price function much like the for congestion in Equation 1.18.

In the first counterfactual Uber is banned from picking up in the green exclusion zone. Because the goal of this exercise is to assess how much Uber relies on Manhattan to "subsidize" its operations in the outer boroughs, I introduce an additional profit condition in the algorithm, described below. The welfare results of the simulation are

Algorithm 3: CF Algorithm Banning Uber from Manhattan

```
while the current average profit of Uber drivers is less than the original average profit do
    Remove 100 Uber drivers
    Shock the system with the regulatory change
    Holding fixed the demand system, allow cabs and Uber to re-optimize their location decisions
    Re-calculate the wait times and congestion based on these decisions
    Feed the wait times and travel times back into the demand system and allow consumers to re-optimize
end
```

reported by density quantile in Table 1.7 as "C4". A naive baseline for the findings would be to claim that 50% of the Uber drivers would exit the market after the implementation of the Manhattan ban, since, as of June 2016, about half of Uber pickups

are in Manhattan. The ban has a further effect, however, as not being able to pick up in Manhattan makes traveling from Bronx to the other boroughs, or vice versa, more expensive in expectation to Uber drivers because of the higher idle time. As a result, Queens and Brooklyn actually enjoy a higher quantity of Uber while Bronx disproportionately suffers. In the end 65% of Uber drivers exit the market. While the result does suggest a disproportionate revenue value from access to the dense Manhattan market, part of the result is mixed in with the value of Manhattan as a low-idle time bridge between the boroughs.[38] More importantly, the high cost to Manhattan consumers — the welfare cost mirrors that in the C2 — suggests there might be room for a tax not born entirely by Uber drivers, as I propose in the final counterfactual.

The final counterfactual implements a tax on Uber for pickups in the green exclusion zone. For a given tax level, I calculate the new equilibrium as in the previous algorithm. I adjust the tax searching over a discrete grid of $0.10 increments starting from $1.00. For simplicity I assume the tax incidence fall on the drivers, that is Uber does not adjust its prices. I continue to raise the tax until the shift revenue of yellow taxis matches the level from June 2013, an approximately $100 difference.[39] The welfare results are reported in Table 1.7 as "C5." Ultimately, the algorithm settles on a tax of $2.50. Like in the previous counterfactual the areas hurt most by the policy are those in the outer areas of the city, even though the tax does not target them. The tax, however, functions similar to a ban in that the revenue value to Uber drivers from access to Manhattan is, by design, dampened.

1.5.4 Concluding Comment

Underpinning all these results is the relative advantage and disadvantage of a hailing taxi's technology to Uber's dispatching system across markets of varying density. Taken together they offer an answer to the question underlying this paper: has Uber's success been born from "technology" or its ability to field supply in a way traditional taxis are often prohibited from doing? In New York City, the answer is clear; it depends on the

[38] Another valuable counterfactual would be to allow Uber drivers to pay a fee for access to Manhattan. This set up parallels the congestion tax in the final counterfactual and also more cleanly assesses the value of this Manhattan subsidy.

[39] An alternative planner might adjust the tax with actual concern for congestion. In the future I could run the same procedure to have traffic speeds reach parity with June 2013 levels. I find average difference of around 3mph.

density of the market. In New York's densest markets, Uber offers little if any advantage with its technology; unobservable consumer preference — which may be indeed be a knowable characteristic Uber manipulates but I cannot measure — drives consumers to its platform. Yellow taxi service could itself yield the same quality rides. Elsewhere, the old hailing technology is simply insufficient to service the market at the same level of efficiency as Uber or similar services; the counterfactual illustrating the much higher quantity of yellow taxis required to match the baseline service demonstrates this result.

The second set of counterfactuals highlight a separate result about the benefits from Uber. First in cities of densities similar to the outer boroughs, assuming all other transit options the same, the benefit from Uber would likely be lower. These cities would not benefit from the high revenue value rooted in denser areas. On the other hand, Uber has the ability to adjust the costs to drivers in each of these markets. Extending these results to other markets thus remains an open problem. The corollary is that the clear cut case of business stealing I presented in Section 1.2 is less evident at the level of the whole market. The rents extracted from old taxi drivers in Manhattan are, in part, transferred to consumers in the outer boroughs.

1.6 Conclusion

Uber is the face of a sea change in the transportation industry. Uber's incredible inroads into the historically stagnant taxi market serves as decent evidence of this perspective. How beneficial the new technology in matching consumers Uber has brought to the market, however, is an open question. On the surface the ability to hail a taxi from a phone is but an incremental change to traditional dispatch services. This paper proposes that the size of the benefit offered by Uber's technology is a function of the density of the market it is serving. In the densest of markets hail taxi services, which match consumers to drivers through physical contact, can actually generate lower waiting times for consumers than Uber's dispatching program. Using New York City as the context, I find that the introduction of Uber to the market has consumer welfare benefits that vary by a factor of ten from the most dense to least dense locations studied.

To study the development of the market I use publicly available trip-level data on the pickups of taxis and for-hire vehicles like Uber. These rich records permit a study

of the New York City market over both space and time. I augment this dataset with two unique sources on consumer wait times for Uber and taxis. In the first I scraped the Uber app on a simulated Android phone to collect wait time and surge price data for that service in 47 locations across the city at different times of day. For taxis I follow Frechette et al. (2016) in using the pickup data to estimate a measure of the time consumers wait for taxis. The data allow me to estimate a discrete choice of model of demand for multiple types of transportation services in the city and imbed it in a spatial equilibrium supply model in which Uber and taxis simultaneously. Controlling for spatial heterogeneity in demand for Uber and taxis and the quality of alternative transit options is critical in separating out the effects of market density from other conflating factors.

The relative technological advantage of Uber in less dense areas manifests itself in several ways. First, the estimated consumer surplus per ride from Uber in the least dense areas of the city outweigh those in the most dense by a factor of ten. In terms of revenue from this rides, the consumer surplus ranges 2% to 10%. Second, if indeed Uber had no technological advantage over taxis, one could remove the supply cap from yellow and green taxis and acquire a similar level of service, as measured by wait times. I explore this possibility in two counterfactuals easing the supply cap on yellow cabs in the absence of Uber. The results from both suggest that a much higher volume of yellow taxi capital would be required in the market to provide the same level of service to the outer boroughs, the less dense parts of NYC, as Uber had as of June 2016.

The paper also opens the door to several more specific questions about Uber but also the innovations in the transportation market at large. Notably, this paper does not account for other technological differences between Uber and alternative services including other app-based competitors. For example, how much does the quick drop-in, drop-out system of supply specifically contribute to the welfare generated by Uber? The advantage from this system may exist independent of density. This paper also skirts around the critical question of the magnitude of scale economies in these markets, as demonstrated by Bian (2017). With more reliable data on the supply side for Uber, the framework developed in this paper could easily be extended to reliably answer that question. How easy it is to achieve scale is relevant to further understanding the nature of free entry in this market. Will new entrants eventually drive platform profits to zero?

Alternatively Uber's early expansion through ample financial capital may have been its true advantage and given the company an insurmountable advantage.

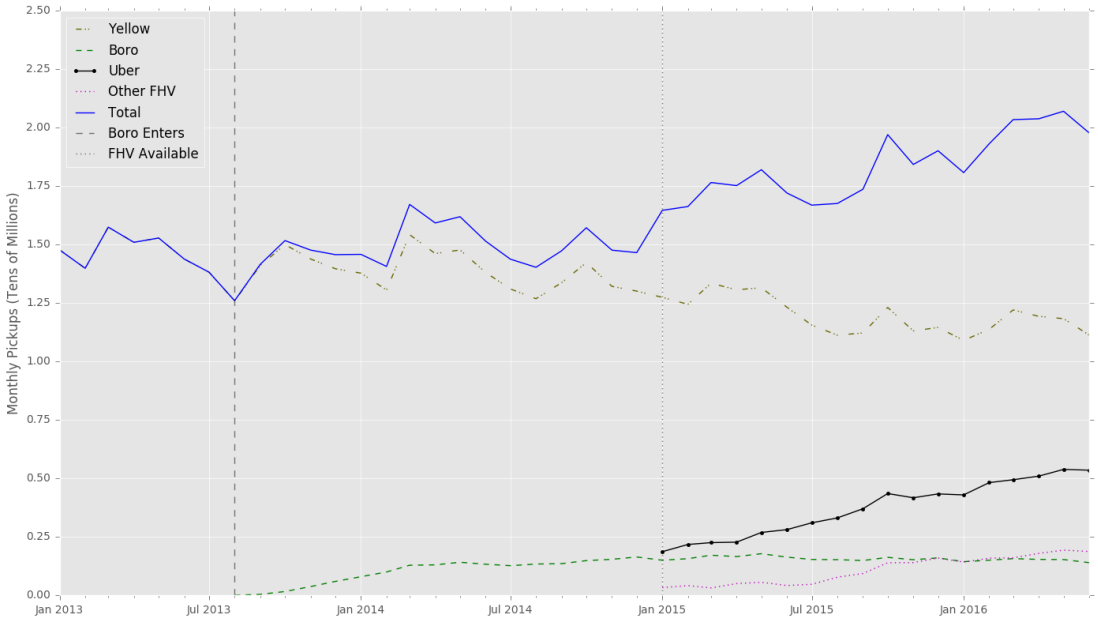
Using the framework developed in this paper, however, a number of additional applications are already under development. The first was accounting for the impact of supply growth on congestion. New York has already begun considering implementing a congestion tax to deal with perceived problems of traffic growth from ride-sharing services in the city. This work includes preliminary results incorporating congestion into service quality through transit times and assessing the impact of such a policy. The second extension is modeling the e-hail technology in use by an increasing number of taxis. In Israel, for example, Gett is exclusively an app that links consumers to existing taxis. As the use of this technology grows, the differences between taxis and Uber diminish, but the latter still has the principal advantage of being a flexible and centralized platform. In the context of what this paper can do, I can isolate the role of surge pricing, a feature taxis could not replicate, in the service quality provided by Uber.

1.7 Figures

Figure 1.1: **Exclusion Zones for Taxi Service in NYC**

Note: Divisions in the map are different taxi zones, the unit of analysis for most of the paper. The exclusion zone for Green cabs is the hatched area over Manhattan. Green cabs additionally cannot serve as "hail" cabs at JFK and LaGuardia. Pickups at these locations must be pre-arranged. In the non-exclusionary zone — save the airports — green cabs can pick up street hails.

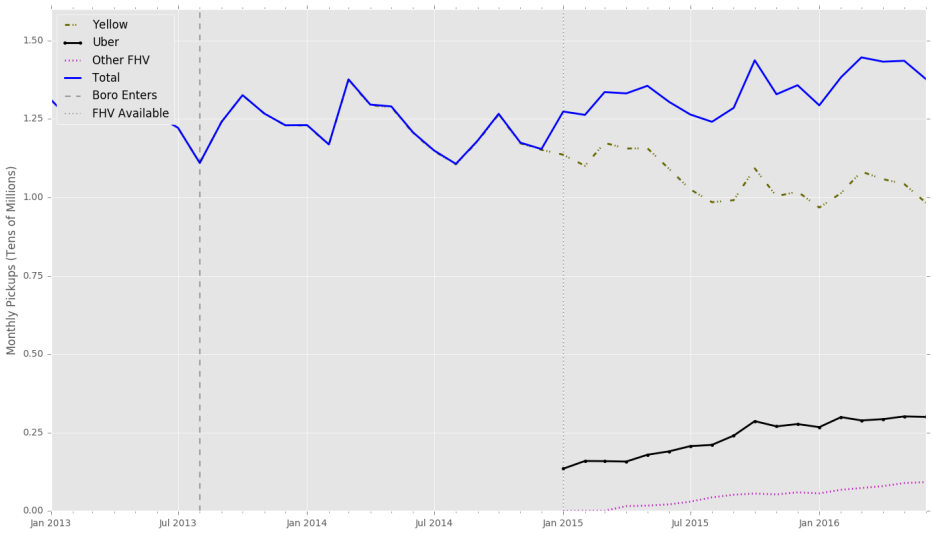
Figure 1.2: **Monthly Pickups by Cab Type**



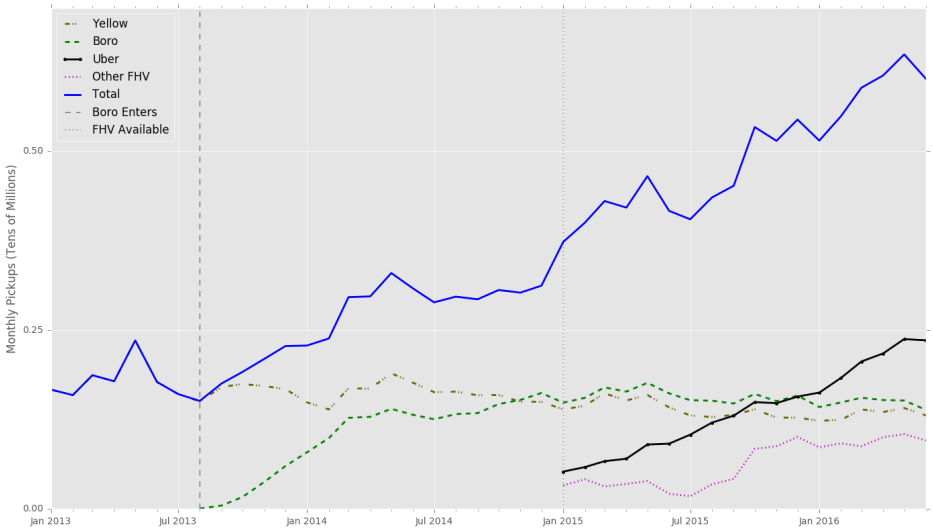
Note: The two vertical lines denote, respectively, the introduction of green taxis in August 2013 and the availability of FHV data in January 2015. Lyft and Uber entered the NYC market before January 2015 but FHV data are only available continuously at this date. Therefore, the total pickups are measured consistently before and after January 2015 but not across that date.

Figure 1.3: **Monthly Pickups by Cab Type and Zone**

(a) Pickups in the Green Cab Exclusion Zone



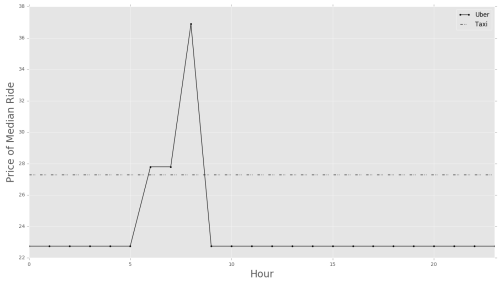
(b) Pickups Outside the Green Cab Exclusion Zone



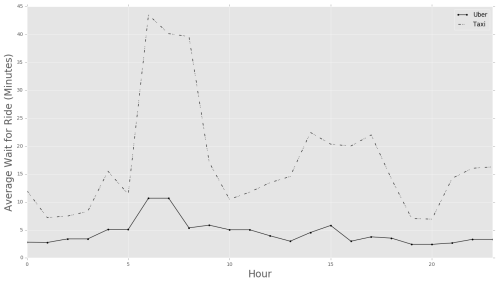
Note: The two vertical lines denote, respectively, the introduction of green taxis in August 2013 and the availability of FHV data in January 2015. Lyft and Uber entered the NYC market before January 2015 but FHV data are only available continuously at this date. Therefore, the total pickups are measured consistently before and after January 2015 but not across that date.

Figure 1.4: **Comparison of Uber and Taxi Price and Wait Times**

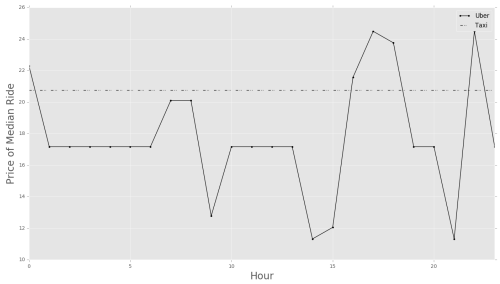
(a) Price, Williamsburg to Times Square



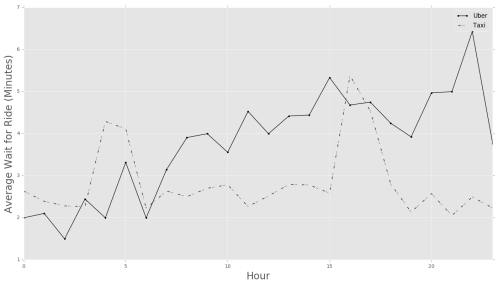
(b) Wait, Williamsburg



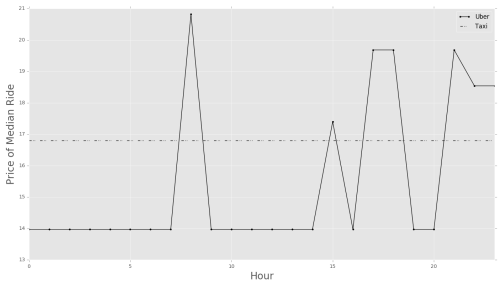
(c) Price, Times Square to JFK



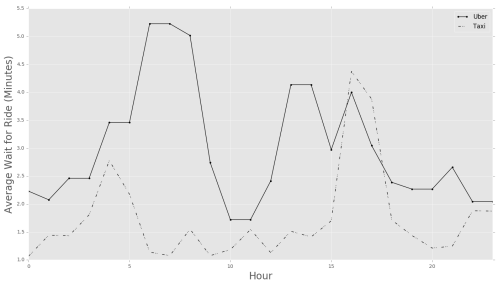
(d) Wait, Times Square



(e) Price, Penn Station to JFK



(f) Wait, Penn Station



Note: The image contrasts prices and wait times for Uber (solid line) and taxis (dotted line) in different areas of the city. Times Square and Penn Station are stand ins for "dense" parts of the city. Williamsburg represents a less dense part of the city. Taxi prices are read from data. Uber prices are estimated using the recorded surge and the median trip time and distance for rides at that time of day. Taxi wait times are estimated via the process described in the text. Uber wait times are the average wait recorded from weekday trips at that time.

Figure 1.5: **Distribution of the Relative Price of Uber and Taxi, 2015**

(a) Rides from Lincoln Square East

(b) Rides from Midtown Center

(c) Rides from Union Square

(d) Rides from the Upper East Side South

Note: Each graph presents the frequency distribution of the ratio of Uber prices to taxi prices for trips originating in the specified location. A dotted line appears at 1.0 where the two prices are equal. The mass of the plot above 1.0 are trips which would have been more expensive to take with Uber than with taxi. For each location Uber prices are estimated using the realized trip distance and time for rides in the taxi trip records. The surge price profile uses averages calculated for each weekday and hour, but from 2016. These estimated prices are compared to total fares, including tip, from realized trips in the taxi trip records.

Figure 1.6: **Distribution of the Relative Price of Uber and Taxi, 2016**

(a) Rides from Lincoln Square East

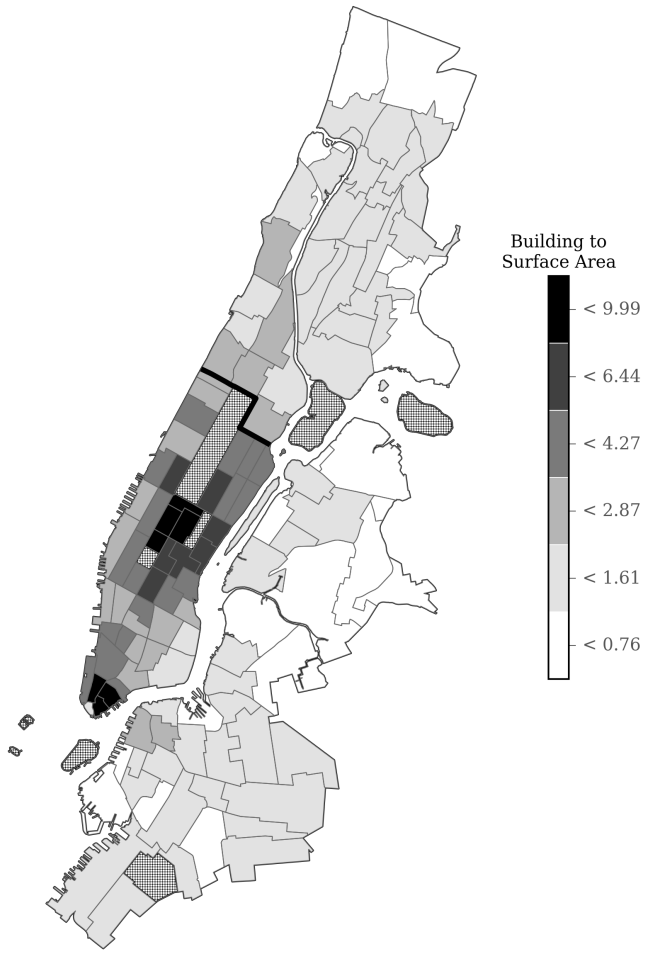
(b) Rides from Midtown Center

(c) Rides from Union Square

(d) Rides from the Upper East Side South

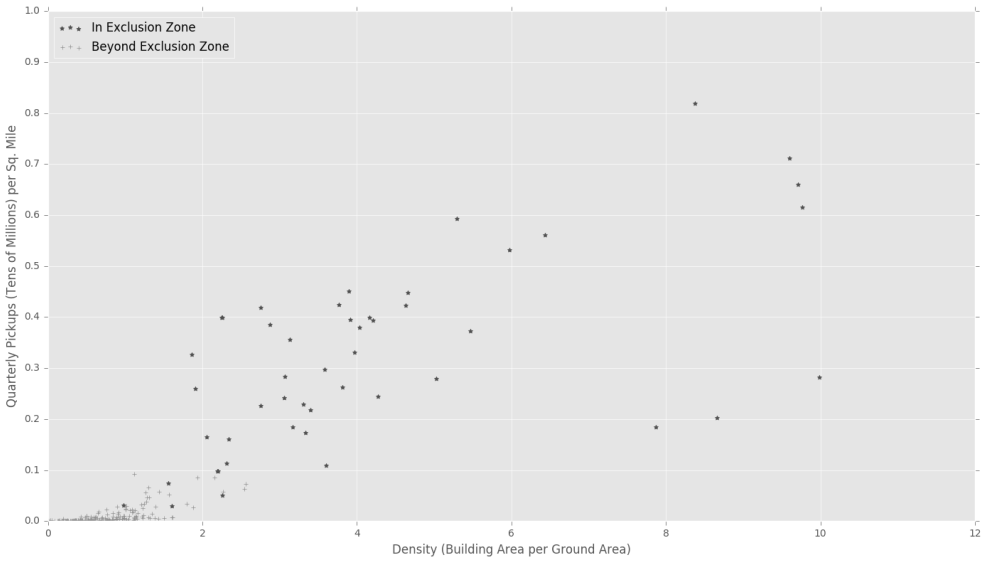
Note: Each graph presents the frequency distribution of the ratio of Uber prices to taxi prices for trips originating in the specified location. A dotted line appears at 1.0 where the two prices are equal. The mass of the plot above 1.0 are trips which would have been more expensive to take with Uber than with taxi. For each location Uber prices are estimated using the realized trip distance and time for rides in the taxi trip records. The surge price profile uses averages calculated for each weekday and hour, but from 2016. These estimated prices are compared to total fares, including tip, from realized trips in the taxi trip records.

Figure 1.7: **Geographic Density Measure over NYC**



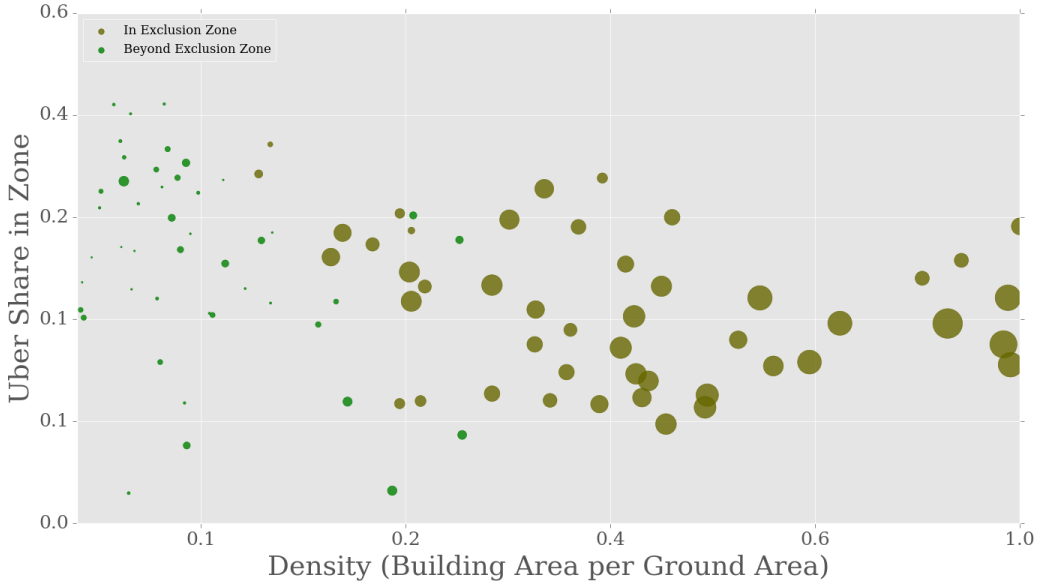
Note: Divisions in the map are different taxi zones, the unit of analysis for most of the paper. The boundary or the exclusion zone is denoted by a thick black line in northern Manhattan, south of which green cabs cannot pick up passengers. Density is measured by the total building area over the ground area in the particular zone. Checkered areas are excluded from analysis in the spatial auto-regressive models in Appendix A.2.1. *Source: NYC Planning PLUTO*

Figure 1.8: **Total Pickups by Taxi Zone, Q1 2016**



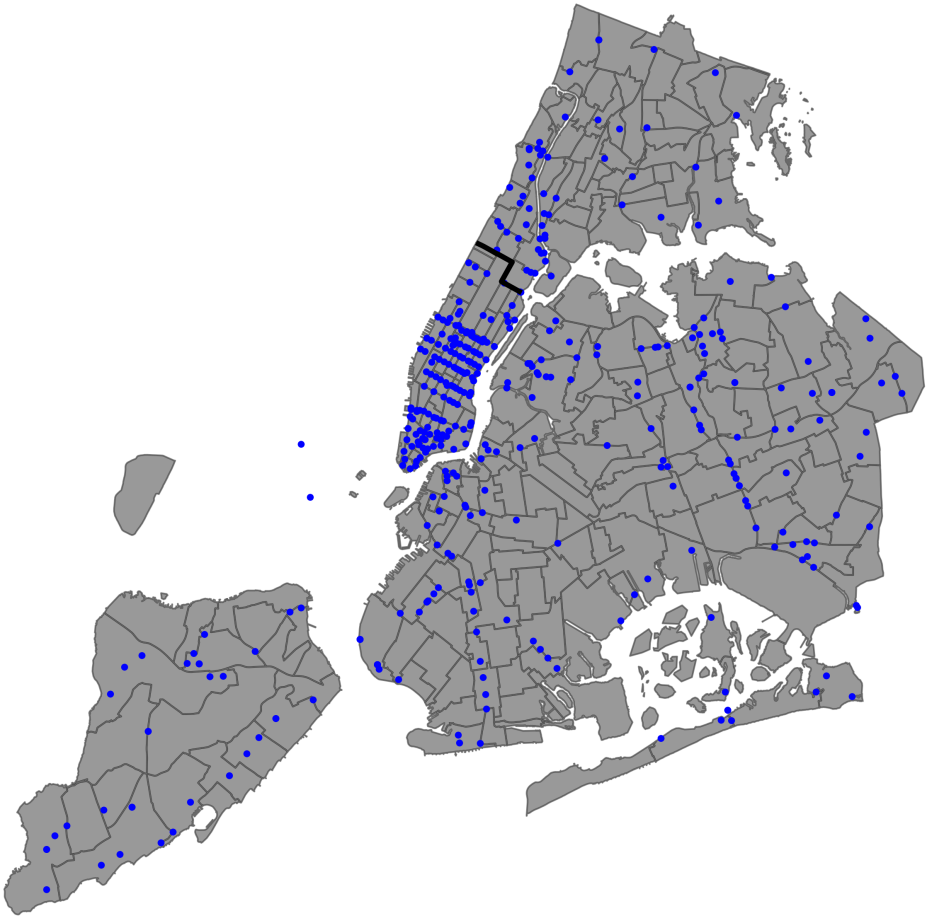
Notes: Observations are organized by density measured as total building area over ground area. Pickups are normalized to pickups per square mile in the zone to account for significant variation in taxi zone size.

Figure 1.9: **Uber Share of Transit Market by Density, Q2 2016**



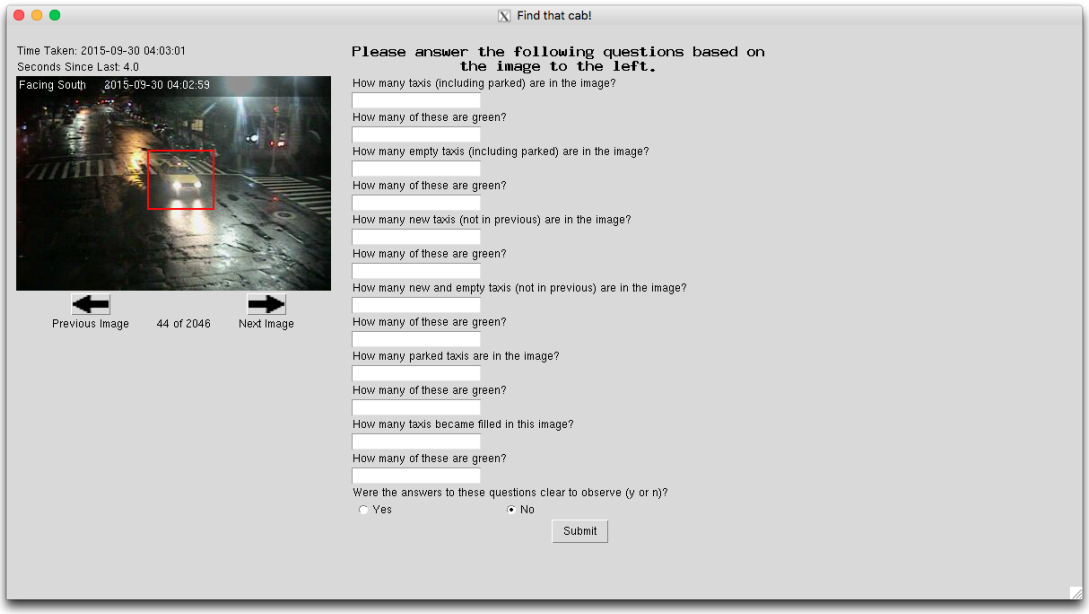
Note: This is a log-log graph of share on density normalized between 0 and 1. Observations are organized by density measured as total building area over ground area at the level of a taxi zone. Share is measured as the share of pickups Uber made over all FHV, yellow taxi, and green taxi trips in the same area for the second quarter of 2016. Point sizes reflect the relative total volume of rides in the period of interest.

Figure 1.10: **Traffic Camera Locations**



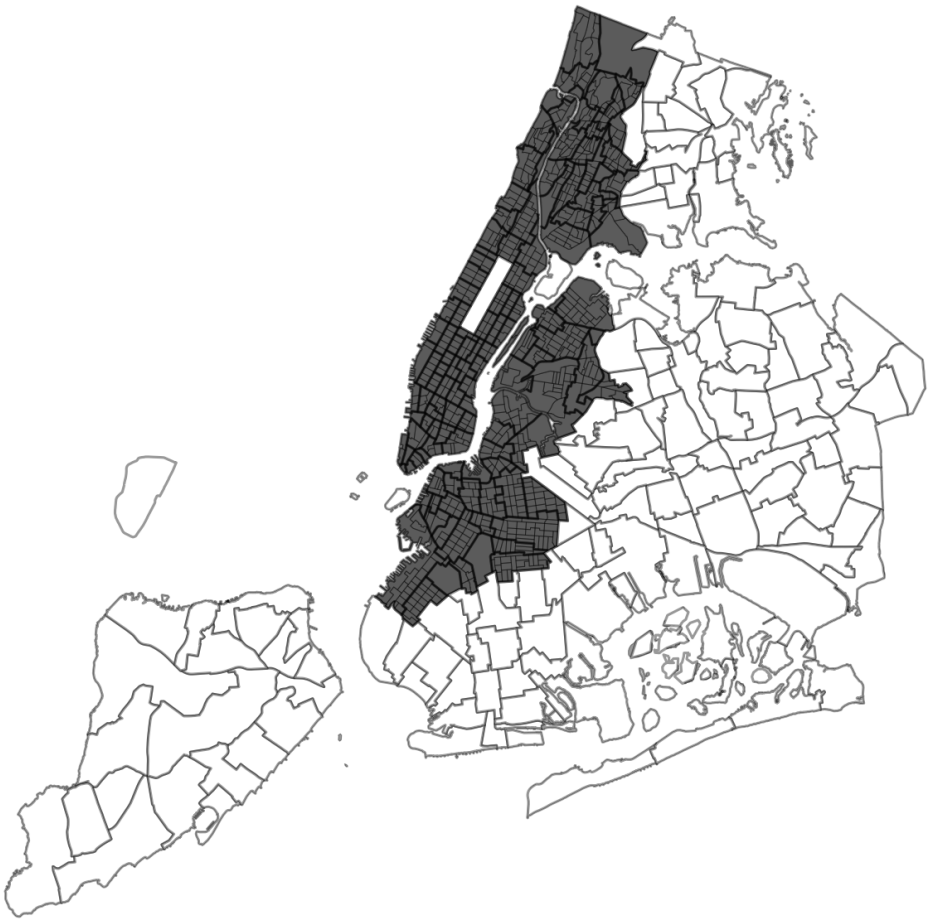
Note: Each blue dot is the location of a traffic camera from which I scraped images in September and December 2015. In this version of the paper not all locations were processed at all times because of manual constraints. Additionally, camera locations in the outer boroughs tend to overlook highways rather than local streets.

Figure 1.11: **Sample of Traffic Camera Image Processing**



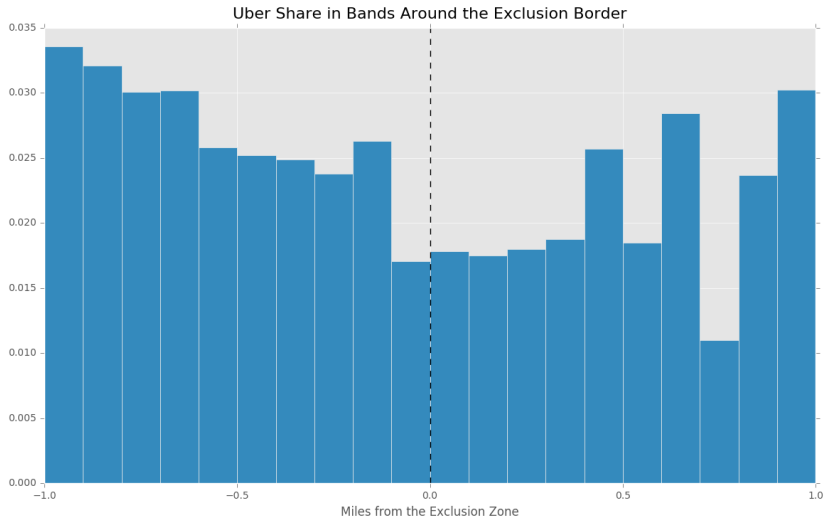
Note: This image is a screenshot of the program developed to process the scraped feeds from New York City traffic cameras. Taxi vacancy is determined by the taxi light atop the cab; hence, processing is more prone to errors in daytime. The red box surrounds the taxi in the image, but it is not part of the original program.

Figure 1.12: **Zones Included in the Estimation**



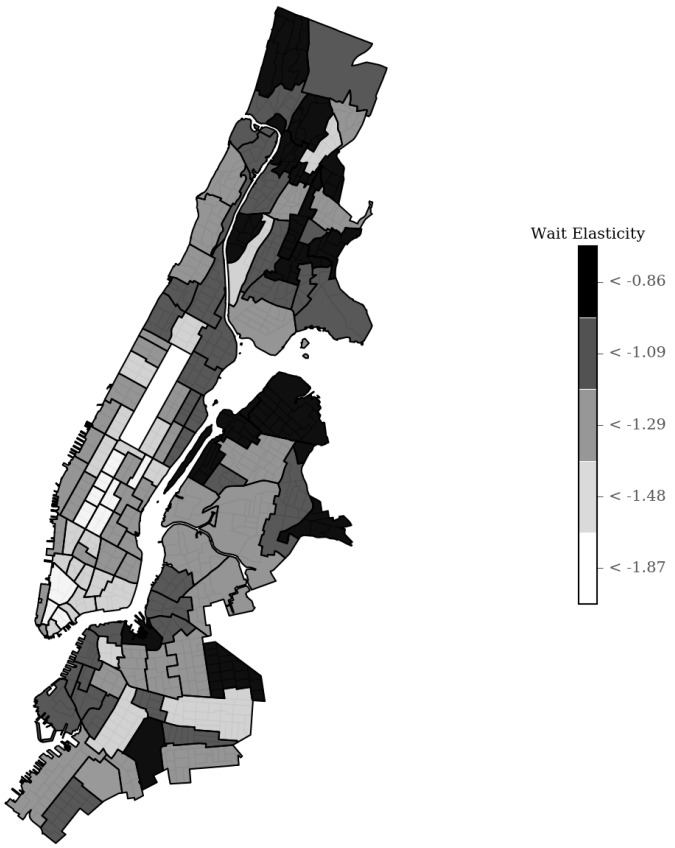
Note: Zones explicitly modeled in the demand analysis are shaded gray. Census tracts are outlined in dimmed gray while taxi zones are separated by dark gray.

Figure 1.13: **Uber's Share of Pickups Around the Exclusion Border, Q3 2014**



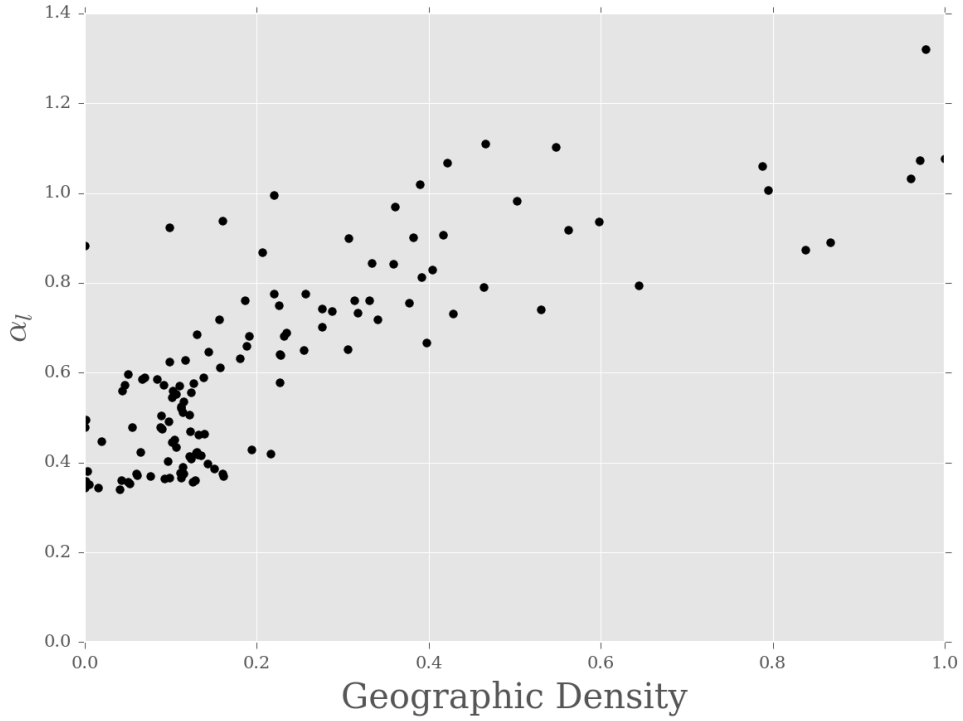
Note: Shares are initially calculated by census tract; data from an earlier period when Uber pickups are available by census tract is used for this figure only. Each tract identified by its distance from the green exclusion border based on its centroid. Shares in each bin are then weighted by the total pickups. Pick up assignment issues at the border precisely likely yield this strange drop off in the bin right before the exclusion zone. It is also possible green taxis are not heavily penalized for infinitesimal infractions.

Figure 1.14: **Average Estimated Taxi Wait Time Elasticity by Area**



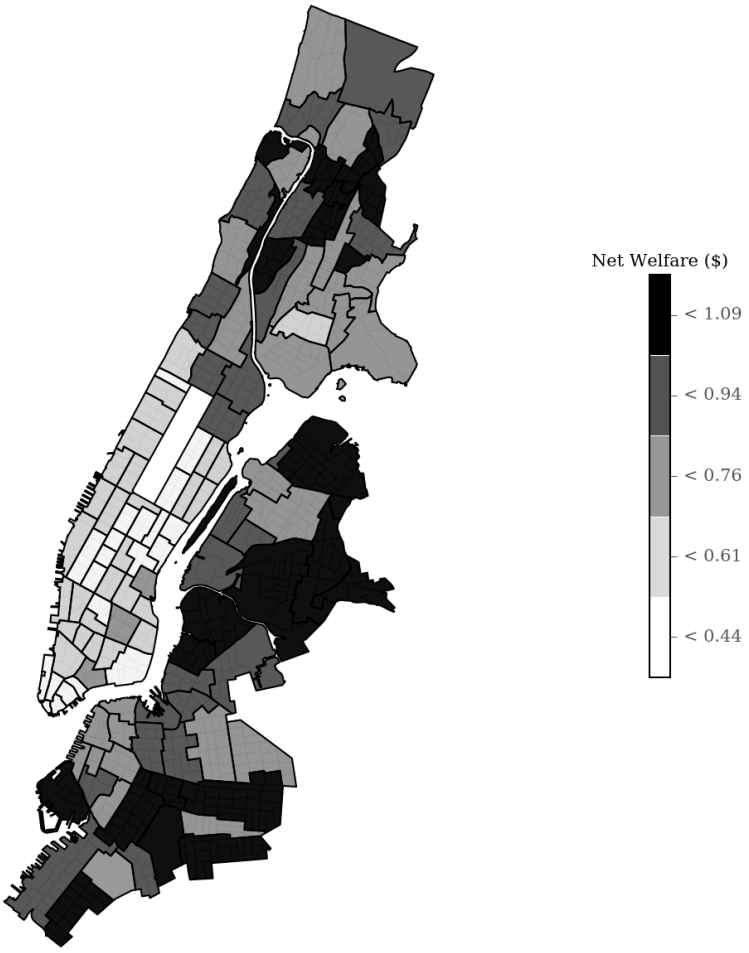
Note: The map measures the average taxi wait time elasticity for consumers included in the demand analysis using the variables' values in June 2016.

Figure 1.15: **Estimated α_l Parameters over Density**



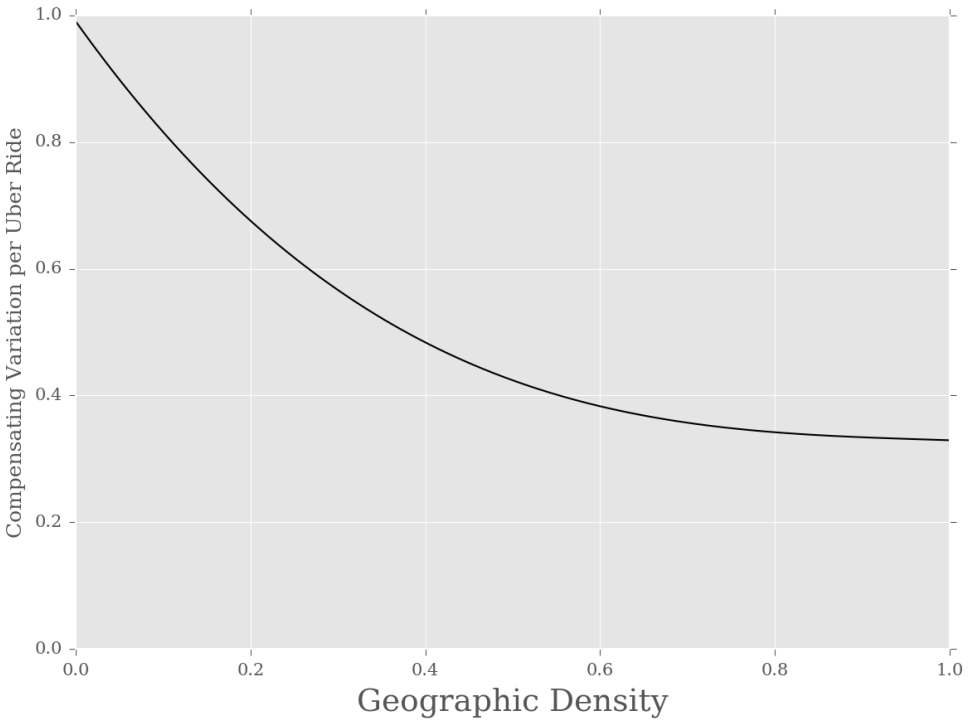
Note: Density is measured by the total building area over the ground area in the particular zone. Here density has been normalized to be between 0 and 1, to emphasize the relative difference across taxis zones. Each dot represents the fitted α_l for the location with that density.

Figure 1.16: **Compensating Variation per Uber Ride, 2013 to 2016**



Note: The map illustrates the average change in consumer welfare (in $) per ride for Uber riders in each of the zones of the map. This change in welfare was measured by calculating the welfare of the rider in 2016 compared to her choice set in 2013, taken as the period "before Uber."

Figure 1.17: **Fitted Compensating Variation over Geographic Density**



Note: The figure rearranges the data from Figure 1.16 along geographic density. The underlying data were fitted using a 3rd-degree polynomial.

Figure 1.18: **Compensating Variation per Uber Ride, After Banning Uber**

Note: The map illustrates the average change in consumer welfare (in $) per ride for Uber riders in each of the zones of the map. This change in welfare was measured by calculating the welfare of the rider in 2016 compared to her choice set after Uber is eliminated from the market. Note, unlike with the first measurement using 2013 data as the "counterfactual", the quality of other non-taxi services do not change here.

1.8 Tables

Table 1.1: **Comparison of Taxi / Uber Platforms**

Type	In Operation	Matching	Supply Cap	Entrance Fee	Geo. Limited
Yellow	–	Hail	Binding	High	No
Green	Aug. 2013	Hail	Not Binding	Low	Yes
Uber	May 2011	Dispatch	No Cap	Lowest	No

Notes: While Uber began operations in New York in May 2011, the scale of its operations only began to close the gap with yellow taxis in late 2014.

Table 1.2: **Data Summary: Transit Choice**

Source	Date Range	Variables	Observations
ACS	2008, 2015	Choice, Time of Departure (ToD)	5600
MTA Travel Survey	05/2008 - 11/2008	Choice, Demographics, ToD	10580
TLC Pickup Data	03/2016 - 06/2016	Choice, Choice Characteristics, ToD	7.5m

Notes: The observations for "ACS" are the number of markets for which shares are generated. There are 8 half-hour time periods for 350 census tracts for two years. In the estimation the tract shares are aggregated to taxi zones.

Table 1.3: **Data Summary: Transit Characteristics**

	Wait Time	Travel Time	Price	Walking Distance
Taxi	Simulated	Data	Data	0
Uber	Data	Simulated	Data, Simulated	0
Walking	0	Simulated	0	Simulated
Public Transit	Simulated	Simulated	Simulated	Simulated

Notes: For each choice and characteristic, I list whether it is read off data or simulated when the choice is an observed choice. Assume whenever a choice is counterfactual, its characteristics are simulated as described in the text. A "0" denotes that field is always assumed negligible.

Table 1.4: **Demand Estimation Results**

Parameter	Estimate	Standard Error
(baseline) α_0	-0.0433	0.0015
$(25K \leq inc < 50K)$ α_2	-0.0428	0.0014
$(inc < 75K)$ α_3	-0.0359	0.0015
$(inc < 100K)$ α_4	-0.0337	0.0019
$(inc < 150K)$ α_5	-0.0287	0.0023
$(inc < 200K)$ α_6	-0.0282	0.0034
$(inc > 200K)$ α_7	-0.0208	0.0043
β_{wait}	-0.0104	0.0012
β_{time}	-0.0006	0.0000
β_{walk}	-0.0001	0.0000

Table 1.5: **Demand Estimation Results**

Parameter	Estimate	Standard Error
σ^2_{tx}	1.224	0.568
σ^2_u	1.315	0.492
σ^2_w	3.145	1.231
$\sigma_{tr,tx}$	0.789	0.346
$\sigma_{tr,u}$	0.894	0.278
$\sigma_{tr,w}$	1.186	0.512
$\sigma_{tx,u}$	1.532	0.167
$\sigma_{tx,w}$	0.632	0.291
$\sigma_{u,w}$	0.831	0.354

Table 1.6: **Supply Estimation Results**

Parameter	Estimate
σ_ε	1.315
V_u	21356
α_{dens}	0.0048

Notes: I currently omit the standard errors because the correct formulation requires accounting for the impact of the first-round (and in the case of α_{dens}, second-round) estimates on the variance of variables estimated later. To calculate the standard errors by bootstrap will be saved for the end of the research process.

Table 1.7: **Summary of Changes in Each Counterfactual, by Density Quartile**

		In Exclusion Zone				Outside Exclusion Zone			
CF	Variable	1Q	2Q	3Q	4Q	1Q	2Q	3Q	4Q
C1	Welfare	0.52	0.52	0.45	0.38	0.94	0.81	0.85	0.79
C2	Welfare	0.65	0.67	0.60	0.57	0.90	0.70	0.74	0.70
	Total Rides	−44.8	−23.9	−7.34	0.53	−96.4	−79.2	−71.3	−52.5
C3	Welfare	0.42	0.44	0.44	0.38	0.51	0.46	0.48	0.41
	Total Rides	−20.2	−15.4	−2.1	6.30	−43.2	−30.1	−33.5	−20.5
C4	Welfare	0.64	0.68	0.60	0.62	0.15	0.13	0.15	0.11
C5	Welfare	0.30	0.28	0.15	0.16	0.05	0.08	0.02	0.07

Notes: Measures are simple, that is unweighted, averages of the average for each zone over the course of a simulated day. Welfare is measured as the compensating variation per ride for an Uber passenger. Changes in rides are reported in average percent changes, again unweighted by initial pickups. All changes are relative to the baseline in June 2016.

Chapter 2

Mitigating Preference Externalities Through Digitization: An Application to Transit Markets

2.1 Introduction

Internet-based services once offered the promise of reducing discrimination in the market through anonymity. In many ways that promise was overstated, as digitization of our lives has given information brokers unprecedented access to consumer preferences.[1] Nor has digitization eliminated incidences of direct social discrimination; Edelman et al. (2017) and Ge et al. (2016) document racial discrimination on Airbnb and the ride-sharing platforms Uber and Lyft, respectively.

The potential remains, however, for digitization to mitigate statistical discrimination in market services. Waldfogel (2003) documents the existence of a type of "discrimination" in radio markets; preference groups that offer businesses lower expected profits

[1] See Shiller (2014).

tend to be underserved.[2] Waldfogel (2017), on the other hand, illustrates how digitization can reduce these externalities by lowering the cost of entry for products targeting previously underserved groups.

In this paper I look for that potential in the rapidly changing taxi market of New York City. A corollary of the results from Chapter 1 is that consumers of less geographically dense markets in the city suffered from negative preference externalities when only hail taxis serviced the city. Digitization, through Uber, mitigated these externalities by lowering the cost of cabs to enter those markets. This paper explicitly documents this argument through two key findings. First, the search behavior of hail taxis, even controlling for the profitability of different search routes, highlights statistical discrimination against certain classes of consumers. Second, the arrival of Uber has mitigated the negative externalities in the cab markets among these consumers. While further work is needed to identify the specific mechanism, a reasonable hypothesis is that Uber's matching technology allows contracts to be made without imposing the cost of undirected searching in previously avoided areas of the city.

Using a rich data set containing the universe of taxi cab transactions in New York City, I estimate a reduced form model of taxi drivers' search behavior throughout the city. Because of the volume of data, I am able to approximate the profitability of the decisions a taxi can make while searching for a consumer. I leverage these estimates to demonstrate taxis are leaving money on the table by not searching in certain areas of the city or serving certain demographics. Finally, using data on Uber pickup locations, I estimate which groups have benefitted most, in terms of ridership, from the entrance of Uber.

This paper ties into several active research areas. The recent availability of rich taxi transaction data has led to a number of papers exploring taxi search behavior. The seminal paper Camerer et al. (1997) reasons that taxi drivers are not profit maximizers but rather have target wages they attempt to hit each day. Recently, this result has come under scrutiny with Farber (2005) and Farber (2015) arguing that taxis do not have reference-dependent search preferences. Farber (2008) and Crawford and Meng (2011) attempt to reconcile the empirical findings of the earlier and more recent research. This paper assumes that throughout the course of a shift taxi drivers maximize their utility

[2] In the case of radio, the scale of the preference group matters.

but not necessarily profitability. This assumption permits estimating the value to these drivers of not traveling to certain areas of the city.

The literature on statistical discrimination of a nature similar to this context extends to Bertrand and Mullainathan (2004). Their research demonstrated discrimination against African American in the labor market, but that discrimination could be mitigated by living in "good", that is wealthy, neighborhoods. Other research has focused on teasing apart "first-degree" discrimination from statistical discrimination. List (2004) tests for both in the sports card trading market. Finally, Edelman et al. (2017) and Ge et al. (2016) both uncover "first-degree" discrimination on new digital platforms. Ge et al. (2016) finds that African American passengers tend to get dropped by Uber and Lyft drivers more frequently than white passengers; they do, however, note that the impact was far less than for traditional taxi companies. In contrast to their work, I focus on statistical discrimination. While digitization may not reduce ingrained prejudice, it does the have possibility to reduce "rational" discrimination against certain groups.

The paper proceeds as follows. Section 2.2 discusses the data and the generation of key variables, such as taxi search choice, used in the estimation. Section 2.2.6 uses the data to illustrate broad characteristics of search patterns in different areas of New York City. Section 2.3 introduces the regression model to derive taxi drivers' search preferences. Section 2.4 presents the results from the estimation and Section 2.5 integrates Uber pickup data to determine which groups have benefitted most from the arrival of Uber into the market. Section 2.6 concludes.

2.2 Data Sources and Construction

In this section I walk through the available data. I then define a taxi cab's search choice set along with its decision nodes. Finally, I explain the procedure for calculating the variables under consideration by a taxi when it makes its search decision.

2.2.1 Data Sources

The primary data source is the New York City Taxi and Limousine Commission (TLC)'s dataset on the universe of yellow taxi trips. An observation in the dataset includes

several pieces of information pertinent to the trip including pickup and drop off locations by longitude and latitude; the pickup and drop off time; fees, broken down into taxes, tolls, tips, and the actual fare; and the trip time and distance. Critically, and unlike in Chapter 1, during the time period studied up to 2013 the data also include identifiers for individual cabs. This additional information simplifies the process of determining how taxis cruise in between a drop off and their next pick up.

Because this data set only provides the locations of pickups and drop offs, I supplement the information with detailed NYC road data from the NYC Department of City Planning LION Database. The database provides a complete mapping of roads and other transportation features in NYC. It will serve many purposes, but the two most important are information on the specific street segment, that is the road between two intersections, of a pickup and the ability to approximate the route a taxi took between a drop off and its next pickup. The process of approximating these routes and notes about its accuracy are described in Appendix B.1.

Demographic and crime data are taken from the 5-year averages of the American Community Survey and the NYC Police Department's crime map. The former are available by census block group, while the latter are available by intersection and aggregated by census block group. Events that occur on the intersections on the corner of multiple census block groups are distributed uniformly across the groups. These census block groups form a partition over New York City and will be denoted by the collection $B = \{B_l\}_{l=1}^{L}$ where L is the total number of census block groups.

2.2.2 Data Generation

Before describing how I determine choice data, I first detail how I take advantage of the ability to track cab drivers over the course of the day. I interpret the TLC trip records as collections of shifts for yellow taxi drivers.[3] The complete set of shifts S partition the data and can be further assigned to their respective drivers $S = \{S_i = \{S_{ik}\}_{k=1}^{N_i}\}_{i=1}^{N}$ where N is the total number of cab drivers and N_i the number of his or her shifts in the dataset. A single observation in the dataset is a specific ride r on one of these shifts. For now I suppress the notation for the specific driver and shift and let S^r for $r \in \{1, \ldots, R_{ik}\}$ denote a specific ride observation given a driver and shift.

[3] See the Appendix B.2 for how shift changes are specifically determined.

Let $X(S^r)$ denote the vector of data available for the particular ride on this shift, including pickup time and drop off time t_b and t_e, respectively, and pickup and drop off coordinates l_b and l_e, respectively. Data available only at aggregated levels, such as demographic data by census block group, are linked to trips through pickup and drop off locations. Hence, the mean population income associated with a ride's pickup point is the income of the census block group such that $l_b(S^r) \in B_l$.

A critical portion of the analysis focuses on the driver's behavior between two rides in a shift. These "between-trip" periods are known as searching trips and are not explicitly in the data set. I infer searching trips from the data of two consecutive rides, e.g. S^r and S^{r+1}. The details of constructing a searching trip are explained in Appendix B.2.

Denote the set of searching rides as C. C_{ik} are the searching rides for shift k of driver i and can be further indexed by $r \in \{0, \ldots, R_{ik}\}$ so $C_{ik} = \{C^r_{ik}\}_r$. Technically, there is a searching trip before the first ride and after the last ride, denoted above by $r = 0$ and $r = R_{ik}$, respectively. Because it is impossible from the dataset to determine when these shifts begin or end in terms of time or location before the first pickup or after the last drop off, the first searching trip considered for each shift is C^1 between S^1 and S^2 and the last is $C^{R_{ik}-1}$ between $S^{R_{ik}-1}$ and $S^{R_{ik}}$.

Because searching trips are not observed, information about them must be gleaned from preceding and subsequent rides. Consider a specific searching trip C^r with preceding ride S^r and subsequent ride S^{r+1}. Key data for this paper include where the searching trip begins and ends and when it begins and ends. Naturally the searching trip begins where the previous ride ends, hence $l_b(C^r) = l_e(S^r)$; likewise $l_e(C^r) = l_b(S^{r+1})$. Similarly the searching trip begins *when* the previous ride ends so $t_b(C^r) = t_e(S^r)$.

2.2.3 Identifying Decision Nodes

With the data set up I proceed to explain what the choice of a taxi driver looks like. Because taxi drivers are obligated to accept fare requests, they will not always make decisions in between rides.[4] To wit, a cab in midtown Manhattan on a weekday will likely find a new fare waiting after the completion of a previous ride. Lest an estimation of the model overstate the propensity of cab drivers to seek fares in high passenger-traffic

[4] They are legally obligated to accept fare requests. Of course this is an assumption, since face-to-face discrimination is still an issue for taxis. See Ge et al. (2016) and Glanville (2015).

areas, it is important to ensure that only non-trivial cruising trips are considered. The following criteria are used to determine whether a cab driver is making a decision about where to search in between rides:

1. An intersection is passed between on a searching trip
2. Over two minutes have passed between rides

If either condition is satisfied, I identify the driver as having made a non-trivial decision. To motivate the first, consider that at an intersection a cab typically has up to three potential options to take: straight, left, and right. I assume any movement at an intersection is a conscious decision on the cab driver's part and ignore the possibility that high traffic precludes making turns if desired. It is possible, however, that a taxi driver circles a block to end up in largely the same location or waits in place. Both are obviously decisions on his or her part but are not explicitly visible in the data. Hence I impose an arbitrary "decision threshold" after 2 minutes, after which it is determined the cab chose to stay in the same location. Further details on the identification of decision points are in Appendix B.3.

To illustrate the frequency of these decision points, I take data from the first six months of 2012 and break down the percentage of cruising trips during which a taxi driver is making a decision. Overall cab drivers cross at least one intersection between 92% of trips. As a check that this metric is negatively correlated with the expected probability of finding a passenger in the same location, I break down the measure by time of day and weekday in Table 2.1 and Figure 2.1. Both figures illustrate that the percentage of searching trips with decision points decreases during rush hour, when it is more likely that a drop off and pickup would occur in the same location.

2.2.4 Taxi Choice Variable

Rather than linking each searching trip featuring a decision node with a complicated sequence of turns at intersections, I summarize the decision a cab makes at these nodes with a simple trajectory variable between its drop off and the subsequent pick up. In Section 2.3 I discuss limitations on the usefulness of this descriptor.

I measure a searching trip's trajectory as the true azimuth, or azimuth with respect to the rotational North Pole, between the start and end point of the trip. The trajectory

is further discretized based on which arc it passes through from the starting point. Figure 2.2 depicts the directions radiating from a starting point as well as the cab's trajectory from the starting point at its last drop off to its next pick up. In this example I identify the cab driver as choosing the 2nd direction. Cab drivers who stay in the same location are labeled as choosing the 0th direction. Denote the direction taken in a specific cruising trip of driver i on shift k as $d(C^r_{ik})$.

2.2.5 Constructing Choice Characteristics

The final step is to link with each choice at a particular point in time and space a series of characteristics. For this analysis I aggregate choices into states s by day of the week wkd; hour group hr, which is six groups of four hours starting at midnight; and, census block group B_l. Let $C(s)$ be the set of searching trips beginning in this state.

I consider two types of outcomes for each of the searching trips in $C(s)$. The first are ride-specific, e.g. fare, trip distance, and demographic outcomes for the ride following the searching trip, i.e./ where the taxi ends up in its next pickup $X(S^{r+1}_{ik})$ for C^r_{ik} in the set. The second is the time between the beginning of the searching trip and the next ride. For notational simplicity I assume this searching time is a characteristic included in $X(S^{r+1}_{ik})$. In a slight abuse of notation, let $X(C^r_{ik})$ designate the collection of these outcomes for the relevant searching trip.

I ultimately also consider the implications on the rest of the shift from a particular searching trip. For example, I want to pick up how costly it would be for a driver to end up dropping off in an outer borough. Let $X(C^r_{ik}; t_r)$ be the outcomes from subsequent rides beginning *before* t_r hours after the beginning of the searching trip, i.e. for trips $S^{r'}_{ik}$ such that $r' > r$ and $t_b(S^{r'}_{ik}) < t_b(C^r_{ik}) + t_r$. It is feasible that the cab driver may have no pickups in the t_r hours after beginning the searching trip. With this notation I can finally describe the construction of the variables used for estimation.

For constructing choice variables I fix the parameter t_r. For a given t_r, constructing $X(C^r_{ik}; t_r)$ requires aggregating outcomes over all subsequent rides within the time period. Depending on the variable, an appropriate aggregation might be summing or averaging variables. In the actual estimation I consider multiple specifications. For now

consider summation. Then

$$X(C_{ik}^r; t_r) = \sum_{r'>r} X(S_{ik}^{r'}) * \mathbb{1}\{t_b(S_{ik}^{r'}) < t_b(C_{ik}^r) + t_r\} * \mathbb{1}\{t_b(S_{ik}^{R_{ik}}) < t_b(C_{ik}^r) + t_r\}$$

The second condition added above requires that the last ride on the shift begins before the end of the time period considered. If this condition is not added, then some observations near the end of shifts might have artificially low outcomes if t_r encapsulates several hours after the beginning of the cruising trip. Finally, given $X(C_{ik}^r; t_r)$, I calculate the *expected* outcomes from choosing direction j for a given state s. The formula is derived as follows

$$X_j(s; t_r) = \frac{\sum_{i,k,r} X(C_{ik}^r; t_r)\mathbb{1}\{C_{ik}^r \in C(s)\} * \mathbb{1}\{d(C_{ik}^r) = j\}}{\sum_{i,k,r} \mathbb{1}\{C_{ik}^r \in C(s)\} * \mathbb{1}\{d(C_{ik}^r) = j\}}$$

Hence $X_j(s)$ is really the *expected* outcome from choosing direction j.

2.2.6 General Patterns

Using these constructed variables, Figures 2.3, 2.4, and 2.5 illustrate the relative values for cabs choosing to move north, which includes directions 8, 1, 2, and 3, versus those that choose to move south. Figure 2.3 depicts the relative share averaged over all states. Areas in the lightest colors (< 1) are those where cabs tend to move south. With the exception of southeast NYC, cabs have a tendency to move south with varying degrees of uniformity. Considering that most pickups in Manhattan are concentrated in the central area south of Central Park, this result is rather unsurprising.

The other two figures serve as references for the estimation. Figure 2.4 maps the relative median income for the next pick up of a taxi that chooses to search north versus south. Figure 2.5 maps the same for race. Race is measure from 0 to 1, where unity indicates the area has a completely African American community. The figure illustrates that moving north nearly everywhere, though, with different degrees of intensity, increases the probability of a pickup in an more African American neighborhood. Linking the pictures from Figure 2.3 and Figure 2.5 provides some preliminary pattern that could result in African American neighborhoods receiving worse taxi service. To control for pecuniary search incentives, however, I turn to the choice model.

2.3 Model and Estimation

2.3.1 Choice Model

I estimate a simple discrete choice logit model of the cab drivers' decisions during *searching trips*. As described in Section 2.2 I consider each cab driver making a simple directional decision from their starting location at the start of the particular searching trip.

To that end I estimate the following decision function

$$f(s, \varepsilon, \xi, \beta) = \arg \max_{j \in C(s)} [u(s, j, \beta) + \xi_j(s) + \varepsilon_{ij}] \tag{2.1}$$

where s, the state includes day of week, hour group, and census block group. $C(s)$ is the cab driver's choice set and will depend only on the location component of the state. The set $C(s)$ and hence the co-domain of f are restricted to a finite set. In the estimation the choice set consists of the four cardinal and four ordinal directions, and the choice to stay in the same location. The choice set can further be constrained by geographical restrictions imposed by starting locations, e.g. driving south from Battery Park in southern Manhattan is not in its choice set.

An important point to highlight is that the directional choices from different starting locations are the same in name only. The choice-specific unobservable $\xi_j(s)$, for example, is not constrained to be the same for the jth choice from different starting states. Hence the unobservable is fit by hour group, census block group, day of week, and direction.

I make the standard assumption that ε_{ij} are i.i.d. choice/driver-specific shock parameters following a type-1 extreme value distribution. In a model without cab-driver specific information, these errors carry the entire weight of differentiating decision outcomes in a given state. This assumption may not be innocuous in this application. Unobservables may be correlated across choices because the simple model does not control for driver. Drivers could be more predisposed to particular neighborhoods closer to their final drop off point or in which they have more familiarity.[5]

The payoff function $u(s, j, \beta)$ is the key object of interest and is assumed to take on

[5] There is some evidence of this behavior in Haggag et al. (2017).

a linear form

$$\begin{aligned} u(s,j,\beta) &= \delta_j(s) \\ &= X_j(s)\beta + \xi_j(s) \end{aligned}$$

where X_j are choice-specific outcomes. I let choice 0, that is the taxi driver staying in the same location, serve as the normalizing choice and denote it by $\bar{j}$. Let $\bar{X}_j(s) = X_j(s) - X_{\bar{j}}(s)$. The probability of a cab choosing a particular direction is then given by the formula

$$\begin{aligned} P(j \mid s,\beta) &= \frac{\exp(\bar{X}_j(s)\beta + \bar{\xi}_j(s))}{1 + \sum_{j' \in C(s)} \exp(\bar{X}_{j'}(s)\beta + \bar{\xi}_{j'}(s))} \\ &= \frac{\exp(\bar{\delta}_j(s))}{1 + \sum_{j' \in C(s)} \exp(\bar{\delta}_{j'}(s))} \end{aligned}$$

2.3.2 Estimation

The high volume of data for Manhattan and the discretization of the choice set allows the choice probabilities to be empirically estimated using calculated shares, denoted $P_j^{data}(s)$.

$$P_j^{data}(s) = \frac{\sum_{i,k,r} \mathbb{1}\{C_{ik}^r \in C(s)\} * \mathbb{1}\{d(C_{ik}^r) = j\}}{\sum_{i,k,r} \mathbb{1}\{C_{ik}^r \in C(s)\}} \tag{2.2}$$

As shown by Berry (1994) I can reduce the estimation of β in the logit model to a simple linear regression with the empirical shares.

$$\begin{aligned} \log(P_j^{data}(s)) - \log(P_0^{data}(s)) &= \bar{\delta}_j(s) \\ &= \bar{X}_j(s)\beta + \bar{\xi}_j(s) \end{aligned} \tag{2.3}$$

β is then backed out in the standard by treating $\bar{\xi}_j$ as an error. This is the key regression model in the analysis.[6]

[6] In the regression the error is clustered by state.

2.3.3 Cleaning Data for Estimation and Endogeneity

To identify the choice set by location I simply observe the number of searching trips starting from state s that have a trajectory in direction j. If $P_j^{data}(s) = 0$ (except for $j = 0$) the direction is excluded from the choice set. Because data is aggregated over such a long time horizon and over thousands of observations, this is the easiest way to detect if a direction is geographically infeasible. Presumably if the direction were geographically feasible at least one cab would choose that direction over the many observations in the dataset.

Second I limit what searching trips I include in the estimation along two criteria. First, only searching trips beginning from Manhattan are used in the estimation. Generally over 90% of trips in the dataset end in Manhattan. This exclusion is required because of the high data requirements in the construction of the variables used in the estimation but simultaneously is not a significant issue since most yellow taxi cab activity is in Manhattan.

The second condition limits the dataset to observations where the directional decision of the cab from its starting point can be gleaned with some confidence. For searching trips the only available locational data is the starting point $l_b(C^r)$ and ending point $l_e(C^r)$. Drawing a vector between the two and labeling that the directional choice can be overly optimistic, particularly if the cab actually meandered significantly in getting from $l_b(C^r)$ to $l_e(C^r_{ik})$. For example, if one point starts at the 5th Avenue Apple Store and ends at Rockefeller Center — roughly 10 blocks apart on 5th Avenue — this analysis would contend the driver chose to search south. Given $t_b(C^r)$ and $t_e(C^r)$, the starting and ending times, I see the cab driver took one hour to complete the cruising trip, however, so it is highly unlikely he or she went straight from the Apple Store to Rockefeller center. He instead might have traveled north, and eventually turned around, or traveled around Central Park. The initial directional decision of this cab driver was actually to travel north, not south.

To address this issue I approximate the time and route taken by the cab drivers taken from $l_b(C^r)$ and $l_e(C^r)$. Details of this approximation are in Appendix B.1. If the actual time and estimated time are within 5 minutes of each other, then the route is considered roughly known and the cab trajectory is taken from $l_b(C^r)$ and $l_e(C^r)$ with

more confidence.[7]

The major concern for the validity of the specification is endogeneity issues that might arise from $\bar{\xi}_j$ being correlated with some components $\bar{X}_j$. One specific instance might be, for example, including expected passenger income in the specification but leaving out crime rates. Generally, the perceived profitability of a choice might be a function of the number of cabs that make the same choice. The channel through which that might occur is by affecting the probability a cab gets a ride after traveling in a certain direction. By calculating the outcome of cruising trips with the time cutoff t_r instead of a fixed number of subsequent rides, this probability is directly taken into account. For example, choosing direction 8 may always lead to a high fare, conditional on the rare event of actually finding a passenger. Direction 4 may always lead to a mediocre fare but finding a passenger is guaranteed. This expected outcome is reflected in the calculation of $X_j(s;t_r)$; after a fixed period of time fewer cab drivers who chose direction 4 will have actually found a fare and hence their assigned outcome $X(C^r_{ik};t_r)$ would be 0.

2.4 Results

Table 2.2 reports the full collection of variables and their method of construction. In this analysis I focus on variables looking forward 2 and 3 hours from the beginning of the searching trip. Variables expressed as percentages, such as race composition, or which do not "accumulate" are aggregated by averaging, whereas fares, waiting times, etc. are aggregated by summation. Previous work established the sensitivity of results to different measures of t_r, that is how far past a searching time cab drivers consider

[7] As might be expected the longer the distance between the starting and ending point of the cruising trip, the more inaccurate the estimated time is. This problem can be dealt with in two ways. First, instead of using a 5 minute error threshold, instead use a threshold that scales with the crow's distance length of the trip. Second, if dropping these observations does not introduce a selection bias, then there is no problem; there is enough data that identifying a single set of "taste" parameters is still feasible with the remaining data. A problem exists if the share of cabs choosing each direction is significantly impacted. While this robustness analysis is not included, it can be by checking the relative size of dropped observations for each state.

outcomes when making a choice.[8] To give a sense of the economic significance of magnitudes that will be reported with the regression results, Table 2.3 reports descriptive statistics for the RHS variables used in the regressions.

The basic model regresses the share of each directional choice on various characteristics of searching trips along that direction for each state. In the different specifications I compare models with and without demographic characteristics and assess fit via likelihood ratio tests. For these regressions I also compare results just using variables accumulated by averaging to avoid consideration of the time horizon over which the cab drivers think about outcomes.

The results suggest that cultural factors have bite in directional decisions. There are also a few curiosities as well as results that help intuitively validate the results observed. First, signs are largely robust across different t_r, which is particularly important for the averaged variables which are less likely to change significantly with larger t_r. Finally, Table 2.5 reports the average change in a direction's share from a one standard deviation increase in each of the eight variables using β from specifications (3) and (6). The output suggests that the most significant economic factor in determining cab's directional choice is, as expected, expected fare.

The critical result for the paper, however, are the terms on demographics. Even controlling for monetary incentives, I find that taxis have a tendency to search in wealthier and whiter neighborhoods. Depending on the robustness of the model, this provides evidence of some externality of living in less wealthy neighborhoods or neighborhoods with more African Americans. In the discussion to follow I then test what communities have most benefitted from the introduction of Uber to the market.

The rest of the section discusses peculiarities of the choice model results. Among the results robust across t_r, there are also three unexpected results. First, crime is somehow an *attractive* property. There are two potential explanations. First, only the location of significant crimes, as listed in Table 2.2, are known and included as part of the crime rate, which Table 2.3 shows, is quite low overall. Cab drivers might be less worried about murder and grand larceny than a passenger not paying.[9] Nonetheless,

[8] Those results suggested drivers are not myopic and initially motivated modeling the dynamic decisions of taxi drivers in Chapter 1.

[9] Unfortunately, an unpaid fare does not show up in the dataset.

the effect is significant, nontrivial in magnitude, and positive in all included specifications. The more logical alternative then is that crime rate is correlated with some other attractive property dropped from the specification. One explanation is that crime rates are higher in commercial areas, which may attract cab drivers because of the high transient population throughout the day. This explanation also serves to deal with the second curiosity that areas where residents claim to take cabs to work are less desirable. Future work may permit coefficient values to changes with time of day.

Finally, fare / time is significantly negative after accounting for fare. This result may again reflect neglecting commercial areas, where fares are high. Commercial areas have generally slower traffic in Manhattan and hence lower fare / time ratios might appear more attractive to cab drivers simply because they have higher expected fares. Indeed both fare / time is positive and significant before accounting for fare.

2.5 Discussion

Given that I have established that taxi drivers are avoiding some demographics in their search behavior, I tie this back to Uber by finding what communities have most benefitted, in terms of total rides, since the entrance of Uber. For this exercise I use information on Uber pickups from the TLC. Because the TLC only provides pickup information by taxi zone, which are partitioned by census blocks, I aggregate all data to the taxi zone level. I then run a simple accounting regression of the total Uber rides per unit area and population on the same characteristics featured in Section 2.4. To control for financial incentives, I use estimates of fares earned from taxis picking up in that zone.

The results of this regression are reported in Table 2.6. In the first specification I regress the total change in rides on the three characteristics. I find that median income is the only significant factor, and in the direction further favoring communities already served well by taxis. Oddly, historic fares are completely irrelevant. This difference might be because the driver does not choose specific pickup locations but is instead directed by the app.

The second specification regresses the total change in pickups from 2016 to 2013, during which time most growth was driven by Uber as established in Chapter 1. In this case the results are more interesting. I find that lower income and more African

American communities were the beneficiaries of new areas serviced by Uber. This is the finding that serves as motivation to further investigate the potential role for Uber in mitigating preference externalities with the old technology.

2.6 Conclusion

This essay has demonstrated the potential importance of non-pecuniary factors in the search decisions of taxis. I leverage these results to demonstrate cabs have abandoned some value by not picking up in certain neighborhoods. From the consumer perspective I interpret these results as the cost of perceived preference externalities or statistical discrimination. I concluded with a simple analysis assessing the characteristics of neighborhoods that have benefitted most from the arrival of Uber into the market. The results suggest the potential for digitization to help mitigate the costs of these externalities.

Further work is needed, however, to isolate the mechanism allowing Uber to serve these neighborhoods better. Chapter 1 of this dissertation would suggest Uber's matching technology enables Uber to pick up in previously underserved neighborhoods at lower cost. Alternatively, other channels might be at play. Taxi drivers might worry about searching in certain neighborhoods because of concerns about payment guarantees, the length of time before minding a match, etc. Different aspects of Uber's dispatch system address these concerns. Credit cards mitigate the worry about a passenger stiffing a driver, while the dispatch system itself hedges against searching in low-demand neighborhoods.[10]

One avenue to directly test this hypothesis is by taking advantage of growth in e-hailing in New York City. With the appropriate data one could compare the e-hailing system to Uber and Lyft's dispatching programs to tease out the marginal impact of these proposed features in reducing perceived preference externalities.

[10] Though, Uber drivers still have concern about lengthy idle times by being matched to consumers far from their location.

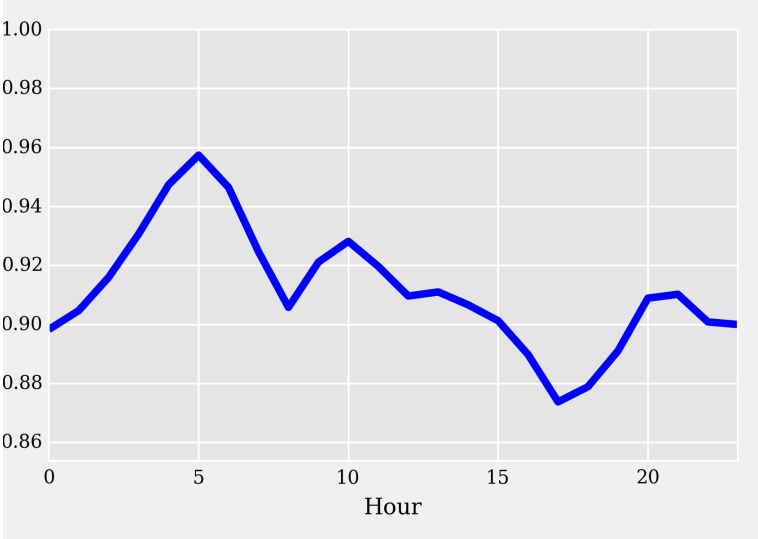
2.7 Tables and Figures

Table 2.1: **Percentage of Searching Trips with Intersection Crossing**

Hour Group	Mean
0	0.9385281
1	0.9616658
2	0.9385808
3	0.901183
4	0.8808307
5	0.9211456

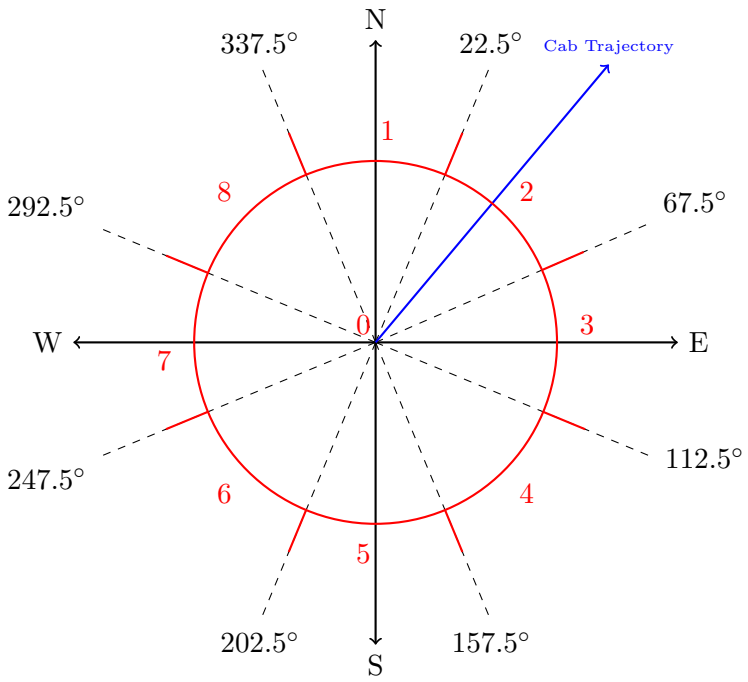
Source: NYC TLC using average weekday data from January to June 2016. Each hour group corresponds to a four hour period starting at midnight.

Figure 2.1: **Percentage of Searching Trips with Intersection Crossing by Hour**



Source: NYC TLC using average weekday data from January to June 2016.

Figure 2.2: **Discretization of Cab Driver's Decision**



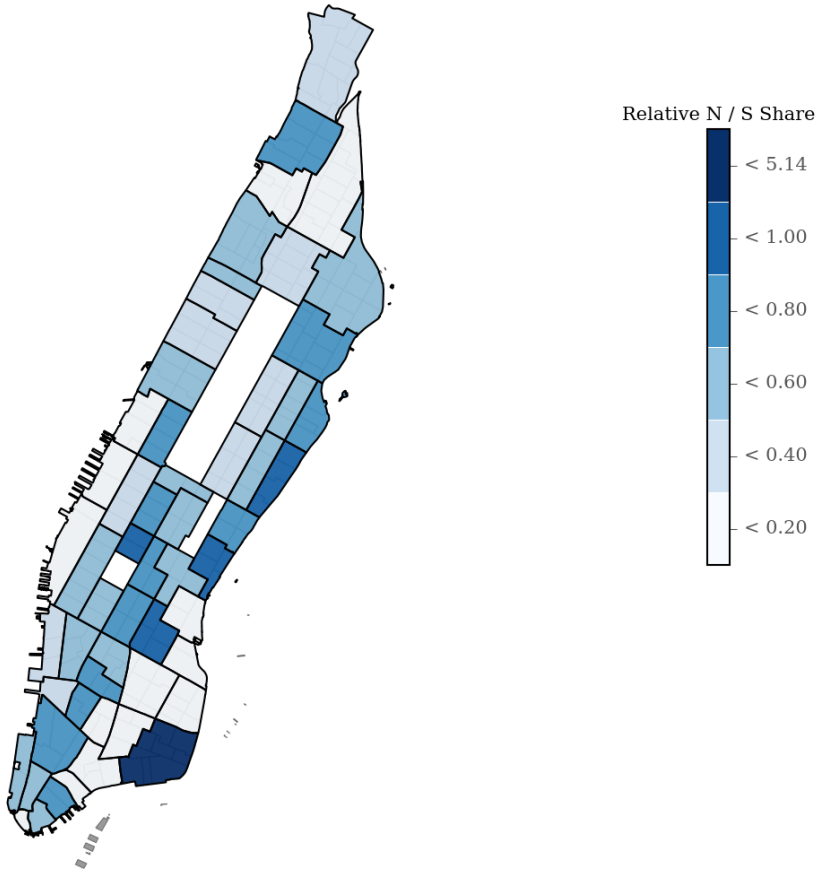
Notes: The figure summarizes the information for a taxi's searching trip. The origin is the location of the drop off, and the trajectory points in the direction of the driver's subsequent pickup. In the discretization of this decision, the decision the driver has made falls into one of X number of discrete categories (here 9) based on the arc the trajectory crosses (here 2).

Table 2.2: **Variables and Aggregation Methods**

Short Name	Description	Aggregation	Source
fare$\{t_r\}$	Total fare less taxes	Sum	TLC
wait$\{t_r\}$	Search time (seconds) between rides	Sum	TLC
time$\{t_r\}$	Time (seconds) elapsed on rides	Sum	TLC
frdist$\{t_r\}$	Fare / distance traveled on rides	Mean	TLC
frtime$\{t_r\}$	Fare / time elapsed on rides	Mean	TLC
race$\{t_r\}$	Percent African American	Mean	ACS
inc$\{t_r\}$	Mean income (1000s USD)	Mean	ACS
taxi$\{t_r\}$	Percent using taxis for morning transit	Mean	ACS
crime$\{t_r\}$	Annual crime rate per million[a]	Mean	NYPD

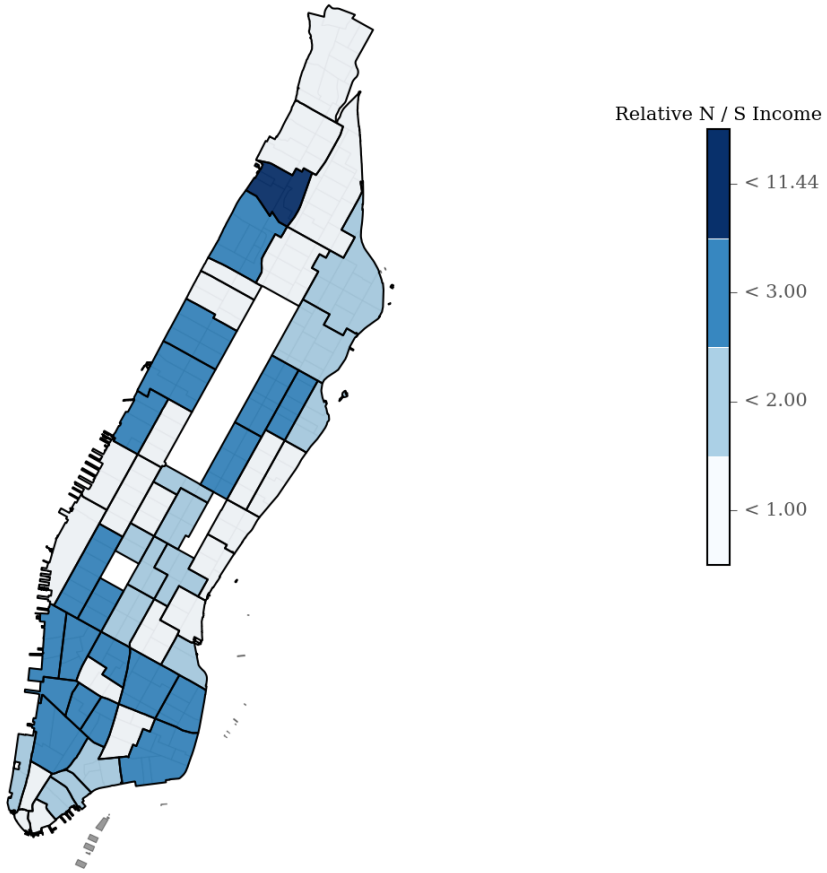
[a] *Crimes include assault, larceny, robbery, burglary, and murder.*

Figure 2.3: **Relative Choice of Taxis to Move North versus South, by Zone**



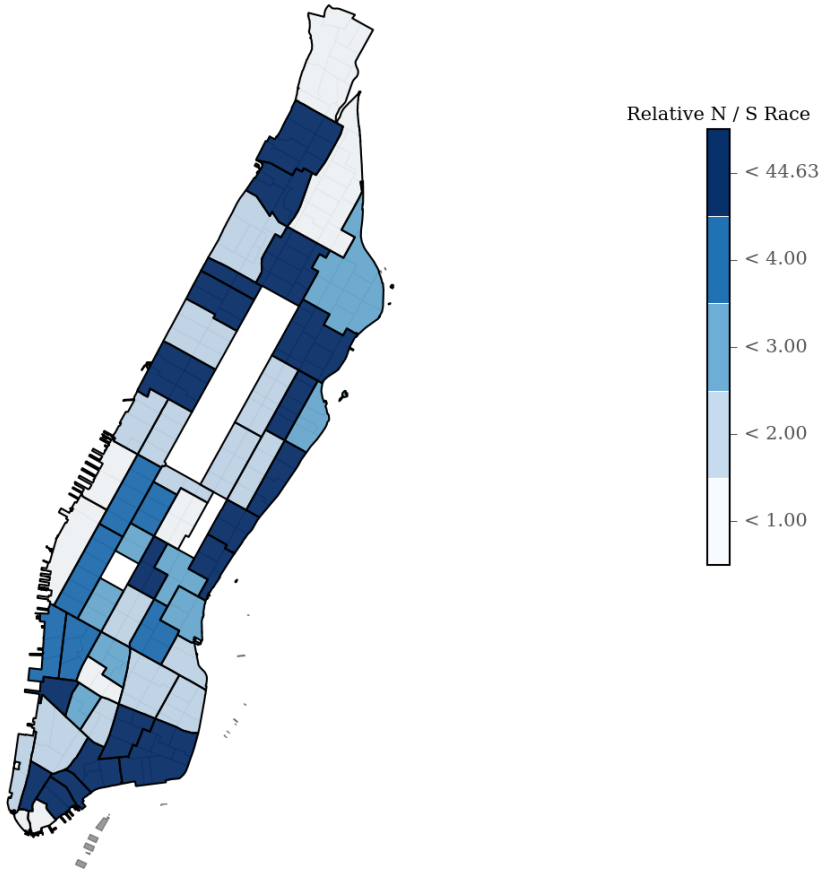
Notes: The share of taxis moving in directions 8, 1, 2, 3 were aggregated by taxi zone and divided by the share of taxis moving in directions 7, 6, 5, 4.

Figure 2.4: **Relative Median Income of Pickups After Moving North**



Notes: The share of taxis moving in directions 8, 1, 2, 3 were aggregated by taxi zone and divided by the share of taxis moving in directions 7, 6, 5, 4.

Figure 2.5: **Relative Race Index of Pickups After Moving North**



Notes: The share of taxis moving in directions 8, 1, 2, 3 were aggregated by taxi zone and divided by the share of taxis moving in directions 7, 6, 5, 4. The underlying measure of race is between 0 and 1, with 1 signifying a completely African American neighborhood.

Table 2.3: **Descriptive Statistics for Regression Variables,** $t_r = 1$

	Observations	Mean	Median	Std. Deviation
Total Fares	119543	28.89	30.93	7.26
Total Delay	119543	2032.57	1992.83	957.39
Fare / Distance	119534	4.666	4.6003	7.8910
Fare / Time	119538	0.0153	0.0143	0.0313
Percent African American	119286	0.0556	0.0462	0.0396
Mean Income (thousands USD)	119513	85.3308	90.9410	23.5723
Taxi Usage	118986	0.0259	0.0254	0.0114
Crime Rate	119504	0.6895	0.3413	1.2211

Notes: Total Fares and Total Delay are calculated by summing from the start of the searching ride over the next hour. The rest of the variables are averaged for that time period. Averages by state are calculated for the first half of 2012.

Table 2.4: **Estimating Economic Versus Social Factors**

	Over the Next 2 Hours ($t_r = 2$)			Over the Next 3 Hours ($t_r = 3$)		
	(1)	(2)	(3)	(4)	(5)	(6)
Fare / Distance	0.0182**	0.0093*	-0.0025	0.0191	0.0034	-0.0026
	(0.0078)	(0.0051)	(0.0048)	(0.0124)	(0.0030)	(0.0016)
Fare / Time	1.9226**	0.7537	-1.5290***	2.8317***	0.2975	-1.2999***
	(0.9776)	(0.6344)	(0.3915)	(0.9414)	(0.3482)	(0.3163)
Race		1.3441***	-2.0091***		6.6901***	-1.2809***
		(0.3350)	(0.3192)		(0.4669)	(0.4597)
Income		0.0386***	0.0233***		0.0573***	0.0234***
		(0.0009)	(0.0010)		(0.0009)	(0.0013)
Taxi Usage		-13.0538***	-8.5087***		-15.8888***	-7.0804***
		(1.3960)	(1.3086)		(2.0096)	(1.8105)
Crime (per 1000s)		0.0646***	0.0517***		0.0775***	0.0517***
		(0.0078)	(0.0073)		(0.0118)	(0.0107)
Fare			0.1963***			0.1331***
			(0.0056)			(0.0032)
Search Time			-0.000021**			0.000092***
			(0.000010)			(0.000026)
Observations	98360	98360	98360	97030	97030	97030
Adj R^2	0.0021	0.0849	0.1715	0.0053	0.2358	0.3197

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Notes: Time and distance are not included here because they are redundant with both fare and fare/time and fare/distance included. Observations are at the state level. Regressors in specifications (1), (2), and (3) are accumulated over two hours from the state and direction according to the method in Table 2.2. Regressors in (4), (5), and (6) are accumulated over three hours.

Table 2.5: **Percent Change in Share from One Standard Deviation in Variable**

	Percent Change in Share
Fare / Distance	-0.0724 %
Fare / Time	-0.1659 %
Percent African American	-2.2692 %
Mean Income	1.8916 %
Taxi Usage	-0.05092 %
Crime Rate	0.5700 %
Total Fares	30.0160 %
Search Time	-0.1726 %

Notes: For each state and direction I change the value of the relevant variable by one standard deviation from the population distribution and calculate the change in share. I then average over all states for the covered area.

Table 2.6: **Areas Benefitting from Ride Growth, June 2013 to June 2016**

	Total Rides: Uber	Total Change
Median Income	0.085***	-0.05***
	(0.018)	(0.011)
Race	1191.7	93.01**
	(4982.1)	(42.0)
Fare	-97.19	174.6
	(378.51)	(241.3)
Pop & Area Control	YES	YES
Observations	63	63
Adj R^2	0.48	0.34

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Notes: Regressors are averaged from census block groups to taxi zones by population weights. The dependent variables are calculated per unit of area in the zone. Race is an index from 0 to 1 with 1 representing an entirely African American community. Fare is the expected fare for a pickup based on historic taxi pickup data.

Chapter 3

Heterogeneous Effects of Subsidy and Infrastructure Investment in Electric Vehicles Adoption

3.1 Introduction

In the wake of mounting attention toward investment in energy efficiency and renewable resources, the US allocated approximately $400 million of the 2009 American Recovery and Reinvestment Act (ARRA) toward investment and research into electric vehicles and the deployment of a charging station infrastructure. The federal and California state governments concurrently rolled out other incentive programs to encourage the adoption of electric cars, including hefty rebates, tax credits, and electricity discounts. Since the initial enthusiasm, however, concerns have grown about the implications for inequality from these subsidies, primarily taken up the wealthy.[1] The dual concerns of inequality and environmentalism led to more targeted subsidy policies that have exposed the heterogeneous efficacy of freely allocated charging stations and flat vehicle subsidies in promoting EV adoption. In this paper we identify important sources of this heterogeneity and develop a model to assess the value, or cost, of this new wave of policies.

[1] See West (2004) and Borenstein and Davis (2016).

Using a new dataset featuring household-level vehicle purchase data with detailed geographic and timing information on the deployment of charging stations, we estimate a rich discrete choice model of demand for the automobile market in California. The dataset permits identification of consumer heterogeneity impossible in the existing literature. We characterize 1) marginal consumers with respect to the EV subsidy along demography-based differences, such as income and factors that correlate with attitudes toward environmental issues and 2) the marginal impact of charging stations along geography-based differences, including employment concentration and the average commuting distance of locals. The key to this empirical analysis is the identification and estimation of price elasticities across these demographic groups and the marginal benefit of non-residential charging stations on EV demand across the different geographies.

We use these dimensions to account for heterogeneity for two reasons. If consumers are heterogeneous in price elasticities, distributing EV subsidies flatly across income groups could generate less EV demand than a tiered subsidy structure. It is not straightforward, however, whether a regressive or progress subsidy would create more EV demand per dollar subsidy. This finding would identify the low-income group as marginal consumers. On the other hand, if consumers consider EVs luxury goods, EVs may face high elasticities for high-income groups. Depending on these competing forces, the same subsidy amount may lead to a greater (lesser) substitution to EVs for gas vehicles among low-income buyers than among high-income, ceteris paribus.

The nature of the current charging technology directs us to characterize the marginal impact of a station by its locations employment concentration and the commuting distance of its locals. Because of the lengthy charge time drivers may most benefit from charging in areas not time costly, that is a place where one would stay regardless of intent to charge the vehicle. Workplaces, much more than other public locations frequented by travelers, satisfy this criterion. These pieces of evidence hint that stations at a location with a large concentration of employees, such as a commercial city center, would have higher benefit to EV drivers, and therefore, higher marginal impact on EV demand.

We use California to estimate a flexible model of vehicle demand as the state offers several advantages. First, California has the largest market for EVs in the United States. Second, it features rich geographic differences across the state. Additionally, regional

entities, from counties and cities to utility districts, have implemented several of their own incentive programs on top of those available federally and from the state. The exogenous variation in resources across regions and over time help identify the different price elasticities and marginal impact of stations across the target dimensions. Not only do these features assist with our identification but also make for an interesting testing ground for alternative funding schemes that would redirect state funds to specific areas. Moreover, the electricity grid in California guarantees that substituting gas vehicles for EVs strictly reduce greenhouse gas emissions, thereby making the maximization of EV adoption a reasonable policy goal for the environmental social planner.

Besides the growing literature on the value of green subsidy programs and the trade-off between this class of subsidies and inequality, a central issue are indirect network effects with respect to the relationship of charging station deployment and electric vehicles. Most recent indirect network literature has emphasized competing platform markets benefiting from the growth of a platform-specific complementary market. Nair et al. (2004), Clements and Ohashi (2005), Dubé et al. (2010), Goolsbee and Klenow (2002), Gandal et al. (2000) analyzed the impact of an indirect network effect from software markets in varying tech industries. This line of research differs from our current project in several ways. First, except for Goolsbee and Klenow (2002), these papers have featured non-compatible platforms competing for market concentration. In this market, with a few exceptions, charging stations have been standardized to work across all PEVs.[2] Our ultimate question of interest is in the growth of the relatively new electric vehicle market space, rather than competition among firms in a mature market. In this respect our setting is similar to Goolsbee and Klenow (2002) which analyzes the diffusion of home computers in the 1990s, where the network size and specific tools, like e-mail, drive adoption more than platform-specific software. Additionally, we may consider the role of a similar peer effect in our setting using the methodology proposed by Bollinger and Gillingham (2012) in which the adoption of solar panels is encouraged by neighbors' adoption.

Two recent papers cover the same market with particular attention toward the angle of the indirect network effect. Li (2017) follows in the tradition of papers highlighting

[2] A notable exception might be Tesla's expanding supercharger network. These chargers are meant for long-distance trips, however, rather than everyday use.

the potential costs and benefits of platform competition. She focuses on the higher end of the charging station market, in which different car companies have an incentive to build car-specific fast charging stations. Our work focuses on the larger, lower end segment of the charging market which does not face these compatibility issues. Springel (2017) studies the effectiveness of charging station build out and vehicle subsidies in the Norwegian market, using county as a geographic market. She finds that while charging station investment is initially effective, there are high diminishing marginal returns. This paper answers a similar question, but incorporates, what we show, are important dimensions of heterogeneity in estimating the value of charging stations, using more granular geographic areas for market definitions. We suggest that by defining a geographic area larger than the effective area the estimated benefit of charging stations will be biased. Section 3.3 will demonstrate how sensitive demand models are to the geography considered for charging stations.

The paper is organized as follows. In Section 3.2 we describe specific obstacles for widespread adoption of PEVs and features of consumers and incentives in California. Section 3.3 offers reduced formed evidence of key heterogeneity in the value of charging stations by location and along consumer demographics. Section 3.5 presents the model of consumer demand and Section 3.4 the data used to estimate the model. Section 3.6 presents the moments we use in a GMM estimation procedure to derive the parameters of the demand model. We outline how features in our data sets allow the estimation procedure to identify the parameters of demand in Section 3.7. We conclude with plans for the future of this paper.

3.2 The Market for Electric Vehicles in CA

Many peculiarities of both electric vehicles and California differentiate its market for electric and gasoline-fueled vehicles and make the region ripe for a focus of this study. We identify two critical issues with the uptake of electric vehicles — their high price and range anxiety — and link these two to the incentive instruments used by governments.

3.2.1 Range Anxiety

Range anxiety when driving an electric vehicle stems from both their relatively short range and the lack of charging stations. Here it is worth distinguishing between the two types of electric vehicles. Plug-in electric vehicles (PEV) rely solely on a battery. A second class of vehicles plug-in hybrid electric vehicles (PHEV) have both a small battery and standard gasoline tank. Range anxiety is more relevant for drivers of the first class of vehicle.

As shown in Table 3.1, the full electric range of PEVs is on average only 30% of the range of gas vehicles. The Nissan LEAF carries a 84-mile range whereas the Chevy Volt has nearly a 400-mile range. The Tesla Model S is one of few fully electric vehicle with comparable range to traditional gasoline vehicles. According to the Center for Sustainable Energy (2013a), nearly 40% of survey respondents were not satisfied with the electric range of their purchased vehicle (see Table 3.3).

The lengthy time to charge these vehicles poses an additional problem for the infrastructure. According to charging station usage statistics of PEV owners by the EV Project (2014), owners leave their car charging for 6.2 to 7.4 hours on private away-from-home Level 2 chargers and 3.5 to 4.9 hours on public away-from-home Level 2 chargers.[3] If a PEV driver is looking for a public charging station as a solution for a drained battery, the low turnover at these stations decrease the likelihood of finding an open charger.

Because every standard electricity outlet is effectively a charging station, albeit slow, for an electric vehicle, the importance of a charging station network may not be immediate despite some evidence on the pervasiveness of range anxiety. Further evidence that suggests charging stations should have little impact on demand comes from CA driver habits. The LEHD LODES dataset shows the average Californian drives approximately 26 miles, well within the single-charge range of the Nissan LEAF or Tesla Model S. Table 3.6 indicates that the drive range for LEAF owners is even shorter (24.8 miles) per day. These figures suggest that charging away from home may be unnecessary for most use cases.

[3] See Table 3.2 for a comparison of the speed standards. Because compared to Level 2 chargers Level 1 chargers are impractical charging solutions for modern electric vehicles, we do not consider them as part of the charging station network in our analysis.

Other evidence, however, reveals these shorter driving ranges may be a symptom of range anxiety. Table 3.6 shows that Volt, which has the range of a standard gasoline vehicle, owners tend to use their vehicles lengthier trips. In addition, despite the short trips PEV owners tend to take every opportunity to charge their vehicles. 27% of charging events were away from home for the LEAF while only 16% were for the Volt in Q4 2013 according to EV Project (2014). Additionally, the survey Center for Sustainable Energy (2013a) noted the importance of having work-place chargers in their decision to purchase their electric vehicle (see Figure 3.3).

More direct evidence of range anxiety comes from the different charging behavior between full electric vehicles and plug-in hybrid electric vehicles. Figure 3.4 reports how much of the battery remains when a charging event takes place, taken from a sample of participants in an EV Project field study (EV Project (2014)). For the range-impaired Nissan LEAF, charging events take place the most when the battery still is 50% full or more. LEAF drivers charge even though the battery is not close to empty. In contrast, the Chevy Volt is charged primarily when the battery is fully exhausted. It reveals that LEAF drivers make a concerted effort to keep the battery at least half full all the time while Volt drivers let them run out. If range were not a concern for LEAF drivers, we should expect to see similar charging event distribution of Volt.

According to the Center for Sustainable Energy (2013a), 71% of the PEV purchasers were reportedly unsatisfied with the public charging infrastructure (see Figure 3.1). Since then, however, the number of charging stations has exploded from 3,000 to over 10,000. Table 3.4 summarizes the number of charging stations over time. This growth has coincided both with a growth in the number of electric vehicles in California and an expansion of the subsidies available for these public stations in California. Early efforts from the EV Project and ChargePoint America installed nearly 2000 charging stations in the years leading to 2013. Table 3.5 has details about the projects. Since then California and local governments have begun offering subsidies for charging stations. Figure 3.2 maps the subsidies available for public charging stations as of December 2016.

To facilitate the expansion of the charging network infrastructure, local governments have afforded several rounds of subsidies. In addition to the nearly 2000 charging stations installed in California by ChargePoint America and the EV Project since 2010, over 2500 chargers have been added to the network. They account for 43% of the total

charging stations in California as of December 2013. Representatives of the Department of Energy have suggested many businesses and workplaces have begun installing stations as a convenience to customers or employees.

3.2.2 High Price

While the market has changed significantly in the past half decade, electric vehicles still tend to have higher upfront prices than gas vehicles. In 2013 the most popular fully electric vehicle, the Nissan LEAF, had a $28,800 baseline price tag in 2013. In contrast the Honda Civic, the most popular vehicle, cost only $16,555. The most expensive electric vehicle, the Tesla Model S, which cost at least $69,900 at the time, while the most expensive gas vehicle among the top 80 CA car models in 2013 cost $52,800.

While PEVs have higher upfront costs, they tend to have lower maintenance and fuel costs than gas vehicles. Without discounting future costs or accounting for changes in the price of oil, the maintenance and fuel costs are about $33,728 for gas vehicle, $20,460 for plug-in hybrid electric vehicle, and $19,344 for full electric vehicles purchased in 2013.[4] Under the assumptions that generate those estimates, the PEV is still overall more expensive than gas vehicles overall.

Both California, via the Clean Vehicle Rebate Program, and the federal government offer tax credits for the purchase of PEVs. The federal government offers up to a $7500 for fully electric vehicles, while California offers up to $2500. Some districts in California offer rebates on top of these two programs. Table 3.7 provides further details on rebate amounts for specific models from federal and state sources. Figure 3.5 illustrates the heterogeneity of local vehicle subsidies across California for a person with a $60,000 income, with the San Joaquin Valley offering the most generous.

Table 3.8 illustrates the practical effect of these cost differences even after accounting for the subsidies. Purchasers of electric vehicles are from a different part of the income distribution than purchases of gas vehicles. For policy makers concerned with

[4] The estimates assume 12,400 annual mileage and a 16-year lifespan. Fuel costs for the gas vehicle are calculated based on 26.7 miles per gallon and a gas price of $3.5 per gallon. The fuel cost for PEV is calculated by assuming $0.13 per kWh, taken from the average CA price as of January 2012, and a mileage of 4.42 miles per kWh. The maintenance and repair costs include battery replacement, engine oil change, tire rotation, etc. The assumptions are taken from Alternative Fuel Life-Cycle Environmental and Economic Transportation (AFLEET) 2013. Numerous data sources include AFDC Price Report, AEO Report, Argonne National Laboratory, and so on. See the User Guide for AFLEET Tool 2013.

the potential for exacerbating inequality issues, these patterns show that subsidy money is primarily funneled to the top of the income distribution. In response, California altered its subsidy system to a regressive system with the largest subsidies available to the lowest income individuals. These income eligibility requirements came into force on March 29, 2016. By November 1, 2016, single households with incomes over 150,000 were ineligible for any subsidy. Table 3.8 shows that this change would have impacted over 50% of historic EV purchases.

Other Incentives

The benefits from the programs described above are included in our analysis. There are some other incentives we do not explicitly model in the paper. For example, limited numbers of PHEVs and fully electric vehicles can gain access to HOV lanes on California highways.[5]

3.3 Electric Vehicle Purchase Patterns

Before laying out the structural model of the paper, we first illustrate some important features about consumers in the market for EVs we have claimed the previous literature has missed. The first is how responsive different income groups are to subsidies on electric vehicles. The second is the effective coverage area of a charging station. That is, unlike a gas station which requires a few minutes to fill up, the specific location of a charging station may be highly important to consumers as charging requires several hours.

Using zip code-quarterly-car model-level data from IHS, we regress a count of each electric vehicle model's purchases in a given zip code on characteristics and median demographics, including distance to work and income bracket, for the zip code. A full description of the source for these variables is in Section 3.4. Each specification counts charging stations in a home and associated work zip codes within varying radii of the zip codes' centroids. See Appendix C.1 for a full description of their construction.

[5] In fact according the Center for Sustainable Energy (2013a) survey of PEV users, 59% claimed HOV access was an "important" consideration in their purchase decision. Sheldon and DeShazo (2017) claims that up to a quarter of registrations in California from 2010 to 2013 could be attributed to this HOV policy.

We report the results of the Poisson regression in Table 3.9. The reported coefficients are the exponent of the coefficient from the Poisson model. Each coefficient can be interpreted as the relative responsiveness from changes in that variable relative to a baseline. For example, a coefficient of 1.786 on 2nd income quantile means that compared to the lowest income quantile, this group is 78% (1.786-1) more likely to purchase an EV. Hence, coefficients less than 1 imply a negative impact of that variable on purchases.

First, we find that results are highly sensitive to the radius used to calculate relevant charging stations and to the location of those charging stations. As we tighten the radius for the calculation, for example, we find that the relationship of home chargers and purchases increases in magnitude but decreases in significance. Work stations meanwhile exhibit the opposite trend. Depending on the selection of this radius, one can overstate the importance of home area chargers and simultaneously miss the relative importance of work-area charging stations.

Second, we find what appears to be the standard result for the income dimension. In the interaction of the income quantiles with price net subsidies, we find higher income zip codes are less responsive to price than lower income are. Simultaneously, from the income quantile terms we see that the "taste" for EVs also grows with income group. To check how the two forces interact, we simulate a new $1000 subsidy on all electric vehicles. We find, according to this simple model, the subsidy would yield 0.04 additional electric vehicles per thousand in the upper quantile and only 0.008 new purchases in the lowest quantile. Compared to the initial distribution of vehicles this represents an increase in 40% for the lowest group but only 18% for the highest income group. For the social planner that cares principally about introducing more EVs to the market, these results suggest the money should be directed toward the highest quantile.

3.4 Data

We draw vehicle purchase data from a number of sources. The primary data source is IHS, which provided us with monthly registration data by car model and fuel type across all California zip codes. This is the dataset used in the generation of the descriptive results of Table 3.9. In total the data cover 1314 unique zip codes from the years 2014

to 2016.

For each vehicle we collect the MSRP of the base model as the base price and various standard characteristics from AutoTrader.com and Edmunds. The characteristics included in the estimation are horsepower-weight, length-width ratio, drive type dummy (2 or 4 wheel base), miles per dollar, driving range, and fuel type. The numerous incentive programs for electric vehicles described in section 3.2 require adjustments to the baseline specifications for these vehicles. In particular, we consider electric vehicles characteristics taking into account 1) the federal and state rebate, 2) the cost of installing a charging station for their home, and 3) miles per dollar using home charging prices at the cheapest price under their energy provider.[6] We get these rates from the dominant energy provider in the county.

Charging station data is collected from the Alternative Fuels Data Center (AFDC), which in turn collects from major charging station operators Blink, SemaCharge, and Chargepoint, and Open Charge Map (OCM). These data sets provide longitude and latitude coordinates thus allowing great flexibility in how we spatially aggregate the data for use in the model. Data on charging station opening dates are provided in older snapshots (before 2014) of the datasets. More recent snapshots do not include opening dates and deduced by assuming stations that appear between snapshots, which we collect twice a week.[7]

Central to identifying heterogenous consumers responses to incentive programs is capturing critical demographics, including income, travel habits, etc. Our first source is survey data from the California Air Resources Board Clean Vehicle Rebate Project (CVRP). Each record in the dataset includes the county of the purchaser, the vehicle make, which typically yields the vehicle model, and the date of purchase. The survey additionally includes a wealth of demographics, including self-reported income brackets. We also utilize the 2010 to 2012 California Department of Transportation CA Household

[6] It might be unreasonable to assume the consumer always charges at the cheapest rate at home, but this method also accounts for the possibility that a lot of charging is actually free from work chargers. We assume consumers install Level 2 chargers at home based on the Center for Sustainable Energy (2013a) survey of CVRP benefactors that found approximately 90% of respondents own a Level 2 home charger for their PEV.

[7] Operators provide the AFDC with opening dates of new stations, but OCM also relies on "community submissions" which might be less reliable than the information provided by operators. One may also worry OCM stations are consistently reported "late" after there is enough of an electric vehicle community in the area to report the station.

Travel Survey (CHTS).[8] In total the 2012 to 2013 survey covered 42,431 households.[9] We use both sets of data to construct aggregated statistics on the income of new vehicle and electric vehicle purchasers. To simulate income data for the model we use American Community Survey (ACS) data on income distributions by census block group.

To control for the travel habits of consumers, we utilize the Longitudinal Employer-Household Dynamics Origin-Destination Employee Statistics (LODES). The dataset, of which we use the 2014 version, lists by home-work census block pairs the number of employees. We use this information to generate information on employee driving distances, since range anxiety is likely exacerbated for those with longer commutes, and to generate counts of charging stations at work locations. The latter point is discussed further in Appendix C.1.

Finally, we control for how environmental concern may increase willingness to pay for clean vehicles. A growing literature documents the role of political preferences in "green consumption." Costa and Kahn (2013) found liberal communities are more likely to participate in "voluntary restraint", that is consume less electricity than more conservative but otherwise identical households.[10]

Therefore, we follow Sexton and Sexton (2014) and others in using political preferences to proxy for concern about the environment. Table 3.11 and Figure 3.6 both provide descriptive evidence regarding how political preferences and income might affect PEV purchases. The second and third columns in Table 3.11 show that more PEVs are sold in high income counties compared to relatively low income counties. However, income is not the only factor explaining the PEV purchase behavior. The fourth column reveals that San Francisco, Marin, and Sacramento counties prefer LEAF over Volt while Orange and Riverside prefer Volt over LEAF. A critical difference in the makeup of these counties are political preferences. The former counties tend to have more Democrats than Republicans whereas the latter counties have more Republicans than Democrats. This is shown in Figure 3.6. The left panel of contains two wealthy counties (San Francisco and Marin and Orange) but with different average political tendencies. The right panel has two less wealthy counties (Sacramento and Riverside) also

[8] See Table 3.10 in Appendix 3.9 for survey size details.
[9] Despite the name of the survey, households were surveyed from 2012 to early 2013.
[10] See also Kahn (2007).

with different political preferences. The statistics suggest that income affects the external margin of the PEV purchase decision while political preference may affect which PEV to purchase.

For our estimation consumers are split into three groups — Republicans, independents, and Democrats — with the latter presumably the most concerned about environmental issues. We use data from the Federal Election Commission (FEC) to develop a distribution of political tastes by zip code. The FEC data provide information on all political contributions over $250. Records are also sufficiently detailed to match most contributions with a recipient political party.

3.5 Demand Specification

In the static model consumers make choices in a specific market m defined by a time q and location g. Our estimation will ultimately consider each zip code-quarter pair in California a specific market.[11] Let $q(m)$ denote the quarter of market m, and $g(m)$ denote the zip code of market m. Each consumer has several relevant characteristics: income y_i, political affiliation $poli_i$, and commuting distance d_i. Additionally, at a specific time, the consumer may take advantage of public or private home and work charging stations. Let $D_i = (y_i, poli_i, d_i, c_i^{pub,h}, c_i^{pri,h}, c_i^{pub,w}, c_i^{pri,w})$. Specific details on how we model which charging stations are relevant to consumers are in Appendix C.1. Consumer i in market m purchases one product from the choice set J_m or an outside good to maximize her utility. While all product characteristics of a vehicle are fixed within a year, the price subsidy available may differ both across quarter and location.

When purchasing a vehicle $j \in J_m$ the utility of consumer i in market m has the following form:

$$
\begin{aligned}
u_{ij} =& -\alpha_0 p_{ij} - \sum_b \alpha_b p_{ij} \cdot 1_{\{y_i \in Y_b\}} + PEV_j \cdot \mu_i' \beta^{PEV} + PHEV_j \cdot \mu_i' \beta^{PHEV} + \delta_{jm} + \varepsilon_{ij} \\
\mu_i' =& [\mathbf{1}_{\{y_i \in Y_b\}}, D_i, d_i \cdot c_i^{pub,w}, d_i \cdot c_i^{prv,w}] \\
\delta_{jm} =& X_j' \beta + \xi_{jm} \\
X_j =& [1,\ hp_j,\ lw_j,\ range_j,\ mp\$_j,\ PEV_j,\ PHEV_j].
\end{aligned}
$$

[11] Since zip codes are not geographically contiguous areas, we specifically use zcta5 designations.

The index b is used to indicate income bin. We choose this specification over the standard log form to permit non-monotonicity in price elasticity over income groups. X_j is a vector of characteristics for vehicle j, including standard quality controls and indicators for whether the vehicle is a fully electric vehicle (PEV_j) or a hybrid electric vehicle ($PHEV_j$). ξ_{jm} is an unobserved product-specific demand shock, which can be correlated with net price p_{ij} and ε_{ij} is an idiosyncratic preference shock. Although we already define vehicle j as specific to a market, we leave the subscript m in to emphasize that a single car model may have characteristics that change with this specific market. The full set of parameters of interest in this utility model are

$$\theta = (\alpha_0, \{\alpha_b\}_b, (\beta^{PEV})', (\beta^{PHEV})', \beta')'.$$

In Equation 3.1 the terms specific to PEVs are highlighted to emphasize additions to the standard models.

$$\begin{aligned} u_{ij} &= \delta_{jm}(\theta) + V_{ij}(\theta) + EV_{ij}(\theta) + \varepsilon_{ij} \qquad (3.1) \\ \text{where} \qquad \delta_{jm} &= X_j'\beta + \xi_{jm} \\ V_{ij}(\theta) &= -\alpha_0 p_{ij} - \sum_b \alpha_b p_{ij} \cdot 1_{\{y_i \in Y_b\}} \\ EV_{ij}(\theta) &= PEV_j \cdot \mu_i' \beta^{PEV} + PHEV_j \cdot \mu_i' \beta^{PHEV} \end{aligned}$$

The interactions in $EV_{ij}(\theta)$ allow additionally flexibility in consumer elasticity to price subsidies and charging stations. As we found in Section 3.3, while lower income individuals are more price elastic, the net effect of a subsidy is higher for high income individuals because of their significantly higher taste for EVs. These interaction terms leave a horse race between price elasticity and tastes open to the empirical findings.

To construct shares from the model we impose the additional that ε_{ij} shocks are i.i.d. type-1 extreme value over products and consumers. Under this assumption the probability consumer i in market m purchases vehicle j takes on the familiar logit form.

$$P_\theta(j|W_{ij}, D_i) = \frac{\exp(\delta_{jm}(\theta) + V_{ij}(\theta) + EV_{ij}(\theta))}{1 + \sum_{j' \in J_m} \exp(\delta_{j'm}(\theta) + V_{ij'}(\theta) + EV_{ij'}(\theta))}$$

where $W_{ij} = (X_j, p_{ij})$. Integrating over all of the individuals in a market yields the

aggregate market share. We allow the distribution of individual characteristics to differ by market according to $f_m(y, poli, d, c)$ and assume the distribution of each component is independent from the others. Hence $f_m(y, poli, d, c) = f_m(y)f_m(poli)f_m(d)f_m(c)$. We now add the assumption that price discounts are the same for any consumers in the same market.[12] This implies $W_{ij} \equiv W_{mj}$ for all i in the same market. Under these assumptions the market share for vehicle j in market m is

$$S_j^m(\theta|W_{mj}) = \int P_\theta(j|W_{mj}, D_i) f_m(y, poli, d, c) d(y, poli, d, c)$$

3.6 Estimation

3.6.1 Reducing the Parameter Space

To simplify the estimation we reduce our parameter space by a standard technique in the literature. Recall from Equation 3.1, $\delta_{jm}(\theta) = X_j\beta + \xi_{jm}$, the product-specific term common to all consumers in a market. Because the mean parameters β are linear in δ, we can back out estimates for them after a nonlinear search over δ. Therefore, we refine $\theta = (\alpha_0, \{\alpha_b\}_b, (\beta^{PEV})', (\beta^{PHEV})')'$. Searching over the J components of δ is expensive, however, so we utilize the share inversion technique introduced by Berry (1994) to "concentrate out" these parameters. This technique requires the restriction that CA-level vehicle shares should match the predicted shares generated by the model at the true parameter values. That is

$$S_j^{DATA} - S_j^m(\delta_{jm}, \theta_0) = 0 \quad \forall j, m \tag{3.2}$$

Given our distributional assumption on consumer tastes, Berry (1994) demonstrates that for each θ, there is a unique $\delta(\theta)$ such that Equation 3.2 holds. This technique proves useful not only as a mechanism to reduce the parameter space but also to mitigate the endogeneity problem with price by conditioning on the component of the error with which it is correlated, i.e. ξ_{jm}. Given this restriction three types of moments identify the parameters θ.

[12] Effectively this is an assumption that everyone will take up the subsidy offered.

3.6.2 Maximum Likelihood Moments

The first set of moments are the score of a maximum simulated likelihood estimator of θ using the IHS micro purchase data $M = \{M_i\}_{i=1}^N$. Estimators using this sample can take advantage of much greater variation in charging stations and purchasing observations across zip codes and quarters than could aggregated data.

Because we observe the time of purchase, model of purchase, and location of individuals in this dataset, it is straightforward to write down a likelihood function. For consumer i in market m let $I_{ij(i)}$ be a vector of length J_m by 1 of purchase indicators (1 for the vehicle the consumer purchased and specially denoted $I_{ij(i)}$). Conditional on all relevant characteristics of person i, i.e. D_i, the log probability of observing I_i under parameter θ is given by

$$\frac{1}{N}\sum_{i=1}^{N} L(\theta; M_i) = \frac{1}{N}\sum_{i=1}^{N} I_{ij(i)} \log(P_\theta(j(i)|W_{jm}, D_i))$$

Following the argument of Goolsbee and Petrin (2004) we claim that conditioning purchase probability calculations on the value of $\delta(\theta)$ described previously eliminates the specification error that might arise because of the endogeneity of price.

However, the available data in M_i for the individual does not include income, political affiliation, or work location so the probability of purchase must still be simulated for each individual by integrating over the distribution $f_m(y, poli, d, c)$. With the simulated probability $\hat{P}_\theta(\cdot|W_{jm}, M_i)$, the sample log likelihood function is

$$\frac{1}{N}\sum_{i=1}^{N} \hat{L}(\theta; M_i) = \frac{1}{N}\sum_{i=1}^{N} I_{ij(i)} \log\left(\hat{P}_\theta(j(i)|W_{jm}, M_i)\right)$$

The score of the log likelihood function generates moment conditions for each element of θ

$$E[\psi_1(\theta_0, M_i)] \equiv E\left[\frac{\partial L(\theta_0; M_i)}{\partial \theta_0}\right] = 0 \tag{3.3}$$

with corresponding sample moment at arbitrary θ, $\frac{1}{N}\sum_{i=1}^{N} \psi_1(\theta, M_i)$.

3.6.3 Matching Income Distribution Moments

The second set of moments takes advantage of aggregated demographic statistics derived from survey participants in the CHTS and a survey of CVRP participants, data sets denoted by $A = \{A_r\}_{r=1}^{R}$ and $B = \{B_l\}_{l=1}^{L}$, respectively. Following the insight of Imbens and Lancaster (1994) these aggregated statistics are simply aggregations of micro data; hence aggregate model predictions should match these statistics. Similar to matching moments considered by Petrin (2002) we consider three conditions matching observed income distributions of various categories of car purchasers against model predictions.

The first condition matches 5 income category densities conditional on purchasing a Tesla or another electric vehicle. Using the data from the CVRP survey and invoking Bayes' rule, i.e. $P(A|B) = P(A,B)/P(B)$, the sample statistics derived are

$$\hat{\mu}(y_i \in Y_b | i \text{ purchases EV}) = \frac{\sum_{i=1}^{L} 1(i \text{ purchases EV}) * 1(y_i \in Y_b)}{\sum_{i=1}^{L} 1(i \text{ purchases EV})}$$

for $b \in \{0, \ldots, 4\}$. We assume that the samples are unbiased so the sampling error is 0 in expectation.[13] We will denote the population version of this probability distribution by $\mu = E[\hat{P}(y_i \in Y_b | i \text{ purchases EV})]$, where the expectation is taken over sampling error.

The corresponding statistics generated by model predictions are

$$P_\theta(y_i \in Y_b | i \text{ purchases EV}) = \frac{\sum_{m=1}^{M} \pi_m \sum_{i=1}^{ns} 1(y_i \in Y_b) * \sum_{j \in \{J_m \cap EV\}} P_\theta(j|D_i)}{\sum_{m=1}^{M} \pi_m \sum_{i=1}^{ns} \sum_{j \in \{J_m \cap EV\}} P_\theta(j|D_i)}$$

where π_m is a population weight for the particular market.

The moment restriction imposes that at the true parameter the model's aggregated statistic and the population statistic should be equal.

$$E[\psi_2(\theta_0, B)] \equiv \mu(y_i \in Y_b | i \text{ purchases EV}) - P_{\theta_0}(y_i \in Y_b | i \text{ purchases EV}) = 0$$

We similarly construct matching moments $E[\psi_3(\theta_0, A)] = 0$ using CHTS data, which

[13] This assumption only requires 1) the CVRP participants taking the survey are not a biased sample of the pool of all CVRP participants and 2) the CHTS sample of households is also not a biased sample of auto purchasers in California.

provides information on gas vehicle purchases.

3.6.4 Subsidy Instrument Moments

The third set of moments are instruments based on two sets assumptions. Following in the spirit Li (2017) we assume that each EV's demand shock, ξ_{jm}, in a market is independent of the price subsidy for that vehicle in the same market. First, federal and state-wide subsidies are set independent of idiosyncratic granular market demand. Second, local subsidies are frequently set to combat pollution issues. For example, the San Joaquin River Valley has the highest local subsidies, as shown in Figure 3.5. It also has the worst pollution in California for reasons only in part from vehicle emissions.[14] Let s^v_{jm} be the subsidy for vehicle j in market m. Then, at true parameter θ_0 we have the moment condition

$$E[\psi_4(\theta_0, M)] \equiv E[\xi_{jm}(\theta_0) s^v_{jm}] = 0$$

To construct the second instrument moments, we assume that the innovation in unobserved demand shocks evolve as a random walk:

$$\nu_{jm}(\theta) = \xi_{jg(m)t(m)}(\theta) - \xi_{jg(m)t(m)-1}(\theta)$$

where $g(m)$ and $t(m)$ are explicitly indexed to note an evolution of the shock in a district over time. To generate the moment condition we assume that the innovation of the shock is uncorrelated with the subsidy for the charging station s^c_{jm}, as a cost shifter, in the market.

$$E[\psi_5(\theta_0, M)] \equiv E[\nu_{jm}(\theta_0) s^c_{jm}] = 0$$

[14] See https://www.citylab.com/environment/2011/09/behind-pollution-californias-central-valley/207/

3.6.5 GMM Estimator

These five sets moments can be stacked into a single vector. Formally, we assume that θ_0 uniquely satisfies the population moment conditions described above.

$$E\left[\psi(\theta_0, U)\right] = E\begin{bmatrix}\psi_1(\theta_0, M)\\ \psi_2(\theta_0, B)\\ \psi_3(\theta_0, A)\\ \psi_4(\theta_0, M)\\ \psi_5(\theta_0, M)\end{bmatrix} = 0 \tag{3.4}$$

with sample analog $\hat{\psi}(\theta, U)$, where $U = (M, A, B)$. The GMM estimator $\hat{\theta}$ solves the following criterion function

$$\hat{\theta} = \arg\min_{\theta} \hat{\psi}(\theta)' W \hat{\psi}(\theta)$$

where W is a weighting matrix. In practice we follow the Hansen (1982) two-step method, by first estimating parameters where W is the identity matrix and then with W as the inverse of the asymptotic variance matrix of the moments derived from the first step.

3.7 Identification

3.7.1 Identifying Heterogeneity

Identifying parameters on terms related to PEVs rely on our data specific to PEV purchases, in particular our rich micro dataset from the CVRP. Because electric vehicles, particularly plug-in hybrid electric vehicles, are similar in performance and range to traditional gas-powered vehicles, differences in shares of these electric vehicles from similar gasoline vehicles must be driven largely by PEV-specific tastes.

We also observe significant variation in purchases of electric vehicles as a category and within that category across counties and time. Across geographic markets consumers differ on three dimensions that can explain purchasing behavior: charging

stations, income, and political affiliation, and importantly subsidies that shift costs differentially across the state. The latter two vary most across geographic markets rather than time, while charging stations vary significantly even over short periods of time as well as across counties. Fixing the number of charging stations at a snapshot in time can thus help pin down income and political orientation parameters.[15]

Significant variation in the number of charging stations across time within markets can pin down parameters relating to charging stations. Because markets are defined by time periods as short as a quarter, we do not worry about general improvements in the perception of electric vehicle usage absorbing most of the change in purchase behavior of PEVs from quarter to quarter. Additionally, given a market, the only area-specific factors that significantly change over this short time period are the number of charging stations.[16]

3.7.2 Endogeneity

Two sources of endogeneity could potentially affect the estimation procedure. The first is the typical assumption that unobserved product characteristics ξ_j are correlated with price p_j. *A priori* these two variables should be positively correlated and generate a positive bias in price elasticity (price elasticity is less negative). Because electric vehicles still tend to carry a higher price than similar gasoline equivalents, the model with bias price would predict consumers are less hesitant to buy high-price electric vehicles than at the true parameter. Ultimately, the bias can force down the magnitude of PEV-specific utility terms to explain the low shares of electric vehicles. As detailed in section 3.6 we address this issue with the standard tactic in this literature by directly specifying the component of the error with which price is correlated (ξ_j) in our calculation of demand.

A more serious issue for analysis is the potential that charging stations are also endogenous. While we expect demand to increase the number of charging stations, it is also reasonable to suspect the number of charging stations in an area are driven by local

[15] See again, for example, Table 3.11.

[16] One concern is that charging stations are always increasing over time; we do not observe many instances of charging stations being shut down. We intend to make our results more robust by considering a more natural definition for charging stations, which considers congestion. Effective charging stations by this definition do not necessarily increase over time hence breaking the potential conflation of time and charging stations.

demand. The final set of moments in the estimation aim to address this issue directly by claiming that local area CS subsidies are independent from temporary demand shocks.

We also argue that special features of the market mitigate these concerns. First, government programs determined the location of many early charging stations in the market. In conversations with representatives involved in these projects, we learned that the government targeted large areas, such as Los Angeles or San Diego, for receipt of charging stations but specific locations were determined independent of demand. The projects placed charging stations wherever willing partners could be found. The potential endogeneity problem can then be mitigated by 1) blunt placement of charging stations and 2) considering consumers at the high level of granularity we do.

Finally, charging stations might trend with market-level shocks favoring electric vehicles. Since ξ_j for PEV vehicles is precisely this shock, specifying this component in the calculation of demand mitigates the endogeneity problem.[17]

3.8 Conclusion

This paper aims to estimate the heterogeneous effect of subsidy and investment on charging station infrastructure and suggests the optimal allocation of funding to promote electric vehicle diffusion. This draft takes a preliminary step toward that goal by setting up a rich discrete choice model of demand with heterogeneous tastes toward the charging station network as well as price elasticities in a static setting. These modeling decisions are informed by the results in Section 3.3, suggesting the importance of these dimensions of heterogeneity. Even this preliminary model can use zip code level vehicle registration data to identify marginal consumers with respect to subsidy, characterized by income, political preference, or residential location. We also specify the marginal benefit of charging stations by their locations.

We believe this paper will offer a number of contributions to the existing literature. First, we can evaluate the potential loss of current incentive programs. Unlike almost all other government subsidy (tax) schedules, most EV incentives do not discriminate

[17] Of course, if the number of charging stations is actually highly correlated with shocks specific to a tight array, specifying ξ_j will not completely eliminate the endogeneity problem. In this case, we would need a more sophisticated model of how charging stations are placed. A future iteration on this paper may revisit that question.

based on demographics and geography. We plan to evaluate the loss from the flat incentive structure compared to a socially optimal schedule, perfect discrimination based on consumer characteristics and location. We can also measure the loss by comparing it to several constrained optima more in line with realistic policy, since discrimination based on political views, for example, is untenable.

Second, we can provide a practical guideline in designing subsidy and investment in infrastructure that can be implemented. The federal tax credit, state rebate programs, and other incentive programs are still ongoing and are constantly changing their design to target different groups of potential EV buyers.

3.9 Tables and Figures

Table 3.1: **Characteristics of Top 4 PEV and PHEV Models in California**

Model	Type	Q[a]	Share	Price	Range	MP$[b]
Tesla Model S	PEV	9013	22.13	64931	256.57	18.95
Nissan Leaf		7046	17.30	27929	84.68	22.44
Tesla Model X		6073	14.91	74000	200.00	18.03
Fiat 500e		5780	14.19	28721	85.80	22.89
Chevrolet Volt	PHEV	9812	34.63	32600	396.67	10.89
Toyota Prius Plug-in		6745	23.80	27484	540.00	13.74
Ford C-Max		5432	19.17	22031	554.00	11.39
Hyundai Sonata Plug-in Hybrid		2115	7.46	34600	595.00	13.29

[a] *Source*: IHS dataset of registrations by car model. Data aggregated over 2014 to 2016.
[b] Miles per dollar (MP$) is calculated assuming average time-of-use rate offered by utility companies in California. Miles per gallon for PEVs is substituted by MPGe, which uses the equivalency 33.7kWh = 1 gallon.

Table 3.2: **Charging Speed by Charger Type**

Charger Type	Vehicle	Level 1	Level 2	DC Fast
Full Charge Time	LEAF	20 hr	8 hr	30 mins[a]
	Volt	10 hr	4 hr	15 mins[a]
Distance with 1 hr of Charge		2-5 mi	10-20 mi	60 mi
Home Installation Fee		Free	∼$1700	N/A

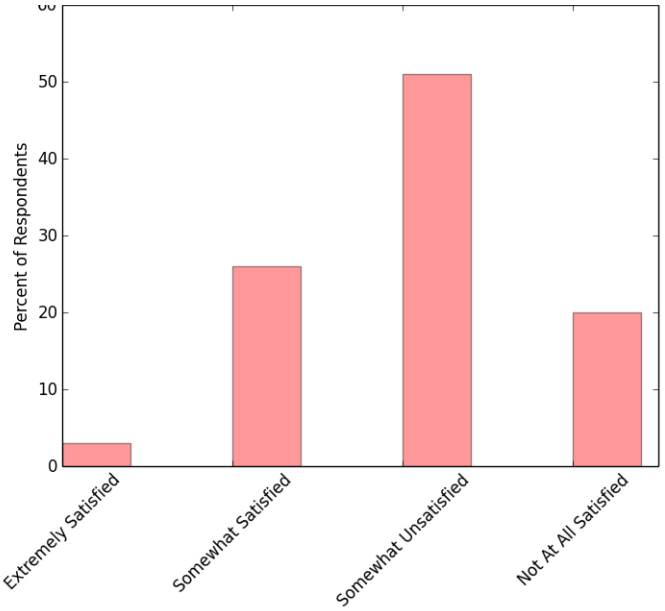
[a] Time for battery to be 80% charged.

Table 3.3: **Desired Range Vs. Actual Range**

	Leaf	Volt	Prius
Desired Electric Range	200	100	50
Actual Electric Range	78	38	12
Dissatisfaction with Public Charging Infrastructure	71%		

Source: California Plug-in Electric Vehicle Driver Survey Results, February 2014

Figure 3.1: **Dissatisfaction with Charging Infrastructure**



Source: California Plug-in Electric Vehicle Driver Survey Results, February 2014

Table 3.4: **Cumulative Charging Stations in California**

Year	Quarter	Cumulative Stations	Cumulative Private	Cumulative Public
2010	1	734	364	370
2010	2	746	370	376
2010	3	752	370	382
2010	4	799	373	426
2011	1	927	384	543
2011	2	1243	486	757
2011	3	1410	531	879
2011	4	1759	569	1190
2012	1	2220	657	1563
2012	2	2770	765	2005
2014	1	4908	992	3916
2014	2	5214	1005	4209
2014	3	5659	1056	4603
2014	4	6355	1124	5231
2015	1	7008	1197	5811
2015	2	7803	1319	6484
2015	3	8792	1377	7415
2015	4	9382	1477	7905
2016	1	10634	1686	8948
2016	2	11294	1719	9575
2016	3	12182	1815	10367
2016	4	12904	1862	11042

Source: AFDC and OpenChargeMap

Table 3.5: **Sponsored Charging Infrastructure Projects**

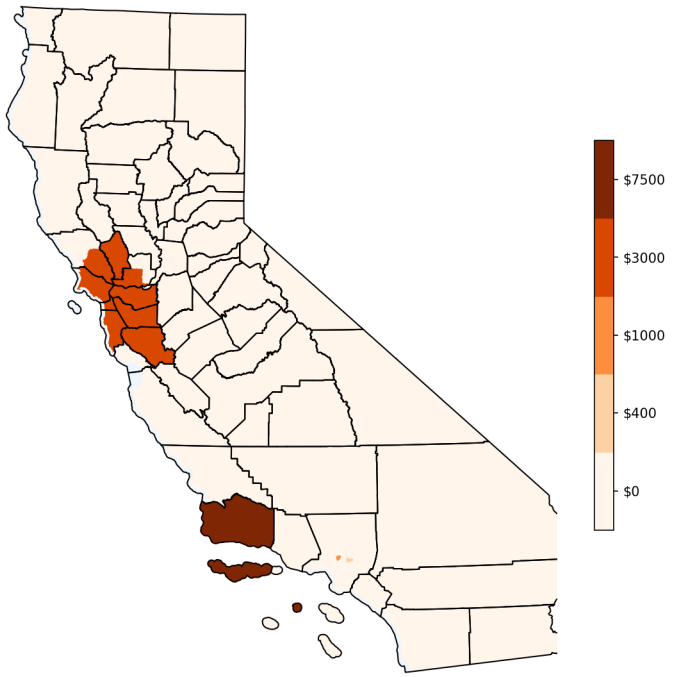
	EV Project	ChargePoint America
Project Period	January 2011 - December 2013	May 2011 - June 2013
Area Covered in CA	Los Angeles, San Diego San Francisco	Los Angeles Sacramento, San Francisco
Charging Network	Blink	ChargePoint
Funding Amount	$130 million dollar[a]	$18.4 million dollar
Total Charging Units[b]	3182	1916
Public Charging Units	933	857

Sources: Project Electric Vehicle Charging Infrastructure Summary Report (Q4 2013), ChargePoint America Vehicle Charging Infrastructure Summary Report

[a] Total budget was $230 million and half of it was funded by the DOE. $130 was allocated to install public or private charging stations. The rest is operational cost and subsidy for residential chargers.

[b] Charging units are counted only in California.

Figure 3.2: **Public Charging Station Subsidy Availability, December 2016**



Source: Local subsidies by zip code are available on driveclean.ca.gov.

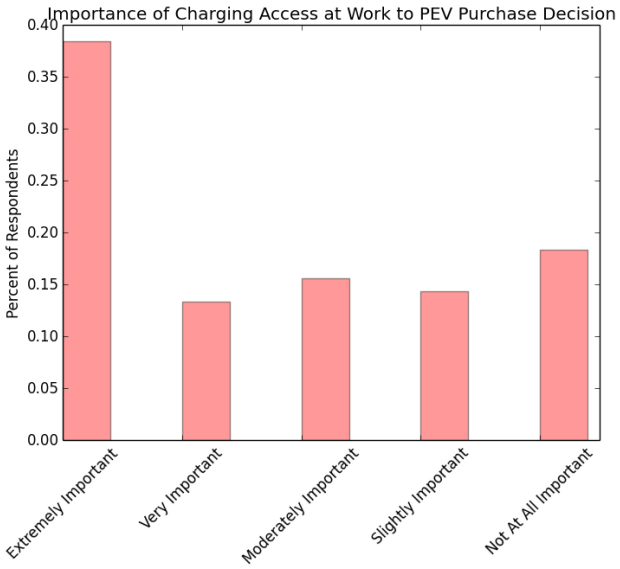
Table 3.6: **Charging and Driving Behavior**

	LA		SD		SF		CA[b]
	Leaf	Volt	Leaf	Volt	Leaf	Volt	Gas Only
Avg Trip Distance[a]	6.4	7.7	6.6	8.2	7.3	9.39	8.4
Avg Distance per Day	24.8	38	26.7	39.6	27.4	41.4	38.48
Avg Number of Trips b/w Charging Events	3.8	3.8	3.7	3.7	3.5	3	
Avg Distance b/w Charging Events	24.2	28.9	24.6	30.1	25.7	28.1	
Avg Charging Events per Day	1	1.3	1.1	1.3	1.1	1.5	
% Charging Events Away from Home	30	23	23	21	28	21	

Sources: EV Project Nissan Leaf Summary Report, Q4 2013; EV Project Chevrolet Volt Summary Report, Q4 2013; 2010-2012 California Household Travel Survey

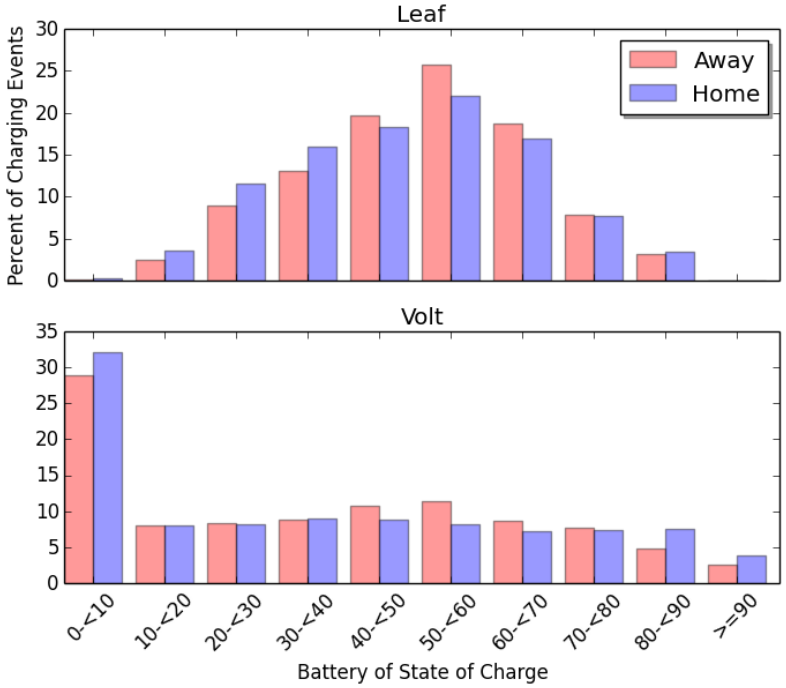
[a] Distance is measured in miles. https://pbs.twimg.com/media/DdQHH6yWAAAmxB6.jpg

Figure 3.3: **Charging Access at Work is Important**



Source: California Plug-in Electric Vehicle Driver Survey

Figure 3.4: **Distribution of Battery Charge at the Start of Charging Events**



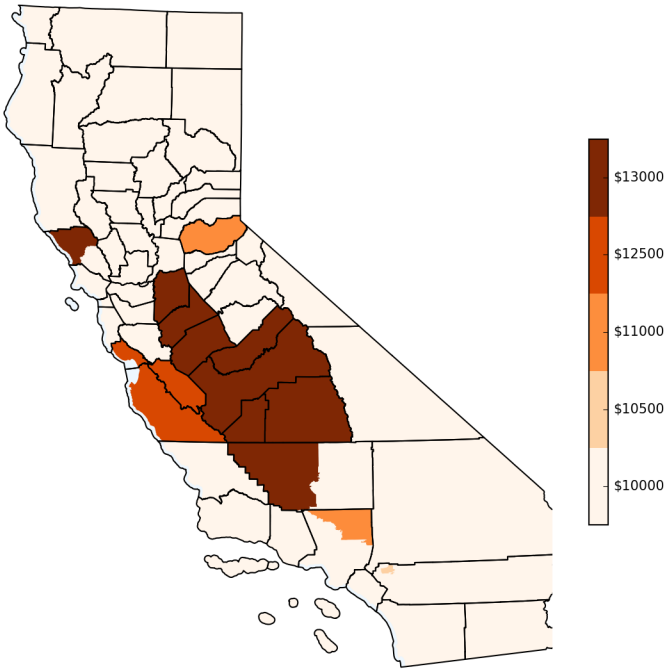
Source: EV Project Electric Vehicle Charging Infrastructure Summary Report, Q4 2013

Table 3.7: **Government Monetary Incentives for PEV**

	Clean Vehicle Rebate Project	Federal Tax Credits[a]
Project Period	2010 - current	2009 - current
Area Covered in CA	All	All
Funding Institute	California Air Resources Board	IRS
Funding Amount	\$790 million dollar[b]	Until manufacturer sells 200,000
Rebate Amount	\$1,500 to \$2,500[c]	\$2,500 to \$7,500
Tesla/Leaf Rebate	\$2,500	\$7,500
Prius Rebate	\$1,500	\$2,500
Eligibility	Zero emission, plug-in hybrid	Zero emission, plug-in hybrid
	Lease $\geq$ 36 months	Battery capacity $\geq$ 5Kilowatt-hour
Total Rebates Issued[d]	236,258	

[a] Plug-in Electric Drive Vehicle Credit (IRC 30D) by IRS
[b] Funding allocation during FY 2010 - 2018.
[c] Starting March 29, 2016 income cap has been implemented and rebates have increased for low and moderate income buyers.
[d] As of February 28, 2018.

Figure 3.5: **Full-Electric Vehicle Subsidy Available, December 2016**



Source: Local subsidies by zip code are available on driveclean.ca.gov.

Table 3.8: **Income Distribution Conditional on a Vehicle or PEV Purchase**

Income 2010-2012	Vehicle Purchase	PEV Purchase
Less than $74,999	0.19	0.07
$75,000 to $99,999	0.16	0.10
$100,000 to $149,999	0.27	0.24
$150,000 to $199,999	0.20	0.23
$200,000 to $249,999	0.09	0.15
More than $250,000	0.10	0.22
Number of Obs	278	92

Source: California Department of Transportation CA Household Travel Survey

Income 2012-2014	Overall	Tesla
Less than $49,999	0.03	0.01
$50,000 to $74,999	0.05	0.02
$75,000 to $99,999	0.10	0.03
$100,000 to $149,999	0.25	0.11
$150,000 to $199,999	0.19	0.13
$200,000 to $299,999	0.20	0.22
$300,000 to $399,999	0.08	0.13
$400,000 to $499,999	0.03	0.07
$500,000 or more	0.07	0.28
Number of Obs	5596	1052

Source: Clean Vehicle Rebate Project EV Owner Demographics and Diffusion Survey

Table 3.9: **Poisson Regression of Electric Vehicles Sold, by Model**

	Radius Around Zip Code Centroid					
	20mi	10mi	5mi	2mi	1mi	.5mi
2nd Inc. Quantile	1.786***	1.800***	1.784***	1.782***	1.780***	1.773***
	(0.250)	(0.250)	(0.242)	(0.240)	(0.240)	(0.237)
3rd Inc. Quantile	2.345***	2.355***	2.342***	2.357***	2.359***	2.347***
	(0.327)	(0.314)	(0.314)	(0.319)	(0.323)	(0.316)
Top Inc. Quantile	3.259***	3.261***	3.259***	3.275***	3.280***	3.247***
	(0.496)	(0.472)	(0.474)	(0.478)	(0.485)	(0.470)
Net Price	0.965***	0.965***	0.966***	0.966***	0.966***	0.966***
	(0.00368)	(0.00373)	(0.00375)	(0.00373)	(0.00373)	(0.00373)
2nd Inc. Quantile x Net Price	1.000	1.000	1.000	0.999	0.999	0.999
	(0.00186)	(0.00188)	(0.00190)	(0.00192)	(0.00192)	(0.00191)
3rd Inc. Quantile x Net Price	1.005***	1.005***	1.005***	1.005***	1.005***	1.005***
	(0.00200)	(0.00203)	(0.00205)	(0.00205)	(0.00205)	(0.00204)
Top Inc. Quantile x Net Price	1.017***	1.017***	1.017***	1.017***	1.017***	1.017***
	(0.00342)	(0.00347)	(0.00349)	(0.00347)	(0.00346)	(0.00345)
Horsepower / Weight	1.33e-07***	1.31e-07***	1.33e-07***	1.34e-07***	1.33e-07***	1.33e-07***
	(2.53e-07)	(2.51e-07)	(2.55e-07)	(2.58e-07)	(2.57e-07)	(2.56e-07)
Length / Width	7.548***	7.545***	7.551***	7.553***	7.550***	7.549***
	(1.856)	(1.856)	(1.862)	(1.871)	(1.870)	(1.870)
Range	1.003***	1.003***	1.003***	1.003***	1.003***	1.003***
	(0.000176)	(0.000178)	(0.000177)	(0.000176)	(0.000176)	(0.000176)
Miles per Dollar	0.983	0.983	0.983	0.983	0.983	0.983
	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)
Distance to Work	0.996*	0.99*	0.996*	0.996**	0.996**	0.996**
	(0.00185)	(0.00187)	(0.00192)	(0.00193)	(0.00196)	(0.00198)
Home Stations within Radius	1.002***	1.003***	1.004***	1.005**	1.004	0.997
	(0.000512)	(0.000521)	(0.00144)	(0.00196)	(0.00411)	(0.00570)
Work Stations within Radius	0.999**	1.000	1.002*	1.011***	1.029***	1.085***
	(0.000528)	(0.000532)	(0.00107)	(0.00294)	(0.00758)	(0.0227)
Observations	175747	175747	175747	175747	175747	175747

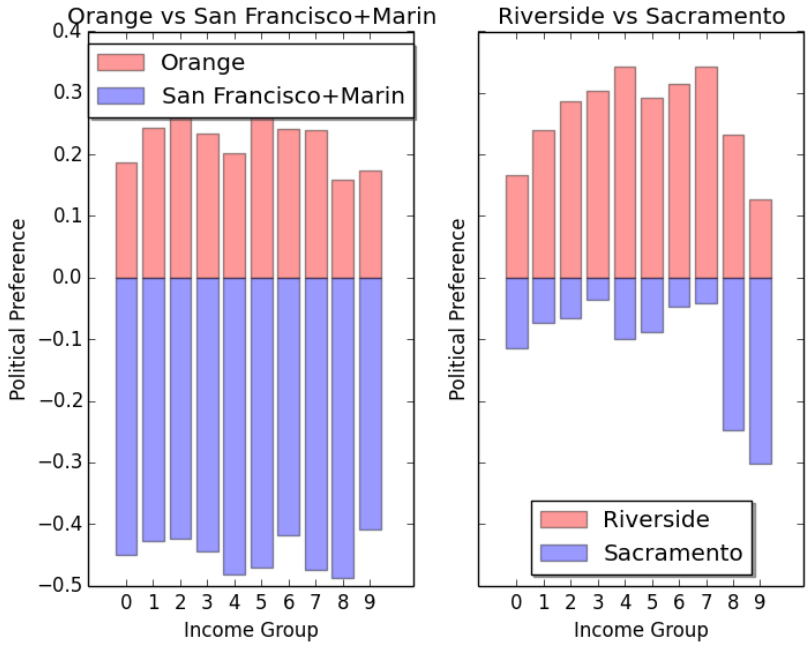
Notes: *** $p < 0.01$, ** $p < .05$, * $p < .1$; Coefficients reported are $\exp(\beta)$. Observations are at the level of zip code-quarter-car model.

Table 3.10: **CHTS and CVRP Survey Summary**

	CHTS[a]	CVRP[b]
Survey Period	2010 - 2012	Oct 2013 - May 2014
Vehicle Purchase Period	1994 - 2012	Sept 2012 - Apr 2014
Number of Respondents	42,431	8,415
Sample Used	1588	6,602
Car Purchase	278	6,602
PEV Purchase	92	6,602

[a] California Department of Transportation CA Household Travel Survey. CVRP surveyed only PEV purchasers who applied for CVRP rebate.
[b] Clean Vehicle Rebate Project EV Owner Demographics and Diffusion Survey.
In the micro income moments we only used the samples which has the income and purchase year are known.

Figure 3.6: **Heterogeneous Income and Political Distribution by County**



Source: Federal Elections Commission
Political preference: -1 if Democrats, 1 if Republicans, 0 if independent
Income group: percentile of FEC donation given county

Table 3.11: **Heterogeneous Income and Political Distribution by County**

County	Income[a]	PEV[b]	LEAF/Volt[c]
Orange	$96,036	8.44	0.72
San Francisco and Marin	$108,690	5.68	1.54
Riverside	$69,835	2.61	0.62
Sacramento	$68,532	2.73	2.15

[a] Average household income in 2013, ACS
[b] Total PEV sold per 1,000 capita until August 2014, Clean Vehicle Rebate Project rebate dataset
[c] Leaf/Volt: ratio between Leaf and Volt demand

Table 3.12: **LODES Residence and Workplace Summary**

Year	Total	Unique Census Blocks			Unique Zip Codes			Distance	
	People	Home	Work	Avg. Link[a]	Home	Work	Avg. Link[a]	Mean	Med
2012	14.59m	380564	235288	1.01	1757	1742	21.26	22.85	5
2013	15.05m	380762	237490	1.01	1760	1745	21.42	23.59	5
2014	15.47m	381181	240242	1.01	1758	1746	18.77	26.61	5
Total	– –	397471	265663	1.01	1760	1752	20.38	24.39	5

Source: 2012-2014 Longitudinal Employer Household Dynamics Origin-Destination Employment Statistics (LODES)
[a] Measured as the average number of people in home area commuting to the work area

References

Ansolabehere, Stephen and Stephen Pettigrew, "Cumulative CCES Common Content (2006-2012)," 2014.

Arnott, Richard, "Taxi Travel Should Be Subsidized," *Journal of Urban Economics*, 1996, *40* (3), 316–333.

Berger, Thor, Chinchih Chen, and Carl Benedikt Frey, "Drivers of Disruption? Estimating the Uber Effect," *Working Paper*, 2017.

Berry, Steven, "Estimating Discrete-Choice Models of Product Differentiation," *The RAND Journal of Economics*, 1994, *25* (2), 242–262.

_ , James Levinsohn, and Ariel Pakes, "Automobile Prices in Market Equilibrium," *Econometrics*, 1995, *63* (4), 841–890.

Bertrand, Marianne and Sendhil Mullainathan, "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination," *American Economic Review*, 2004, *94* (4), 991–1013.

Bian, Bo, "Search Frictions, Network Effects, and Spatial Competition : Taxis versus Uber," *Working Paper*, 2017.

Bimpikis, Kostas, Ozan Candogan, and Saban Daniela, "Spatial Pricing in Ride-Sharing Networks," *SSRN Electronic Journal*, 2016, pp. 1–40.

Bloomberg, Michael R. and David Yassky, "2014 Taxicab Factbook," 2014.

Bollinger, Bryan and Kenneth Gillingham, "Peer Effects in the Diffusion of Solar Photovoltaic Panels," *Working Paper*, 2012.

Borenstein, Severin and Lucas Davis, "The Distributional Effects of U.S. Clean Energy Tax Credits," *NBER Tax Policy and the Economy*, 2016, *30* (1), 191–234.

Buchholz, Nicholas, "Spatial Equilibrium, Search Frictions and Efficient Regulation in the Taxi Industry," *Working Paper*, 2017, pp. 1–60.

Cairns, Robert D. and Catherine Liston-Heyes, "Competition and Regulation in the Taxi Industry," *Journal of Public Economics*, 1996, *59* (1), 1–15.

Camerer, Colin, Linda Babcock, George Loewenstein, and Richard Thaler, "Labor Supply of New York City Cab Drivers: One day at a time," *The Quarterly Journal of Economics*, 1997, *112* (2), 407–441.

Castillo, Juan Camilo, Dan Knoepfle, and E Glen Weyl, "Surge Pricing Solves the Wild Goose Chase," *Working Paper*, 2017.

Center for Sustainable Energy, "California Plug-in Electric Vehicle Driver Survey Results May 2013," Technical Report May 2013.

_ , "Clean Vehicle Rebate Project Fiscal Year 2012-2013 Final Report," Technical Report 2013.

_ , "California Air Resources Board Clean Vehicle Rebate Project, Rebate Statistics," 2014.

Chen, M. Keith, Judith A. Chevalier, Peter E. Rossi, and Emily Oehlsen, "The Value of Flexible Work: Evidence from Uber Drivers," *Working Paper*, 2017.

Clements, Matthew T. and Hiroshi Ohashi, "Indirect Network Effects and the Product Cycle: Video Games in the US, 1994 - 2002," *The Journal of Industrial Economics*, 2005, *53* (4), 515–542.

Cohen, Peter, Robert Metcalfe, Josh Angrist, Keith Chen, Joseph Doyle, Hank Farber, Alan Krueger, Peter Cohen, and Robert Hahn, "Using Big Data To Estimate Consumer Surplus: The Case of Uber," *Working Paper*, 2016.

Costa, Dora L. and Matthew E. Kahn, "Do Liberal Home Owners Consume Less Electricity? A Test of the Voluntary Restraint Hypothesis," *Economics Letters*, may 2013, *119* (2), 210–212.

Cramer, Judd and Alan B. Krueger, "Disruptive Change in the Taxi Business: The Case of Uber," *American Economic Review*, 2016, *106* (5), 177–182.

Crawford, Vincent P. and Juanjuan Meng, "New York City Cabdrivers' Labor Supply Revisited: Reference-Dependence Preferences with Rational-Expectations Targets for Hours and Income," *American Economic Review*, 2011, *101* (5), 1912–1932.

Diakopoulos, Nicholas, "How Uber Surge Pricing Really Works," apr 2015.

Dubé, Jean-Pierre, Günter J. Hitsch, and Pradeep K. Chintagunta, "Tipping and Concentration in Markets with Indirect Network Effects," *Marketing Science*, mar 2010, *29* (2), 216–249.

Edelman, Benjamin, Michael Luca, and Dan Svirsky, "Racial Discrimination in the Sharing Economy: Evidence from a Field Experiment," *American Economic Journal: Applied Economics*, 2017, *9* (2), 1–22.

EV Project, "EV Project Electric Vehicle Charging Infrastructure Summary Report," Technical Report 2014.

Farber, Henry S., "Is Tomorrow Another Day? The Labor Supply of New York Cab Drivers," *Journal of Political Economy*, 2005, *113* (1), 46–82.

_ , "Reference-dependent Preferences and Labor Supply: The Case of New York City Taxi Drivers," *The American Economic Review*, 2008, *98* (3), 1069–1082.

_ , "Why You Can't Find a Taxi in the Rain and Other Labor Supply," *Quarterly Journal of Economics*, 2015, *130* (4), 1975–2026.

Fisman, Ray and Michael Luca, "Fixing Discrimination in Online Marketplaces," 2016.

Flores-Guri, Daniel, "An Economic Analysis of Regulated Taxicab Markets," *Review of Industrial Organization*, 2003, *23* (3-4), 255–266.

Frechette, Guillaume R., Alessandro Lizzeri, and Tobias Salz, "Frictions in a Competitive, Regulated Market: Evidence from Taxis," *Working Paper*, 2016.

Gandal, Neil, Michael Kende, and Rafael Rob, "The Dynamics of Technological Adoption in Hardware / Software Systems: The Case of Compact Disc Players," *The RAND Journal of Economics*, 2000, *31* (1), 43–61.

Ge, Yanbo, Christopher Knittel, Don MacKenzie, and Stephen Zoepf, "Racial and Gender Discrimination in Transportation Network Companies," *NBER Working Paper Series*, 2016, (22776), 1–38.

Glanville, Doug, "Why I Still Get Shunned by Taxi Drivers," *The Atlantic*, oct 2015.

Goolsbee, Austan and Amil Petrin, "Consumer Gains from Direct Broadcast Satellites and the Competition with Cable TV," *Econometrica*, 2004, *72* (2), 351–381.

_ **and PJ Klenow**, "Evidence on Learning and Network Externalities in the Diffusion of Home Computers," *Journal of Law and Economics*, 2002, *XLV* (October 2002).

Gowrisankaran, Gautam and Marc Rysman, "Dynamics of Consumer Demand for New Durable Goods," *Journal of Political Economy*, 2012, *120* (6), 1173–1219.

_ **, Minsoo Park, and Marc Rysman**, "Measuring Network Effects in a Dynamic," 2014.

Graham, Matthew R., Mark J. Kutzbach, and Brian McKenzie, "Design Comparison of LODES and ACS Commuting Data Products," *U.S. Census Bureau*, 2014.

Haggag, Kareem, Brian Mcmanus, and Giovanni Paci, "Learning By Driving - Productivity Improvements by New York City Taxi Drivers," *American Economic Journal: Applied Economics*, 2017, *9* (1), 70–95.

Hall, Jonathan V. and Alan B. Krueger, "An Analysis of the Labor Market for Uber's Driver-Partners in the United States," *Working Paper 587*, 2015, (January), 1–28.

_ **, Jason Hicks, Morris M. Kleiner, and Rob Solomon**, "Occupational Licensing of Uber Drivers," *Working Paper*, 2017.

_ **, John J. Horton, and Daniel T. Knoepfle**, "Labor Market Equilibration: Evidence from Uber," *Working Paper*, 2017.

Hansen, Lars Peter, "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 1982, *50* (4), 1029–1054.

Imbens, Guido W. and Tony Lancaster, "Combining Micro and Macro Data in Microeconometric Models," *The Review of Economic Studies*, 1994, *61* (4), 655 – 680.

Kahn, Matthew E., "Do Greens Drive Hummers or Hybrids? Environmental Ideology as a Determinant of Consumer Choice," *Journal of Environmental Economics and Management*, sep 2007, *54* (2), 129–145.

Kleiner, Morris M., "Regulating Access to Work in the Gig Labor Market : The Case of Uber Regulating Access to Work in the Gig Labor Market The Case of Uber," 2017, *24* (3), 4–6.

Lagos, Ricardo, "An Analysis of the Market for Taxicab Rides in New York City," *International Economic Review*, 2003, *44* (2), 423–434.

Lee, Myoungjae, "Treatment Effects in Sample Selection Models and their Nonparametric Estimation," *Journal of Econometrics*, apr 2012, *167* (2), 317–329.

Li, Jing, "Compatibility and Investment in the U.S. Electric Vehicle Market," *Working Paper*, 2017.

Li, Shanjun, Lang Tong, Jianwei Xing, and Yiyi Zhou, "The Market for Electric Vehicles: Indirect Network Effects and Policy Impacts," *Journal of the Association of Environmental and Resource Economists*, 2017, *4* (1), 89–133.

List, John A., "The Nature and Extent of Discrimination in the Marketplace: Evidence from the Field," *The Quarterly Journal of Economics*, 2004, *119* (1), 49–89.

Nair, Harikesh, Pradeep K. Chintagunta, and Jean-Pierre Dubé, "Empirical Analysis of Indirect Network Effects in the Market for Personal Digital Assistants," *Quantitative Marketing and Economics*, 2004, *2*, 23–58.

Petrin, Amil, "Quantifying the Benefits of New Products : The Case of the Minivan," *The Journal of Political Economy*, 2002, *110* (4), 705–729.

_ **and Boyoung Seo**, "Identification and Estimation of Discrete Choice Demand Models When Observed and Unobserved Product Characteristics Are Correlated," 2014.

Poulsen, Lasse Korsholm, Daan Dekkers, Nicolaas Wagenaar, Wesley Snijders, Ben Lewinsky, Raghava Rao Mukkamala, and Ravir Vatrapu, "Green Cabs vs. Uber in New York City," *Proceedings - 2016 IEEE International Congress on Big Data, BigData Congress 2016*, 2016, pp. 222–229.

Schaller, Bruce, "Entry Controls in Taxi Regulation: Implications of US and Canadian Experience for Taxi Regulation and Deregulation," *Transport Policy*, 2007, *14* (6), 490–506.

Schroeter, John R., "A Model of Taxi Service under Fare Structure and Fleet Size Regulation," *The Bell Journal of Economics*, 1983, *14* (1), 81–96.

Sexton, Steven E. and Alison L. Sexton, "Conspicuous Conservation: The Prius Halo and Willingness to Pay for Environmental Bona Fides," *Journal of Environmental Economics and Management*, 2014, *67* (3), 303–317.

Sheldon, Tamara and J.R. DeShazo, "How Does the Presence of HOV Lanes Affect Plug-in Electric Vehicle Adoption in California? A Generalized Propensity Score Approach," *Journal of Environmental Economics and Management*, 2017, *85*, 146–170.

Shiller, Benjamin Reed, "First-Degree Price Discrimination Using Big Data," 2014.

Springel, Katalin, "Network Externality and Subsidy Structure in Two-Sided Markets: Evidence from Electric Vehicle Incentives," *Working Paper*, 2017.

Sundararajan, Arun, *The Sharing Economy*, 1st ed., Cambridge, MA: MIT Press, 2017.

Waldfogel, Joel, "Preference Externalities: An Empirical Study of Who Benefits Whom in Product-Differentiated Markets," *RAND Journal of Economics*, 2003, *34* (3), 557–568.

_ , "How Digitization Has Created a Golden Age of Music, Movies, Books, and Television," *Journal of Economic Perspectives*, 2017, *31* (3), 195–214.

_ **and Steven Berry**, "Free Entry and Social Inefficiency in Radio Broadcasting," *RAND Journal of Economics*, 1999, *30* (3), 397–420.

Weintraub, Gabriel Y., C. Lanier Benkard, and Benjamin Van Roy, "Markov Perfect Industry Dynamics With Many Firms," *Econometrica*, 2008, *76* (6), 1375–1411.

West, Sarah E., "Distributional Effects of Alternative Vehicle Pollution Control Policies," *Journal of Public Economics*, 2004, *88* (3-4), 735–757.

Appendix A

Appendix to Chapter 1

A.1 Data Sources and Construction

This appendix goes into more detail about the collection and scope of raw data sources. The following appendix explains how these data are transformed for use in the model estimation.

A.1.1 Collecting Uber and Lyft Characteristic Data

To collect information on the characteristics of Uber services (and Lyft), I emulated two Android phones set up with the applications (app) for Uber and Lyft from March through June 2016. Figure A.1 depicts the Uber app from the time of collection.[1] I automated the applications to feed the apps the locations depicted in Figure A.2 at specific times throughout the collection period. The Uber app ran every hour and a half, and the Lyft every forty-five minutes. Locations were sampled in the same sequence for every run but in an order to minimize the time between collecting information for each general area.

For Uber I collected data on UberPOOL, UberX, UberXL, Uber BLACK, and Uber SUV, though I use the low-cost UberX as the de facto Uber choice. Specifically, I scraped the information visible in Figure A.1. The ETA, i.e. "wait time" was of particular

[1] The Uber has updated their application for significant changes since mid 2016. A critical, and perhaps behavior-changing, update was to report estimated ride prices before confirming the process. At this time the application only reported the minimum fare, seen in the figure, and the surge price multiplier.

interest. Once the automator "clicked" `SET PICKUP LOCATION`, the app revealed the surge multiplier.

I also collected data on Lyft Standard, Lyft Plus, and Lyft Courier, using Lyft Standard as the de facto Lyft choice. I managed to capture the web traffic from the Lyft app in which details on the pricing scheme for the particular ride — Lyft had no visible surge multiplier like Uber — and wait time were available.

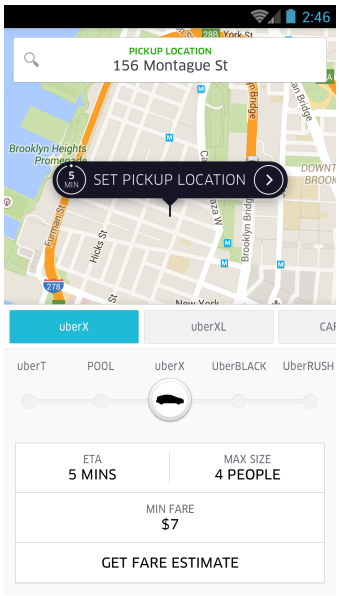


Figure A.1: **Image Captured from Uber Application**

A.1.2 Constructing Wait Time Data

Before delving into the specific method to simulate wait times, I first illustrate how I would use information on cab movements and images from traffic cameras in an idealized setting to deliver a probabilistic count of the taxi in each location. Figure A.3 depicts a stylized city with one taxi. The cab starts on green (node L) at time 0, after completing a drop off, and ends at red (node A) at time 5 for its next pick up. Traversing any street segment takes one time period.

The first step is to consider the potential streets the taxi could have crossed. Given I know travel time, the start, and the end of the trip, the cab could have taken the following routes: LKJGDA, LIFCBA, LIFEDA, LIFEBA. I assume at first each of

Figure A.2: **Locations Sampled for Scraping in NYC**

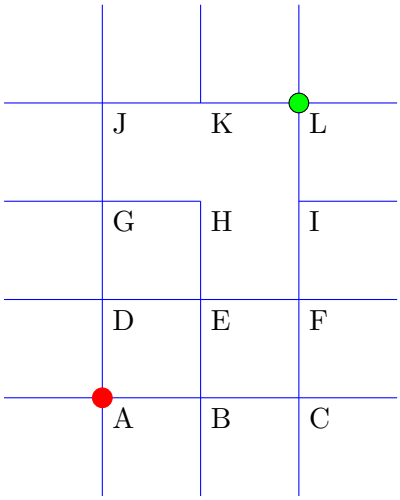


Figure A.3: **Stylized City with One Taxi**

these time-consistent routes is equally likely. Therefore, for example, the probability a cab was on street segment LI at this time is 75%. With additional taxis, I could develop an estimate of the expected time between two cabs passing on segment LI. This would be the construction of the wait time for a consumer on street segment LI. In the simplest case, suppose I observe a repeat of the same trip every six periods. The expected wait

time in periods comes from the survival analog of an exponential distribution with rate parameter $.75 * 1/6$. Similarly, the expected wait time on segment LK would be 24 periods.

The addition of traffic cameras helps pin down what route the taxi is taking. Suppose I now have access to cameras at every intersection in the city. At $t = 4$ I find a taxi crossing node D. This additional information narrows down the possible routes this cab could have taken down to only LKJGDA. I can in turn use this information to refine my estimate of the wait time for passengers throughout the city.

Taking this procedure to the full data steps quite far from the idealized example, but the intuition of converting the probabilistic positions of taxis into consumer wait times remains the same. Three particular sets of features distinguish reality from the idyllic city in Figure A.3. First, many thousands of taxis are simultaneously searching the city. Second, since 2013, the taxi pick up and drop off data from the TLC does not permit tracking taxis over the course of the day. This is irrelevant in the simple example, but potentially important in a city with more than two taxis. Third, the roads of New York City are significantly more complicated and permit far more route permutations between two points in a given time.

The following algorithm details the procedure by which I estimate wait times. In brief the procedure works by checking how far a cab can get without picking up another passenger. Once a potential cab is a assigned a next pickup, the latter of which I observe in data, probability weights are assigned to segments in the fashion described in the simple example. The role of the camera images are similar to how the pickups are used. Both sets of data essentially tag locations and times for when a cab must be in that location.

The algorithm accelerates time t by one minute through every iteration. Once a cab c makes a drop off, it is assigned a vintage starting at zero. The next period it is assigned to all street segments accessible within one minute. Naturally, the areas it could be in any period explode quickly, so two assumptions help with tractability: cabs cannot double back and are disappeared after 20 periods if they are not assigned pickup or a camera tag. The object V, a matrix of size L, the number of locations, by 20 keeps track of a count of assigned cabs by vintage for each location. For each active cab in the algorithm, I track c_s, the starting location, and vintage c_v. These two are

sufficient information to determine where the cab could have traveled in the time period. Finally, P is matrix of size L by T, the number of time periods, holding the probabilistic count of cabs in each location at each time in the day. After the algorithm runs for the desired period of time (4 AM to 4 AM in my sample), wait times throughout the day are constructed using the probabililistic locations of cabs through an approximation of the survival function). The algorithm "burns in" using an arbitrary day to generate a starting C. I typically choose Sunday to prepare the algorithm to run for the weekdays.

Algorithm 4: Taxi Wait Time Algorithm

```
Set t = 0
while t < T do
    Check drop offs at time t and update the first column of V
    Assign all pickups to an arbitrary taxi in the location with priority to the
      oldest vintage
    Subtract these cabs from V
    Update P by evenly weighting the routes the cab could take from c_s(l) to c_v(l)
    Assign camera images to an arbitrary taxi with priority
    Update P by evenly weighting the routes the cab could take from c_s(l) to c_v(l)
    Reset these cabs as vintage 0 and reset its starting location
    Increment cab vintages and update V
    Remove all vintage 21 cabs
    Set t = t + 1
end
```

With P, it is straightforward to calculate wait times with an approximated survival function. The calculation follows formula in Equation 1.12 with the probability of a cab passing through the segment in any t using the weights in P and checking up to $x = 60$.

One issue with extending this algorithm to other time periods is that I only have camera images processed from the end of 2015. To extend the value of the camera information, I assume that the marginal impact of the camera data on P uncovers a propensity for cabs to travel certain routes and street segments.[2] To capture this effect, I run the algorithm twice in the relevant months of 2015; I run it once with the camera data and once without. For each time period, the relative difference in the probability counts of each street segment I take as something analogous to a fixed effect

[2] I am implicitly, and now explicitly, assuming that these favored routes do note change with the entrance of competitors.

for that time period. I then apply this same weighting to P for periods when the camera data is not available.

One concern about this procedure is that it lacks any measure of external validity. This is an important issue to revisit in future iterations of this work.

A.1.3 Classifying For-Hire Vehicle Data

The pickup data for FHVs provided by the TLC does not explicitly list which company the consumer contracted for the ride. It does, however, list the base station linked to the particular ride. The TLC provides separate documentation with the name of the company operating each base station as well the name ("doing business as") under which that station operations. I separate Uber pickups from pickups of other FHV companies on the criterion that "Uber" appears in the DBA name of the linked base station.

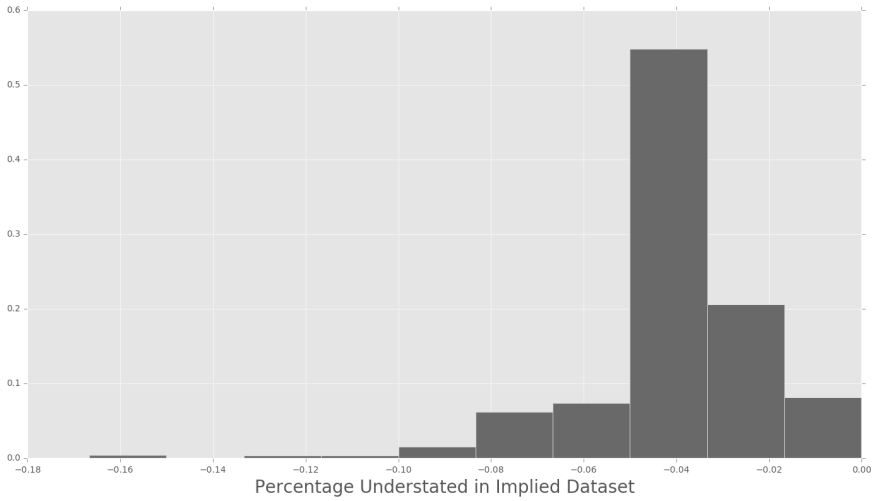
A lucky quirk in the TLC release of data allows me to check the veracity of this methodology. The TLC released a separate dataset of exclusively Uber pickups for only the first six months of 2015, a time period also covered by my standard FHV dataset. For the purpose of this comparison, I will call the Uber data from the full FHV dataset "implied Uber" and the other dataset "true Uber."

The first check compares the total Uber and implied Uber pickups on any given day. On average the implied Uber dataset yields 3.6% fewer rides than the true dataset. At worst it missed 4.1% and at best 3.2%. Hence, in a given day the implied Uber dataset tends to understate the total pickups. In the demand estimation this should manifest as a slightly underestimated unobserved quality term for Uber. But, one might worry still that the bias is not random with respect to time of day or location. The next two tests check this issue.

Unsurprisingly, for both the location- and time-based comparisons the implied Uber dataset uniformly understates the total pickups. By location the implied dataset on average yields 3.9% fewer rides than the true dataset. At best the two datasets perfectly match up, but at worst the implied dataset misses up to 16.7% of the rides in the true dataset. Figure A.4 shows the distribution of the understatement. Fortunately the vast majority of locations feature deviations of a similar magnitude. Finally, I checked the difference between the two datasets by half-hour segments of the day. On average the implied dataset gives 3.6% fewer rides than the true dataset. The range of the deviations

is otherwise tight; at best the understatement is 3.3% and at worst 4.2%.

Figure A.4: **Deviation of Implied Uber Dataset from True Dataset**



Note: The frequency distribution is calculated using deviations at locations as individual observations.

A.1.4 Determining Morning Taxi Commutes

For a fine granularity, e.g. census tracts, the American Community Survey does not report taxi usage separately from other means of transit. Along with the fact that the ACS data is taken over a long period of time and the personal vehicle segment of the transportation market is rapidly changing, I instead opt to construct commuting data from the rich TLC data set. Unfortunately, the dataset does not distinguish between the purpose of various trips. This section describes the procedure to extract commuting data from the raw TLC trip information.

I use the an algorithm to sort through the taxi commute observations and determine which should be classified as work rides based on two criteria:

1. the "same" ride appears in the dataset within γ_1 days; and
2. the "same" ride appears on that day within γ_2 hours of the first day.

To be explicit designate a trip record as $r_{ll'}^{dt}$, where l designates the identified starting census tract, l' designates the ending tract, d is the date, and t is the time of day. Only

rides associated with the same starting and ending tract pair are compared so we can drop the ll' notation. Let R be the set of all trip records which will be compared. For each record $r^{dt} \in R$, the algorithm extracts a subset $R^{dt} = \{r^{ij} \in R : i \in [d - \gamma_1, d + \gamma_1], j \in [t - \gamma_2, t + \gamma_2]\}$. If R^{dt} is not the empty set, record r^{dt} is designated a commuting trip.

Ultimately, "typical" commuting choices are of interest, rather than day-to-day decisions. To convert the trip-level commuting designations into a tract-level measure of cab usage comparable to the data available from the census, a measure of "typical" cab usage is constructed. I simply take the average number of commute trips per day over an extended period of time.

To calibrate the parameters γ_1 and γ_2, I match predicted taxi commute usage to PUMS-calculated usage for 2009 to 2012, prior to the introduction of green cabs, by year and PUMA in lower Manhattan, i.e. the yellow-taxi exclusive zone depicted in Figure 1.1. Designate $\mathcal{L} = \{\bar{l}\}$ the set of PUMAs, which is itself a collection of tracts. Let $f_t(l, \gamma_1, \gamma_2)$ be the number of typical taxi commuters identified by the algorithm for tract l in year t and $f_t^{Data}(\bar{l})$ the commuters identified by PUMS data for PUMA $\bar{l}$. γ_1 and γ_2 are the solution to the following minimization problem.

$$\begin{aligned} \underset{\gamma_1, \gamma_2}{\text{minimize}} \quad & \sum_t \sum_{\bar{l}} \left(\sum_{l \in \bar{l}} f_t(l, \gamma_1, \gamma_2) - f_t^{Data}(\bar{l}) \right)^2 \\ \text{subject to} \quad & \gamma_1 \in \mathbb{Z} \end{aligned} \tag{A.1}$$

The algorithm works on the assumption that people generally leave for work at the same time of day while not necessarily taking a cab every day.

For Uber, I conduct the same exercise using 2014 data in which the TLC-released Uber data is comparable to that available for yellow taxis. I then lump Uber rides in with those of traditional taxis and re-solve Equation A.1 matching 2014 PUMS data. Because later Uber data does not include the destination of the trip, I am unable to use these parameters in later years to tease out which Uber trips are allegedly for morning commutes. Instead I calculate the fraction of trips identified in 2014 as commuting trips and assume that fraction holds over time.

A.1.5 Routing

An important feature of the choice dataset is the time each option takes and how much walking is required as part of that choice. Unfortunately, little of this information is available in the datasets. The one partial exception is that the TLC data does note the time taken to complete realized taxi trips. For walking, transit, driving, and for-hire vehicles in areas with few recorded trips, I use a mix of OpenTripPlanner — an open-source route planning program, much like Google Maps — and a separately built graph of NYC's road network, based on the LION geographic database of NYC streets. [3]

Mechanically OpenTripPlanner functions similarly enough to Google Maps that it is not worth detailed explanation. A key difference, however, is that the former can be used with arbitrary public transit schedules stored in the General Transit Feed Specification (GTFS) format. For 2016, for example, I utilize the transit schedules published for 2016. While historic, off-schedule delays could also be incorporated, they have not at this time.[4]

Three pieces of data are fed into the routing program: starting location, ending location, and time of departure. Simulated and survey individuals assigned to a census tract are assumed to start from the centroid of their tract and travel to the centroid of their work tract. Individuals in the travel survey start their trip at the time indicated in the survey while simulated individuals leave at the start of their designated half hour slot. One concern might be that the rough departure assignment overstates the waiting time. For example, I assume a commuter leaves her house at 9:00AM for a proximal train leaving at 9:15AM. Like Google Maps Directions, the wait time at the beginning of the trip is shaved off. Only the time waiting for transfers counts toward the total wait time for that trip.

Vehicle trips are routed through a custom-made graph of New York City's street network. Simulated and survey individuals are again assumed to start from the centroid of their tract and travel to the centroid of their work tract. Unlike OpenTripPlanner, however, traffic speeds are approximated using data on cab travel time from when the trip was taken.

[3] From source http://www1.nyc.gov/site/planning/data-maps/open-data/dwn-lion.page

[4] On average one rests on the hope that the official schedules are correct, but the NYC MTA is becoming more notorious for its delays.

A.2 General Appendix

A.2.1 Uber's Expansion Pattern Across Geographies

This appendix expands on Section 1.2 by linking changes in the NYC taxi market more closely to density in parts of the city. Again, I use the ratio of building area to ground area in the various taxi zones of the city as the measure of density. Figure 1.7 depicts density using this metric. Ultimately, this density measure is inappropriate for areas where people congregate in wide space, e.g. parks, and where people magically appear, e.g. airports and train terminals. These zones are thus eliminated from the analysis in this section and noted by a checkered overlay in Figure 1.7.

I use this measure of density to illustrate two key facts about Uber's expansion from 2014 to 2016. The first is that market activity in area has grown in terms of total pickups with the sparsity of that area. In the densest areas Uber simply substitutes the existing incumbents. The second notes that Uber dominates in the sparsest areas and most of its growth since 2014 has been redirected to these locales. These results are preliminary evidence of the relative technology position of Uber and incumbents across areas of different density.

The analysis proceeds under the framework of spatial autoregressive models with a spatial lag. The model is attractive because of several issues present which exacerbate the potential for bias introduced by spatial correlation. First, incumbent taxis and Uber both benefit from local scale and the primary units of analysis, the taxi zones, are not isolated from each other. Unobserved characteristics impacting the variables of interest surely spill over these geographic boundaries. Second, the key regressor, density, is also highly spatially correlated. The general model for these regressions is

$$Y_i = \beta X_i + \rho W Y_i + \varepsilon_i \tag{A.2}$$

The elements in the neighborhood matrix W are defined by

$$w_{ij} = \begin{cases} 1 & \text{if } i \text{ borders } j \\ 0 & \text{if } i = j \\ 0 & \text{all other cases} \end{cases} \tag{A.3}$$

with ρ serving as the spatial correlation parameter. The unit of analysis is all taxi zones not excluded by the criterion mentioned above for each month January to June 2015 and 2016. Table A.1 offers relevant summary statistics for the regressions in and out of the green exclusion zone.

Table A.1: **Summary Statistics for SAR**

	Inside the Exclusion Zone			Outside the Exclusion Zone		
	Mean	5th Pctile	95th Pctile	Mean	5th Pctile	95th Pctile
Density	4.212	1.579	9.651	0.687	0.007	1.492
Daily Rides[a]	36945	4861	82343	934	6	5162
Y-Y Growth[b]	5.08%	-2.76%	19.01%	233%	4.28%	1136%
Uber Share[c]	23.5%	14.8%	39.0%	69.6%	25.2%	98.1%

[a] Total daily rides per sq mile by zone, as of March 2016
[b] Percent change in total rides from March 2015 to 2016
[c] Uber share of pickups per zone as of March 2016

The second two rows of Table A.1 describe the regressands for the two regression specifications. Beside the spatial lag of the regressand, each of the regressions additionally fit a quadratic function of density, after density has been normalized to be between 1 and 0, and a dummy for whether the zone is in the green exclusion zone. I fit a quadratic of density rather than a single term because patterns between the variables of interest and density tend to break down in the most extreme dense and least dense zones. The results or all specifications are reported in Table A.2.

The first specification is reported in the first column of Table A.2. The regression establishes the link between zone density and growth. Less dense markets have enjoyed greater growth over the time period. These results hold accounting both for the starting level of rides ("Lag Total"), which appears to be irrelevant, along with the exclusion zone dummy. The fitted function on density is a U shape but for densities lower than .15 (roughly 1.5 in unadjusted terms) the predicted growth is higher than the densest area. As clear from Table A.1, most zones in the outer boroughs fit into this category.

The second specification examines how tightly the substitutions between Uber with yellow and green taxis (called incumbents in Table A.2 over time follows from the density of the area. The dependent variable in this set of regressions is the total change — not percent — in Uber rides from 2015 to 2016 in the given month and taxi zone.

The additional regressands are the change in yellow and green taxis over the same time period and that term interacted with density. The key variables of interest in this regression are these two terms. If Uber and taxis were perfect substitutes in all locations, the regression would yield -1 on the "incumbent change term". If instead substitution increases with density the interaction term should be negative. A coefficient of -1 on this term would imply that Uber and taxis approach perfect substitutes in the densest market. Column 2 in Table A.2 shows the latter is indeed the result. In the least dense locations of NYC the interaction term is dominated by the incumbent change term and Uber tends to growth with taxis. In the most dense locations incumbent and Uber growth tend to move in opposite directions with an expected rate of substitution of .67 taxis for each new Uber in the most dense market.

Table A.2: **SAR Regressions**

	Y-Y Total Growth	Uber Pickup Change
Density	-16.635**	34305.88***
Density^2	14.455**	-14361.21
Exclusion Zone	1.192	-4320.52**
Lag Total	0.000	
Taxi Change		0.314***
Taxi*Density		-0.982***
ρ	0.698***	0.446***
N	1536	1536

Notes: *** $p < 0.01$, ** $p < .05$, * $p < .1$; Observations are at the month- and taxi zone-level for January to June over the period 2015 to 2016. The measure of density defined as zone building area to ground area has been normalized to be between 0 and 1 or these regressions. Taxis are defined as yellow or green taxis in these regressions.

A.2.2 Transportation Habits

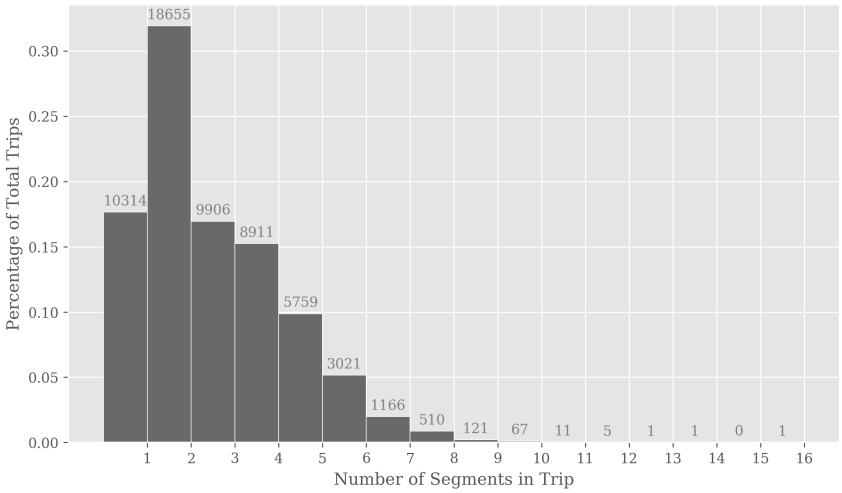
An important assumption in the consumer choice model is that consumers choose a single mode of transit to travel from their origin to their destination. The assumption guarantees a potential consumer's choice set can be reasonably modeled as a single form

of transportation.[5]

To test this assumption I use a transportation survey the Metropolitan Transit Authority conducted over a random sample of NYC residents in mid-to-late 2008. The samples were taken to roughly match demographics and residence distributions across the five boroughs in NYC and within community districts, with locations verified by the participants' home addresses. Over 13,000 households, or 16,000 people, were interviewed with each person offering details on an average of 2 trips, including transit method(s) and destination and origin. These households also provided demographic details and other relevant information, e.g. what kind of MTA card they hold.

Most importantly, the transit survey allows participants to list up to 16 transit segments used to get from origin to destination; a segment change might occur when a passenger switches from one mode of transit to a second or simple at an event like a bus transfer. Figure A.5 groups the 58,452 recorded trips by the number of segments.

Figure A.5: **Recorded Trips Grouped by Number of Segments**



This figure includes segments where passengers, for example, walk from home to their bus stop. Walking to transit nodes is explicitly modeled so Figure A.6 presents the same data without walking segments. Nearly 90% of trips feature only one or two

[5] One issue that arises from multi-modal transit is the question where mode transitions occur in a trip and the implications for cost, travel time, etc.

segments. The problem of modeling a consumer's choice set is further reduced by noting that indeed most trips are monomodal. Table A.3 illustrates that monomodal, exclusive walking, make up almost 85% of the total sample.

Figure A.6: **Recorded Trips by Number of Segments, Excluding Walking**

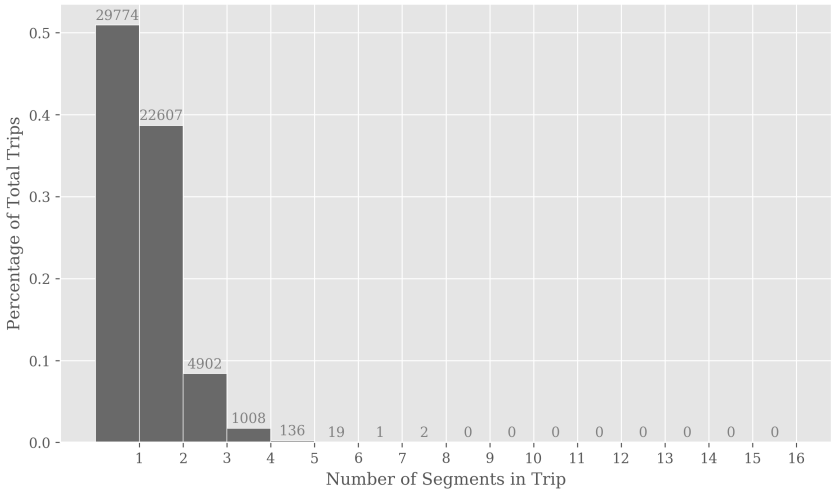


Table A.3: **Count by Number of Modes**

	Unrestricted	Without Walking
1	24594	48434
2	24545	5543
3	8144	493
4+	1169	73

Appendix B

Appendix to Chapter 2

B.1 Route Planning and Timing

Using road data and information on cabs' drop off and pickup locations, it is possible to determine the shortest street routes between any two locations by speed or distance. Calculated speeds for neighborhood areas — Manhattan is broken up into 29 geographically contiguous neighborhoods — are approximated by hour and day using known travel distances and times from the TLC dataset for rides that start and end in the same neighborhood. For Manhattan the volume of rides permits reasonable approximation of these speeds by day and hour for most hours of every day. For neighborhoods with less traffic or hours in lower traffic I aggregate up to blocks of four hours and / or by weekday.

The routing algorithms presume the driver is on a direct path between his / her starting and ending location. In the event that the actual times between a drop off and pickup are roughly consistent with the times implied by the shortest or fastest route, in practice a 5-minute tolerance is permitted, I assume I know what path the cab driver took between pickup and drop off. This condition is important for estimation as the decision of the drivers modeled presupposes some understanding of what happened between trips. If the routing program predicted a five-minute time between drop off and pickup and the actual time was 1 hour, the driver could have taken many routes or even been on a break between in the interim. Figure B.1 shows one such accurate mapping (with the start at green, end at red) superimposed on a map of the area in

Manhattan. Note the algorithm respects one-way versus bidirectional roads. The red dotted line denotes the trajectory that would be used in the estimation.

Figure B.1: **Sample of Accurate Route Estimate**

Determining Route Estimation Accuracy

To test how accurate these routing algorithms are for relatively direct routes, I check the consistency of estimated times and distances against known values for rides with fares.[1] In a random sample taken of nearly 300 thousand trips, the best estimated trip distance — the best match to actual distance using the short and fast algorithm — is "extremely" incorrect for only 2.9% of trips. Conditional on not being "extremely" incorrect, that is the estimation error is off by less than 100%, Figure B.2 illustrates the error distribution.

The averages (mean and median) of the distribution are both significantly above 0, 7.3% and 6.7% respectively, meaning that the trip distance is underestimated. This bias may be born from a yet-to-correct issue with the routing that only plans the trip of a cab from the nearest intersection to the intersection closest to its final destination. There is no correction for the distance from the actual starting location of the cab to

[1] The TLC dataset provides both actual travel distance and time for rides.

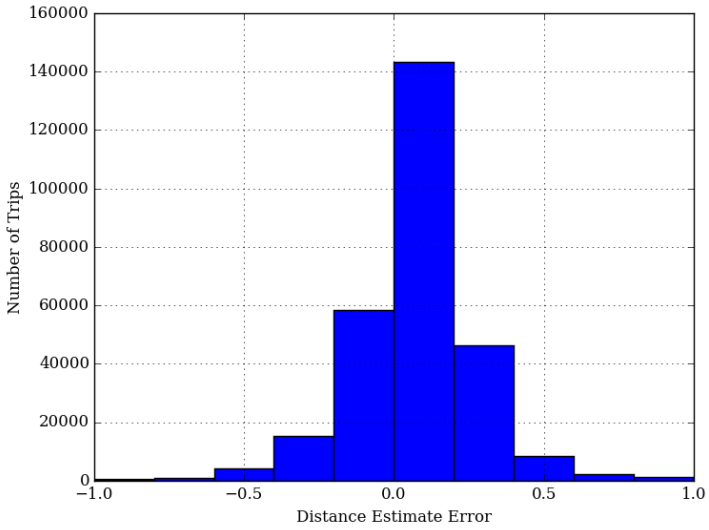


Figure B.2: **Percent Error in Estimated V. Actual Trip Distance**

the first intersection on this route. The result of this issue on the error, however, is ambiguous since the direction of bias will depend on whether this algorithm puts the cab further along and further behind the actual route used.

Whether these routing algorithms are actually useful for determining the real route taken by the cab is suspect. For shorter routes, small deviations from the actual in the actual will yield large errors. In large trips, even larger deviations would be needed to yield significant errors. Unfortunately, the correlation between trip distance and error are positively correlated. For longer trips claims about actual routes would be inaccurate. For shorter trips it may be possible to determine a small set of the potential routes quite well if future work requires it.

The purpose of this exercise was to demonstrate the accuracy of estimated route planning for cabs between rides by comparing known and estimated travel distance. But, one assumption skirted above was in determining which of the two routing algorithms is considered best. An alternative is not to select the best algorithm based on estimated versus actual distance but based on estimated versus actual time. Redoing the analysis with this definition of best, the mean error in estimated distances is only 3% in the not-extremely-incorrect sample (within 300 seconds).

B.2 Constructing Shifts

Shifts do not appear explicitly appear in the dataset. The information used to determine shifts is depicted in Table B.1. The data have been sorted by medallion and time; medallion is a cab identifier while the hack license is the cab driver identifier.

Medallion	Hack License	Pickup Time	Dropoff Time	Shift
A	1	4:30:00	4:36:00	1
A	1	4:40:00	5:00:00	1
A	2	6:00:00	6:14:00	2
A	2	6:20:00	6:26:00	2
A	2	10:18:00	10:30:00	3
A	2	10:42:00	10:48:00	3
B	2	13:10:00	13:22:00	4
B	2	13:24:00	13:30:00	4

Table B.1: **Sample Data for Determining Shifts**

The table depicts a few common situations that identify a change in shifts. Some are obvious while others require discretion. If any of the following conditions are satisfied then a change in shift is noted (again note that the data has been sorted by medallion and pickup time by this point):

1. The cab driver (hack license) changes between subsequent observations

2. The cab driver does not change between subsequent observations BUT more than 3 hours have elapsed

It also implicitly assumed that if a cab driver jumps cabs, his or her activity in the first cab is a different shift than in the second cab. Assuming this cab driver rents from a fleet, which is likely if he or she is operating more than one cab, he or she would have to pay a separate lease fee for activity in the first and second cab so splitting them into shifts seems reasonable. In Table B.1 condition 1 determines the break between shift 1 and 2, condition 2 between shifts 2 and 3, and this implicit assumption between shifts 3 and 4.

The first condition is a clear indicator that shifts have changed. The second is completely based on discretion and is typically required for cabs not managed by a

fleet. Three hours seemed like a reasonable length of inactivity to presume the cab driver stopped working for that particular shift, though even longer lengths might be more appropriate. As a "safety net" for errors in discretion, the distribution of shift lengths in terms of number of rides from known shifts, i.e. those determined by condition 1, is used to cut off shifts determined by condition 2 running abnormally long.

Table B.2 relates this exercise to the determination of searching trips. Technically, each cab driver must be searching at least before his or her first pickup on a shift and potentially after the last observed drop off. Unfortunately, there is no way to determine the activity of the cab driver before and after each shift without assuming starting location and time. Cab A, for example, would be considered out of service between 5:00:00 and 6:00:00, the last drop off of shift 1 and the first pickup of shift 2. Hence searching trips are restricted to trips between observed rides in the data.

Medallion	Hack License	Search Start	Search End	Shift
A	1	4:36:00	4:40:00	1
A	2	6:14:00	6:20:00	2
A	2	10:30:00	10:42:00	3
B	2	13:22:00	13:24:00	4

Table B.2: **Searching Trips for Sample Data**

B.3 Identifying Decisions and Intersection Passing

Consider Figure B.3, a sample of what GPS road data and starting / ending location data for three searching trips, denoted A, B, and C. There are two simple ways to try to link cab location data to road data. One way is to link start and end points to intersections (nodes); the other is to link them to road segments between intersections. In both cases the easiest way to link off road grid GPS data to road data is via some minimum distance function (detailed later). The figure shows in all three cases that linking GPS data to nodes (red dotted lines) yields ambiguous results as to whether the vehicle passed through an intersection. In cases A and B, the cab moves through an intersection but both starting and ending points would be linked to the same node. In case C the cab does not move through an intersection, but the starting and ending points are associated to different nodes.

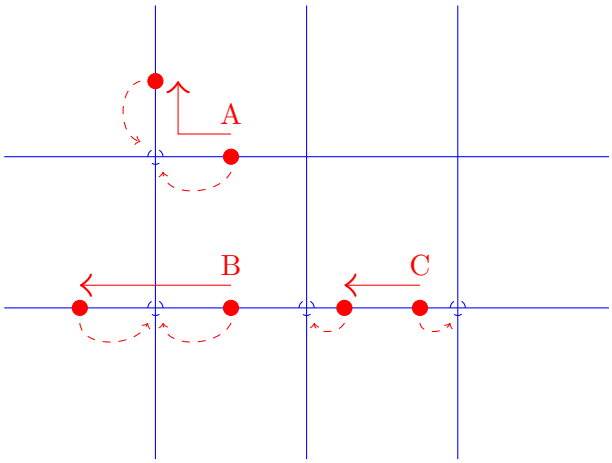


Figure B.3: **Determining Choice from Data**

The easy alternative is to link off road GPS data to edges, i.e. road segments. If the linked edges are different for the starting and end point, the cab must have traversed through an intersection. This method correctly identifies whether the cab has moved through an intersection in all three cases depicted in the figure. To link GPS data to edges, I find which edge is the shortest distance from the point and project the point onto each line segment. The second condition requires that the projected point falls in between the end points of the edge segment.

Appendix C

Appendix to Chapter 3

C.1 Construction of Charging Station Variable

The full structural model and the descriptive Poisson model in Section 3.3 both take advantage of our highly granular data on the opening time and locations of charging stations in California. We construct two sets of charging stations counts for consumers living in a particular zip code: the charging stations near that home zip code and an estimate of the charging stations near their work zip code. While home-area zip code charging stations may seem like the most natural measure, home chargers for PEVs appear to mitigate the need for other proximal charging stations. According to the Center for Sustainable Energy (2013a) survey, approximately 90% of PEV owners own Level 2 home chargers.

To construct a measure of home charging stations at a particular point in time, we count the open charging stations within an X-mile radius of the centroid of every census block group; this X can be flexible, as in the regressions in Table 3.9. To assign a count to a particular zip code, we weight census block zip codes by the relative contribution of the block to the zip code's.[1]

To assign a count of work charging stations to a home zip code, we use additional information from the Longitudinal Employer Household Dynamics Origin-Destination Employment Statistics (LODES). LODES links workers to employers and further allows us to observe a worker's residential and employer census blocks. Employer location

[1] Note that census blocks do not partition zip codes.

is collected from the Quarterly Census of Employment and Wages (formerly ES-202) and employee residence location is collected from the Composite Person Record by the Census Bureau.[2] See Table 3.12 in Appendix 3.9 for more information.

We use the information from LODES to determine the relevant work locations for a given home zip code. To construct the count of work charging stations, we weight the number of open charging within an X-mile radius of the work census block's centroid by the home zip code's population working in that census block. In the structural model we can integrate simulated consumers over their using these weights as a probability mass function.

For each home zip code, there are, on average, 11.89 associated work place zip codes. Most home-work zip code links feature few unique households because of the sample size and level of granularity. On average 1.78 households make the trip between the home-work zip code, though some feature as many as 100.

[2] See Graham et al. (2014).