Frictions in a Competitive, Regulated Market
Evidence from Taxis

Guillaume Frechette (NYU) Alessandro Lizzeri (NYU)

Tobias Salz (Columbia University) *

December 15, 2016

Abstract

This paper presents a dynamic general equilibrium model of a taxi market. The model is estimated using data from New York City yellow cabs. Two salient features by which most taxi markets deviate from the efficient market ideal is the need of both market sides to physically search for trading partners in the product market as well as prevalent regulatory limitations on entry in the capital market. To assess the relevance of these features we use the model to simulate the effect of changes in entry and an alternative search technology. The results are contrasted with a policy that improves the intensive margin of medallion utilization through a transfer of medallions to more efficient ownership. We use the geographical features of New York City to back out unobserved demand through a matching simulation.

Keywords: Frictions, regulation, labor supply, industry dynamics.

*We are extremely grateful to Claudio T. Silva, Nivan Ferreira, Masayo Ota, and Juliana Freire for giving us access to the TPEP data and their help and patience to accustom us with it. Jean-Francois Houde, Myrto Kalouptsidi, Nicola Persico and Bernardo S. da Silveira provided very helpful conversations and feedback. We gratefully acknowledge financial support from the National Science Foundation.

1 Introduction

This paper estimates a dynamic general equilibrium model of the New York City (NYC) taxi-cab market. The estimated model is used to assess the importance of regulatory entry restrictions and of search frictions. The ability to overcome these barriers to trade has been a key element for the success of new technology entrants such as Uber, which have expanded supply in many local markets and introduced a novel dispatch technology to reduce search frictions. Our counterfactual results isolate the relative importance of these two effects. We also show that segmenting the taxi market between Uber and traditional taxis could lead to a reduction in market thickness that worsens search frictions in the aggregate.

Taxi services offer limited room for product differentiation and markups. Moreover, a firm in this market is of relatively low organizational complexity. In its simplest form it consists of a unit of capital (a car) plus the labor (a driver) needed to operate it. The labor skill requirements are relatively modest and the capital is in vast supply. Finally, in a city like NY, taxi drivers take many decisions independently, with little real-time information about aggregate conditions. Thus, absent regulatory interventions, this market could serve as a textbook example of a "perfectly competitive" industry with many firms making decisions independently, jointly affecting aggregate market conditions. It therefore presents an interesting case study of an important benchmark.

We first document some important patterns in this market. In most cities taxi markets are subject to stringent regulations on entry and fares. NYC is no exception. Under the current system at most 13,520 yellow cabs can serve the market, and we provide evidence that these restrictions are strongly binding. If there were no other friction, one might therefore expect all taxis to be utilized at least during the day-time. However, activity is often well below capacity, highlighting the importance of understanding the intensive margin in labor supply decisions.

Because of regulations, there is no price flexibility in this market. This, together with the limits in capacity, means that regular patterns of variation in demand for rides during the day (e.g., rush hours) lead to large variations in the delays for matches between passengers and taxis. Drivers' earnings and the number of active taxis vary during the day depending on how long drivers need to spend searching for their passengers. The average search time for an active taxi between dropping off a passenger and picking up the next one ranges between 5 and 20 minutes depending on the time of day. Absent price adjustments, passenger wait times and taxi search times serve as the market clearing variables.[1]

In our model, drivers make daily entry and hourly stopping decisions. Medallions are scarce, so entry is only possible for inactive medallions. Hourly profits are determined by the number of matches between searching taxis and waiting

[1] This form of rationing is common in other markets.

passengers. Ceteris paribus, increasing the number of taxis increases the search time for a driver to encounter the next passenger and drives down expected hourly income. The number of taxis is determined endogenously as part of the competitive equilibrium in this market. Stopping (exit) decisions are determined by comparing a random terminal outside option with hourly earnings minus a marginal cost of driving that is increasing in the length of a shift. Starting (entry) decisions result from a comparison between an outside option and the expected value of a shift (given optimal stopping behavior).

To estimate the model we make use of rich data on the NYC taxi market from the years 2011 and 2012. This data includes every single trip of the yellow cab fleet in this time span. The data entry of a trip includes the fare, tip, distance, duration as well as geo-spatial start and end points of the trip. We have a panel identifier for the medallion as well as the driver. This allows us to account for an important source of heterogeneity in drivers' characteristics (namely, owner-operators vs. fleet drivers) that affects the intensity of utilization.

On the demand-side, we face a challenge because neither the passengers' waiting time nor the number of hailing passengers is observed in the data. However, we can take advantage of the geographical nature of the search process to recover how many people must have been waiting for a cab given the number of passenger pickups we observe (successful matches), how long taxis search for passengers, the number of cabs on the street, and the speed at which traffic is flowing; all of which are variables we observe in our data. While the empirical literature on search and matching typically uses known inputs and observed number of matches to infer the functional form of the matching function, we go the opposite direction and use a specific matching process as well as observed matches to infer one of the inputs to the matching function.

With the recovered demand data in hand, we proceed to estimate a demand function in terms of the expected waiting time for a cab (recall that fares are fixed). We find relatively modest elasticities of demand with respect to waiting time, but this elasticity does play a significant role in the counterfactuals.

Our first counterfactual evaluates the effects of additional entry. An increase in the number of medallions of 10 percent leads to an increase of 7.5 percent in the number of active taxis.[2] The reason for the less than proportionate increase is that drivers respond to reduced earnings by choosing shorter shifts, highlighting the importance of modeling the intensive margin on the supply side. However, the increase in activity would be a lot smaller if we did not incorporate in the model the dependence of passenger demand on expected wait times. The increase in the number of taxis leads to a reduction in wait time, which leads to an increase in the number of passengers, which in turn moderates the reduction in earnings caused by the increase in the number of medallions.

[2]There is some difference in the extent of the increase in activity at different hours of the day. This number is an average across the day for a typical weekday.

Our second counterfactual considers an improved matching technology in line with the dispatch system of the Uber platform and other startup ride hailing services. We first consider a polar opposite of the NYC decentralized decision making model by introducing a centralized dispatcher for the entire fleet. We show relatively large gains for both sides of the market due to reductions in wait times for both passengers and taxis. Interestingly, the number of active taxis increases by almost the same amount as in the counterfactual with 10 percent more medallions, despite the fact that the number of medallions is left unchanged. The number of matches increases by 13%, almost double the increase in the activity of taxis because frictions drop significantly.

We then consider the consequences of partial coverage by a dispatcher, with the remainder of the market functioning with the traditional street-hailing system. This is analogous to partial market penetration by an entrant such as Uber which operates on a separate dispatch platform. We show that there is a large liquidity effect that emerges from the market segmentation on different platforms.[3] Partial coverage by a dispatcher has two effects: there is improvement of matching in the covered market but, there is a segmentation of the market that makes both segments thinner, with the consequence of longer average distances between a random taxi and a random passenger. When we consider the case of a 50-50 split between dispatch and decentralized platforms, we find that the second effect dominates, and therefore, aggregate outcomes become worse than in the baseline case. Interestingly, the effects are quite different during the daytime relative to night-time hours, reflecting the importance of the initial thickness of the baseline market environment. Finally, we discuss a case which combines entry with the dispatch technology: 10% additional taxis, all of which are on the dispatch platform. In this counterfactual, waiting times improve relative to the baseline, but not relative to the case of the same amount of additional entry without the separate dispatch technology. This is again due to market segmentation.

Lastly, we use the estimated model for a policy evaluation specific to NYC that is of some interest because it highlights the importance of firm organization even in environments where the production process is simple. A peculiarity of the regulation in NYC is the restriction that about 40% of all medallions have to be owned and operated by individuals. The remaining medallions are unrestricted and are operated by several dozen companies which are known as minifleets. These vary in size, with the largest operating hundreds of taxis. We find sizable differences in the utilization rates across medallion types. Owner-operated medallions have significantly lower utilization rates and markedly slower transitions between shifts. The city has recently presented a proposal to convert owner operated medallions to regular medallions without these additional restrictions.[4]

[3] Uber seems to be well aware of this effect and therefore subsidizes drivers, especially when they first enter a city.

[4] See, http://www.nyc.gov/html/tlc/downloads/pdf/proposed_rule_omd_repeal.pdf and

This policy change leads to an increase in consumer surplus that is approximately half of the one that we computed for the case of a 10% increase in the number of medallions. This policy change would also lead to gains in the value of medallions and depress driver wages only modestly. As such, it seems a politically more feasible policy relative to an increase in the number of medallions.

2 Related Literature

This project combines elements from the entry/exit literature, neoclassical labor supply models and search. Structural estimation of entry and exit models goes back to Bresnahan and Reiss (1991). This entry/exit perspective on the problem is motivated by the fact that drivers in the New York Taxi industry, like in many other cities, are private independent contractors and decide freely when to work subject to the regulatory constraints. The labor supply decisions of private contractors have for example been studied in Oettinger (1999), who uses data from stadium vendors. In the spirit of the entry/exit literature we recover a sunk cost, which is in our case the opportunity cost of alternative time use from observed entry and exit decisions and their timing. One distinguishing feature of our work relative to the typical I.O. literature is that our market contains tens of thousands of entrants. Entrants therefore are competitive and only keep track of the aggregate state of the market, which is summarized in the hourly wage that is determined in equilibrium as a function of aggregate entry and exit decisions. Another distinguishing feature is that previous papers on the topic, see for instance Bresnahan and Reiss (1991), Berry (1992), Jia (2008), Holmes (2011), Ryan (2012), Collard-Wexler (2013), and Kalouptsidi (2014), feature relatively *long-term* entry decisions (building a ship, building a plant, building a store, etc.) making both entry and exit somewhat infrequent. In our setting, entry and exit decisions are made daily creating a closer link between realized payoffs and expected payoffs.

A direct application of spatial search to the taxi market is provided in Lagos (2003), which calibrates a general equilibrium model (with frictions) of the taxicab market, and includes some heterogeneity among locations. However, he assumes that all medallions are active throughout the day and thus does not model the labor supply decision nor does he allow demand to be elastic to wait time.[5] Using the model, he quantifies the impact of policies increasing fares and the number of medallions.

Buchholz (2015) also estimates a structural model of the NYC yellow cab market but focuses on the spatial dimension of the drivers' choice, while taking the intertemporal supply of taxis as exogenous. Buchholz (2015) relies on spatial vari-

http://www.nydailynews.com/new-york/tlc-plans-nix-owner-drive-rule-article-1.2543202

[5] Lagos (2003) does not have data on hourly or daily decisions by taxi drivers.

ation in fares due to the the two components of fares, i.e., a fixed fare and a fare that is variable in the travel distance, to estimate demand as a function of price. In contrast, absent aggregate intertemporal variation in fares, we allow demand to depend on the expected passenger waiting time. While we do not study the spatial dimension of drivers' choices, taxi activity is endogenous in our model and we attempt to match the patterns of daily activity allowing for different behavior for fleet versus owner operators. Hence, while his paper can explore counterfactuals with respect to fare structure our paper ours can investigate ownership structure as well as relaxing entry restrictions. Similarly, although one of our counterfactual is related, namely our analysis of a (fully) centralized dispatcher, the responses in Buchholz (2015)'s model are entirely spatial, whereas in ours total taxi activity also responds.

Some earlier papers have used NYC trip sheet taxi data to investigate individual labor supply decisions. Camerer et al. (1997) find a sizable negative elasticity of daily labor supply and they argue that this is inconsistent with neoclassical labor supply analysis. This interpretation has been challenged by Farber (2008). Crawford and Meng (2011) estimate a structural model of the stopping decision by a taxi driver allowing for a more sophisticated version of reference-dependent preferences. They do not consider the entry decision by a cab driver and do not analyze the industry equilibrium. We opted to stay within the neoclassical framework to study the general equilibrium of the taxi market, in contrast to these papers that all focus on the intensive margin of daily individual labor supply decisions. We note that although it may very well be that some drivers do not fit this assumption, the aggregate patterns are consistent with a standard model, and it fits the data quite well. Hence, a standard model of labor supply seems to be a reasonable starting place.[6] This perspective is also supported by the new evidence in Farber (2014), who uses the TPEP data and shows that only a small fraction of drivers exhibit negative supply elasticities.[7]

3 Industry Details and Data

3.1 Industry Details

Operating a yellow cab in NYC requires ownership of a medallion. In the time

[6] Note also that our models fits aggregate patterns relatively well, hence any gains from allowing for a richer labor supply decision would be small in aggregate.

[7] Other papers are not directly relevant, for instance Haggag and Paci (2014) that study the impact of suggested tips in the NYC taxi driver payment screen for clients on the realised tip. Haggag et al. (2014) who study how taxi drivers learn driving strategies based on the experiences. Finally, Jackson and Schneider (2011) find evidence of moral hazard in the behavior of taxi drivers and document that this problem is moderated if drivers lease from fleets owned by someone in their social network.

period covered by the data only yellow cabs are allowed to pick up street-hailing passengers.[8] This differentiates them from other transportation services such as black limousines for which rides have to be pre-arranged via a phone call or the internet.[9] Yellow cab rides cannot be ordered via phone or internet unlike in many other American cities. The yellow cab market is regulated by New York's Taxi and Limousine Commission (TLC), which sets rules for most aspects of the market such as the fare that drivers can charge, the qualifications for a taxi-driver license, the insurance and maintenance requirements, and restrictions on the leasing rates (daily or weekly) that medallion owners can charge drivers.

Approximately 40% of the medallions, called owner-operated, require that the owner of the medallion drive the taxi for at least 210 shifts in a year. The remaining 60% of medallions, which are called minifleets, are operated by approximately 70 fleet companies that operate an average of 115 taxis each, although some fleet owners operate more than one thousand taxis.[10] Fleet companies therefore manage many medallions and rent taxis out to drivers on a daily or weekly basis.[11] The presence of owner-operated medallions prevents concentration of ownership and therefore guarantees a fraction of "small businesses" in the industry. We later show that the requirement of owner-operated medallions leads to less flexible rental arrangements and lower utilization, implying that it is important to allow for heterogeneity among ownership types in our estimation.

The TLC imposes several restrictions on the terms of the leases between medallion owners and drivers. Leases can either be for a shift or an entire week. A rental for a shift has to last twelve consecutive hours and a weekly lease seven consecutive days. Minifleets must operate their cabs for a minimum of two nine hour shift per day every day of the week.[12] The TLC also specifies a cap on the price that medallion owners can charge that varies with the time of the lease and the type of vehicle.[13] The fixed fare for transporting a passenger is $2.50 and the fare for an additional unit is $0.4.[14]

[8] In the time period that we consider, Uber was not yet a significant presence

[9] The TLC recently established a possibility to hail cabs via a a mobile app, which was not possible during our observation period.

[10] Source: http://www.nycitycab.com/Services/AgentsandFleets.aspx

[11] Fleet companies not only operate medallions that they own for themselves but might also operate medallions for medallion agents who lease them to the fleet companies.

[12] See TLC Rules and Regulations, paragraph 58-20 (a) (1) http://www.nyc.gov/html/tlc/html/rules/rules.shtml.

[13] The leasing rate caps are between $115 and $141 for a day-shift depending on the Weekday, the vehicle type, as well as whether it is a night-shift or day-shift. Rate caps for weekly rentals vary between $690 and $812.

[14] What constitutes a unit depends on the speed of driving. If the cab is slower than 12 mph a unit is 60 seconds and above that speed it is 1/5 of a mile. On weekdays there is an additional fixed surcharge of $ 1 for trips between 4:00 pm and 8:00 pm and $0.50 for trips between 8:00 pm and 6:00 am. Trips from JFK airport to Manhattan are subject to a flat fare of $45.00 while those to other borrows and trips from La Guardia are still governed by a variable fare.

3.2 Data

Our main data source is the TLC's Taxicab Passenger Enhancements Project (TPEP) which creates an electronic record of every yellow cab trip. For each trip it records a unique identifier for the driver as well as the medallion; the length, distance, and duration of the trip, the fare and any surcharges, and the geo-spatial start and endpoint of the trip. The TPEP data can be obtained from the TLC. In this project we only use a subset of the data from 2011 and 2012, which ranges from the October 1st, 2011 to November 22nd, 2011; and August 1st, 2012 to September 30th, 2012. The data from 2012 encompasses the time at which the unit charge was increased from $ 0.4 to $ 0.5. During the time spanned by our data we see the universe of 13,520 medallions. We also observe all 37,406 licensed drivers that have been active in that period. We complement this data with information about the medallion type (minifleet or owner-operated), and the vehicle type.

We will focus most of our analysis, including all of the counterfactuals, on Monday through Thursday. The average activity of these days looks almost identical whereas Friday, Saturday, and Sunday each have some peculiarity. The reason we do not further differentiate between weekdays is that it would then be computationally prohibitive to obtain counterfactuals

4 Descriptive Evidence

We now provide some background information and descriptive evidence about the functioning of the market. We also offer evidence for each of the following features of the market that are later incorporated in the model and will be addressed in the counterfactual calculations: (1) entry restrictions, (2) daily patterns of activity, and (3) search frictions.

4.1 Entry Restrictions

As we mentioned in the introduction, there are tight entry restrictions in most taxi markets. In NYC, the number of medallions during the time-period of our data is 13,520. This is an absolute limit on the number of possible taxis on the street at any moment in time. Prices of medallions are an indicator of the quantitative importance of the entry restriction.[15] During the period we consider these prices exceed half a million dollars. Of course, such medallions would not be valuable in a market with no entry restriction. We have also verified that the percentage of medallions that are driven at least once a day is close to 97% from Tuesday to Thursday, and 92% even on Sunday. Given that there will be some natural failure

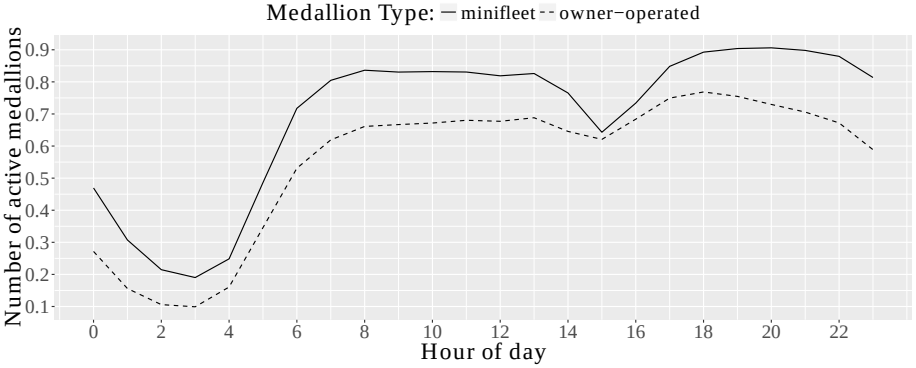
[15] Medallions are often traded in auctions and prices are public data.

rate of vehicles and other idiosyncratic reasons why taxis may fail to be utilized, this seems to be a very intense rate of utilization.[16]

4.2 Daily Patterns of Activity

Figure 1 displays the fraction of taxis that are active at each hour of the day for a typical (Monday-Thursday) weekday, distinguishing whether it is a minifleet or an owner-operated medallion.[17]

Figure 1: Comparing Activity of Owner-operated and Fleet Medallions.



We wish to draw attention to several features of this figure. First, fleet medallions are more intensely utilized for every hour of the day. To place this in perspective the difference in utilization is larger than the difference between Uber and taxis reported by Cramer and Krueger (2016). Second, there is substantial intra-day variation in activity, but this variation does not seem to fully reflect expected patterns of intra-day variation of demand, despite the fact that activity is well below capacity for the entire day.[18] Third, there is a large reduction in activity precisely during the evening rush hour. This is known in NYC as the witching hour. This drop in activity is stronger for fleet medallions.

Our model of the supply-side of the market will incorporate features that allow it to match all these data patterns of daily activity. In particular, these patterns imply the need to take into account the intensive margin of supply and its variation during the day, as well as the importance of allowing for differences between fleet and owner-operated medallions.

[16] As we will see, this is quite different from the utilization rates for a typical hour, which also depends on drivers hourly stopping behavior.

[17] An owner is required to drive at least 210 shifts (nine-hour minimum) per year. Thus, for this type of medallion, one owner cannot manage multiple medallions. However, they can lease the taxi to another driver for the shifts he does not himself drive.

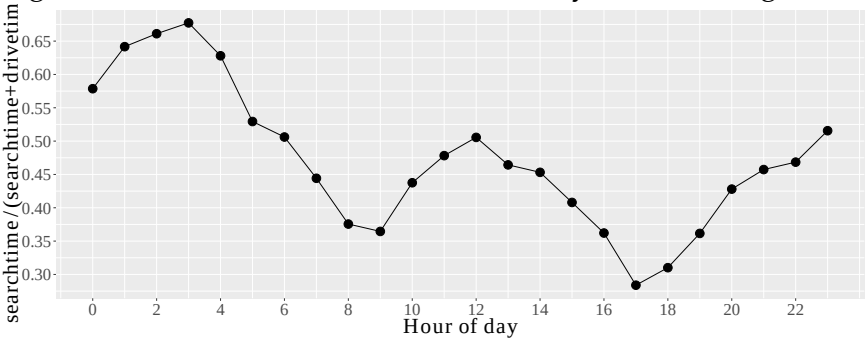
[18] Below we provide evidence that activity is indeed less variable than demand.

One of our counterfactuals aims to understand the quantitative importance of the restriction on owner-operated medallions. We will discuss other features of this difference when we present that discussion.

4.3 Search Frictions

An important friction in this market relative to the ideal of a Walrasian market arises from the fact that drivers and passengers have to physically search for trading partners. Figure 2 describes the fraction of time that an average taxi spends

Figure 2: Search-time Relative to Delivery Time During the Day



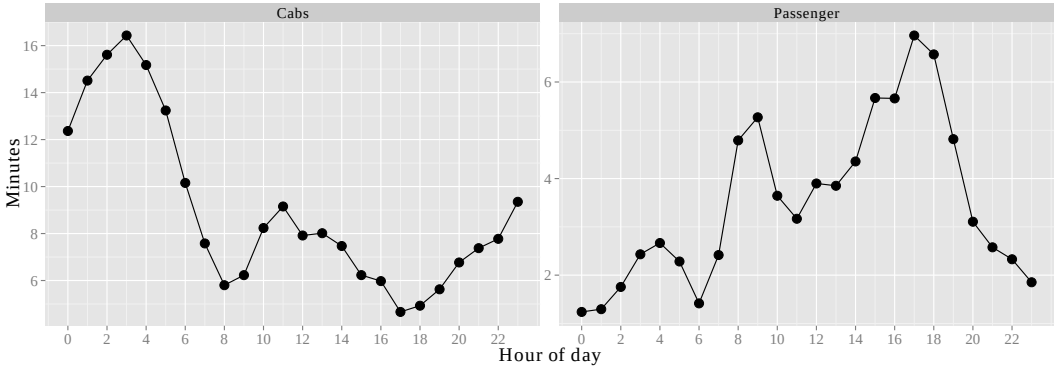
Notes: This plot shows the ratio of time a taxi spends searching for a passenger as a percentage of total time spent driving. For the plot we take the average over those ratios using data from Monday to Thursay. The plot shows that search time is highest in the nighttime hours and lowest during the "witching" hour when demand picks up for the evening rush hour and many medallions are transitioned between shits.

searching for passengers relative to the total time it is active, i.e., the unemployment rate for taxis.

Two notable features are the following: First, the fraction of time taxis spend searching is almost never lower than thirty percent and shows substantial variation throughout the day, going as high as 65% at some points of the day. It is important to note that, under the current system of fixed fares, most inter-temporal variation in driver profits and customer welfare are created by variation in delays in finding a partner. Indeed, a simple linear model reveals that variation in taxi search time explains about 60% of the variation in drivers' hourly wages.[19] The low point of the time taxis spend searching is reached at 5PM when many medallions become inactive because of shift changes. One interesting observation from Figure 3 in conjunction with Figure 2 is that waiting time for both passengers and taxis increases during the night although the ratio of passengers to taxis is relatively stable (the wait time for passengers is inferred from our simulation, which

[19]Most of the remaining variation is explained by trip-length and the rate that is charged per minute of driving. This rate varies with the speed of traffic due to the mixture of time-based and distance-based metering.

Figure 3: Search Time for Taxis (from data) and Wait Time for Passengers (from simulation)



Notes: The left panel shows search time for taxis in minutes, averaged for each hour of the day. The right panel shows waiting time for passengers as recovered by our simulation, again averaged for each hour of the day. Only data from Monday to Thursday is considered.

is described in detail later). This illustrates an "economy of density" that implies that both market sides benefit from the fact that density facilitates the matching process.

Additional evidence for the presence of search frictions can be obtained by comparing the actual travel time between observed drop-off location and pick-up locations (the start and endpoint of the search process) with the travel time of the fastest route between these points. To obtain the latter we query Google's distance *API* for the travel time between 1500 randomly selected drop-off and pick-up locations from our data. For each of these observations we computed the ratio of the actual time taxis spent traveling between the two points over the suggested fastest time and find that taxis spend on average 220% more time to travel between these points.

5 Model and Estimation

5.1 Demand Side Model

In the estimation we focus on Manhattan, which, with 93% of all trips in the data accounts for most of the activity.

Since the NYC taxi market operates under a fixed fare system, the endogenous variables of interest that adjust to clear the market are the wait time w_t for passengers to find a taxi, and the search time s_t for taxis to find a passenger.

There are two separate challenges regarding the estimation of the demand function. The first problem is that, while we observe the number of matches

between passengers and taxis, we do not directly observe either the number of waiting passengers (demanded quantity in equilibrium) or their waiting time (the price). The second problem is the issue of simultaneity, which is the typical challenge in the estimation of demand. In this section, we explain how we deal with the first problem to recover the demand data that we need to estimate a demand function. In section 5.1.3, we then describe the instrumental variable approach to deal with endogeneity concerns.

Given a number of searching taxis (which we observe), the average time that a taxi spends searching (which we also observe) reveals information about the number of passengers that must have been waiting on the street. To be more concrete, imagine two scenarios, both have the same number of searching cabs $c_1 = c_2$, but the time they search is higher in scenario one, $s_1 > s_2$. Assume also that all other relevant factors (such as the speed of traffic) are the same in the two scenarios. Then, more passengers must have been waiting on the street in scenario two. Our approach uses this basic intuition.

Let $i \in \{1, ..., I\}$ be the index of an area in the city. The total number of waiting passengers is d_t and they are split up across areas of the city according to proportions $\{p_i^d | i = 1, ..., I\}$.[20] We denote the vector of waiting passengers $\mathbf{d}_t = (d_t \cdot p_1^d, ..., d_t \cdot p_I^d)$. The matching process is captured by a function g that maps a vector of waiting passengers $\mathbf{d}_t$, and searching taxis $\mathbf{c}_t$=$(c_{1t}, ..., c_{It})$ as well as other exogenous time varying variables ϕ_t into an aggregate search time s_t and wait time w_t:

$$\begin{pmatrix} s_t \\ w_t \end{pmatrix} = g(\mathbf{d}_t, \mathbf{c}_t, \phi_t) \tag{1}$$

If we knew $g(\cdot)$, then, for given values of $\mathbf{c}_t$, and ϕ_t, inverting this function from s_t allows us to infer d_t as long as the search time s_t itself is decreasing everywhere in d_t. Without knowledge of the shape of $g(\cdot)$, this inversion is, of course, not feasible. However, we can use our knowledge about the *geographical nature* of the matching process to infer the form of $g(\cdot)$. In particular, we simulate the matching process of waiting passengers and searching taxis on a grid that represents an idealized version of the Manhattan street grid.

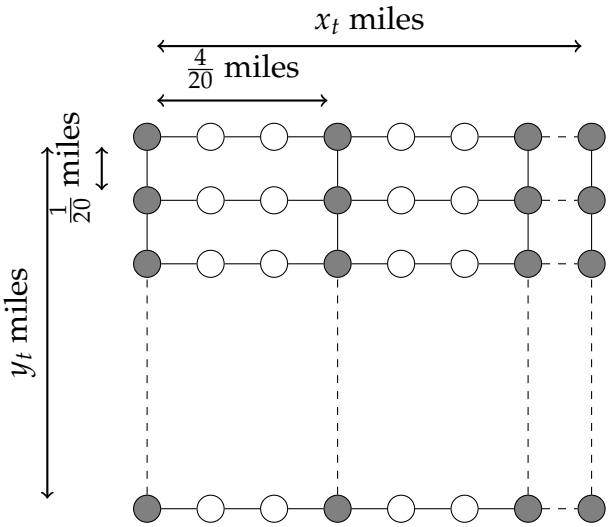
5.1.1 Implementing the Simulation

In the simulation, which provides an approximation of $g(\cdot)$, we assume that passengers are waiting at fixed locations on a two-dimensional grid. The map consists of nodes whose spacing is proportional to $1/20^{th}$ of a mile. Each of these nodes serves as a spot at which passengers potentially wait. Street blocks are assumed to be 40^{th} of a mile wide (east-west) by $1/20^{th}$ of a mile long (north-south),

[20] These proportions are inferred from the data, see 5.1.1 below for explanation

which corresponds to the approximate block size in Manhattan. Figure 4 shows the structure of the resulting grid. Gray nodes represent intersections between streets and avenues, at which cabs can change direction of travel. Turns at (gray) nodes are random with equal probability for each feasible travel direction.

Figure 4: Schematic of the Simulation Grid



Notes: Grey nodes represent street corners, but all nodes (grey and white) are potential passenger locations.

We assume that the effect of time varying factors can be summarized by the speed mph_t at which the traffic flows and the average distance $miles_t$ to deliver a passenger from his fixed position on the grid to the destination. Both factors are included in ϕ_t, and both are directly observed in our data by using the average hourly speed of the entire taxi fleet as well as the average distance of all trips on an hourly basis. For each combination of variables that we feed into g, we simulate the resulting average waiting time for passengers and search time for taxis over an hour long time interval. Every ten minutes $d/6$ potential passengers are born and placed on the map for a total of d passengers during the hour.

To account for the fact that the number of trips originating from different parts of the city varies, we divide Manhattan in eight equally spaced areas (see Figure 18 for details). Passengers appear on the corresponding parts of the grid in proportion to the observed pick-up probabilities of those areas and cabs reappear according to the observed drop-off probabilities. In other words, for each of the passengers placed on the map, we first randomly determine an area according to a multinomial distribution with probabilities $\{\hat{p}_i^d | i = 1, ..., 8\}$ and then a node within an area, where each node has equal probability. The probabilities $\{\hat{p}_i^d | i = 1, ..., 8\}$ are estimated as the fractions of trips originating in these areas. Similarly, cabs re-appear on the map in area i according to multinomial probabilities $\{\hat{p}_i^c | i = 1, ..., 8\}$ and with equal probability on each node within an area.

The probabilities $\{\hat{p}_i^c | i = 1, ..., 8\}$ are measured as the frequencies of drop-offs in those areas.[21] While in this way we partially account for heterogeneity across city locations, we do not endogenize location choices as in Buchholz (2015).

Note that except for the multinomial probabilities none of the steps so far required the use of observed data. This simply generates a representation of $g()$ for any point in its domain.[22] This procedure gives us a simulated version $\hat{g}$ of the matching function. As long as the simulation approximates the true matching process closely enough, we can use $\hat{g}$ to back out d_t for each combination of hourly averages of search time s_t, number of taxis c_t, traffic speed, and trip distance observed in the data. Once the number of passengers is known we can insert it into $\hat{g}$ to determine wait-time w_t as well. Additional details on the simulation, including the algorithm, are provided in Appendix B.

5.1.2 Properties of the Matching Function

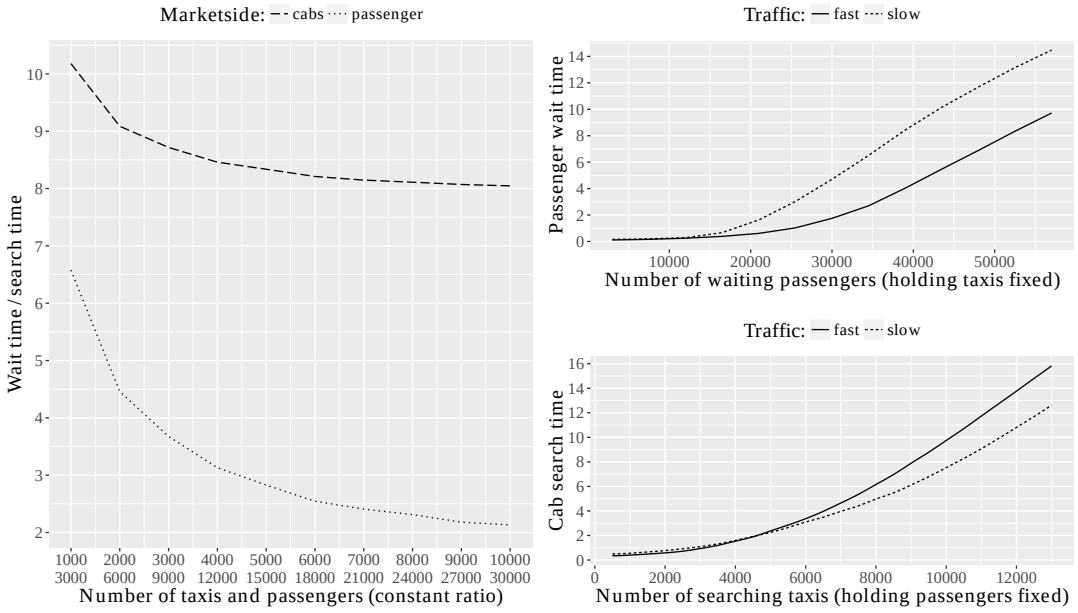
Figure 5 graphically illustrates properties of the matching function. The figure on the left shows what happens to wait time and search time as the market becomes thicker. On the horizontal axis the number of taxis and passengers vary while keeping constant the ratio between the two. Both passenger wait-time and search-time decrease as the market becomes thicker, but the improvements decrease relatively quickly. It should be noted that the average number of daily taxis is about 7800. At this number of taxis and passengers, additional returns to market thickness are fairly small, especially for drivers. However, this would not be the case in more sparsely populated cities.

The figures on the right give a sense of how market tightness affects outcomes for both market sides. On the top right panel, the number of passengers is increased while holding fixed the number of taxis at the median observed in the data. In the bottom right panel, the number of taxis varies, holding fixed the number of passengers at the median. Both figures also display level changes due to differences in traffic speed, one of the exogenous inputs to the matching func-

[21] Note that conditional drop-off and pick-up probabilities would only be of interest if the identity of drivers were to make a difference. For the purpose of this simulation, however, driver identities are irrelevant. We will hold those probabilities fixed throughout, since we treat them as an exogenous process.

[22] Due to computational limitations, it is not feasible to repeat this simulation for each point in the domain of g. If, for example, we assume that in an hour there are at most $70,000$ passengers waiting, and multiply this by the maximal number of medallions, we would already obtain 945 million different points in the domain without even considering variation in ϕ. We therefore simulate g for a lower number of grid points and we interpolate linearly between those points to obtain the image for points in between. For each of the four independent variables of $g(.)$, we pick eight different evenly spaced grid points. Because the outcome of the matching process is random, we have to repeat the simulation multiple times for each of those points. In practice we have found that the average of these simulations does not change much after ten iterations, which is therefore what we use to produce this average.

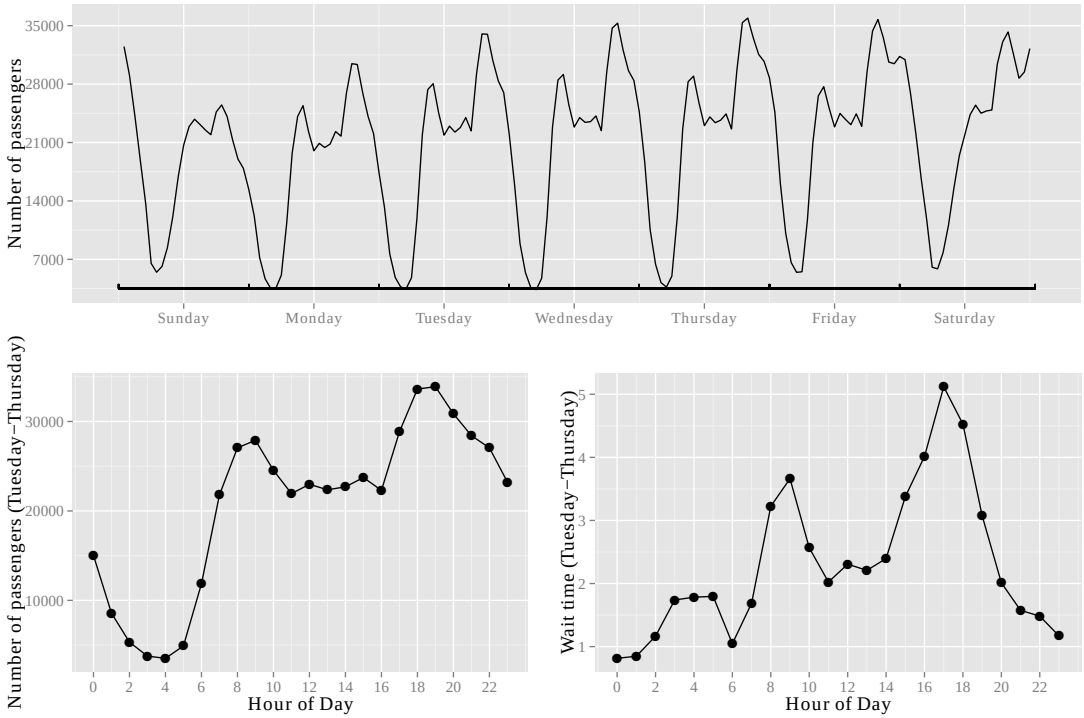
Figure 5: Graphical Illustration of Matching Function



tion. The dotted line is obtained under a traffic speed that is one standard deviation below the median and the solid line is for traffic speed that is one standard deviation above the median. While faster traffic speed is unequivocally good for passengers, there are two contrasting effects for cabs. On the one hand, faster traffic speed allows an individual cab to more quickly reach a waiting passenger, thereby reducing search time. On the other hand, faster traffic speed speeds up aggregate deliveries of passengers, and therefore leads to an increase in competition, effectively increasing the number of available cabs. At the median number of observed passengers, the latter effect outweighs the former at about 4000 cabs. We will make use of this effect in our demand instrument.

Figure 6 shows the wait time and the number of passengers for an average weekday as well as the number of passengers for an entire average week based on an inversion of our simulated $\hat{g}(.)$. The passenger figures confirm an expected pattern of strong rush hour demand in the morning and evening hours on all weekdays. On Sundays we find that demand is lower than during weekdays, and there is also no clear division between morning and evening rush hours. This, again, appears to be reasonable. One can see that the wait time in the morning rush hour spikes exactly when demand peaks between the hours of 7AM to 10AM. This stands in contrast to the peak in wait time in the evening that occurs at 5PM, before the spike in demand occurring at 7PM. The reason for this pattern is the coordinated shift change (witching hour) that leads to the more unfavorable ratio of active cabs to searching passengers.

Figure 6: Graphical Results of Demand and Wait Time



5.1.3 Estimating the Demand Function

Now that we have obtained the number of passengers and the wait time, we need to estimate a demand function that relates these two variables. We assume a constant elasticity demand function of the following form:[23]

$$d_t = exp(\beta_0 + \sum_{h_t} \beta_{h_t} \cdot \mathbf{1}\{h_t\} + \mathbf{x} \cdot \beta_x) \cdot w_t^{\eta} \cdot exp(\xi_t). \tag{2}$$

The multiplicative component $exp(\beta_0 + \sum_{h_t} \beta_{h_t} \cdot \mathbf{1}\{h_t\} + \mathbf{x} \cdot \beta_x)$ captures observed exogenous factors that may shift demand, as well as persistent unobserved components through dummy variables and ξ_t captures *unobserved* time varying conditions that shift demand. The main parameter of interest is η, the elasticity of demand with respect to waiting time. Taking logs, demand can be estimated as a linear model:

$$log(d_t) = \beta_0 + \sum_{h_t} \beta_{h_t} \cdot \mathbf{1}\{h_t\} + \mathbf{x} \cdot \beta_x + \eta \cdot log(w_t) + \xi_t. \tag{3}$$

[23]Since we break up the data down to hourly levels and only use a subset of weekdays, we are left with slightly more than 700 observations for this estimation. The assumption of log-linearity is common in such a case, see for example Kalouptsidi (2014).

A potential problem is that the wait time itself is a function of the number of passengers as well as the number of cab drivers. In particular, unobserved factors that shift demand will directly appear in w_t. Furthermore, drivers may condition their decisions on factors included in the error term ξ_t, which would lead to a decrease in waiting time. For both of these reasons we have to expect that the error term ξ_t and the wait time w_t are correlated. This would, of course, introduce a bias in the estimation of η. To address this concern we instrument for the wait time. We need a variable that is correlated with wait time and that affects demand only through wait time. A supply shifter satisfies this requirement. In our preferred specification, we instrument for the wait time with the traffic speed *outside* of Manhattan.

Table 1: Different Specifications for Demand Estimation.

	First Stage $log(w_t)$	First Stage $log(w_t)$	Second Stage $log(d_t)$	Second Stage $log(d_t)$
$log(mph_t)\|OutsideManhattan$	0.800**			
	(0.167)			
$log(mph_t)\|InsideManhattan$	-3.252**	-2.373**	-2.597**	-2.379**
	(0.223)	(0.113)	(0.571)	(0.104)
Shift Instrument		0.392**		
		(0.0165)		
$log(w_t)$			-0.596*	-0.396**
			(0.237)	(0.0442)
Observations	714	714	714	714
Hour FE				
2-Hour FE				
R^2	0.923	0.870	0.965	0.957
F	4168.9	2019.8		

Note: $+\ p < 0.10, *\ p < 0.05, **\ p < 0.01$, All regressions are based on our subset of 2011 and the August 2012 trip sheet data (we only use the 2012 data before the fare change), excluding Fridays, Saturdays and Sundays. An observation is comprised of an hourly average over all trips in that hour. Standard errors clustered at the date level.

The first stage shows that, controlling for traffic speed inside of Manhattan, the sign of traffic speed outside of Manhattan is positive. Traffic speed has two opposing effects on drivers earnings, on the one hand it increases search time (since trips are finished faster and more taxis therefore search, see Figure 5). On the other hand, earnings per minute of driving time are higher on faster trips due to the structure of metered fares. The latter effect outweighs the former if travel distances are longer, and trips outside of Manhattan are on average much longer. It seems plausible that high traffic speed makes it differentially more attractive to serve the outer boroughs, reducing the number of cabs in Manhattan.[24] We have

[24] A substantial fraction of taxi rides outside Manhattan are airport rides. We have in fact veri-

explored an alternative instrument based on the shift transition of taxis. Appendix D provides an institutional explanation for why shift transitions can be considered a good instrument for supply shifts. Since this is a coarse instrument, it limits the extent to which this specification allows us to control for persistent unobservables. For instance, under this specification, we cannot control for time of day dummies at the hourly level. Partly for this reason, we prefer the specification with the traffic speed instrument. In any event, the two instruments lead to estimates of demand elasticities that are of similar magnitude. Table 1 shows the results of our demand estimation. The first two regressions are the first stages for both respective instruments. In our preferred specification we use the traffic speed outside of Manhattan as a supply shifter. The latter two specifications are the second stages. All second stage regressions include traffic speed in Manhattan as a control. Our preferred specification leads to an estimated elasticity of approximately -0.6, whereas the specifications in which we use the shift change as an instrument reduces this estimate to approximately -0.4. It is also worth pointing out that the high R^2 in the specifications with dummy variables highlight the fact that most of the hourly variation in the market is captured by these relatively parsimonious specifications.

Figure 7: Demand Function Evaluated at the Mean of Wait-time

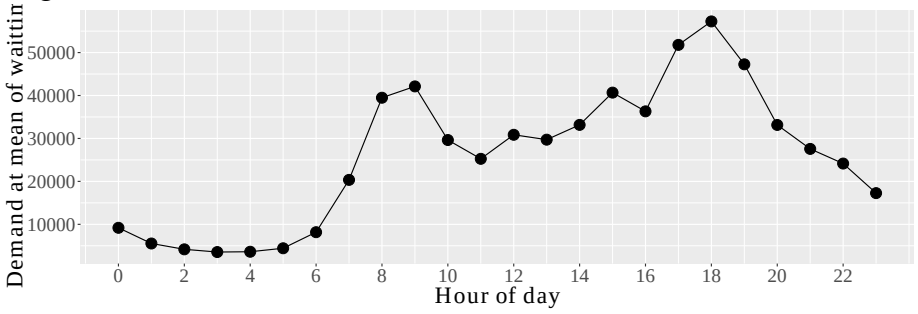


Figure 7 shows what demand would be if the waiting time was held fixed at the daily mean throughout the day. This highlights how the inelastic portion of demand is varying throughout the day. All else being equal, demand would be highest in the evening hours between 5PM and 7PM and lowest in the morning hours from 2AM to 7AM.

fied that the fraction of taxis that wait for rides at JFK airport is in fact increasing in traffic speed from the airport into Manhattan.

Note, however, that the sign of traffic speed outside of Manhattan is only positive after controlling for traffic speed inside of Manhattan. This could be because traffic speed, in general, is also a proxy for unoberved demand shocks and also might shift the relative attractiveness of ride hailing transportation relative to the subway.

5.2 Supply Side Model

In modelling the driver's problem, we wish to incorporate the regulatory and organizational constraints imposed by the medallion system that we discussed in section 3. In our model agents make their decisions in discrete hourly intervals. We interpret those as representative hours of a weekday.

At each hour t there are N_t active and M_t inactive medallions, which sum up to the total number (13,520) of medallions issued by the city. Each hour an empty medallion is probabilistically filled with a driver, depending on the outside option of drivers and their optimal strategy. Active drivers make an hourly stopping decision. As mentioned above, we focus on Monday-Thursday. Agents form expectations about earnings for each weekday hour. We let h_t be an index denoting hours from the set $H = \{1, ..., 24\}$.

In section 3 and section 4 we highlighted the fact that minifleet medallions are more heavily utilized than owner-operated medallions, and are also more likely to transition between shifts at 5PM, and therefore over-proportionally contribute to the supply shortage at that time of the day. Because of these differences between types of medallion, we allow for two types of heterogeneity: the first captures the medallion type, the second captures the time at which a medallion typically transitions between shifts. For a driver i the index $z_i \in \{Fleet, Own\}$ denotes the ownership of the medallion and takes on those values depending on whether it is a minifleet or an owner-operated medallion. We allow the model parameters to be different for minifleet and owner-operated medallions. The index $k_i \in K$ captures the time at which the medallion transitions between day-shift and night shift. We allow for four possible transitions in the morning and four in the evening. Drivers have to pay fines when they return the case after the end of their shift. The transition type determines when those fines need to be paid. If, for example, the most common transition time in the morning (as measured by the mode) is 5PM, we assume that a driver has to pay a fine for handing in the medallion later than this.[25] To sum up, the model allows medallions to vary along two dimensions, one is the type (minifleet vs. owner operated) and the second is the transition time.

We first discuss the optimal stopping decision for a driver who is already on a shift, then we discuss the decision to start a shift. The value of a shift conditional

[25] Assuming such fixed times for fines partially endogenizes transition time and gives the model some flexibility to match patterns where drivers drive longer than in normally the case. However, transition times are in principle endogenous and one could imagine that drivers take into consideration daily variation in the value of day-shifts and night-shifts to time the transition in a negotiation process. Such a fully endogenous model is currently outside the scope of this paper. However, as we argue in Appendix D, the current regulatory arrangements limits the flexibility of transitions.

on state $(\boldsymbol{x}_{it}, \pi_t, \boldsymbol{\epsilon}_{it})$ is given by:

$$\begin{aligned} V(\boldsymbol{x}_{it}, \pi_t, \boldsymbol{\epsilon}_{it}) = max\{\epsilon_{it0}, \pi_t - C_{z_i,h_t}(l_{it}) - f(h_t, k_i) + \epsilon_{it1} \\ + \mathbb{E}_{\pi_{t+1}, \boldsymbol{\epsilon}_{i(t+1)}}[V(\boldsymbol{x}_{i(t+1)}, \pi_{t+1}, \boldsymbol{\epsilon}_{i(t+1)})|h_{t+1}]\} \end{aligned} \tag{4}$$

At the beginning of each hour the driver decides whether she wants to collect the flow payoff from driving plus the continuation value of an active shift, or the random value of the outside option.[26] This expression depends on an observable state vector $\boldsymbol{x}_{it} = (h_t, l_{it}, z_i, k_i)$, the realization of hourly earnings π_t (drawn from an endogenous distribution $\mathcal{F}_\pi(.|h_t)$ that depends on the hour). as well as an idiosyncratic unobservable vector $\boldsymbol{\epsilon}_{it} = (\epsilon_{i0}, \epsilon_{it1})$, assumed to be distributed according to i.i.d. $T1EV$ distributions with scale parameter σ_ϵ. The cost of driving $C_{z_i,h_t}(l_{it})$ is a function of the length l_{it} of the shift. It allows for an increasing (effort) cost of driving as the shift gets longer. The parameters of this function are indexed both by the medallion type z_i and by the hour-weekday combination h_t. We interpret this cost function as a combination of the hourly opportunity cost of driving, which may vary throughout the day, as well as the disutility of driving. We assume that the cost function takes the following form:

$$C_{z_i,h_t}(l_{it}) = \lambda_{0,z_i,h_t} + \lambda_{1,z_i} \cdot l_{it} + \lambda_{2,z_i} \cdot l_{it}^2.^{27}$$

While the fixed cost components can depend on the hour, we only allow for two different values of λ_0 for each 12-hour shift.

The term $f(h_t, k_i)$ is a fine that has to be paid if the medallion is delivered late to the next driver. Such fines are very common to insure that drivers do not operate the cab longer than contractually specified. Since fines are not directly observable we estimate them as parameters. We use the fact that we can see the same medallions operate over a long period of time to classify each into a category k_i of morning and evening transition times. For each medallion we compute the most common starting hour (the mode) in the morning as well as in the evening. We then assume that a morning driver has to pay a fine f_{z_i} whenever he goes past the common night shift starting time and analogously for a night driver. The fine is again indexed by z_i to account for the fact that owner-operated medallions seem to have less stringent transition times. We denote by h_{k_i} the set of hours for which a driver of medallion k_i is subject to the fine when driving. We therefore have: $f_{z_i}(h_t, k_i) = f_{z_i} \cdot \mathbf{1}\{h_t \in h_{k_i}\}$.

We now discuss the decision of whether to start a shift. For each hour t dur-

[26] We assume that there is no discounting since these are hourly decisions.

In the estimation the maximal shift length at we allow is 13 hours. This is longer than the regulatory maximum of 12 hours and it is surpassed only in a very small number of cases.

[27] We have also explored a specification with higher order polynomials but this does not make a difference in the results.

ing which a medallion is inactive, a driver i has an opportunity to be matched with this medallion. He decides to enter if the value of driving is higher than his outside option over the expected optimal length of a shift. If he decides not to enter, then the medallion will be available to another potential driver the following hour. We assume that the utility from the outside option is comprised of a fixed value μ_z that depends on the medallion type and on whether it is a day or night shift, as well as an idiosyncratic random component v_{it0}. The utility of driving depends on the entire expected value of a shift (described below) as well as an idiosyncratic component v_{it1}.[28] We assume that v_{it0} and v_{it1} are i.i.d. random variables distributed according to a Type 1 Extreme Value ($T1EV$) with scale parameter σ_v.[29] Drivers also have to pay a daily rental fee r that depends on whether they drive during the day shift or the night shift. If they own the medallion they have an opportunity cost of driving equal to r . We set r_{h_t} equal to the rate caps, which, according to anecdotal evidence, were always binding during the data period.[30] The deterministic part of the state vector is $\boldsymbol{x}_{it} = (h_t, l_{it}, z_i, k_i)$. Note that h_t as well as l_{it} progress deterministically and that z_i and k_i do not change over time.

To summarize, the utility of the outside option is given by:

$$u_{it0} = \mu_{h_t,k} + v_{it0},$$

and the utility of starting a shift is given by:

$$u_{it1} = \mathbb{E}_{\pi_{t+1},\boldsymbol{\epsilon}_{i(t+1)}}[V(\boldsymbol{x}_{i(t+1)}, \pi_{t+1}, \boldsymbol{\epsilon}_{i(t+1)})] - r_{h_t} + v_{it1}.$$

As is well known, a convenient feature of assuming a $T1EV$ distribution for the error terms is the closed form expression for the choice probabilities. Denoting by $q(\boldsymbol{x}_{it})$ the probability that an inactive driver starts a shift at time t, conditional on state $\boldsymbol{x}_{it}$, we obtain:

$$q(\boldsymbol{x}_{it}) = \frac{exp((\mathbb{E}_{\pi_{t+1},\epsilon_{i(t+1)}}[V(\boldsymbol{x}_{t+1}, \pi_{t+1}, \epsilon_{i(t+1)})|h_{t+1}] - r_{h_t})/\sigma_v)}{exp((\mathbb{E}_{\pi_{t+1},\epsilon_{i(t+1)}}[V(\boldsymbol{x}_{t+1}, \pi_{t+1}, \epsilon_{i(t+1)})|h_{t+1}] - r_{h_t})/\sigma_v) + exp(\mu_{h_t}/\sigma_v)}.$$

[28] Of course, all that matters for drivers' choices is the difference between v_{it0} and v_{it1}.

[29] The scale parameter σ_v is identified because $\mathbb{E}_{\pi_{t+1},\epsilon_t} V(\boldsymbol{x}_{t+1}, \pi_{t+1}|h_{t+1})$ is a given value from the stopping problem and not pre-multiplied by any parameter.

[30] Recall that the sample period is before UBER became important in NYC. In 2015 it had already become clear that rental rate caps were no longer binding.

5.2.1 Equilibrium Definition

Hourly earnings π_t are determined by the equilibrium distributions of active medallions $\mathcal{F}_c(.|h_t)$ and searching passengers $\mathcal{F}_d(.|h_t)$. Drivers forecast their earnings according tothe distribution of earnings $\mathcal{F}_\pi(.|h_t)$. The distributions $\mathcal{F}_c(.|h_t)$ and $\mathcal{F}_d(.|h_t)$ determine a distribution of wait times $\mathcal{F}_w(.|h_t)$ and search time $\mathcal{F}_s(.|h_t)$ in a way that depends on the matching process.

Definition 1 *A competitive equilibrium in the taxi market is a set of distributions* $\{\mathcal{F}_s(.|h_t), \mathcal{F}_w(.|h_t), \mathcal{F}_c(.|h_t), \mathcal{F}_d(.|h_t), \mathcal{F}_\pi(.|h_t) : h_t \in H\}$*, such that:*

1. $\mathcal{F}_d(|h_t)$ *results from the demand function* d_t (w_t) *under the distribution of waiting times* $\mathcal{F}_w(.|h_t)$ *and demand shocks* ξ_t*.*
2. $\mathcal{F}_s(.|h_t)$ *and* $\mathcal{F}_w(.|h_t)$ *result from* $\mathcal{F}_d(.|h_t)$ *and* $\mathcal{F}_c(.|h_t)$ *under the matching function* $g(.)$*.*
3. $\mathcal{F}_c(.|h_t)$ *results from optimal starting and stopping under* $\mathcal{F}_\pi(.|h_t)$*.*
4. $\mathcal{F}_\pi(.|h_t)$ *results from the distribution of search times* $\mathcal{F}_s(.|h_t)$*.*

It is worth pointing out that the main sources of aggregate uncertainty in the model are shocks to demand, traffic speed, and trip length.[31] The idiosyncratic uncertainty on the supply side, such as the random outside options, averages out across the large number of taxis. We do not allow for autocorrelation in the shocks or earnings. An inspection of the demand regressions in Table 1 shows that most of the variation is explained by hourly fixed effects (which enters the structural model through our specification for the demand function). Most of the variation in search and wait times is alsocaptured by hour of day fixed effects. Allowing for autocorrelation in earnings or idiosyncratic shocks would therefore change results only slightly albeit adding significantly to computational cost.

5.3 Identification of Supply Side Parameters

In this section we briefly discuss how the primitives of the model are identified. The cost function has multiple components: **(1)** There is a term that varies with the duration of the shift, **(2)** an hourly fixed component, **(3)** the fines, **(4)** the standard deviation of the hourly outside option, (**5**) the mean of the daily outside option, and (**6**) the standard deviation of the daily outside option. The identification of **(1)** can be best understood by using backwards induction for the driver's

[31] Hourly earnings are determined as $\frac{e(miles,mph)}{e(miles,mph)+s(d,c,miles,mph)} \cdot 60 \cdot \pi^0$, where e is the expected trip length and π^0 is the rate that drivers earn per minute of driving. Search time s is determined under the matching function $g(d, c, miles, mph)$.

decision problem. At the maximum allowed shift length (drivers are not allowed to drive more than 12 consecutive hours), the continuation value is zero; thus, the driver only compares expected income in that last hour against the cost of driving. The value of the cost function for l_{max} is therefore determined to match the expected income in the last shift hour, which is a data object. Once the value of the cost function in the last hour is identified, it determines the continuation value from the perspective of the preceding hour. Hence, the second to last value of the cost function is identified: earnings in that hour and the continuation value are composed of data and identified objects. We can repeat the argument until we reach the first hour of the shift. However, **(2)** is also dependent on the hour of the day. This part of the cost function is identified by systematic inter-temporal variation in the stopping probabilities throughout the day even after conditioning on shift-length and earnings. For example, the stopping probability increases sharply after 12pm even though there is not a contemporaneous sharp decline in the earnings. This kind of variation in the data identifies the differences in the λ_0 values. **(3)** is identified by the increase in the stopping probabilities at those times, again after conditioning on shift-length and expected earnings. **(4)** is identified by the variation in the earnings π_t. This concludes the identification of the value function, which can be treated as a known object for the discussion of the primitives of the entry decision. The varying values throughout the day of $\mathbb{E}_{\pi_{t+1},\boldsymbol{\epsilon}_{i(t+1)}}[V(\boldsymbol{x}_{i(t+1)}, \pi_{t+1}, \boldsymbol{\epsilon}_{i(t+1)})] - r_{h_t}$, which is composed of data and identified objects, identify the different values of μ_{h_t} (**5**) and their dispersion, i.e., the value of σ_v (**6**).

5.4 Estimation

5.4.1 Constructing the Data for Supply Side Estimation

To estimate the model, the trip based TPEP data has to be transformed into shift dataset where the unit of observation is a medallion-hour combination. For estimation we use data from 2011 as well as the August data from 2012.

Shifts are defined following Farber (2008) who determines them as a consecutive sequence of trips where breaks between two trips cannot be longer than five hours. This definition might sometimes lead to long breaks within a shift if there is a long interval between two trips. This conflicts with our assumption that drivers plan with the conditional steady state distribution of wages $\mathcal{F}_\pi(.|h_t)$ for each hour of their shift. Since we do not model breaks we instead assume them to be an exogenous process. To that end we estimate the likelihood of a break for each hour conditional on the state and compute hourly earnings as the expected wage that is earned while searching for passengers multiplied by the probability that the driver is not on a break. Formatting the data this way leads to 9,562,892 medallion-hour observations during which medallions have been active in a shift as well as 5,747,837 medallion-hour observations during which medallions have

been inactive. From this data we drop shifts that are only one hour long, which make up less than 0.3% of the active shift data. The reason for this is that the chance of stopping is slightly higher after the first hour than after the second hour, whereas it is monotonically increasing afterwards. This indicates that these disparate hours might be part of an interrupted longer shift and that this is not captured by the shift definition used here.

The wait time for passengers links the aggregate market conditions to the hourly earnings potential of drivers. Remember from the discussion above that earnings are a result of a combination of time-based and mile-based metering. We first calculate the *actual hourly based rate* π_t^0 for each trip by dividing the total fare of each trip by the duration of a trip. These rates also include the tip that drivers earn. Since tips are only recorded for credit card transactions, we impute tips for trips that have been paid in cash.[32] For each hour we also compute the average search time for a taxi to find a passenger as well as the average trip length. Before we compute these averages all variables are winsorized at the 1% level to avoid averages being driven by large outliers in the data. Based on these hourly averages we can then compute a realization of the hourly wage rate as $\pi_t = \pi_t^0 \cdot (e_t/(e_t + s_t))$, i.e. the actual hourly rate times the fraction of the time the driver is delivering a passenger as opposed to searching.

5.4.2 Estimation Procedure

For the estimation of supply side parameters we make use of the fact that, given a known set of distributions for hourly equilibrium earnings, $\mathcal{F}_\pi(.|h)\ \forall h \in H$, we can compute the supply side problem as if it were a single agent decision problem against these equilibrium earnings. In other words, since we observe equilibrium earnings directly in the data, there is no need to compute equilibria in the estimation. For the estimation of the supply side problem we also make use of the fact that the dynamic decision problem can be formulated as a constraint on the likelihood for starting and stopping probabilities of drivers. This approach is known as mathematical programming with equilibrium constraints (MPEC).[33]

In our case this constraint comes from the assumption that the data is generated by a model of optimal starting and stopping decisions. Since the latter is a dynamic decision problem, one would normally iterate on the contraction map-

[32] We first run a regressions with hourly dummy variables predicting the tip rate for each hour of the day. Predicted rates are then used to impute the tips for trips where the tip is not observed. About 47% of all transactions are paid by creadit card.

[33] Su and Judd (2012) demonstrates the computational advantage of MPEC over a nested fixed point computations (NFXP) in the classical example of Rust's bus engine replacement problem, Rust (1987). Applications of MPEC to demand models and dynamic oligopoly models can be found in, for example Conlon (2010) and Dubé et al. (2012). An intuitive explanation for the computational advantage of MPEC is that the constraints imposed by the economic model are not required to be satisfied at each evaluation of the objective function.

ping to solve for the value function for each parameter guess $\hat{\theta}$.[34] MPEC allows the constraint imposed by the value function to be slack during the search but makes sure that they are satisfied for the final set of recovered parameters.[35]

We specify a likelihood objective function. Following the model discussion above, we allow all parameters to be different for minifleet and owner-operated medallions. For a driver j we allow for two different daily outside option μ_{z_j}, a parameter from 5AM to 5PM and one for the remaining time. We also allow the fine f_{h_t,z_j} to depend on whether the driver is on a night f_{0,z_j} or day shift f_{1,z_j}.

The constant part of the cost function is allowed to vary by hour in the following way: from 12AM to 5AM, from 5AM to 12PM, and from 5PM to 12AM. The other two parameters of the cost function λ_{1,z_j} and λ_{2,z_j} are assumed to be time invariant. The remaining parameters are the standard deviation of the idiosyncratic shocks to the starting decision σ_{v,z_j} and the stopping decision σ_{ϵ,z_j}. We will refer to the combined vector of parameters as θ.

5.5 Parameter Estimates

Table 2 gives an overview of the estimated parameters. Results are shown separately for minifleet and owner-operated medallions. Standard error calculations are bootstrapped: we drew 50 samples with replacement at the medallion level.

Table 2: Parameter Estimates (standard errors in parentheses)

parameter	**description**	**minifleet** $(z_j = F)$	**owner-operated** $(z_j = NF)$
$\mu_{z_j,0}$	outside-option, 6pm-4am	354.16 (15.79)	476.34 (20.45)
$\mu_{z_j,1}$	outside-option, 5am-5pm	372.23 (16.44)	472.49 (20.58)
f_{0,z_j}	fine (nightshift)	95.68 (3.61)	105.92 (4.0)
f_{1,z_j}	fine (dayshift)	94.92 (3.3)	88.92 (2.99)
$\lambda_{0,z_j,0}$	fixed cost (1am-5am),	38.9 (0.51)	59.92 (1.14)
$\lambda_{0,z_j,1}$	fixed cost (6am-12pm),	14.42 (0.69)	22.41 (0.67)
$\lambda_{0,z_j,2}$	fixed cost (1pm-5pm),	0.0 (1.03)	17.85 (0.73)
$\lambda_{0,z_j,3}$	fixed cost (6pm-12am),	10.06 (0.94)	16.39 (1.01)
λ_{1,z_j}	linear cost coefficient	0.0 (0.0)	0.0 (0.0)
λ_{2,z_j}	quadratic cost coefficient	0.5 (0.01)	0.58 (0.02)
σ_ϵ	sd iid hourly outside option	56.82 (2.0)	80.0 (2.78)
σ_v	sd iid daily outside option	59.5 (2.09)	61.75 (2.28)

[34] Note that there are other suggestions in the literature that would avoid the nested fixed point computation, such as Bajari et al. (2007) where value functions are forward simulated.

[35] A second advantage of MPEC is that it provides a convenient way of specifying an optimization problem in closed form, which allows the use of a state of the art non-linear solver. In this paper we use the JuMP solver interface (Lubin and Dunning (2013)), which automatically computes the exact gradient of the objective function as well as the exact second-order derivatives. JuMP also automatically identifies the sparsity pattern of the Jacobian and the Hessian matrix.

The mean values of the outside option for night ($\mu_{z_j,0}$) and daytime ($\mu_{z_j,1}$) are estimated to be \$354.16 and \$372.23 for minifleet drivers and \$476.34 and \$472.49 for owner-operated drivers. As we discussed in the identification section, these values are pinned down by the values of starting a shift. The value for minifleets is therefore lower at all times of the day, consistent with the descriptive evidence that minifleet shifts last longer. The estimated fines f_{0,z_j} and f_{1,z_j} for violating the shift constraint are \$95.68 (nightshifts) and \$94.92 (dayshift) for minifleet medallions. For owner-operated medallions the night-shift fine is \$105.92 and the dayshift fine is \$88.92. The fixed part of the cost function parameters for minifleets are estimated at \$38.9 from 1am to 5am, \$14.42 from 6am to 12pm, \$0.0 from 1pm to 5pm and \$10.06 from 6pm to 12am. For owner-operated medallions the corresponding values are \$59.92, \$22.41, \$0.0, and \$0.58. The linear parameter of the hourly increase in cost is estimated to be zero for both minifleet and owner-operated cabs and the quadratic parameter are 0.5 for minifleets and 0.58 for owner-operated medallions. The standard deviations of the hourly outside option (conditional on driving) is 56.82 for minifleet medallions and 80.0 for owner-operated medallions. The standard deviations of the daily outside option is 59.5 for minifleet medallions and 61.75 for owner-operated medallions.

5.5.1 Discussion of Parameter Estimates

A few observations about the estimates are worth highlighting. As shown in Figure 1, minifleet medallions follow the 5AM to 5PM shift pattern much more stringently than owner-operated taxis. This is reflected in the estimates. Owner-operated medallions have a higher standard deviation in the hourly error terms, which are the random parts that determine stopping behavior. This leads to their stopping behavior being "smoother", i.e. having a higher percentage of short shifts. Ceteris paribus, a larger standard deviation moves the stopping probabilities towards one-half as can be seen by inspecting Equation 8. To induce stopping of medallions near the end of the shift, the cost function and the fines therefore have to be higher for owner-operated medallions compared to minifleets, which is indeed the case. Lastly, we see that the daily outside option for owner-operated medallions is larger than for minifleets. This outside option captures all the surplus from driving, which is the wage plus the continuation value minus the cost of driving. This surplus is larger for owner-operated medallions because the cost function is steeper (Figure 13 in Appendix A), and the expected random term for continuing to drive is larger because of the larger standard deviation of the hourly outside option.

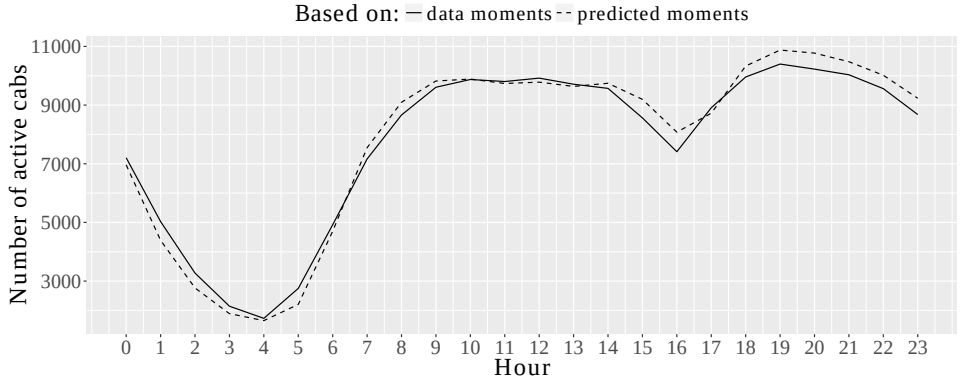
5.5.2 Model Fit

To evaluate the model fit we transform the drivers decision problem into a law

of motion for medallions. Medallions that are "available" for starting drivers are those that are unutilized or in the last hour of their shift. The probability that an active medallion becomes inactive is the probability that the driver who utilizes stops and no other driver decides to utilize it in the same hour: $\hat{p}^M(t) = \hat{p}(h_t) \cdot (1 - \hat{q}(h_t))$. The probability that an inactive medallion becomes active is the probability that a driver starts utilizing an inactive medallion $\hat{q}^M = \hat{q}(h_t)$. The stopping probabilities unconditional on the shift length are obtained from the conditional stopping probabilities. Let N_t be the number of inactive medallions and M_t be the number of active medallions. The law of motion for medallions discretized into hourly intervals is: $N_t = (1 - p_t^M) \cdot N_t + q_t^M \cdot M_t$ and $M_t = (1 - q_t^M) \cdot M_t + p_t^M \cdot N_t$.

The model fit for the law of motion is presented in Figure 8, which shows that we are able to replicate the daily pattern of supply activity quite well.

Figure 8: Model Fit



6 Counterfactual Experiments

As we argued earlier, two important inefficiencies in the taxi market are regulatory entry barriers and search frictions. Part of the success of startup ride hailing services can be attributed to the fact that they address both of these inefficiencies.[36] Our first set of counterfactuals tries to separate out the effects of additional entry and of more efficient matching. We also investigate whether market segmentation between different operators may lead to inefficiencies: the introduction of a dispatch system à la Uber that only covers part of the market may reduce overall market thickness. Our last counterfactual is motivated by a recent policy proposal of the TLC, which wants to lift the requirement of owner

[36] An additional effect is that supply is made more responsive through surge pricing. Since we have no estimate of price elasticities we cannot address this issue here.

operation. The model allows us to look at a policy that assigns all medallions to minifleet companies, which utilize medallions more, as we show in Figure 1.

For all counterfactuals we highlight changes in the average number of active cabs, the number of passengers, hourly driver revenues, discounted medallion revenues, as well as consumer surplus (measured in wait time), which we compute as

$$CS(w^*) = \sum_{h \in \{0...23\}} \int \int_0^{d_h(w_h^*)} (w_h(d_h) - w_h^*)) dd_h dF(d_h). \tag{5}$$

Medallion revenue streams are obtained via simulation: we compute the number of times a medallion can be rented out in a year and then use this to compute the present discounted revenue under an annual interest rate of 3%. Revenues are averaged over the different types of medallions according to their observed fractions in the data.

All counterfactuals are computed in two steps. We first compute an equilibrium in which we do not allow demand to expand. This scenario is then compared to the full counterfactual in which demand is allowed to adjust.

In the estimation there was no need to compute market equilibria since the supply side parameters were estimated using the observed process of hourly earnings. For counterfactuals we have to address the challenge of equilibrium computation, which is demanding since all endogenous objects vary by hour of day. Instead of an exact solution, which would require iteration over 48 distributions (one demand distribution and one supply distribution for each hour of day), we approximate these by normal distributions. Thus, for each hour one needs to keep track of the first two moments of each endogenous distribution, leaving us with 94 endogenous variables. We iteratively solve for the supply and demand side parameters, holding the variables of the respective other side of the market fixed until updates on both sides of the market fall below a threshold. The exact algorithm is described in detail in item C.

Table 3: Counterfactual Results

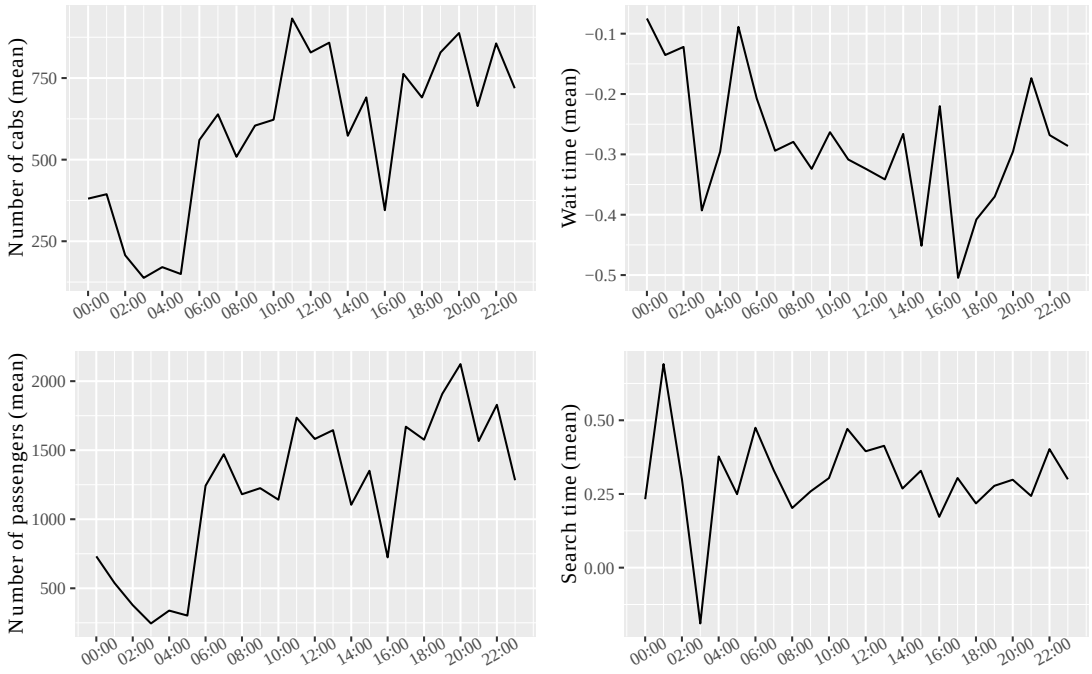
	hourly active cabs	**hourly demand**	**passenger waittime**	**taxi searchtime**	**hourly taxi revenues**	**consumer surplus (minutes)**	**number of trips per day**	**medallion revenue**
Baseline	7759.0	25111.0	4.01	7.62	40.14	2.67 Million	556987.0	2.54 Million
Entry	8343.0	26315.0	3.7	7.93	39.5	2.85 Million	589536.0	2.49 Million
Δ%	7.52	4.79	-7.53	4.09	-1.59	6.85	5.84	-1.69
Entry **(PE)**	8156.0	25111.0	3.54	8.33	38.76	2.97 Million	565454.0	2.44 Million
Δ% **(PE)**	5.12	0.0	-11.73	9.33	-3.44	11.29	1.52	-3.79
Dispatcher	8314.0	27440.0	3.48	6.63	42.34	2.97 Million	629090.0	2.67 Million
Δ%	7.15	9.28	-13.17	-12.94	5.48	11.14	12.95	5.4
Dispatcher **(PE)**	7927.0	25111.0	3.21	7.45	40.83	3.15 Million	578569.0	2.57 Million
Δ% **(PE)**	2.17	0.0	-19.85	-2.24	1.73	18.16	3.87	1.46
Dispatcher (%50)	7663.0	24492.0	4.19	7.8	39.9	2.56 Million	545466.0	2.5 Million
Δ%	-1.23	-2.47	4.55	2.46	-0.59	-3.91	-2.07	-1.2
Dispatcher (%50) **(PE)**	7799.0	25111.0	4.25	7.63	40.1	2.55 Million	558378.0	2.53 Million
Δ% **(PE)**	0.52	0.0	5.99	0.12	-0.1	-4.29	0.25	-0.2
Dispatcher (%10)	8335.0	25968.0	3.8	7.94	39.78	2.78 Million	589018.0	2.5 Million
Δ%	7.43	3.41	-5.15	4.27	-0.88	4.15	5.75	-1.47
Dispatcher (%10) **(PE)**	8241.0	25111.0	3.66	8.28	39.18	2.88 Million	573257.0	2.46 Million
Δ% **(PE)**	6.22	0.0	-8.72	8.65	-2.4	7.87	2.92	-3.12
Ownership	8104.0	25814.0	3.83	7.84	39.72	2.75 Million	575678.0	2.62 Million
Δ%	4.45	2.8	-4.35	2.94	-1.05	3.09	3.36	3.26
Ownership **(PE)**	7969.0	25111.0	3.8	8.04	39.32	2.79 Million	560287.0	2.59 Million
Δ% **(PE)**	2.71	0.0	-5.15	5.55	-2.03	4.74	0.59	2.05

Note: The changes are a mean over all 24 hours of the day. The wait-time and search time averages over hours are weighted by the number of trips and the hourly driver profits are weighted by the number of active drivers across hours. **PE** means partial equilibrium and holds demand fixed to give a sense of how much the demand expansion changes counterfactual results. The percentage changes Δ% are the changes of the means over all hours compared to the baseline. Consumer surplus is computed under the assumption that the demand function is truncated above the maximal waiting time observed in the data. The reason is that for our parameter specifications consumer surplus would be infinite is we integrated over all waiting times. This issue results from the assumption of constant elasticity, log-linear demand. A similar issue arises, for example, in Wolak (1994), who also truncates the demand distribution. Note, however, that except for the limit case, the absolute difference in consumer surplus will be well defined and the same, no matter how high we choose the truncation point.

6.1 Relaxing Entry Restrictions

To explore how additional entry affects the market, we increase the number of medallions by 10%, from 13,500 to 14,850. To put this policy change in perspective, Uber served 4% of the total number of trips in 2014 and 13% at the beginning of 2015.[37] Figure 9 compares the counterfactual outcomes to the baseline

Figure 9: Entry Counterfactual Results as Compared to Baseline



for each hour for four key quantities, while Table 3 gives a more detailed account aggregated at the daily level. The figures show that the number of cabs on the street expands throughout the day, with larger changes during the day time, and only moderate ones during the night. On average, taxi activity expands by 8.36%, less then proportionally to the medallion increase, because earnings drop, causing drivers to work less. This increase in available taxis reduces wait time for passengers and, as a result, demand expands moderately. On average, the expansion of demand is 7.28%, while the average reduction in wait-time for passengers over all hours is 7.5%.[38] The demand expansion moderates the drop in taxi earnings caused by increased supply. This is a sizable effect: if demand were held fixed, supply would only expand by 5.12%. The demand expansion almost completely compensates drivers for the additional competition: taking the

[37] See http://fivethirtyeight.com/features/uber-is-taking-millions-of-manhattan-rides-away-from-taxis/. This is based on data that the city obtained from Uber for a traffic study.

[38] All averages across hours are weighted in proportion to the number of trips taken in each hour.

mean over all changes in hourly income, the hourly wage of drivers would be decreased by only 1.6%. This small change is partly explained by demand expansion: wages would fall by 3.5% if demand were held constant. The increase in the number of matches (or total number of trips) is 5.84%, not far from the increase in demand, but lower because additional entry increases frictions overall. The present discounted revenue stream from a medallion decreases by 1.7%, much less than if demand did not increase (in which case it would be 3.8%). Consumer surplus would be higher with no demand response.[39] Note first that wait time drops a lot more when demand does not adjust than when it does. Second, the additional gain in the number of serviced trips when demand adjusts is relatively small, partly because of the modest demand elasticity. Thus, the gain in the inframarginal trips that are already served in the scenario with no demand adjustment.

Our model does not account for the traffic externality and adverse environmental effects due to additional taxis on the street. While environmental damage is hard to assess, the traffic externality has recently been investigated by the city. Their finding was that ride-hailing services are only a minor contributing factor to the recent decline in NYC traffic speed.[40]

Table 4: Entry Counterfactual

	Baseline	**Entry**
Total Consumer Surplus (per day)	2.67 Million Minutes	2.85 Million Minutes
Driver Revenue (hourly income)	\$40.14	\$38.76
Medallion Revenue (present value)	\$2.54 Million	\$2.49 Million

6.2 Improved Matching

6.2.1 Universal dispatcher

We now consider a counterfactual in which a dispatcher is available to match each empty (searching) cab with a waiting passenger. The dispatcher matches a passenger with the closest empty cab, if one is available within a one-mile radius.[41] This matching process is a natual alternative to the street hailing system and approximates the one used, for instance, by Uber. The reason we restrict matches to a one-mile radius is the following. Ideally, a matching algorithm would not only search across empty taxis but would also take into account the possibility that a soon to be empty taxi may be closest to a passenger. Such an algorithm would

[39] This is also the case for other counterfactuals. The reason is the same as the one outlined here.

[40] See http://www1.nyc.gov/assets/operations/downloads/pdf/For-Hire-Vehicle-Transportation-Study.pdf

[41] Passegers who do not find an immediate match wait for other opportunities to match. Taxis that are unmatched drive randomly until matched.

be difficult to compute so we focus on a simpler case. But, since our dispatcher does not optimize in a forward looking way, it will in general not be efficient to allow for matches of passengers and cabs that are too far apart. The restriction to one-mile radius alleviates this problem.[42]

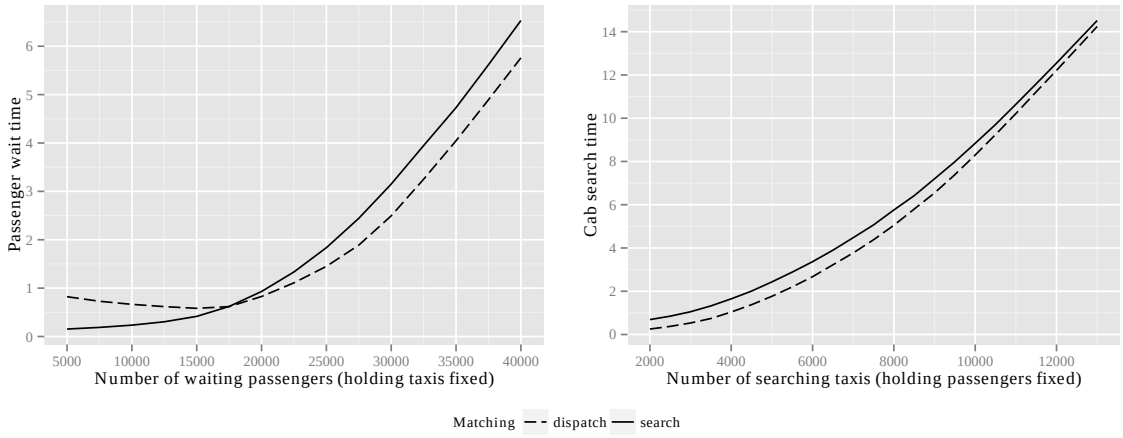
Table 5: Dispatch Counterfactual

	Baseline	**Dispatch**
Total Consumer Surplus (per day)	2.67 Million Minutes	2.97 Million Minutes
Driver Revenue (hourly income)	\$40.14	\$42.34
Medallion Revenue (present value)	\$2.54 Million	\$2.67 Million

Before presenting the counterfactual results, it is useful to simulate the dispatch matching function over a range of outcomes as in Figure 10 and to contrast it with the decentralized search process that we have assumed so far. The figure on the left shows that wait time for passengers increases as the market tightens. The figure on the right repeats the exercise for taxis, holding passengers fixed at the median number observed in the data. We can see that the dispatch system leads to lower search times for cabs of about half a minute on average. For passengers, wait times can become longer (relative to search) when the ratio of cabs to passengers is high. The reason for this potentially surprising outcome is related to the argument detailed above: passengers may be better off waiting for a random cab that is currently delivering a passenger and therefore not available for dispatch. However, this only takes place for numbers that are low relative to daytime traffic in NYC. We now describe results under the different dispatch scenarios, starting with the extreme case of an entire market operating under the dispatcher. Figure 11 gives an overview of the changes across different hours. The panel for each respective variable shows the difference between what happens in the baseline case minus what happens in the counterfactual. We see that both demand and supply expand at all hours of the day, with wait times and search times going down, particularly at night, demonstrating the higher returns to a dispatcher under a more sparsely populated map. In this counterfactual there is an increase in the number of active taxis (7.15%), and a substantial reduction in the search time for taxis (-13%). (If we didn't allow for an expansion of demand in response to reduced wait times, the supply increase would only be 2.17%.) Passenger wait time is reduced by 13.17%, consumer surplus increases

[42]The following extreme-case illustrates why a pure spatial global search may not be optimal. Consider a scenario in which there is only one passenger left waiting, only one empty cab searching, and they are on opposite ends of the grid. The remaining cabs are delivering passengers. If the dispatcher were to commit them to a match, there is a high chance that a better outcome could be obtained by waiting. The passenger is likely to obtain a faster match by waiting for one of the busy cabs to finish its trip and become available. Analogously, for the empty cab a new passenger may appear on the map closer to its position. A dynamic algorithm that searches spatially as well as across time could account for these better match opportunities.

Figure 10: Comparing Matching Functions



by 11.14%, and the number of trips increases more than proportionally to demand by 12.95% (because frictions were also reduced). The value of medallions would increase by 5.4%. In this counterfactual scenario, all stakeholders benefit from the introduction of the dispatch technology.

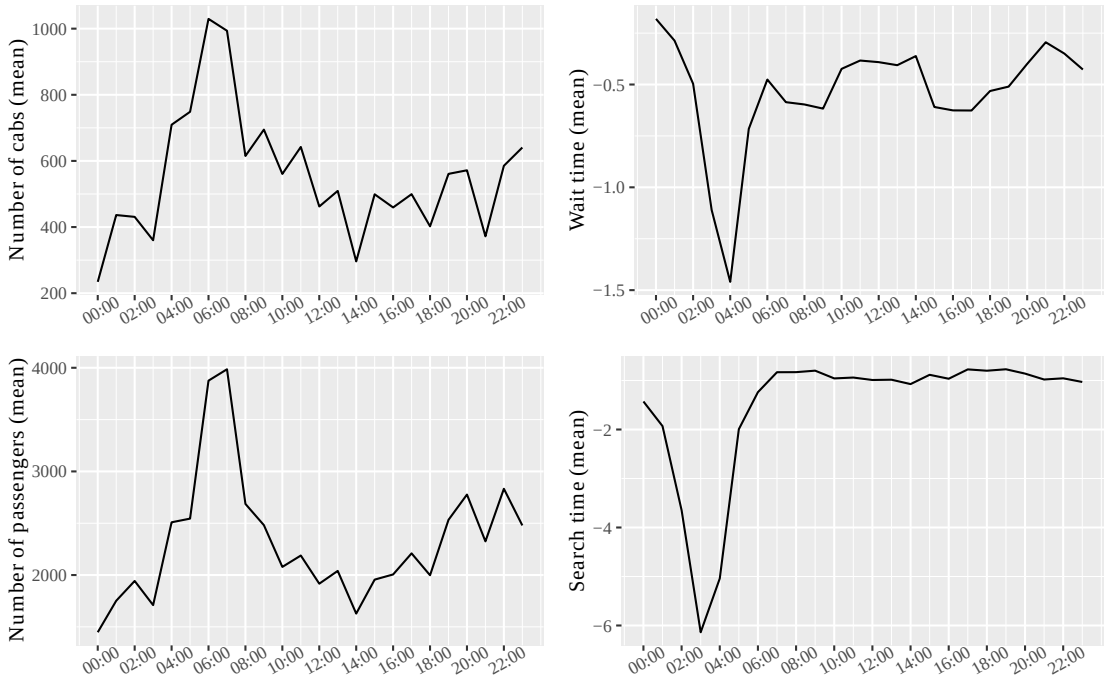
6.3 Partial Dispatching and Liquidity Externality

A potentially important consequence of the introduction of a more efficient matching technology is the resulting segmentation of the market. If consumers are divided between competing platforms, both segments of the market become thinner. The resulting reduction in market thickness could potentially lead to larger losses than the improvements due to better matching within the dispatch platform. To explore whether this is a plausible outcome, we first look at a scenario where we keep the total number of medallions fixed, and half of the medallions operate on the search platform (baseline), half on the dispatch platform. We also divide up demand across platforms, by simply pre-multiplying the constant elasticity demand function $d(w_t)$ by the shares of 0.5. This ensures that, if the wait-time in the two platforms were the same, total demand would add up the baseline total demand.[43] We do not allow passengers to choose among platforms.[44] In particular, we assume that both the driver and the passenger are committed to this match: neither can cancel should another match option become avail-

[43] For the results we weight wait-times, driver earnings, etc. by the respective hourly number of trips taken on both market sides and report those averages. The exception are medallion values, where we directly multiply the results by the respective share of issued medallions, in this case 50% of each type.

[44] In our current setup one platform would generically dominate, so we would observe a tipping phenomenon if we allowed for a choice of platforms. A richer model would allow for heterogeneity among passengers, as well as additional heterogeneity among platforms.

Figure 11: Counterfactual Results as Compared to Baseline, Full Dispatch



able sooner.[45] Figure 15 in Appendix A shows that the dispatch platform always serves more than 50% of the rides and that this share is highest during the night, where it can account for more than 60% of all rides.[46] This is because the dispatch platform has lower wait times, and therefore, higher demand. Figure 16 in Appendix A, however, illustrates that, in the aggregate, compared to the baseline, the reduction in market thickness has negative consequences that are larger than the better matching in the segment of market served by the dispatcher. Regardless of the hour, there are fewer active cabs and lower demand. Wait time and search time are marginally larger during the day, while search time decreases for some night-time hours. This shows that when the dispatch platform accounts for significant market share, all stake-holder may be harmed, despite the technological improvement. Consumer surplus decreases by about 110,000 Minutes per day, hourly driver income decreases by about 0.59%, and medallion revenue decreases by $40,000.

[45] For example it might happen that a waiting passenger, who is promised to a cab, encounters a cab that was not previously available before the promised cab arrives.

[46] This advantage of the dispatch platforms in hours with lower demand density is related to the findings in Cramer and Krueger (2016). They show that Uber's advantage in capacity utilization than taxis (defined as the fraction of time delivering passengers) is relatively minor in NYC but large in other cities which are less dense.

Table 6: Segmented Market 50%

	Baseline	**Dispatch**
Total Consumer Surplus (per day)	2.67 Million Minutes	2.56 Million Minutes
Driver Revenue (hourly income)	$40.14	$39.9
Medallion Revenue (present value)	$2.54 Million	$2.50 Million

In NYC, no ride hailing service has achieved a market share close to 50%, and the entry of ride hailing serviced has been associated with by-passing the medallion system, and therefore increased entry. To understand how important the trade-off between matching efficiency and market thickness is for a more realistic market share we now combine the entry counterfactual with the dispatch counterfactual: we add 10% new medallions that operate under the dispatch system. As in the previous case we split demand by multiplying the estimated demand function by 0.1 for the dispatch market and 0.9 for the remaining market.

During the day, fewer than 10% of the trips are served by the dispatch platform. The reason for this is that the thinness of the dispatch market overwhelms the better matching technology. However, as the number of passengers decreases during the night, the relative advantage of the dispatch solution increases and, starting from 11PM, the dispatch platform starts serving more than 10% of the market, reaching more than 35% market share at 4AM. Overall, in this scenario we observe an increase in daily consumer surplus, of the order of 110,000 minutes. For drivers the increase in competition due to the additional entry outweighs the matching efficiencies and leads to a slight decrease in earnings of about 0.88%. The decreased incentives for drivers to rent medallions is reflected in the loss of medallion revenues, of about $40,000, or 1.5%.

Table 7: Segmented Market, 10 %

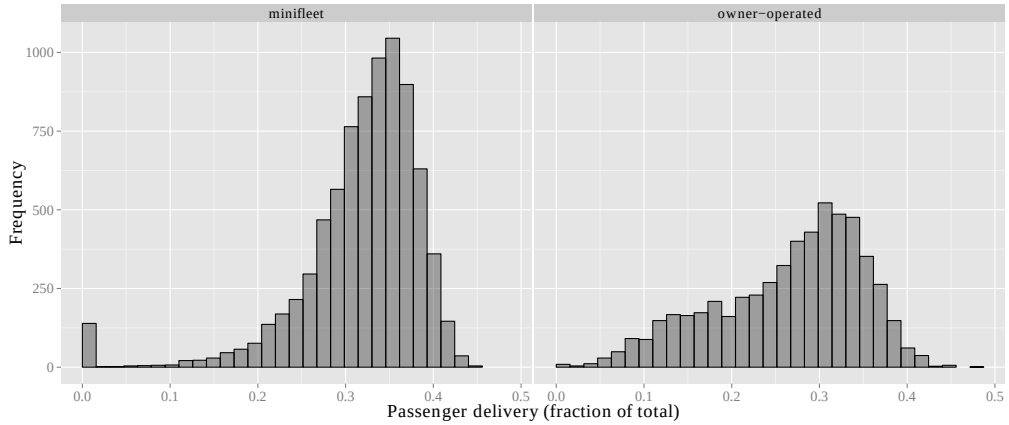
	Baseline	**+10% Medallions**	**+10% dispatched Medallions**
Total Consumer Surplus (per day)	2.67 Million	2.85 Million	2.78 Million
Driver Revenue (hourly income)	$40.14	39.5	$39.78
Medallion Revenue (present value)	$2.54 Million	$2.49 Million	$2.50 Million

6.4 Removing Ownership Restrictions

The final counterfactual is motivated by the observation that, as discussed above (see Section 3), fleet-owned medallions are more heavily utilized and have more coordinated shift changes. We now discuss additional features of the differences between fleet and owner-operated medallions.

A natural question is whether owner-operated medallions are managed as efficiently as minifleet medallions, whose owners specialize in managing other drivers and may benefit from scale economies of managing multiple medallions.

Figure 12: Histogram of Utilization Separated by Medallions



Notes: Each observation in these histogram is a medallion-average of the fraction of time that this medallion spends delivering a passenger out of the total time we observe these medallions. Note that the rest of the time the medallion could either be searching for a passenger or be idle and not on a shift at all. The histograms shows stark differences between owner-operated and minifleet medallions. The lower tail of low utilization is much thicker for owner-operated medallions.

Figure 12 shows the cross-sectional distribution of the fractions of time a medallion spends delivering a passenger out of the total time that we observe a medallion. The distribution for owner-operated cabs displays a much thicker left tail of low utilization rates and is overall more dispersed.[47]

The left panel of Figure 14 in Appendix A shows the length of time a medallion is *inactive* conditional on the stopping time of the last shift. Since most day shifts start around 5AM and most night shifts around 5PM, the time of non-utilization is minimized for stops that happen right around these hours, while a stop at any other time causes the medallion to be *stranded* for a longer time period. We see that minifleets typically return a medallion to activity faster after each drop-off. This difference is particularly large after the common night shift starting times (6PM and later), which suggests that minifleets have access to a larger pool of potential drivers, and this in turn makes it easier for them to find a replacement for someone who does not show up at the normal transition time. In the structural model we allow for a different set of parameters for minifleets and owner-operated to capture these differences. The right panel of Figure 14 in Appendix A shows the number of shift that end conditional on the hour. We see that minifleet medallions have a more regular pattern with most day-shifts ending at 4PM. This is also reflected in Figure 1 which shows a stronger supply decrease for minifleets before the evening shift relative to owner-operated medallions.

[47]The observed differences might be due to the fact that minifleets enables a more efficient utilization; but another plausible argument is a selection-effect.

We now simulate what would happen if all medallions were operated by minifleets (fleet counterfactual). An actual implementation of this policy would be to allow fleets to purchase owner-operated medallions, which currently sell at a substantial discount. The counterfactual is computed by changing the primitives of the model for the fraction of owner-operated medallions to those that we estimated for fleets (as shown in Table 2).

Taxi activity in this counterfactual increases by 4.45%, which compares to 2.7% if we did not allow demand to adjust. Demand in this counterfactual would increase by 2.8%, wait time goes down by 4.35%. Search time for cabs increases by 3%. The hourly income for drivers is reduced by 0.18%, which means that drivers are again nearly fully compensated by the demand adjustment compared to the case where demand would not go up, which imply a reduction in wages of about 1.5%. Taking everything into account consumer surplus rises by 2.81%. Medallion revenues go down by 1% and the number of trips rises by 3.4%.

7 Conclusion

This paper develops and estimates a dynamic general equilibrium model of the NYC Taxi market, which we use to understand the magnitude of the effects of entry restrictions and matching frictions. Drivers hourly revenue is determined by the equilibrium number of searching cabs and waiting passengers mediated by the time it takes to find the next passenger. Passengers' demand is affected by the waiting time for a cab. To estimate the model we back out unobserved demand by making use of the geographical nature of the matching process.

Counterfactual results from the model show that an improvement in the matching technology leads to substantial increases in consumers welfare as well as drivers' earnings. However, our results also point to the fact that competition among dispatch platforms can lead to decreases in welfare because it leads to market segmentation and lower market thickness. Our analysis of segmented platforms is only suggestive as the model does not allow for any additional heterogeneity among the platforms and assumes exogenous assignments of passengers to platforms. Including such richness is not within the scope of the current paper but it would be interesting the extend the analysis to study this issue in more depth.[48] We have found that more efficient utilization of existing medallions due to the elmination of favored treatment for owner-operators can lead to comparable gains as a policy that allows for a substantial number of additional entrants. This points to the potential importance of regulations favoring small firms. Such regulations exists in many industry and countries. Few studies have explored and quantified the potential impact of such restrictions.

[48] Cantillon and Yin (2008) study a related question in their analysis of competition among financial exchanges that trade the same securities.

We have not considered the issue of surge pricing. In order to study this question, one would need to estimate a richer demand system that allows for dependence on both prices and wait time, while recognizing that the likely correlation between consumers responsiveness to wait time and to prices. Uber data may be helpful for studying such a question.

References

Bajari, Patrick, C Lanier Benkard, and Jonathan Levin, "Estimating dynamic models of imperfect competition," *Econometrica*, 2007, *75* (5), 1331–1370.

Berry, Steven T, "Estimation of a Model of Entry in the Airline Industry," *Econometrica: Journal of the Econometric Society*, 1992, pp. 889–917.

Bresnahan, Timothy F and Peter C Reiss, "Entry and competition in concentrated markets," *Journal of Political Economy*, 1991, pp. 977–1009.

Buchholz, Nicholas, "Spatial Equilibrium, Search Frictions and Efficient Regulation in the Taxi Industry," 2015.

Camerer, Colin, Linda Babcock, George Loewenstein, and Richard Thaler, "Labor supply of New York City cabdrivers: One day at a time," *The Quarterly Journal of Economics*, 1997, pp. 407–441.

Cantillon, Estelle and Pai-Ling Yin, "Competition between Exchanges: Lessons from the Battle of the Bund," 2008.

Collard-Wexler, Allan, "Demand Fluctuations in the Ready-Mix Concrete Industry," *Econometrica*, 2013, *81* (3), 1003–1037.

Conlon, Christopher T, "A Dynamic Model of Costs and Margins in the LCD TV Industry," *Unpublished manuscript, Yale Univ*, 2010.

Cramer, Judd and Alan B Krueger, "Disruptive change in the taxi business: The case of Uber," *The American Economic Review*, 2016, *106* (5), 177–182.

Crawford, Vincent P and Juanjuan Meng, "New york city cab drivers' labor supply revisited: Reference-dependent preferences with rationalexpectations targets for hours and income," *The American Economic Review*, 2011, *101* (5), 1912–1932.

Dubé, Jean-Pierre, Jeremy T Fox, and Che-Lin Su, "Improving the numerical performance of static and dynamic aggregate discrete choice random coefficients demand estimation," *Econometrica*, 2012, *80* (5), 2231–2267.

Farber, Henry S, "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 Lessons from Cab Drivers," Technical Report, National Bureau of Economic Research 2014.

Haggag, Kareem and Giovanni Paci, "Default tips," *American Economic Journal: Applied Economics*, 2014, *6* (3), 1–19.

_ , Brian McManus, and Giovanni Paci, "Learning by Driving: Productivity Improvements by New York City Taxi Drivers," 2014.

Holmes, Thomas J, "The Diffusion of Wal-Mart and Economies of Density," *Econometrica*, 2011, *79* (1), 253–302.

Jackson, C Kirabo and Henry S Schneider, "Do social connections reduce moral hazard? Evidence from the New York City taxi industry," *American Economic Journal: Applied Economics*, 2011, *3* (3), 244–267.

Jia, Panle, "What Happens When Wal-Mart Comes to Town: An Empirical Analysis of the Discount Retailing Industry," *Econometrica*, 2008, *76* (6), 1263–1316.

Kalouptsidi, Myrto, "Time to build and fluctuations in bulk shipping," *The American Economic Review*, 2014, *104* (2), 564–608.

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

Lubin, Miles and Iain Dunning, "Computing in operations research using Julia," *arXiv preprint arXiv:1312.1431*, 2013.

Oettinger, Gerald S, "An Empirical Analysis of the daily Labor supply of Stadium Venors," *Journal of political Economy*, 1999, *107* (2), 360–392.

Rust, John, "Optimal replacement of GMC bus engines: An empirical model of Harold Zurcher," *Econometrica: Journal of the Econometric Society*, 1987, pp. 999–1033.

Ryan, Stephen P, "The costs of environmental regulation in a concentrated industry," *Econometrica*, 2012, *80* (3), 1019–1061.

Su, Che-Lin and Kenneth L Judd, "Constrained optimization approaches to estimation of structural models," *Econometrica*, 2012, *80* (5), 2213–2230.

TLC, "2011 Anual Report of the TLC," 2011.

Wolak, Frank A, "An econometric analysis of the asymmetric information, regulator-utility interaction," *Annales d'Economie et de Statistique*, 1994, pp. 13–69.

A Additional Figures

Figure 13: Cost functions at different times of day

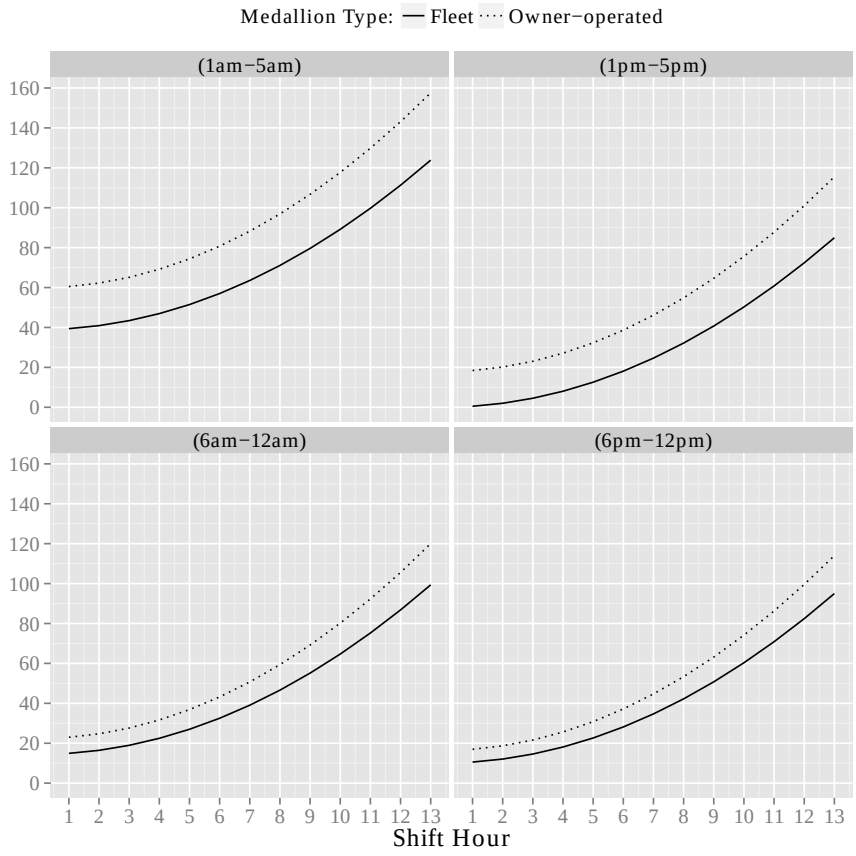


Figure 14: Time medallion is unutilized conditional on hour of drop off.

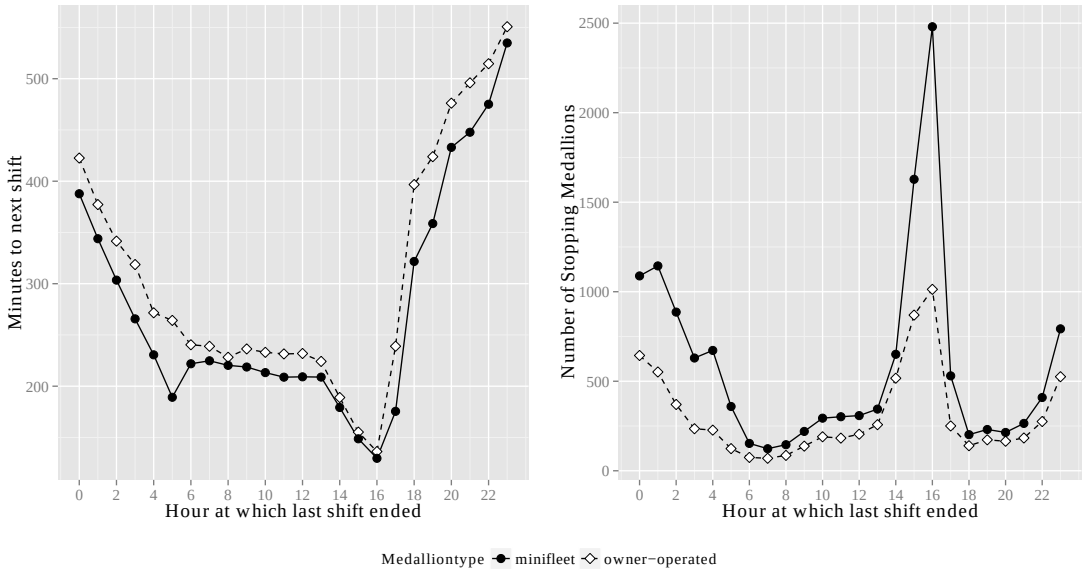


Figure 15: Fraction of Trips Served by Dispatcher (50% Dispatch)

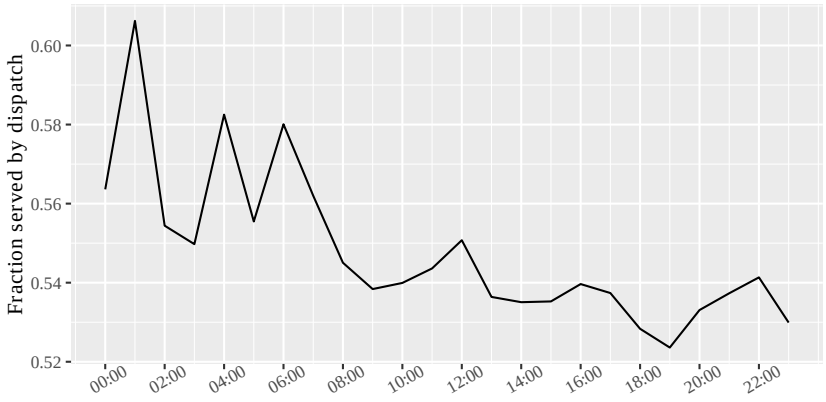


Figure 16: Counterfactual Results as Compared to Baseline, 50% Dispatch

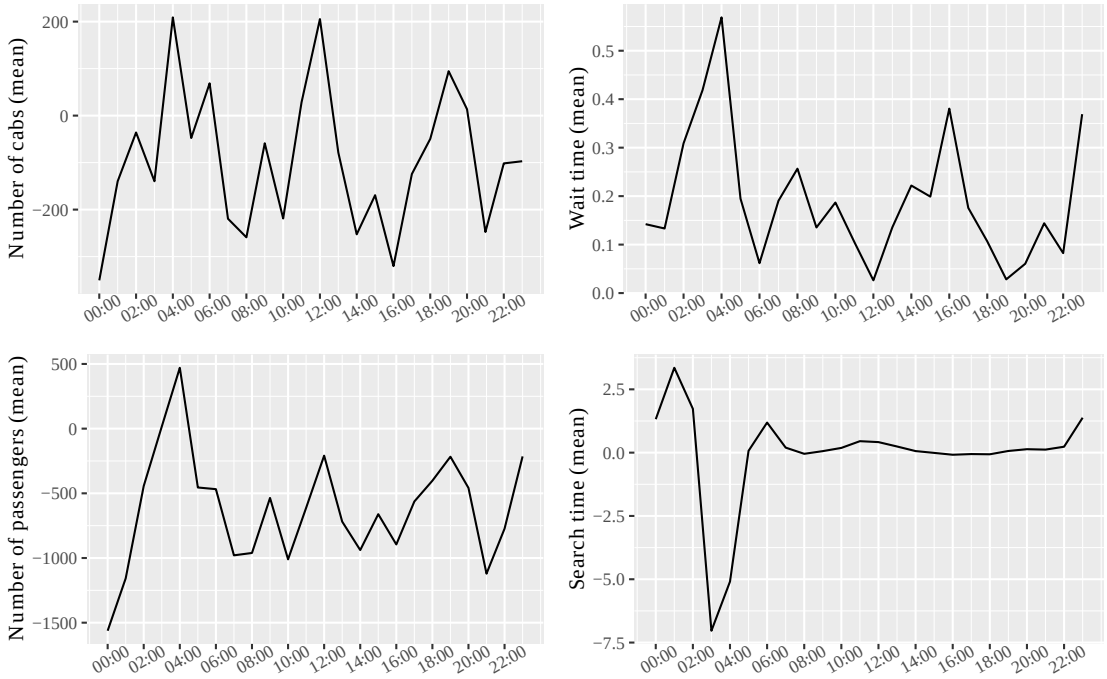
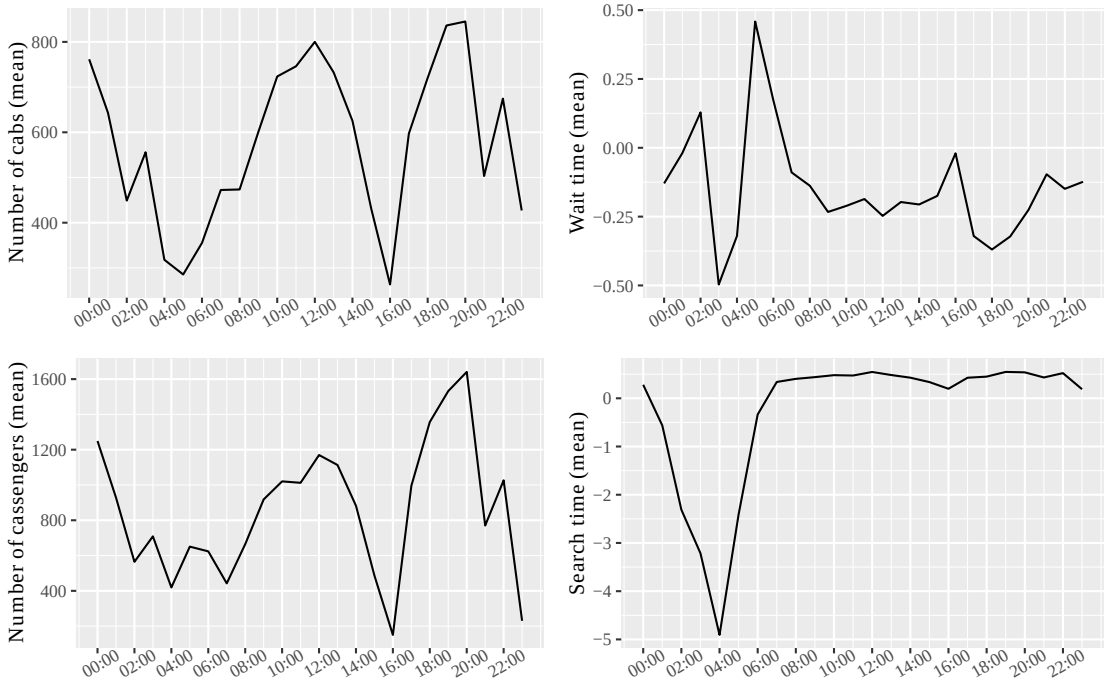


Figure 17: Counterfactual Results as Compared to Baseline (10% additional medallions, Dispatch)



B Details on Simulation

The goal of the simulation is to obtain a mapping of the number waiting passengers and searching cabs within an hour to the waiting-time and search-time of those passengers and cabs. The mapping is used to infer the number of waiting passengers from observed number of active cabs and their search time. Waiting and search-time are also influenced by other exogenous factors, which therefore need to be arguments of the matching function. These factors are the speed mph_t at which the traffic flows, the average trip length $miles_t$ requested by passengers. Table 8 provides an overview both over the taxi search time observed in the data as well as the observed inputs to the matching function.

Table 8: Summary Statistics for Simulation Variables.

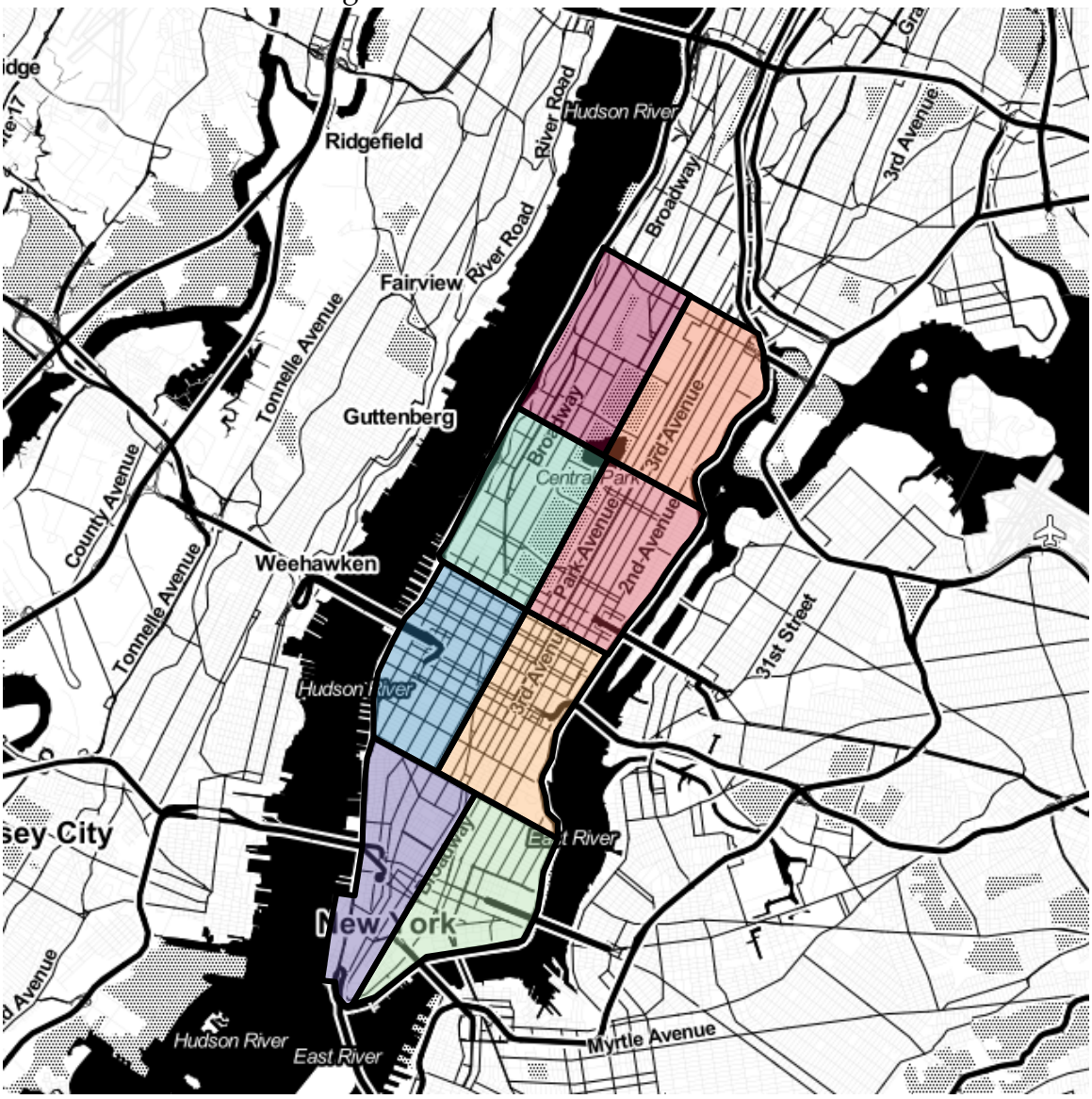
Variable	Mean	SD	Median	Min	Max
Miles per Hour	13.6	3.6	12.4	8.4	22.8
Average Trip Length (Miles)	3.0	0.5	2.8	2.3	4.9
Number of Cabs	7789.3	3084.2	9109.5	1262	11448
Average Wait time for Cabs	13.3	7.0	10.9	5.1	32.0

Note: Based on the available 2012 trip sheet data excluding the days Friday to Sunday. Statistics are reported after winsorizing variables at the 0.01 and the 0.99 percentile to account for some nonsensical outliers.

The baseline simulation is performed under the assumption that cabs search randomly for passengers. The search is performed on an idealized map of Manhattan. Figure 4 provides a schematic of the grid that we use for the simulation. In line with the topography of Manhattan we require the area to be four times as long in north-south direction (y_t) than wide in east west-direction (x_t). Cabs are moving on nodes that are $1/20$ mile segments apart from each other, which is based on the average block length in north-south direction. In the north-south direction they can turn at each node whereas in the east-west direction they can only turn at every fourth node. Figure 4 highlights nodes on which cabs can turn as gray. This corresponds to the block structure of Manhattan where a block is approximately $1/20$ miles long in north-south and $4/20$ miles wide in east-west direction. Under the random search assumption cabs take random turns at nodes with equal probability weight on each permissible direction. However, we assume that they never turn back to the direction from which they were coming (i.e. no U-turns).

Since we only model the Manhattan market (below 128th street), our grid corresponds to an area of 16 square miles. Figure 18 shows the modeled part on the map in its division in the eight equally sized different areas for which we separately compute the pick up and drop-off probabilities. Correspondingly, our grid is divided into eight equal parts, which correspond to those areas.

Figure 18: Division of Manhattan



Each node on the grid is a possible passenger location. For each hourly simulation $\frac{d_t}{6}$ passengers are placed in ten minute intervals randomly on the map. Those $\frac{d_t}{6}$ are divided up and placed in proportion to the corresponding (observed) pickup probabilities on the eight areas on the grid. Within those areas passengers appear with equal probability on each node.

If a cab hits a node with a passenger a match occurs. There are no additional frictions on a node (which corresponds to a street corner) and the number of matches is the minimum of the number of passengers and the numbers of taxis on the node, which corresponds to the assumption of a Leontieff matching function on each node. Once the match takes place the cab is taken of the grid for $60 \cdot \frac{miles_t}{mph_t}$ minutes, i.e. the average measured delivery time from the data, after which is has delivered the passenger and is again placed randomly on the map with a random travel direction. Cabs reappear in locations on the grid in proportion to observed drop-off locations of the eight areas (Figure 18).

The full algorithm is described in pseudo-code below. It takes the following inputs: the number of cabs c, the number of passengers d, the trip length $miles$, and the trip speed mph. A unit of time in the algorithm is scaled so that it always represents the time it takes a cab to travel from one node to the next since there is no need for a smaller time unit. Passengers are added to the map for one hour in ten minute intervals. Since nodes are spaced $1/20$ miles apart, the last time passengers are added to the map is at $\bar{t} = 20 \cdot \text{mph}$. The set of times at which new passengers arrive is given by: $\{\bar{t}/6 \cdot k | k = 1, ..., 6\}$. The following additional variables are used to describe the algorithm: $npick$ refers to the number of matches that have already taken place, $deliverytime_i$ to the remaining delivery time of taxi i, and $searchtime_i$ to the time that taxi i has spent searching since the last delivery, $total\_searchtime$ refers to the total time that taxis have spent searching for passengers, and $total\_waittime$ to the total wait time that passengers have been waiting.

while $npick < d$ **do**
  $t = t + 1$ (time units represent travel time from one node to the next, scales with mph);
  if $t \in \{\bar{t}/6 \cdot k | k = 1, ..., 6\}$ **then**
    add $d/6$ passengers to random nodes on map, stratified by eight areas;
  end
  for $i = 1 : c$ **do**
    if $deliverytime_i = 0$ *(cab i is not occupied)* **then**
      update the node of cab i. Cabs only take turns on gray nodes (Figure 4) and do not make u-turns. All feasible travel directions are chosen with equal probability;
      if *new node of cab i has a passenger* **then**
        cab becomes occupied, set $deliverytime_i$ to $20 \cdot miles$, and add $searchtime_i$ to *total_searchtime*;
      else
        $searchtime_i = searchtime_i + 1$
      end
    else
      $deliverytime_i$ = $deliverytime_i$ - 1;
      if $deliverytime_i$ == 0 **then**
        place cab in random area on map according to observed drop-off probabilities (all nodes within area equal probability), give cab random feasible travel direction;
      end
    end
  end
  Add one to *total_waittime* for each passenger that is on the map;
end
Result: Use *total_waittime* and *total_searchtime* to compute the average search time for taxis and average wait time for passengers.

In the dispatcher simulation we assume that each cab - as soon as the previous passenger has been delivered - is matched with the closest passenger available. We also assume that neither the driver nor the passenger has an option to cancel this match for another match option. It might for example happen that a waiting passenger, who is promised to a cab, encounters a cab that was not previously available before the promised cab arrives. The option to cancel might in some instances be beneficial to a market-side because our search for the optimal match

is only over the currently available cabs and passengers and does not take into account cabs and passengers that will soon appear somewhere close on the map.

It is not feasible to perform the simulations for each point in the domain of the matching function. We therefore perform them for the Cartesian product of the sets: $c \in \{500, 1000, ..., 17000\}$, $d \in \{3000, 6000, ..., 75000\}$, $miles \in \{1, 2, ..., 7\}$, $mph \in \{4, 8, ..., 24\}$. To obtain the search-time and wait-time for other points we interpolate linearly between the grid points.

B.1 Details on Estimation.

As defined in the text, let π_t be a realization of the earnings and $\boldsymbol{x}_{it}$ denote the other part of the observable state (the shift length, the hour of the day as well as the medallion invariant characteristics). Let $p(\boldsymbol{x}_{it}, \pi_t)$ be the theoretical probability that an active medallion/driver i stops at time point t and $q(\boldsymbol{x}_{it})$ be the probability that a inactive medallion/driver i starts at t. Correspondingly, let d^A be the indicator that is equal to one if an active driver stops and d^I be an indicator that an inactive driver starts. Using this notation we maximize a constrained log-likelihood that we formulate as an MPEC problem. MPEC does not perform any intermediate computations, such as value function iterations, to compute the objective function. It instead treats these objects as parameters. This means that the solver will be maximizing both over the parameters of interest θ and an additional set of parameters δ. The parameter vector δ consists of all $p(\mathbf{x}_{it}, \pi_t)$, $q(\mathbf{x}_{it})$, $\mathbb{E}_{\pi,\epsilon}[V(\mathbf{x}_{i(t+1)}, \pi_{t+1}, \epsilon_{i(t+1)})|h_t]$ for $\mathbf{x}_{i(t+1)} \in \mathbf{X}$, $\pi_t \in supp(F_\pi(.|h_t))$. In other words, δ consists of expected values and choice probabilities for each point in the observable state space. Note, however that π_t follows a continuous distribution and it is therefore not possible to specify a constraints for each value in the support of its distribution. We instead approximate the distribution of π_t with a discrete number of nodes $\tilde{\pi}$ and weights using gauss-hermite integration.[49]

With this notation in place we can express the maximization problem as follows

$$\begin{aligned}\min_{\theta, p(\mathbf{x}_{it},\pi_t), q(\mathbf{x}_{it}), \mathbb{E}_{\pi,\epsilon}[V(\mathbf{x}_{i(t+1)},\pi_{t+1},\epsilon_{i(t+1)})|h_t]} & \sum_{j\in J}\sum_{t\in T_j} d^A_{it}\cdot log(p(\mathbf{x}_{it},\pi_t)) + (1-d^A_{it}) \\ & \cdot(log(1-p(\mathbf{x}_{it},\pi_t))) + d^I_{it}\cdot log(q(\mathbf{x}_{it})) + (1-d^I_{it})\cdot(log(1-q(\mathbf{x}_{it})))\end{aligned} \tag{6}$$

subject to:

[49] We use six nodes.

$$\mathbb{E}_{\pi,\epsilon}[V(\mathbf{x}_{i(t+1)}, \pi_{t+1}, \epsilon_{i(t+1)})|h_t] = \sigma_\epsilon \cdot log\left(exp\left(\frac{1}{\sigma_\epsilon}\right)\right.$$
$$\left. + exp\left(\frac{\pi_t - C_{z_i,h_t}(l_{it}) - f(h_t,k_i) + \mathbb{E}_{\pi_{t+1},\boldsymbol{\epsilon}_{i(t+1)}}[V(\boldsymbol{x}_{i(t+1)}, \pi_{t+1}, \boldsymbol{\epsilon}_{i(t+1)})|h_{t+1}]}{\sigma_\epsilon}\right)\right.$$
$$+ \gamma * \sigma_\epsilon \forall \boldsymbol{x}_{it} \in \boldsymbol{X}, \quad \forall \pi_t \in \tilde{\pi} \quad (7)$$

$$p(\boldsymbol{x}_{it}, \pi_t) = \frac{exp\left(\frac{1}{\sigma_v}\right)}{exp\left(\frac{1}{\sigma_v}\right) + exp\left(\frac{\pi_t - C_{z_i,h_t}(l_{it}) - f(h_t,k_i) + \mathbb{E}_{\pi_{t+1},\boldsymbol{\epsilon}_{i(t+1)}}[V(\boldsymbol{x}_{i(t+1)}, \pi_{t+1}, \boldsymbol{\epsilon}_{i(t+1)})|h_{t+1}]}{\sigma_v}\right)}$$
$$\forall \boldsymbol{x}_{it} \in \boldsymbol{X}, \quad \forall \pi_t \in \tilde{\pi} \quad (8)$$

$$q(\boldsymbol{x}_{it}) = \frac{exp\left(\frac{\mathbb{E}_{\pi_{t+1},\boldsymbol{\epsilon}_{i(t+1)}}[V(\boldsymbol{x}_{i(t+1)}, \pi_{t+1}, \boldsymbol{\epsilon}_{i(t+1)})|h_{t+1}] - r_{h_t}}{\sigma_v}\right)}{exp\left(\frac{\mathbb{E}_{\pi_{t+1},\boldsymbol{\epsilon}_{i(t+1)}}[V(\boldsymbol{x}_{i(t+1)}, \pi_{t+1}, \boldsymbol{\epsilon}_{i(t+1)})|h_{t+1}] - r_{h_t}}{\sigma_v}\right) + exp\left(\frac{\mu_{h_{t+1}}}{\sigma_v}\right)} \quad \forall \boldsymbol{x}_{it} \in \boldsymbol{X} \quad (9)$$

The constraint given by equation Equation 7 ensures that the starting and stopping probabilities obey the intertemporal optimality conditions imposed by the value functions. The log-formula is the closed form expression for the expectation of the maximum over the two choices of stopping and continuing, which integrates out the T1EV unobserved valuations. Equation 8 and Equation 9 are again the closed form expressions for the choice probabilities under extreme value assumption. We also restrict the search for the cost-functions to the domain of increasing functions by requiring $\lambda_{0,z_j,0}$, $\lambda_{1,z_j,0}$ and $\lambda_{2,z_j,0}$ to be larger than zero.

C Details on the Computation of Counterfactuals

Define the follwoing six steps as **Block1(i)** for iteration i.

1. *For each hour simulate values from* $\mathcal{N}(\alpha^{c,i}, \psi_h^{c,i})$ *and* $\mathcal{N}(\alpha^{d,i}, \psi_h^{d,i})$ *as well as the observed empirical distributions of speed of traffic flow and the length of re-*

quested trips to determine the distributions of search time for taxis $\mathcal{F}_s^i(.|h)$, $h \in \{0,\ldots,23\}$ *under* $g(.)$.

2. *Simulate drivers earnings* $\mathcal{F}_\pi^i(.|h)$, $h \in \{0,\ldots,23\}$ *from the ratios of passenger delivery time over delivery and search time (computed in step 2) and rate earned per minute of driving. Simulate new distribution of passengers* $\mathcal{F}_d^i(.|h)$, $h \in \{0,\ldots,23\}$ *from the distribution of waiting times and the estimated demand function* $d(w_t)$.

3. *Compute the optimal starting and stopping probabilities* $p^i(\mathbf{x},\pi;\theta)$, $q^i(\mathbf{x};\theta)$ *under the new distribution of earnings (computed in step 3). The distribution of earnings is approximated using gauss-hermite integration in the stopping problem of drivers.*

4. *Use* $p^i(\mathbf{x},\pi;\theta)$ *and* $q^i(\mathbf{x};\theta)$ *to simulate a new distributions* $\mathcal{F}_c^i(.|h)$, $h \in \{0,\ldots,23\}$. *For each medallion type (z,k) we simulate thirty medallions, where each medallion starts inactive at 12PM and iterate forward for 48 hours. Across these thirty medallions we then compute the fraction of times the medallion has been active in this hour (using only the last 24 hours) and multiply this by the total number of medallions. We repeat this 30 times and then compute the average and the standard deviations across these simulations.*[50]

5. *Compute* $\alpha_h^{c,i}$ *as the first and* $\psi_h^{c,i}$ *as the second moment from* $\mathcal{F}_c^i(.|h)$.

6. *Compute* $sumsq_1 = \sum_h(\alpha_h^{c,i} - \alpha_h^{c,(i-1)})^2 + \sum_h(\psi_h^{c,i} - \psi_h^{c,(i-1)})^2$.

Define the following four steps as **Block2(i)** for iteration i.

1. *For each hour simulate values from* $\mathcal{N}(\alpha^{c,i},\psi_h^{c,i})$ *and* $\mathcal{N}(\alpha^{d,i},\psi_h^{d,i})$ *as well as the observed empirical distributions of speed of traffic flow and the length of requested trips to determine the distributions of waiting time* $\mathcal{F}_w^{ij}(.|h)$, $h \in \{0,\ldots,23\}$ *for passengers.*

2. *Simulate new distribution of passengers* $\mathcal{F}_d^i(.|h)$, $h \in \{0,\ldots,23\}$ *from the distribution of waiting times and the estimated demand function* $d(w_t)$.

3. *Compute* $\alpha_h^{d,i}$ *as the first and* $\psi_h^{d,i}$ *as the second moment from* $\mathcal{F}_d^i(.|h)$.

4. *Compute* $sumsq_2 = \sum_h(\alpha_h^{d,i} - \alpha_h^{d,(i-1)})^2 + \sum_h(\psi_h^{d,i} - \psi_h^{d,(i-1)})^2$.

[50] We have also experimented with different numbers in this step, for example simulating each medallion for more than 48 hours or increase the number of simulations. For the final counterfactual results this does not seem to make a large difference.

Using those definitions, the algorithm can be described as follows:

```
while STOP ≠ 1 do
    Compute sumsq_1 using Block1(i).
    if sumsq_1 < tol then
        STOP=1
    else
        while sumsq_1 > tol do
            Block1(i)
        end
    end
    Compute sumsq_2 using Block2(i).
    if sumsq_2 < tol then
        STOP=1
    else
        while sumsq_2 > tol do
            Block2(i)
        end
    end
    i = i + 1
end
```

D Shift Transition Instrument

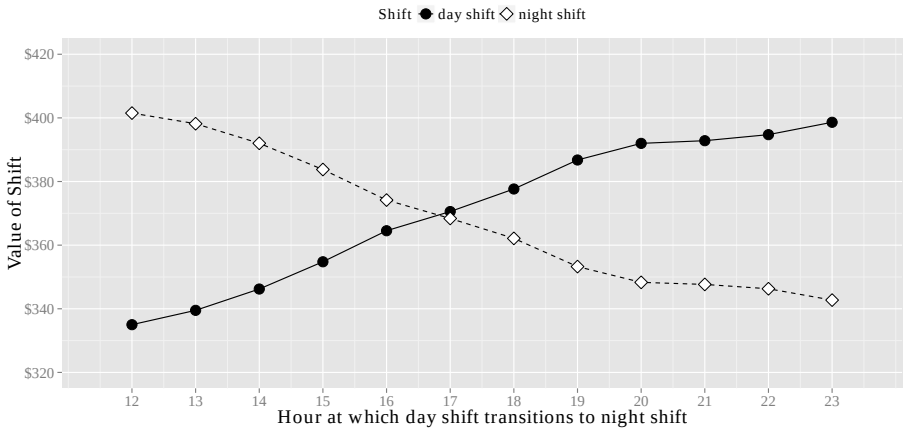
We now argue that the timing of the shift transition is such a supply side driven shifter. The kink in the number of active taxis in the later afternoon hours is clearly visible in Figure 1 and **??** shows that this is due to the transitioning of shifts.[51] New Yorker's refer to this as the *witching hour*. There may be multiple reasons that lead to most shifts being from 5AM to 5PM and 5PM to 5AM, but the data (and the rules) suggests that some factors are key.

First, the rules are such that minifleets can only lease for exactly two shifts per day: they must operate a medallion for at least two shifts of 9 hours and the lease must be on a per day or per shift basis.[52] Second, there is a cap on the lease price for both day and night shifts. Anecdotal evidence from the TLC and individuals in the industry suggests that these lease caps are binding. Given those rules, minifleets may try to equate the earning potentials for the day and night shifts, as a way to ensure they will get similar number of drivers willing to drive each shifts. A similar argument applies for owner drivers that might want to ensure they always find a driver for the second shift, which they do not drive themselves. Figure 19 shows the earnings for night and day-shifts under different hypothetical shift divisions. The x-axis shows each potential division-point, i.e. each point

[51] Shifts are defined following the definition used by Farber (2008) who determines them as a consecutive sequence of trips where breaks between two trips cannot be longer than five hours.

[52] See section 58-21(c) in TLC (2011).

Figure 19: Earnings of Day and Night Shift for Different Split Times



Notes: This graphs shows the average earnings that would accrue to the night-shift and day-shift driver for each possible division of the day. The x-axis shows the end-hour of the day shift and the start-hour of the night shift. Since these earnings are a function of the current equilibrium of the market, they have to be understood as the shift-earnings that one deviating medallion would give to day and night-time drivers. The graph shows that earnings are almost equal at 5PM, the prevailing division for most medallions.

at which a day shift could end and a night shift start. The y-axis reports the earnings for the day-shift (black dots) and night-shift (white diamonds).[53] As can be seen, the 5-5 division creates two shifts with similar earnings potential. Combined with the above observation, the difference in rate caps for day and night shifts may reflect different disutility from working at night. Hence, requiring two shifts and imposing a binding cap on the rates results in most medallions having shifts that start and end at the same time. Since transitions do not happen instantaneously, this correlated stopping therefore leads to a negative supply shock at a time of high demand during the evening rush hour. We use the interaction term between the traffic flow and shift transition times as a supply shifter. Since taxis are transitioned at predefined locations, variation in the traffic creates variation in the time needed to transition cabs and how long they "disappear"

[53]Clearly this comparison ignores any equilibrium effects of changing the sifts structure. The graph can therefore be understood as the earnings that one deviating medallions could have under the current system.