Frictions in a Competitive, Regulated Market
Evidence from Taxis
(Preliminary Draft)

Guillaume Frechette (NYU)  Alessandro Lizzeri (NYU)
Tobias Salz (NYU) *

August 19, 2015

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. Nicola Persico and Bernardo S. da Silveira provided very helpful conversations and feedback. We gratefully acknowledge financial support from the National Science Foundation.
This paper estimates a dynamic general equilibrium model of the New York City (NYC) taxi-cab market. It assesses the relevance of regulatory and matching frictions that are simultaneously present in this market and evaluates (via counter-factuals) the relative importance of these frictions. In the taxi market, capital and skill requirements are modest and natural entry barriers are low. Taxi services also 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 is just a unit of capital (a car) plus the drivers needed to operate it. Finally, in NYC, taxi drivers take many decisions independently, with little real-time information about aggregate conditions. Thus, absent regulatory intervention this market could serve as a textbook example of a “perfectly competitive” industry with many firms making decisions independently, and therefore an interesting case study of an important benchmark.

In most cities taxi markets are subject to stringent regulations on entry and fares. NYC is no exception. Under the current medallion system at most 13,520 yellow cabs can serve the market. 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 all operated by mini fleets, the largest of which operates hundreds of taxis. There are interesting organizational consequences of these differences in medallion ownership. The management aspect of operating a taxi includes timing shift-transitions efficiently and finding replacement drivers for shifts not operated by the owner. We find sizable differences in the utilization rates across medallion types, and faster shift transition for corporate medallions. There are also regulations on fares, which interact with the search process of taxis and passengers. Since rates are essentially fixed during the day, variation in the relative demand and supply schedule leads to large variation in waiting time for passengers and taxis. Even though fares are regulated, drivers’ earnings and the number of active taxis vary during the day depending on how long drivers need to spend searching for their next passengers. The average search time that an active taxi incurs between dropping off a passenger and picking up the next one ranges between 5 minutes at 5PM to about 20 minutes in the early morning. Absent price adjustments, waiting time serves as the market clearing variable. This form of rationing is common in other markets, such as health care.

In order to quantify the effects of some of these regulatory and matching frictions, we estimate a model in which drivers make both entry and stopping decision. Medallions are scarce so entry is only possible for inactive medallions. Hourly profits are determined by the number of matches between the number of searching taxis and the number of waiting passengers. Ceteris paribus, increasing the number of taxis increases the search time for drivers to encounter the next match and drives down expected hourly income. The number of taxis is determined endogenously as part of the competitive equilibrium in this market. Stopping decisions are determined by comparing hourly earnings with the combination of a marginal cost of driving that is increasing in the length of a shift and a random terminal outside option. Starting (entry) decisions are determined by a comparison between the expected value of a shift given optimal stopping behavior and the value of a random outside option.

To estimate the model we make use of rich data on the New York City 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 can also identify medallion owners as well as individual drivers which allows us to control for important sources of heterogeneity in drivers’ decisions.

1With an average daily volume of $ 6.4 Million, street hailing is an important piece of New York’s infrastructure.
On the demand-side, we face a challenge because neither the passengers wait-time nor the number of hailing passengers is observed in the data. However, we can take advantage of the geographical nature of the search by taxis to recover how many people must have been waiting for a cab given the number of pickups we observe, 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 way and use a known 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). In estimating the demand function we face the classical simultaneity problem which leads the price (i.e. waiting time) to be potentially correlated with unobserved shocks to demand. To overcome this endogeneity problem we make use of a feature of the market which is known as the “witching hour” to New Yorkers and refers to the disappearance of roughly a quarter of taxis during the evening rush hour. The “witching hour” arises due to a combination of regulatory constraints and the way drivers’ earnings accrue during the day: There are daily (and nightly) lease caps and usage rules that limit the ability and the incentive for medallion holders to make use of taxis that are returned unexpectedly early by drivers. Since medallion owners do not marginally participate in drivers earnings they have an incentive to simply maximize the probability that they can rent out the medallion for two shifts and therefore want to make both shifts equally attractive to drivers. The only way to insure this is to give the evening rush hour to night-shift drivers which leads to this seemingly coordinated transitioning of shifts at 5PM. Since transitions leave the cab unutilized for some time, there is a withdrawal of supply just when demand picks up. We interact the occurrence of the “witching hour” with traffic conditions to instrument the waiting time for a cab.

We present counterfactuals relaxing entry and ownership restrictions. These demonstrate the importance of modeling the demand side as well as the intensive margin on the supply side. Interestingly, the change in ownership leads to gains that are of similar magnitude to the change in entry, while being easier to implement politically because it harms operators a lot less. We also present a counterfactual that changes the matching technology between cabs and passengers. We move to the polar opposite of the NYC decentralized decision making model to study a centralized dispatcher who sends empty cabs to the closest passenger. We show large gains in efficiency and reductions in wait times for both passengers and taxis. We have also started to study a more realistic counterfactual, with partial coverage by a dispatcher. This allows us to study one of the effects of traditional taxi dispatchers as well as entry of new operators such as Uber. Our preliminary results suggest that there is a large network effect: unless penetration covers almost the full market, wait time for passengers and taxis are worse than in the baseline with no dispatcher. Partial coverage by a dispatcher has two effects: 1) improvement of matching in the covered market but 2) segmentation of the market with the consequence of longer average distances between a random taxi and a random passenger.

This project combines elements from the entry/exit literature, neo-classical labour 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 labour 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.

The spatial search problem of taxis for passengers plays an important role in determining drivers wages and passenger welfare through differences in waiting time. Spatial search processes have received some attention in the labor literature. Manning and Petrongolo (2011), for example, estimate a job search model for the UK labor market that allows for spill-over effects across wards. 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. Using the model, he quantifies the impact of policies increasing fares and the number of medallions.

Some earlier papers have used NYC 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 point. 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.

2Lagos (2003) does not have data on hourly or daily decisions by taxi drivers.

3Note also that our models fit aggregate patterns relatively well, hence any gains from allowing for a richer labor supply decision would be small in aggregate.

4Other papers are not directly relevant, for instance Haggag and Paci (2014) that study the impact of suggested tips in the New York city taxi driver payment screen for clients on the realized tip. Haggag et al. (2014) who study how taxi drivers learn driving strategies based on the experiences. Finally, Jackson and Schneider 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.
1 Industry Details and Data

1.1 Industry Details

Operating a yellow cab in NYC requires a medallion. In the time period covered by the data only yellow cabs are allowed to pick up street-hailing passengers. 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. 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 that medallion owners can charge drivers. The TLC periodically auctions off new medallions which may come with certain restrictions, such as the requirement to operate a hybrid or wheelchair-accessible vehicle. Approximately 60% of all medallions can be operated by minifleets that typically manage several medallions and rent them out to drivers along with the vehicle. In total there are about 70 fleet companies. The remaining 40% are owner-operated medallions. These require that they be owned by individual drivers who have to operate a taxi with this medallion for a minimum amount of time. Owner-operated medallions must be driven by the owner for at least 210 shifts (at least 9-hour long) in a year.

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.

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. Table 1 provides an overview of the different lease caps. The table reveals that the lease caps take into account the fact that the value of driving varies throughout the day due to differential demand conditions and that a hybrid vehicle is more fuel efficient.

The fixed fare for transporting a passenger is $2.50 and the fare for an additional unit is $0.4. 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.

1.2 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

---

5Fleet 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.

6Source: http://www.nycitycab.com/Services/AgentsandFleets.aspx (Accessed on 01.31.15)

7As more and more medallions in the market have been purchased by larger fleet companies, the TLC wanted to insure that 40% of medallions are owned and operated by the same person.

Table 1: Leasing rate caps

<table>
<thead>
<tr>
<th>Type of lease</th>
<th>Standard</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td>12 hour dayshift</td>
<td>115</td>
<td>118</td>
</tr>
<tr>
<td>12 hour nightshift Sun/Mon/Tue</td>
<td>125</td>
<td>128</td>
</tr>
<tr>
<td>12 hour nightshift Wed</td>
<td>130</td>
<td>133</td>
</tr>
<tr>
<td>12 hour nightshift Thur/Fri/Sat</td>
<td>139</td>
<td>141</td>
</tr>
<tr>
<td>one week dayshift</td>
<td>690</td>
<td>708</td>
</tr>
<tr>
<td>one week nightshift</td>
<td>797</td>
<td>812</td>
</tr>
</tbody>
</table>

the geo-spatial start and endpoint of the trip. The TPEP data from 2009 to 2013 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 see 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.

2 Descriptive Evidence for Market Frictions

We now provide some background information and descriptive 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) search frictions, (3) ownership requirements.

2.1 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.

Figure 1 shows a time series of recent medallion auctions. Prices of medallions are very high, and have increased 1,030% between 1980 and 2011, exceeding returns on many other forms of investment. For example, the rise in housing prices over the same time frame has been only 210%. This high medallion prices can be seen as evidence of the distorting effect of entry restrictions relative to an unregulated market. If the medallions are priced correctly they should reflect the discounted expected value from the profit stream of continued medallion rental. These profits should be zero in a market where entry is not restricted. The utilization rates of medallions provides further evidence and reveals that the medallion

constraint is binding on an average weekday. Figure 2 shows the percentage of medallions (out of the total of 13520) that are driven at least once a day, averaged over weekdays. This percentage is always above 94% except for Sundays and close to 97% from Tuesday to Thursday. Given that there will be some natural failure rate of vehicles, this means that all medallions are used during a typical weekday.

### 2.2 Search Frictions

An important friction in this market relative to the ideal of a Walrasian market arises from the fact that drivers and passenger have to physically search for trading partners. Figure 3 describes the fraction of time that an average taxi spends searching for passengers relative to the total time it is active (search time plus time spent delivering a passenger to destination). Two notable features of this figure 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, ranging up to 65%. It is important to note that under the current system of essentially fixed fares, most inter-temporal variation in driver profits and customer and driver-welfare are created by variation in waiting time for a trading partner. A simple linear model reveals that waiting time for passengers explains about 60% of the variation in hourly wages for drivers.\(^\text{10}\) The low point of the time spent searching is reached at 5PM where cab that is still on the street profits from the fact that the large majority of medallions is transitioning between shift at this time. Another interesting observation can be made from Figure 4 in conjunction with Figure 3.\(^\text{11}\) Both for passengers and taxis waiting time increases during the night although the ratio of passengers to taxis is relatively stable. This illustrates that both market sides profit when the city is more populated.

---

\(^{10}\)Most of the remaining variation is explained by trip-length and the rate that is charged per minute of driving. Note that this rate is varying with the speed of traffic due to the mixture of time-based and distance-based metering. There are also modest adjustments of the fixed surcharge during the day, but the fare regulation is still in stark contrast to a system of variable fares that encourages entry through high fares at times where supply is low.

\(^{11}\)Note that the wait time for passengers is inferred from our simulation, which is described in detail below.
Figure 2: Number of average medallions active (at some point during the day)

Notes: This figure shows the utilization rates of medallions on an average weekday. Utilization means that at least one trip was made with this medallion.

Figure 3: Searchtime 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 raitos using data from Monday to Thursday. 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 shifts.
which facilitates the matching process.

2.3 Inefficient Utilization due to Ownership Requirements

A peculiarity of the regulatory framework in New York is the restriction that 40% of medallions have to be owner-operated. An owner is required to drive at least 210 shifts (nine-hour minimum) per year. This implies that one owner cannot manage multiple owner-operated medallions. A natural question is therefore 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.

Comparing the behavior of owner-operated and minifleets reveals a number of stark differences. Figure 5 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 of owner-operated displays a much thicker left tail of low utilization rates and is overall more dispersed.\footnote{Note that in all that follows we are not arguing that there is a causal relationship from the type of medallion on the observed behavior that we document. The observed differences might be due to the fact that minifleets enables a more efficient utilization but another plausible argument is a selection-effect.}

The left panel of Figure 6 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 in general manage to 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 set 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 will allow for a different set of parameters for minifleets and owner-operated to capture these differences. The right panel of Figure 6 shows the number of shift ends conditional on the hour. We see that minifleet medallions have a more
Figure 5: Histogram of utilization separated by medallions

<table>
<thead>
<tr>
<th>Passengers delivered (fraction of total)</th>
<th>Minifleet</th>
<th>Owner-operated</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>1000</td>
<td>0</td>
</tr>
<tr>
<td>0.1</td>
<td>750</td>
<td>250</td>
</tr>
<tr>
<td>0.2</td>
<td>500</td>
<td>500</td>
</tr>
<tr>
<td>0.3</td>
<td>250</td>
<td>750</td>
</tr>
<tr>
<td>0.4</td>
<td>0</td>
<td>1000</td>
</tr>
<tr>
<td>0.5</td>
<td>250</td>
<td>250</td>
</tr>
</tbody>
</table>

Notes: Each observation in these histograms 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 show stark differences between owner-operated and minifleet medallions. The lower tail of low utilization is much thicker for owner-operated medallions.

regular pattern with most day-shifts ending at 4PM. This is also reflected in Figure 7 which shows a stronger supply decrease for minifleets before the evening shift relative to owner-operated medallions.

3 Model and Estimation

3.1 Demand Side Model

Since the New York City taxi market operates under a fixed fare system, the endogenous variables of interest that adjusts to clear the market is the wait time for taxi rides as well as the search time for passengers. There are two separate challenges regarding the estimation of the demand function. The first and more fundamental problem is that the traded quantity (i.e. rides taken by passengers) and the price (i.e. waiting time for passengers) are both not observed in our data. The second problem is that even if we knew equilibrium quantities and prices, a regression of quantities on prices would not yield consistent demand estimates due to endogeneity, which is a typical problem in the literature on demand estimation. In this section we explain how we deal with the former problem and recover the missing data we need to estimate a demand function. In subsection 3.4 we then explain the instrumental variable approach to deal with the endogeneity concern.

The idea behind our approach to recover the demand data is that the number of searching taxis together with the average time that a taxi spends searching reveals information about the number of passengers that must have been waiting on the street. Both the number of searching taxis and the average time they search for a match are observed in our data. To give the intuition: imagine two scenarios, both have the same number of searching cabs $c_1 = c_2$ but the time they spent on average to produce a match is higher in scenario one, $s_1 > s_2$. Assume also that other relevant factors are identical.
Figure 6: Time medallion is unutilized conditional on hour of drop off.

![Graph showing the relationship between time medallion utilization and hour of drop off.](image)

**Medallion type**
- minifleet
- owner-operated

Figure 7: Comparing activity of owner-operated and fleet medallions.

![Graph comparing activity of owner-operated and fleet medallions.](image)

**Medallion type**
- minifleet
- owner-operated
in both scenarios, such as the speed at which traffic is moving. Then it must be the case that there were more passengers waiting on the street in scenario two relative to scenario one. Our approach follows this basic intuition.

Define the following function \( g \), which for a given number of waiting passengers \( d_t \), searching taxis \( c_t \) as well as other exogenous time varying variables \( \phi_t \) determines the search time \( s_t \) and waiting time \( w_t \):

\[
\begin{pmatrix}
  s_t \\
  w_t
\end{pmatrix} = g(d_t, c_t, \phi_t; \theta)
\] (1)

For a given guess of \( \theta \) as well as \( c_t \) and \( \phi_t \) one could invert this function from \( s_t \) to infer \( d_t \) as long as the search time \( s_t \) itself is at least weakly decreasing everywhere with regard to the number of passengers waiting. However, without knowledge of the parameter \( \theta \) this inversion is in general not feasible as, for example, the scale of \( d_t \) might not be separately identified from the value of the parameters. Instead of using some parametric form for \( g \) we therefore use our knowledge about the geographical nature of the matching process to infer the functional 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. In the simulation, which provides an approximation of the true \( g(\cdot) \), we assume that passengers are waiting at fixed locations on a two-dimensional grid. The map consists of nodes whose spacing is proportional to \( \frac{1}{20} \) of a mile. Each of these nodes serves as a potential spot at which passengers wait. Street blocks are assumed to be \( \frac{4}{20} \) of a mile wide (east-west) by \( \frac{1}{20} \) of a mile long (north-south), which corresponds to the approximate block size in Manhattan. This means that cabs can change their direction of driving at every fourth node in the east-west direction and at every node in the north-south direction.

![Figure 8: Schematic of the grid that is used for simulation](image)

We assume that for this mapping, the effect of all relevant time varying factors can be summarized by the speed \( mph_t \) at which the traffic flows and the average distance \( miles_t \) it takes to deliver a passenger from his fixed position on the grid to the destination. Both of these factors are included in \( \phi_t \), 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. We also have to take into account that the area of potential locations at which taxis and passengers search for each other is changing throughout the day. To measure how large the area is on which matching takes place we divide the city map into census tracts and for each hour in the data sum up the area of all census tracts on which a match occurred. For each hour of the day we then take the average of all those measurements. Figure 14 shows those hourly averages and Figure 15 a map with census tracts areas from which those values have been derived. These hourly averages show that the matching occurs on a relatively smaller area at night- then at daytime. In particular, the average covered square miles are particularly low during the hours from 2AM to 4AM. Based on this observation we perform the simulation separately for the average sized map during night-time (2AM to 4AM) and day-time hours (the remaining hours).

For each combination of input variables to $g$ we simulate the resulting average waiting time for passengers and search time for taxis over an hour long time interval that represents an hour in real time. Every ten minutes $d/6$ potential passengers are born and placed on the map at random locations (with each node having equal probability) for a total of $d$ passengers during the hour. Passengers request trips to a random node where each node has again equal probability weight. A node can host multiple passengers. The initial locations of cabs are also randomly assigned (again with equal probability weights on each node). Note that in this step we are not yet using any observed data. We merely want to obtain a representation of $g()$ for any point in its domain.

However, due to computational limitations it is not feasible to literally 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 with the maximal number of medallions it would already amount to 945 Million different points in the domain without even considering variation in $\phi$. We therefore simulate $g$ for a lower number of grid points and we choose to interpolate linearly between those points to obtain the image for points inbetween. For each of the four independent variables of $g(\cdot)$, we pick eight different evenly spaced grid points. Note that 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 is not changing much more after ten iterations, which is therefore what we pick to produce this average.

Using this approximation one can invert the approximated $\hat{g}$ and back out $d_t$ for each combination of hourly averages of $s_t$, traffic speed, and trip distance observed in the data. Once the number of passengers is known we can use this to determine their wait-time $w_t$ as well. Additional details on the simulation are provided in Appendix A.

Figure 9 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 approximation $\hat{g}(\cdot)$. The passenger graphs 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 which occurs at 5PM, which is well before the spike in demand occurring at 7PM. The reason is the coordinated shift change leading to the more unfavorable ratio of active cabs to searching passengers.
3.1.1 The uniform Grid Assumption

The simulation assumes that the grid is uniform and that search is random, thereby ruling out directed search across neighborhoods. A first important fact to support this assumption is that 93% of all trips in the Market happen in Manhattan, which is dense and geographically relatively homogeneous. Moreover, if we wanted to model directed search with heterogeneous neighborhoods, we would need to make additional assumptions both on how passengers are differentially located on the map during different times of the day and how drivers search for them. Another complication would be that the division of the map into different parts would need to be relatively fine grained because the average search time is below ten minutes during the day, which means that drivers search local, if at all. Implementing such a local simulation would be too computationally intensive to accommodate.

3.2 Supply Side Model

We model a driver’s decisions to work as a function of earnings as well as the regulatory and organizational constraints imposed by the medallion system that we discuss in section 1. We estimate a structural dynamic model in which driver’s optimally time their shifts under these constraints. In each time period (hour) $t$ there are $N_t$ active and $M_t$ inactive medallions, which sum up to the total number of medallions issued by the city.

In section 1 and section 2 we highlight the fact that minifleet medallions are overall more 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. Because of these differences, we allow medallions in the model to differ in two dimensions: the first dimension captures the ownership and the second dimension the time at which a medallion typically transitions between shifts. The index \( z_j \in Z \) denotes the ownership of the medallion and takes on two values depending on whether the medallion is a minifleet or an owner-operated medallion. We allow for a different set of parameters for minifleet and owner-operated medallions. The index \( k_j \in K \) captures at which time the medallion transitions between day-shift and night shift.

For each hour \( t \) during which a medallion \( j \) is inactive, a driver \( i \) arrives and decides to start driving when the value of driving is higher than his outside option over the expected length of a shift. The hourly wage is denoted as \( \pi_i \) and follows an endogenous distribution \( F_\pi(.|h_t) \). We assume that the utility from the outside option is comprised of a fixed value \( h \mu_{h_i,z_j} \) as well as an idiosyncratic component \( v_{it0} \). The utility of driving depends on the entire expected value of a shift \( EV(x_{t+1},\pi_{t+1}|h_{t+1}) \) as well as an idiosyncratic component \( v_{it1} \). Drivers also have to pay a rental fee \( r_{ht} \) unless they own the medallion, in which case, they have an opportunity cost of driving equal to that value \( r_{ht} \). We set \( r_{ht} \) equal to the time varying rate caps shown in Table 1, which according to anecdotal evidence are always binding. 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 \( \sigma_v \).

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

\[
u_{it0} = \mu_{h_i,k_i} + v_{it0},\]

and the utility for starting a shift is given by:

\[
u_{it1} = EV(x_{t+1},\pi_{t+1}|h_{t+1}) - r_{ht} + v_{it1}.
\]

As is well known, the convenient feature of T1EV distribution for the error terms is that it allows us to obtain a closed form for choice probabilities. Denoting by \( p^l \) the probability that an inactive driver starts a shift, we obtain:

\[
p^l = \frac{\exp((EV(x_{t+1},\pi_{t+1}|h_{t+1}) - r_{ht})/\sigma_v))}{\exp((EV(x_{t+1},\pi_{t+1}|h_{t+1}) - r_{ht})/\sigma_v)) + \exp(\mu_{h_i}/\sigma_v)}
\]

The value of a shift under an optimal shift length \( l_\beta \) is given by:

\[
V(x_{t},\pi_{t},e_{it}) = \max\{e_{it0},\pi_{t} - C_{z,j,h_{t}}(l_{it}) - f(h_{t},k_{j}) + e_{it1}
\]

\[+\beta \cdot E_{\epsilon_{it}}[V(x_{t+1},\pi_{t+1},e_{i(t+1)}|h_{t+1})]\]

This expression depends on an observable state vector \( x_{t} = (h_{t},l_{t},z_{j},k_{j}) \) as well as an idiosyncratic unobservable vector \( e_{it} = (\epsilon_{it0},\epsilon_{it1}) \), assumed to be distributed according to i.i.d. T1EV distributions with scale parameter \( \sigma_e \). The driver’s problem takes the form of an optimal stopping problem. Each period the driver decides whether he wants to collect the flow payoff from driving plus the continuation value of an active shift or the random value of the outside option. The cost of driving \( C_{z,j,h_{t}}(l_{it}) \) is a

\[\text{The scale parameter } \sigma_v \text{ is identified because } EV(x_{t+1},\pi_{t+1}|h_{t+1}) \text{ is a given value from the stopping problem and not pre-multiplied by any parameter.}\]
function of the length $l_t$ of the shift. The parameters of this function are indexed both by the medallion type $z_j$ as well as 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:

$$C_{z_j,h_t}(l_t) = \lambda_0, z_j, h_t + \lambda_1, z_j \cdot l_t + \lambda_2, z_j \cdot l_t^2.\text{14}$$

The term $f(h_t, k_j)$ is a fine that has to be paid if the medallion is not transitioned to the next driver in time. Anecdotal evidence suggests that such fines are very common to insure that drivers do not operate the cab longer than contractually specified. Since fines are not observed we estimate them as parameters.

We use the fact that we can see medallions over a long period of time to classify each medallion into a category $k_j \in K$ 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 $f_{z_j}$ whenever he goes past the common night shift starting time and vice versa. The fine is again indexed by $z_j$ to account for the fact that owner-operated medallions seem to have less stringent transition times as suggested by Figure 7. We denote by $h_{k_j}$ the set of hours at which a medallion of type $k_j$ transitions.

We therefore have:

$$f_{z_j}(h_t, k_j) = f_{z_j} \cdot \mathbb{1}\{h_t \in h_{k_j}\}$$

Equation 3 says that a driver of type $k_j$ must pay a fine that depends on the hour $h_{k_j}$ in which he typically transitions.

### 3.2.1 Equilibrium Definition

The descriptive evidence presented in section 2 suggests that the market operates under a regular pattern of intra-daily activity. Regressions of the number of active taxis on hour-weekday dummies confirm that most of the variation can be explained this way. Based on this observation we suggest a model in which agents make their decisions in discrete hourly intervals, which we interpret as representative hours of a weekday. In the estimation of the model, which we discuss in detail below, we use data from Monday to Thursday, which look identical (Figure 9). Agents in the model build their expectations for each of these weekday hours. To formalize this notion let $h_t$ be an index from the set $H \times D$, where $H = \{1, \ldots, 24\}$ and $D = \{1, \ldots, 7\}$, i.e. for each time period $t$, $h_t$ tells us the hour-weekday combination.

Hourly earnings $\pi_t$ of drivers are determined by the equilibrium distribution of active medallions and searching passengers of such a steady state equilibrium and drivers forecast their earnings according to $\mathcal{F}_\pi(\cdot|h_t)$.

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

1. $\mathcal{F}_d(\cdot|h_t)$ results from the demand function $d_t$ under the distribution of waiting times $\mathcal{F}_s(\cdot|h_t)$

---

14We have also explored a specification with higher order polynomials but this does not make a difference in the results.
2. \( \mathcal{F}_s(\cdot | h_t) \) and \( \mathcal{F}_w(\cdot | h_t) \) result from \( \mathcal{F}_d(\cdot | h_t) \) and \( \mathcal{F}_c(\cdot | h_t) \) under the matching function \( g(\cdot) \).

3. \( \mathcal{F}_c(\cdot | h_t) \) results from optimal starting and stopping under \( \mathcal{F}_\pi(\cdot | h_t) \).

4. \( \mathcal{F}_\pi(\cdot | h_t) \) results \( \mathcal{F}_s(\cdot | h_t) \) under the rate that drivers.

3.3 Identification of Supply Side Parameters

In this section we briefly discuss how the primitives of the model are identified. Due to the short time horizon of hourly stopping decisions we assume that the hourly discount factor \( \beta \) is equal to one. The cost function has multiple components: (1) There is a term that varies with the duration of the shift, (2) there is an hourly fixed component, (3) there are the fines, and, finally (4) there is the standard deviation. The identification of (1) can be best understood using backwards induction for the driver’s 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 \( l_{\text{max}} \) 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. It can be observed, for example, that the stopping probability increases sharply after 12pm even though there is not an equal sharp decline in the earnings. This kind of variation in the data identifies the differences in the \( \lambda_0 \) 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 \( \pi_t \). 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 of \( EV(x_{t+1}, \pi_{t+1} | h_{t+1}) - r_{ht} \), which is composed of data and identified objects, throughout the day identify the different values of \( \mu_{ht} \) and the dispersion the value of \( \sigma_{vt} \).

3.4 Estimation

The estimation is performed only using the days from Monday to Thursday. The average activity of these days looks almost identical whereas Friday nights, for example, are very different from weekday nights. The reason we do not further differentiate between weekdays is that it would be computationally too hard to obtain counterfactuals in this case. Note that already in this reduced case we have to solve for 96 endogenous variables when computing an equilibrium, the means and variances for taxis and passengers at each hour of the day. All our results are therefore interpreted as pertaining to an “average” weekday.
### 3.4.1 Estimating the Demand Function

The simulation of passengers and their wait time provides us with the ingredients to estimate a demand function. We assume a constant elasticity demand function of the form:

\[
d_t = \exp(a + \sum_h \beta_h \cdot 1\{h_t\}) \cdot w_t^n \cdot \exp(\xi_t) \tag{4}
\]

The multiplicative component \( \exp(a + \sum_h \beta_h \cdot 1\{h_t\}) \) captures observed exogenous factors that may shift demand, such as weather conditions or the rush hours, and \( \xi_t \) captures unobserved conditions that shift demand. The main parameter of interest is \( \eta \), the constant elasticity of demand with respect to waiting time. The parameters \( a \), the \( \beta_h \) and \( \eta \) can be estimated using a simple log-linear transformation of Equation 4:

\[
\log(d_t) = a + \sum_h \beta_h \cdot 1\{h_t\} + \eta \cdot \log(w_t) + \xi_t \tag{5}
\]

A potential problem with this specification 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 \( \xi_t \) which would lead to a decrease in waiting time. For both of these reasons we have to expect that \( \text{cov}(\xi_t, w_t) \neq 0 \), which would introduce a bias in the estimation of \( \eta \). To address this concern and assess the severity of the potential problem we provide both OLS specifications and specifications in which we instrument for the wait time. We need a variable that is correlated with wait time and which affects demand only through the waiting time. A pure supply shifter satisfies this requirement. 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 7 and Figure 6 shows that this is due to the transitioning of shifts.\(^\text{15}\) 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.\(^\text{16}\) 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 10 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 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).\(^\text{17}\)

\(^{15}\)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.

\(^{16}\)See section 58-21(c) in TLC (2011).

\(^{17}\)Clearly this comparison ignores any equilibrium effects of changing the shifts structure. The graph can therefore be understood as the earnings that one deviating medallions could have under the current system.
Figure 10: Earnings of day and night shift for different split times

Notes: This graph 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.

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 “dissappear” (in Appendix C we also present the regressions of Table 2 where we just use the transition times as dummies which leads to almost identical negative and significant demand elasticities).
Table 2: Different specifications for demand estimation.

<table>
<thead>
<tr>
<th>Variable</th>
<th>(1 OLS)</th>
<th>(2 OLS)</th>
<th>(3 IV)</th>
<th>(4 IV)</th>
<th>(5 IV)</th>
<th>(6 IV)</th>
<th>(7 IV)</th>
<th>(8 IV)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>log(d₁)</td>
<td>log(d₂)</td>
<td>log(w₁)</td>
<td>log(d₁)</td>
<td>log(d₂)</td>
<td>log(w₁)</td>
<td>log(d₁)</td>
<td>log(d₂)</td>
</tr>
<tr>
<td>log(w₁)</td>
<td>-0.160*</td>
<td>-0.0837**</td>
<td>-1.045**</td>
<td>-0.805**</td>
<td>-1.922**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0738)</td>
<td>(0.00283)</td>
<td>(0.0449)</td>
<td>(0.0432)</td>
<td>(0.344)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>shift ins.</td>
<td></td>
<td>0.239**</td>
<td></td>
<td>0.137**</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.00693)</td>
<td></td>
<td>(0.00767)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>miles inst.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>714</td>
<td>714</td>
<td>714</td>
<td>714</td>
<td>714</td>
<td>714</td>
<td>714</td>
<td>714</td>
</tr>
<tr>
<td>2-Hour FE</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>R²</td>
<td>0.0124</td>
<td>0.909</td>
<td>0.288</td>
<td>.</td>
<td>0.648</td>
<td>0.794</td>
<td>0.613</td>
<td>0.165</td>
</tr>
</tbody>
</table>

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.

Table 2 shows the results of our demand estimation. There are two OLS specifications and two IV specifications. Our instrument consists of dummy variables for the times at which most medallions enforce their shift transition (4AM, 5AM, 4PM and 5PM). For the regressions there is a trade-off in how finely we control for recurring patterns of demand through time fixed-effects. Since demand is extremely predictable, allowing for hourly controls effectively absorbs most of the variation in demand. However, the fact that demand is predictable does not mean that it is also inelastic. In the case of hourly controls our instrument would also be collinear with fixed effects. To decide on how fine-grained time-fixed-effects should be we run a series of IV regressions starting from a model without any time controls moving towards specifications with finer controls. Table 6 in Appendix C shows the results of these regressions. The elasticity is relatively stable when moving to finer controls while the percentage of explained variation increases. Table 2 shows the two extreme cases of the IV specifications, the one without time controls and the one with two-hourly FE. Importantly, the elasticities of both specifications are almost the same: −1.045 versus −0.805. Of course, the latter specification explains a lot more of the demand variation. We use this specification with two-hourly fixed effects as the basis for our structural model. The table shows that the OLS specification produces estimates that are biased towards zero.

Figure 11 shows what demand would be if the waiting time was constant (at the overall mean) throughout the day and therefore 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 6PM and 7PM and lowest in the morning hours from 2AM to 7AM.

3.4.2 Validity of Instrument and an Alternative

The potential concern with our instrument is that, although we interact the occurrence of the witching hour with traffic condition, there is not enough un-anticipated variation to make up for passengers adjustment to the high waiting time during the “witching hour”. To adress this concern we explore the average miles of a trip as an alternative instrument. The intuition behind this instrument is that
the average mile length of a trip captures road conditions, such as closures of bridges or tunnels which lead to longer trips and therefore less supply as taxis spend more time delivering passengers. For the instrument to be valid it would need to be the case that trip length is uncorrelated with unobserved shocks to demand. The last two specifications of Table 2 show these alternative demand regressions. This specification leads to a larger elasticity of $-1.922$. While our current structural specification uses the elasticity obtained by the shift change instrument, we are planning to implement an alternative set of counterfactuals with the larger elasticity.

3.4.3 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 the data from 2011 as well as the August data from 2012. The model is estimated for a typical weekday so only the data from Monday to Thursday is used.

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. This definition might sometimes lead to long breaks within a shift if there is a long break 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 times 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 this rate is a combination of time-based and mile-based metering. As most fares are the result of a mixture between these two types of metering we first calculate the actual hourly based rate $\pi_t^0$ 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.\(^{18}\) For each hour we also compute the average search time for the next 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 now compute a realisation 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.

As we discuss in section 2 transitions from day shift to night shift (d-to-n) and night shift to day shift (n-to-d) occur at concentrated times around 5AM and 5PM. In the data we find that even for very short shifts, drivers have a much higher likelihood of stopping around these times. This indicates that drivers make a constrained stopping decision, which is due to the obligation to hand the car to the next driver. In the model we estimate the fine $f$ to quantify this institutional constraint. To take into account that different medallions have different arrangements for transitions, we create an indicator for each medallion, which identifies the hour at which most of the (d-to-n) and (n-to-d) occur for this medallion. In the estimation we then assume that the fine applies whenever the night shift driver drove at or after the (d-to-n) hour and vice versa for the night shift driver.

3.4.4 Estimation Procedure

The equilibrium that we assume above is a stationary one. For the estimation of supply side parameters we make use of the fact that given a known set of distributions for hourly equilibrium earnings, $F(\pi_t|h) \forall h \in H$, 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). 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).\(^{19}\) 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.

In our case this constraint is coming 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 mapping to solve for the value function for each parameter guess

\(^{18}\)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 credit card.

\(^{19}\)Nested estimation refers to the necessity to compute value functions or the full equilibrium for each parameter guess.
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.\textsuperscript{21}

We specify a likelihood objective function. Following the model discussion above, we allow all parameters to be different for minifleet and owner-operated medallions. We allow the daily outside option $\mu_{z_j}$ to be different from 5AM to 5PM versus the rest of the time. We allow the fine $f_{h_0,z_j}$ 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 24PM to 5AM, from 5AM to 5PM and 5PM to 24PM. The other two parameters $\lambda_{1,z_j}$ and $\lambda_{2,z_j}$ of the cost function are assumed to be time invariant. The remaining parameters are the standard deviation of the idiosyncratic shocks to the starting decision $\sigma_{v,z_j}$ and the stopping decision $\sigma_{e,z_j}$. We will refer to the combined vector of parameters as $\theta$.

As discussed above, we allow all parameters to be different for minifleet and owner-operated medallions. We allow the daily outside option $\mu_{z_j}$ to differ for the time from 5am to 5pm (day-shift) versus the rest of the time (night-shift). The fine $f_{h_0,z_j}$ also depends 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 time of the day in the following way: from 1am to 5am, 6am to 12am, 1pm to 5pm and 6pm to 12pm. The other two parameters $\lambda_{1,z_j}$ and $\lambda_{2,z_j}$ of the cost function are assumed to be time invariant. The remaining parameters are the standard deviation of the idiosyncratic shocks to the starting decision $\sigma_{v,z_j}$ and the stopping decision $\sigma_{e,z_j}$. As defined above, let $\pi_t$ be a realization of the earnings and $x_{jt}$ denote the other part of the observable state (the shift length, the hour of the day as well as the medallion invariant characteristics). And let $p(x_{jt}, \pi_t)$ be the theoretical probability that an active medallion $j$ stops at time point $t$ and $q(x_{jt})$ be the probability that an inactive medallion $j$ 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 constraint log-likelihood that we as an MPEC problem. MPEC does not perform any intermediate computations, such as value function iterations, to compute the objective function. It therefore treats objects as parameters that would normally be a known input for the formulation of the objective function. This means that the solver will be maximizing both over the parameters of interest $\theta$ and an additional set of parameters $\gamma$. The parameter vector $\gamma$ consists of all $p(x_{jt}, \pi_t), q(x_{jt}), EV(x_{jt+1}, \pi_t|h_t)$ for $x_{jt} \in X, h_t \in H, \pi_t \in supp(F_{\pi}(\cdot|h_t))$. In other words, $\gamma$ consists of expected values and choice probabilities for each point in the observable state space. Note, however that $\pi_t$ follows as 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 $\pi_t$ with a discrete number of nodes $\hat{\pi}$ and weights using gauss-hermite integration.\textsuperscript{22}

\textsuperscript{20}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.

\textsuperscript{21}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.

\textsuperscript{22}We use six nodes.
With this notation in place we can express the maximization problem as follows

$$
\min_{\theta, p(x, \pi), q(x)} \sum_{j \in J} \sum_{t \in T} \left( d^A_{jt} \cdot \log(p(x_{jt}, \pi_t)) + (1 - d^A_{jt}) \cdot \left( \log(1 - p(x_{jt}, \pi_t)) \right) \right)
$$

subject to:

$$
E_{\epsilon} V(x_t, \pi_t | h_t) = \sigma_{\epsilon} \cdot \log \left( \frac{1}{\sigma_{\epsilon}} \right) + \exp \left( \frac{\pi_t - C_{x_t}(x_t) - f(x_t) + E_{\epsilon, \pi} V(x_{t+1}, \pi_{t+1} | h_{t+1})}{\sigma_{\epsilon}} \right)
$$

$$
p(x_{jt}, \pi_t) = \frac{\exp \left( \frac{1}{\sigma_{\epsilon}} \right)}{\exp \left( \frac{1}{\sigma_{\epsilon}} \right) + \exp \left( \frac{\pi_t - C_{x_t}(x_t) - f(x_t) + E_{\epsilon, \pi} V(x_{t+1}, \pi_{t+1} | h_{t+1})}{\sigma_{\epsilon}} \right)}
$$

$$
q(x_{jt}) = \frac{\exp \left( \frac{E_{\epsilon, \pi} V(x_{t+1}, \pi_{t+1} | h_{t+1}) - r_{x_t}}{\sigma_{\epsilon}} \right)}{\exp \left( \frac{E_{\epsilon, \pi} V(x_{t+1}, \pi_{t+1} | h_{t+1}) - r_{x_t}}{\sigma_{\epsilon}} \right) + \exp \left( \frac{\mu_{x_t}}{\sigma_{\epsilon}} \right)}
$$

The constraint given by equation Equation 7 insures that the $EV(x_t)$ obeys the intertemporal optimality condition. 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. Table 3 shows the resulting parameter estimates. 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.

### 3.5 Parameter Estimates

Table 3 gives an overview over the estimated parameters. Results are shown separately for minifleet and owner-operated medallions. Standard error calculations are bootstrapped where we drew 50 samples with replacement at the medallion level.
Table 3: Parameter Estimates (standard errors in parentheses)

<table>
<thead>
<tr>
<th>parameter</th>
<th>description</th>
<th>minifleet ((z_j = F))</th>
<th>owner-operated ((z_j = NF))</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\mu_{z,j,0})</td>
<td>outside-option, 6pm-4am</td>
<td>129.75 (1.43)</td>
<td>242.32 (2.61)</td>
</tr>
<tr>
<td>(\mu_{z,j,1})</td>
<td>outside-option, 5am-5pm</td>
<td>132.39 (1.4)</td>
<td>234.86 (2.5)</td>
</tr>
<tr>
<td>(f_{0,z,j})</td>
<td>fine (nightshift)</td>
<td>52.17 (0.85)</td>
<td>69.65 (1.23)</td>
</tr>
<tr>
<td>(f_{1,z,j})</td>
<td>fine (dayshift)</td>
<td>55.06 (0.59)</td>
<td>58.33 (0.59)</td>
</tr>
<tr>
<td>(\lambda_{0,z,j,0})</td>
<td>fixed cost (1am-5am),</td>
<td>20.28 (0.21)</td>
<td>25.87 (0.4)</td>
</tr>
<tr>
<td>(\lambda_{0,z,j,1})</td>
<td>fixed cost (6am-12am),</td>
<td>8.57 (0.2)</td>
<td>1.33 (0.36)</td>
</tr>
<tr>
<td>(\lambda_{0,z,j,2})</td>
<td>fixed cost (1pm-5pm),</td>
<td>0.0 ((\leq 1e^{-6}))</td>
<td>0.0 ((\leq 1e^{-6}))</td>
</tr>
<tr>
<td>(\lambda_{0,z,j,3})</td>
<td>fixed cost (6pm-12pm),</td>
<td>8.22 (0.16)</td>
<td>0.0 (0.31)</td>
</tr>
<tr>
<td>(\lambda_{1,z,j})</td>
<td>linear cost coefficient</td>
<td>0.0 ((\leq 1e^{-6}))</td>
<td>0.0 ((\leq 1e^{-6}))</td>
</tr>
<tr>
<td>(\lambda_{2,z,j})</td>
<td>quadratic cost coefficient</td>
<td>0.28 ((\leq 1e^{-6}))</td>
<td>0.36 ((\leq 1e^{-6}))</td>
</tr>
<tr>
<td>(\sigma_{\epsilon})</td>
<td>sd iid hourly outside option</td>
<td>32.99 (0.36)</td>
<td>57.6 (0.45)</td>
</tr>
<tr>
<td>(\sigma_{v})</td>
<td>sd iid daily outside option</td>
<td>32.89 (0.38)</td>
<td>39.19 (0.54)</td>
</tr>
</tbody>
</table>

The mean values of the outside option for night \((\mu_{z,j,0})\) and daytime \((\mu_{z,j,1})\) are estimated to be $129.75 and $132.39 for minifleet drivers and $242.32 and $234.86 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 of a minifleet is therefore higher at all times of the day, consistent with the descriptive evidence that minifleet shifts last longer. Comparing parameter values for fleets and non-fleets also reflects the observation that fleets operate under a more stringent shift-system. The estimated fines \(f_{0,z,j}\) and \(f_{1,z,j}\) for violating the shift constraint are $69.65 (nightshifts) and $58.33 (dayshift) for owner-operated medallions. For minifleets the nightshift fine is $52.17 and the dayshift fine is $55.06. The fixed part of the cost function parameters for minifleets are estimated at $20.28 from 1am to 5am, $8.57 from 6am to 12am, 0.0 from 1pm to 5pm and 8.22 from 6pm to 12pm. For owner-operated medallions these values are 25.87, 1.33, 0.0 and .0 respectively. The linear part of the hourly increase in cost is estimated to be zero for both minifleet and owner-operated cabs and the quadratic parts are 0.28 for minifleets and 0.36 for owner-operated medallions. The standard deviations of the hourly outside option (conditional on driving) is 32.99 for minifleet medallions and 57.6 for owner-operated medallions. The standard deviations of the daily outside option is 32.89 for minifleet medallions and 39.19 for owner-operated medallions.

3.5.1 Discussion of Parameter Estimates

A few observations about the estimates are worth highlighting. Figure 7 has shown that minifleet medallions follow the 5AM to 5PM shift pattern much more stringently. This is reflected in the estimates. owner-operated have a higher standard deviation in the hourly error terms, which are the random parts that determine stopping behavior. Because the standard deviation for owner-operated medallions is so much higher their stopping behavior is “smoother”, i.e. there is 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 then 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 12) and the expected random term for continuing to drive is larger because of the larger standard deviation of the hourly outside option.

Figure 12: cost functions at different times of day

3.5.2 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: $p^M(t) = p(h_t) \cdot (1 - q(h_t))$. 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:

\[
N_t = (1 - p^M_t) \cdot N_t + q^M_t \cdot M_t
\]

and

\[
M_t = (1 - q^M_t) \cdot M_t + p^M_t \cdot N_t
\]

The model fit for the law of motion is presented in Figure 13, which shows that we are able to replicate the daily pattern of supply activity relatively well with few parameters.

Figure 13: Model Fit

4 Counter-factual Experiments

We now use the model to perform counter-factual experiments to explore the relative quantitative importance of the product market search friction as well as the capital market constraints on medallion ownership. We first look at a policy that increases the number of medallions by 10%, from 13500 to 14850. In a second counter-factual we explore how the market equilibrium changes if all medallions were utilized by the relatively more efficient minifleet companies. Third, we investigate a policy that improves the matching technology between cabs and passengers. For the latter case we explore both what would happen if the entire market operates under a more efficient matching technology.

For all counter-factuals the following equilibrium objects are documented: the hourly aggregate number of active cabs \( \{c_{h_t} | h_t \in \{0, ..., 23\}\} \), expected hourly driver revenue \( \{\pi_{h_t} | h_t \in \{0, ..., 23\}\} \), expected hourly passenger demand \( \{d_{h_t} | h_t \in \{0, ..., 23\}\} \), hourly passenger wait time \( \{w_{h_t} | h_t \in \{0, ..., 23\}\} \).
and hourly cab search time \( \{ s_{ht} | h_t \in \{0, \ldots, 23\} \} \). We also quantify the changes (measured in time units) in overall consumer surplus:

\[
CS = \sum_t \int_0^{d_{ht}^*(w_{ht}^*)} (w_{ht}^*(d_{ht}) - w_{ht}) dd_{ht} dF(d_{ht})
\]

In addition, we compute the counter-factual number of serviced trips and the counter-factual revenue stream of a medallion. The latter is obtained via simulation where 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 stream under an annual interest rate of 3%. For the latter we also assume that the cap on the rental price of a medallion is held fixed. Revenues are averaged over the different types of medallions according to their observed fractions in the data. All counter-factuals are computed in two steps. We first explore a new equilibrium in which we do not allow demand to expand and only look at changes due to the supply side. This scenario is then compared to the full counter-factual in which demand is allowed to adjust.

4.1 How to compute Equilibria

In the estimation there was no need to compute market equilibria since the dynamic problem of drivers was estimated against the observed process of hourly earnings. Solving for equilibria is challenging due to the large number of endogenous variables. There are four endogenous variables for each hour of the day, which makes for a total of 96 endogenous variables. We found that solving for an equilibrium over all variables at once hardly ever let the solver to converge. Instead we iteratively solve for the supply and demand side parameters, holding the variables of the respective other market side fixed. We found that this procedure always converged to the same equilibrium. The details of the algorithm are as follows, where \( i \) is the index of the outer loop and \( j \) the index of the inner loop:

1. Under the assumption that the hourly distribution of passengers and cabs can be approximated by a normal distribution, guess means \((\alpha_{h}^{c,0}, \alpha_{h}^{d,0}, h \in \{0, \ldots, 23\})\) and variances \((\psi_{h}^{c,0}, \psi_{h}^{d,0}, h \in \{0, \ldots, 23\})\) for taxis and passengers for each hour.\(^{23}\)

2. Supply Side:
   
   2.1. Compute initial deviation \( \text{sumsq}_{i0} = \sum_h (\alpha_{h}^{c,i0} - \alpha_{h}^{c,(i-1)0})^2 + \sum_h (\psi_{h}^{c,i0} - \psi_{h}^{c,(i-1)0})^2 \).
   
   2.2. For each hour simulate values from the distribution of taxis and passengers (using latest guess) as well as the observed empirical distributions of speed of traffic flow and the length of requested trips to determine the distributions of search time for taxis \( F_{ht}^s(.|h), h \in \{0, \ldots, 23\} \) under \( g(.) \).
   
   2.3. Simulate drivers earnings \( F_{ht}^d(.|h), h \in \{0, \ldots, 23\} \) 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 \( F_{ht}^w(.|h), h \in \{0, \ldots, 23\} \) from the distribution of waiting times and the estimated demand function \( d(w_t) \).

\(^{23}\)Allowing for a non-parametric distribution would not be feasible. Already under this simplifying assumption of a normal distributions we have to solve for 96 endogenous values, which takes several days to solve for.
2.4. Compute the optimal starting and stopping probabilities \( p_{ij}(x, \pi), q_{ij}(x) \) under the new distribution of earnings (computed in step 3) and the estimated parameters \( \theta \). The distribution of earnings is approximated using gauss-hermite integration in the stopping problem of drivers.

2.5. Use the starting and stopping probabilities (computed in step 4) to simulate a new distribution of taxis for each hour \( F^{ij}(.|h), h \in \{0, \ldots, 23\} \). 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.\(^{24}\)

2.6. Compute the sum of squared deviations between the old and the newly obtained means and variances for taxis: \( \text{sumsq}_{ij} = \sum_h (\alpha^{c,ij}_h - \alpha^{c,ij}_{h(1)})^2 + \sum_h (\psi^{c,ij}_h - \psi^{c,ij}_{h(1)})^2 \).

2.7. Iterate on steps 2.1 to 2.6 until convergence.

3. **Demand Side:**

3.1. Compute initial deviation \( \text{sumsq}_{i0} = \sum_h (\alpha^{d,0}_h - \alpha^{d,0}_{h(1)})^2 + \sum_h (\psi^{d,0}_h - \psi^{d,0}_{h(1)})^2 \).

3.2. For each hour simulate values from the distribution of taxis and passengers (using latest guess) 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 \( F^{wij}(.|h), h \in \{0, \ldots, 23\} \) for passengers.

3.3. Simulate new distribution of passengers \( F^{wij}(.|h), h \in \{0, \ldots, 23\} \) from the distribution of waiting times and the estimated demand function \( d(wt) \).

3.4. Compute the sum of squared deviations between the old and the newly obtained means and variances for passengers: \( \text{sumsq}_{ij} = \sum_h (\alpha^{d,ij}_h - \alpha^{d,ij}_{h(1)})^2 + \sum_h (\psi^{d,ij}_h - \psi^{d,ij}_{h(1)})^2 \).

3.5. Iterate on steps 3.2 to 3.5 until convergence.

4. Iterate on steps 2 to 3 until both initial deviations (supply and demand side) fall below some threshold.

\(^{24}\)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.
Table 4: counter-factual results

<table>
<thead>
<tr>
<th></th>
<th>hourly active cabs</th>
<th>hourly demand</th>
<th>passenger waittime</th>
<th>taxi searchtime</th>
<th>hourly wages</th>
<th>consumer surplus</th>
<th>number of trips per day</th>
<th>medallion revenue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>7937.0</td>
<td>23378.0</td>
<td>1.57</td>
<td>8.11</td>
<td>42.28</td>
<td>1.77 * 10^6</td>
<td>560939.0</td>
<td>2.6 * 10^6</td>
</tr>
<tr>
<td>Entry</td>
<td>8601.0</td>
<td>25080.0</td>
<td>1.44</td>
<td>8.25</td>
<td>42.06</td>
<td>1.84 * 10^6</td>
<td>604301.0</td>
<td>2.58 * 10^6</td>
</tr>
<tr>
<td>Δ%</td>
<td>8.36</td>
<td>7.28</td>
<td>-8.29</td>
<td>1.76</td>
<td>-0.5</td>
<td>4.18</td>
<td>7.73</td>
<td>-0.86</td>
</tr>
<tr>
<td>Entry (PE)</td>
<td>8171.0</td>
<td>23378.0</td>
<td>1.44</td>
<td>8.66</td>
<td>41.24</td>
<td>1.85 * 10^6</td>
<td>562177.0</td>
<td>2.47 * 10^6</td>
</tr>
<tr>
<td>Δ% (PE)</td>
<td>2.94</td>
<td>0.0</td>
<td>-8.54</td>
<td>6.86</td>
<td>-2.46</td>
<td>4.51</td>
<td>0.22</td>
<td>-5.04</td>
</tr>
<tr>
<td>Ownership</td>
<td>8382.0</td>
<td>24497.0</td>
<td>1.46</td>
<td>8.2</td>
<td>42.2</td>
<td>1.82 * 10^6</td>
<td>590688.0</td>
<td>2.6 * 10^6</td>
</tr>
<tr>
<td>Δ%</td>
<td>5.6</td>
<td>4.78</td>
<td>-7.02</td>
<td>1.17</td>
<td>-0.18</td>
<td>2.81</td>
<td>5.3</td>
<td>-0.13</td>
</tr>
<tr>
<td>Ownership (PE)</td>
<td>8070.0</td>
<td>23378.0</td>
<td>1.51</td>
<td>8.43</td>
<td>41.64</td>
<td>1.8 * 10^6</td>
<td>561409.0</td>
<td>2.51 * 10^6</td>
</tr>
<tr>
<td>Δ% (PE)</td>
<td>1.67</td>
<td>0.0</td>
<td>-4.01</td>
<td>4.03</td>
<td>-1.5</td>
<td>1.96</td>
<td>0.08</td>
<td>-3.42</td>
</tr>
<tr>
<td>Search</td>
<td>8502.0</td>
<td>24783.0</td>
<td>1.35</td>
<td>7.35</td>
<td>43.76</td>
<td>1.88 * 10^6</td>
<td>621310.0</td>
<td>2.73 * 10^6</td>
</tr>
<tr>
<td>Δ%</td>
<td>7.11</td>
<td>6.01</td>
<td>-13.81</td>
<td>-9.33</td>
<td>3.51</td>
<td>6.34</td>
<td>10.76</td>
<td>4.82</td>
</tr>
<tr>
<td>Search (PE)</td>
<td>8191.0</td>
<td>23378.0</td>
<td>1.29</td>
<td>7.81</td>
<td>42.79</td>
<td>1.92 * 10^6</td>
<td>586003.0</td>
<td>2.66 * 10^6</td>
</tr>
<tr>
<td>Δ% (PE)</td>
<td>3.19</td>
<td>0.0</td>
<td>-18.04</td>
<td>-3.67</td>
<td>1.22</td>
<td>8.72</td>
<td>4.47</td>
<td>2.18</td>
</tr>
</tbody>
</table>

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 counter-factual 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 1.5 times the maximal waiting time observed in the data. The reason is that for our parameter specifications consumer surplus would be infinite if 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.
4.2 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 (entry counter-factual). Table 4 shows the resulting changes in variables along with the baseline model results and the results of the other counter-factuals. Taking the average over all hourly changes, the supply expands by 8.36%, less then proportional to the medallion increase. This compares to only 2.94% if demand were held fixed under the new policy. The average expansion of demand is about 7.28%. The average reduction in wait-time for passengers over all hours is 8.29%, which compares to minus 8.54% in the case were demand would be held fixed. Overall, this means that consumer surplus would increase by about 4.18%. Interestingly, the demand expansion almost completely compensates drivers for the additional competition, which can be seen both examining the search-time for taxis as well as their hourly income. Taking the mean over all changes in hourly income, the hourly wage of drivers would be decreased by only 0.5% compared to a reduction of 2.46% when demand is held constant. The number of serviced trips is approximately proportional to the demand increase at 7.73%. The present discounted revenue stream from a medallion would decrease by .86%, which less than 5.04% in the case without an increase in demand. The average hourly standard deviation of cabs in the baseline is 442 and rises to 490. The average hourly standard deviation for passengers is 1766 and rises to 1971. Figure 16 shows the hour by hour changes in the number of cabs, passenger, the consumer surplus, the search time for, the wait time for passenger and the hourly income for drivers. We can see that the increase in supply and demand is relatively proportional to the baseline across hours, which means the absolute change is largest in the afternoon hours and during the evening rush hour. Figure 16 provides the same overview for the counter-factual without demand response.

4.3 Removing Ownership Restrictions

The second counter-factual 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 simulate what would happen if all medallions were operated by minifleets (fleet counter-factual). An actual implementation of this policy could be implemented by simply allowing fleets to purchase individual medallions which currently sell at a substantial discount. The counter-factual is computed by changing the primitives of the fraction of owner-operated medallions to those that we estimated for fleets as shown in Table 3. Note that part of the explanation for the difference of minifleets and owner-operated medallions might simply be that different driver’s select into these different rental agreements or that the owners driving with their own medallions are different from drivers that rent. To the extend that this is true we assume that under this counter-factual, the composition of drivers for all medallions equals the composition that lead to our results for minifleets.Supply in this counter-factual increases by 5.6%, which compares to 1.67% if we did not allow demand to adjust. Demand in this counter-factual would increase by 4.78%, wait time for passengers goes down by 7.02% as compared to a reduction of only 4.01% if demand didn’t adjust. Search time for cabs increases by 8.2% if demand adjusts and 8.43% if demand is held fixed. 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 .13% and the number of trips up by 5.3%. A complete overview is again provided in Figure 18 and Figure 19. A notable feature in the hour by hour
change of supply in Figure 18 is that there is no increase during the witching hour which leads to a relative exacerbation of the decrease in supply during this time. This stands in contrast to the entry counterfactual results and is due to the more regular 5AM to 5PM shift behavior of minifleets.

4.4 Informed Search

We now consider a counter-factual in which we assume that each empty (searching) cab is matched with the closest available passenger. We also assume that both the driver and the passenger are committed to this match: neither can cancel should another match option become available sooner. We wish to emphasize that this matching process is not the most efficient method that could be used because it only allocates empty taxis and does not take into account the possibility that a soon to be empty taxi may be closer to a passenger than a currently empty one. In fact, if we assume that there are fewer passengers, we have computed scenarios in which this counter-factual allocation leads to worse outcome than the baseline, which is impossible if we used the optimal algorithm. We believe that the dispatch scenario that we currently model is interesting in itself. It is also much harder to compute the optimal algorithm, although we hope to find ways to do so in future work. In this counterfactual there is an increase in the number of active taxis (7.11%), and a substantial reduction in the search time for taxis (−9.33%). If we didn’t allow for an expansion of demand the supply increase would be 3.19%. Passenger wait time is reduced by 13.81% in this scenario. Consumer surplus increases by 6.34% and the number of trips increases more than proportional to demand by 10.76%. The present discounted value of medallions would increase by 4.82%. This counter-factual is therefore the only one that benefits all stakeholder. Figure 20 and Figure 21 provide an overview over the hour by hour changes and reveal that this policy also narrows the supply gap during the witching hour.

4.5 Discussion of Counterfactual Results

One of the most commonplace interventions in taxi markets and other industries is a limitation on the number of firms that can enter the market. In our counter-factual calculations we compare a partial change to such a regulation with two alternative policies, one that allows medallions to be bought by more efficient owners and another policy that reduces inherent search frictions in the product market for taxi rides. The results of all three policies highlight the importance to account for a supply response to such policies. More medallions, or more effectively operated medallions, lead to a decrease in wait times for consumers and thereby increase demand. This effect partially offsets the reductions in earnings that drivers have to suffer from an increase in competition in those first two scenarios. A policy that allows medallions to be bought by professional management (as opposed to individuals) leads to almost the same gains in consumer surplus and modest losses in wages for drivers. Such a policy also highlights the importance of intensive margin gains, which is surprising given the scarcity of medallions. It is likely that a new policy lifting the ownership restriction is also easier to implement than a policy that allows for more extensive margin competition as it potentially faces less backlash from existing owners. Individual medallion holders will be able to sell their medallions with capital gain and corporate medallion holders will be compensated for their modest losses in existing medallion revenues through the possibility of

---

25For 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.
such purchases. Note, however, that as can be seen in the counter-factuals, such an ownership change exacerbates the witching hour. Lastly, a counter-factual that gives taxis information about the closest possible match outperforms the previous two policies in terms of gains in consumer surplus. More importantly, under such a policy, both market sides would gain in contrast to the previously discussed policies where drivers and medallion holders suffer modest losses. Medallion prices would increase by almost 5%.

5 Conclusion

This paper estimates a general equilibrium model of the New York Taxi market, which we use to explore imposed frictions in the capital market and inherent matching frictions in the product market. 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 in terms of the waiting time for a cab. To estimate the model we back out unobserved demand where we make use of the fact that the geographical nature of the search process allows us to assume a known process for the matching function. This approach can be regarded as a counterpart to existing empirical work on matching where observed inputs and matches are used to back out a functional form. Counter-factual results from the model show that more efficient utilization of existing medallions can lead to comparable gains as a policy that allows for a substantial number of additional entrants. An improvement in the matching technology leads to large increases in consumers welfare and at the same time increases drivers wages. The model also highlights that it is important to account for the demand expansion effect in such policies.
References


A Details on Simulation

The goal of the simulation is to obtain a function that computes the waiting-time and search-time for each possible combination of waiting passengers and searching cabs within an hour. The function is used to infer the number of waiting passengers using observed search time and observed searching cabs. 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 as well as the size of the area, \( area_t \), on which passengers and cabs are searching for each other. The data tells us for which values in the domain we have to simulate the function. The average number of cabs, the speed of traffic and the requested trip length can be measured at hourly intervals. Table 5 provides an overview.

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

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.

To measure the area we match census-tract shape files to the pickup coordinates. Census tracts are relatively evenly spaced areas that mostly cover only a few blocks. For each hour we then count the total pickup area as the sum of the square-miles of all census-tracts in which at least one pickup happened. Figure 15 shows the long-run average number of pickups for each census tract. Based on the observation that the size of the area on which cabs and passengers match drops off steeply at night we divide the day in two parts (Figure 14), performing the simulation with a night-time area and a daytime area. During the hours 2AM, 3AM and 4AM we simulate on an area corresponding to 10.24 square miles and during the daytime on an area corresponding to 23 square miles. Because a simulation is computationally costly, we have to limit the simulation to a selected set of points in the space of the remaining variables \( d_t, c_t, mph_t \) and \( miles_t \). For the number of cabs we go in steps of 1000 from 1000 to 15000. For the speed of driving we go in steps of 3 \( mph_t \) from 8 to 23. For the number of miles requested from 2 to 5 in steps of 0.3. For the number of passengers, the only unknown variable in the domain, we simulate from 1000 to 60000 in steps of 5000. To obtain the search-time and wait-time for other points in the domain we interpolate linearly.

The baseline simulation is performed under the assumption that cabs search randomly for passengers. The search is performed on an idealized map of Manhattan with the aforementioned sizes for day and night-time. Figure 8 provides a schematic of the grid that we use for the simulation. In line with the

---

\(^{26}\)The actual average night-time area is 11 square miles and the average day-time area is 23.6 square miles. The reason that the areas used in the simulation are smaller is that we also want to satisfy two additional constraints. The first constraint it that the area is that the north-south expansion of the area is four times the east-west expansion. Secondly, to take into account the block-structure the city we want that the number of 1/20 mile segments in each direction is a multiple of four.
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. In north-south direction they can turn at each node whereas in east-west direction they can only turn at every fourth node. Figure 8 highlights notes on which cabs can turn as grey. 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 towards the direction from which they were coming (i.e. no u-turns).

Each node is a possible passenger location. For each hourly simulation $\frac{dt}{5}$ passengers are placed in ten minute intervals randomly on the map with equal probability weight assigned to each node. If a cab hits a node with a passenger a match occurs. In this case the cab is taken off the map for $60 \cdot \frac{\text{miles}}{\text{mph}} \cdot dt$ minutes after which it has delivered the passenger and is again placed randomly on the map with a random travel direction.

In the informed simulation we assume that each cab - as soon as it is available from the last passenger delivery - 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.

**B Additional Figures**
Figure 15: long-run average pickups (census-tract)

![Map showing number of pickups by census tract, with different shades indicating different ranges of pickups.]
Figure 16: Entry counter-factual results as compared to baseline
Figure 17: Entry counter-factual (no demand response) results as compared to baseline.

The graphs display the impact on various metrics such as consumer surplus, hourly income, wait times for both cabs and passengers, and medallion rank distribution over the course of a day (Hour of day). The data is compared between baseline and two different scenarios, showcasing how the absence of demand response affects different aspects of the transportation system.
Figure 18: Ownership counter-factual results as compared to baseline.
Figure 19: Ownership counter-factual (no demand response) results as compared to baseline

- **cabs consumer surplus**
- **hourly income cabs**
- **passenger waittime cabs**
- **waittime passenger**

<table>
<thead>
<tr>
<th>Hour of day</th>
<th>scenario</th>
<th>baseline</th>
<th>fleet_nd</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hour of day</th>
<th>scenario</th>
<th>baseline</th>
<th>fleet_nd</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 20: Search counter-factual results as compared to baseline

- **cabs**
- **consumer surplus**
- **hourly income cabs**
- **passenger waittime cabs**
- **waittime passenger**

Hour of day

**scenario** — baseline — search
Figure 21: Search counter-factual (no demand response) results as compared to baseline

- cabs
- consumer surplus
- hourly income cabs
- passenger
- waittime cabs
- waittime passenger

Hour of day
scenario baseline search_nd
## C Additional Specifications for the Demand Estimation

Table 6: Demand regressions with successively more narrow time controls.

<table>
<thead>
<tr>
<th></th>
<th>(1 IV FS)</th>
<th>(2 IV)</th>
<th>(3 IV FS)</th>
<th>(4 IV)</th>
<th>(5 IV FS)</th>
<th>(6 IV)</th>
<th>(7 IV FS)</th>
<th>(8 IV)</th>
<th>(9 IV FS)</th>
<th>(10 IV)</th>
<th>(11 IV FS)</th>
<th>(12 IV)</th>
</tr>
</thead>
<tbody>
<tr>
<td>log($w_t$)</td>
<td>0.239**</td>
<td>0.243**</td>
<td>0.212**</td>
<td>0.209**</td>
<td>0.196**</td>
<td>0.137**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.137**</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.00693)</td>
<td>(0.00673)</td>
<td>(0.00731)</td>
<td>(0.00673)</td>
<td>(0.0102)</td>
<td>(0.00767)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.00767)</td>
<td></td>
</tr>
<tr>
<td>log($d_t$)</td>
<td>-1.045**</td>
<td>-0.659**</td>
<td>-0.445**</td>
<td>-0.874**</td>
<td>-1.144**</td>
<td>-1.144**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-0.805**</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0449)</td>
<td>(0.0236)</td>
<td>(0.0218)</td>
<td>(0.0317)</td>
<td>(0.0536)</td>
<td>(0.0432)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>714</td>
<td>714</td>
<td>714</td>
<td>714</td>
<td>714</td>
<td>714</td>
<td>714</td>
<td>714</td>
<td>714</td>
<td>714</td>
<td>714</td>
<td>714</td>
</tr>
<tr>
<td>2-Hour FE</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>3-Hour FE</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>4-Hour FE</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>6-Hour FE</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>8-Hour FE</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.288</td>
<td>0.320</td>
<td>0.402</td>
<td>0.420</td>
<td>0.787</td>
<td>0.466</td>
<td>0.546</td>
<td>0.479</td>
<td>0.593</td>
<td>0.648</td>
<td>0.794</td>
<td></td>
</tr>
</tbody>
</table>

**Note:** $+ p < 0.10$, $* p < 0.05$, $** p < 0.01$, All regressions are based on the October and November 2011 trip sheet data. An observation is comprised of an hourly average over all trips in that hour. Standard errors clustered at the date level.