Frictions in a Competitive, Regulated Market
Evidence from Taxis

Guillaume Frechette (NYU) Alessandro Lizzeri (NYU)

Tobias Salz (Columbia University) *

May 6, 2018

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 are: first, the matching frictions created by the need for both market sides to physically search for trading partners, second, the regulatory limitations on entry in the capital market. To assess the importance of these features we use the model to simulate the effect of changes in entry, alternative search technologies, and different market density. We use the geographical features of the matching process to back out unobserved demand through a matching simulation. We find that the matching function exhibits increasing returns to scale.

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

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

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

Taxi services offer limited room for product differentiation and markups. Moreover, a firm in this market is of relatively low organizational complexity. In its simplest form it consists of a unit of capital (a car) plus the labor (a driver) needed to operate it. The labor skill requirements are relatively modest and the capital is in vast supply. Finally, in a city like NY, taxi drivers take many decisions independently, with little real-time information about aggregate conditions. Thus, absent regulatory interventions, this market has several of the features of textbook examples of a “perfectly competitive” industry with many firms making decisions independently, jointly affecting aggregate market conditions. It therefore presents an interesting case study of an important benchmark. However, even absent regulatory restrictions there are inherent matching frictions that create barriers to trade. Furthermore, most taxi markets are subject to licensing restrictions that restrict entry and to regulations of fares.

We first document some important patterns in this market that will then motivate some of the key ingredients of our model. We first provide evidence that entry restrictions are strongly binding in NYC. If there were no other frictions, one might therefore expect all taxis to be utilized at least during the day-time. However, activity is often well below capacity, highlighting the importance of understanding the labor supply decisions of taxi drivers. Labor supply cannot instantaneously adjust to market conditions because drivers operate on a two-shift system, leading to indivisibilities in labor supply.

Because of regulations, there is no price flexibility in this market. This, together with capacity limits, means that regular patterns of variation in demand for rides during the day (e.g., rush hours) lead to large variations in costly delays for matches between passengers and taxis. Drivers’ earnings and the number of active taxis vary during the day depending on how long drivers need to spend searching for their passengers. The average search time for an active taxi between dropping off a passenger and picking up the next one ranges between 5 and 20 minutes depending on the time of day. In order to appreciate the magnitude of these numbers, note that the average duration of a trip is twelve minutes. We
report that the fraction of time taxis are “unemployed” ranges between 30% and 70% depending on the hour of day. Passengers also wait to obtain a taxi and this wait time varies during the day. Absent price adjustments, passenger wait times and taxi search times serve as (wasteful) market clearing variables.

In our model, drivers make daily entry and hourly stopping decisions. Medallions are scarce, so entry is only possible for inactive medallions. Hourly profits are determined by the number of matches between searching taxis and waiting passengers. Ceteris paribus, increasing the number of taxis increases the search time for a driver to encounter the next passenger and reduces expected hourly earnings. The number of taxis is determined endogenously as part of the competitive equilibrium in this market. Stopping (exit) decisions are determined by comparing a random terminal outside option with continuation values determined by expected hourly earnings net of a marginal cost of driving that is allowed to depend on how long the driver has been driving. Starting (entry) decisions result from a comparison between an outside option and the expected value of a shift (given expected optimal stopping behavior).

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

Prior papers have used the taxi market as a useful environment to study labor supply decisions since drivers have more flexibility in deciding when to stop than workers employed in firms. However, we show that there are important rigidities tied to the two-shift structure of operation. Our model delivers responses to earnings shocks that strongly depend on their timing and persistence. Specifically, we contrast the response to a uniform shock throughout the day to an hour specific shock for each hour. The former results in an elasticity estimate of 1.3 which, interestingly, is remarkably close to the number (1.2) reported by Angrist et al. (2017) from an experiment on Uber data. The latter results in estimates that vary from 0.7 to 1.5 depending on the hour, with beginning and end of shift elasticities that are lower than those for the middle of the shift.

On the demand-side, we face a challenge. While we observe the number of matches, neither the passengers’ waiting time nor the number of hailing passengers is directly observable in the data. However, we are able to recover these by using information about other observables as well as the nature of the matching process. Our first step is to obtain the matching function that maps the number of taxis and passengers (in addition to other observables such as traffic speed) into

---

1 This form of rationing is common in other markets.
2 These identifiers are no longer available for more recent data from NYC.
the number of matches as well as values for search time for taxis and wait time for passengers. We develop an explicit description of the geographical nature of the matching process, and we then recover, via simulation, a numerical representation of the matching function. We can then invert the function to recover how many people must have been waiting for a cab given the number of passenger pickups that we observe (successful matches), how long taxis search for passengers, the number of cabs on the street, and the speed at which traffic flows; all of these are variables that we observe in our data. The empirical literature on search and matching typically uses known inputs such as the number of job vacancies and unemployed workers as well as the observed number of matches to estimate parameters of an assumed functional form (e.g., Cobb-Douglas) for the matching function.\footnote{See Petrongolo and Pissarides (2001)} We proceed in the other direction by using a specific matching process that defines a matching function, and we then use the observed matches and the number of active taxis to infer the other key input to the matching function, i.e., the number of waiting passengers. Interestingly, the matching process displays increasing returns. These are very modest for the range of values observed during the daytime but are more substantial for the smaller scale during the nighttime. This feature of the matching process has potentially important welfare consequences.

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

Our first counterfactual evaluates the effects of additional entry. An increase in the number of medallions of 10 percent leads to an increase of 7.5 percent in the number of active taxis.\footnote{There is some difference in the extent of the increase in activity at different hours of the day. This number is an average across the day for a typical weekday.} The reason for the less than proportionate increase relative to the additional entry is that drivers respond to reduced earnings by choosing shorter shifts, highlighting the importance of modeling the intensive margin on the supply side. However, the increase in activity would be a lot smaller if we did not incorporate in the model the dependence of passenger demand on expected wait times. The increase in the number of taxis leads to a reduction in wait time, which leads to an increase in the number of passengers, which in turn moderates the reduction in earnings caused by the increase in the number of medallions.

Our next set of counterfactuals concerns the magnitude of matching frictions and possible ways to ameliorate them. Specifically, we consider an improved matching technology in line with the dispatch system of the Uber platform and other ride-hailing services. We first consider a polar opposite of the NYC decentralized decision making model by introducing a centralized dispatcher for the
entire fleet. We show relatively large gains for both sides of the market due to reductions in wait times for both passengers and taxis. Interestingly, the number of active taxis increases by almost the same amount as in the counterfactual with 10 percent more medallions, despite the fact that the number of medallions is left unchanged. The number of matches increases by 11%, a larger amount than the increase in the number of taxis. The difference is due to the fact that matching frictions are reduced by the dispatch system.

We then consider what happens in the more realistic case in which the dispatch platform only achieves partial market penetration, with the remainder of the market functioning according to the traditional street-hailing system. We show that market segmentation on different platforms creates an inefficiency that is due to a reduction in market thickness for both platforms. Partial coverage by a dispatcher has two effects compared with a market with no dispatcher: on the one hand, there is improvement of matching in the covered market; on the other hand, there is a segmentation of the market that makes both segments thinner, with the consequence of longer average distances between a random taxi and a random passenger. When we consider the case of a 50-50 split between dispatch and decentralized platforms, we find that the second effect dominates, and therefore, aggregate outcomes become worse than in the baseline case. Interestingly, the effects are quite different during the daytime relative to night-time hours, reflecting the importance of the initial thickness of the baseline market environment.

Finally, we consider the effects of density: we simulate a city that is otherwise identical to Manhattan, but is 1/3 as dense, i.e., the same number of potential passengers is spread over a larger territory covered by the same number of potential taxis. We find that our model predicts dramatic losses in efficiency due to lower density. However, these inefficiencies are substantially alleviated by a dispatch platform whose performance is (comparatively) much better in a less dense environment.

2 Related Literature

This project combines elements from the entry/exit literature, neoclassical labor supply models and search. Structural estimation of entry and exit models goes back to Bresnahan and Reiss (1991). This entry/exit perspective on the problem is motivated by the fact that drivers in the New York Taxi industry, like in many other cities, are private independent contractors and decide freely when to work subject to the regulatory constraints. The labor supply decisions of private contractors have for example been studied in Oettinger (1999), who uses data

\(^5\)Uber seems to be well aware of this effect and therefore subsidizes drivers, especially when they first enter a city.
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. Brancaccio et al. (2017), considers the related application of exporters matching to ships. It develops and estimates a model features geography, search frictions, and forward-looking optimizing ships and exporters.

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.

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

Some earlier papers have used NYC trip sheet taxi data to investigate individual labor supply decisions. Camerer et al. (1997) find a sizable negative elasticity of daily labor supply and they argue that this is inconsistent with neoclassical

\[^6\] Lagos (2003) does not have data on hourly or daily decisions by taxi drivers.
labor supply analysis. This interpretation has been challenged by Farber (2008). Crawford and Meng (2011) estimate a structural model of the stopping decision by a taxi driver allowing for a more sophisticated version of reference-dependent preferences. They do not consider the entry decision by a cab driver and do not analyze the industry equilibrium. We opted to stay within the neoclassical framework to study the general equilibrium of the taxi market, in contrast to these papers that all focus on the intensive margin of daily individual labor supply decisions. We note that although it may very well be that some drivers do not fit this assumption, the aggregate patterns are consistent with a standard model, and it fits the data quite well. Hence, a standard model of labor supply seems to be a reasonable starting place.\(^7\) 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.\(^8\)

\section{Industry Details and Data}

\subsection{Industry Details}

Operating a yellow cab in NYC requires ownership of a medallion. In the time period covered by the data only yellow cabs are allowed to pick up street-hailing passengers.\(^9\) This differentiates cabs 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.\(^10\) The cab market is regulated by New York’s Taxi and Limousine Commission (TLC), which sets rules such as the fare that drivers can charge, the qualifications for a taxi-driver license, the insurance and maintenance

\(^7\)Note also that our models fits aggregate patterns relatively well, hence any gains from allowing for a richer labor supply decision would be small in aggregate.

\(^8\)Other papers that use this data are less directly related. For instance, Haggag and Paci (2014) that study the impact of suggested tips in the NYC taxi driver payment screen for clients on the realised tip. Haggag et al. (2014) who study how taxi drivers learn driving strategies based on their experiences. Buchholz et al. (2016) argue that both “behavioral” and “neoclassical” wage responses are present in the data, with the behavioral income-targeting story explaining shorter shifts, and the standard neoclassical model explaining longer shifts. Methodologically, their paper uses a different estimation procedure from ours: they develop a new closed-form estimator for the stopping problem. Thakral and Tô (2017) find reductions in cabdriver labor supply in response to higher accumulated daily earnings and stronger effects for more recent earnings. They argue that the income effect is inconsistent with the neoclassical model and the non-fungibility of daily income rejects models invoking daily income targets. Finally, Jackson and Schneider (2011) find evidence of moral hazard in the behavior of taxi drivers and document that this problem is moderated if drivers lease from fleets owned by someone in their social network.

\(^9\)In the time period that we consider, Uber was not yet a significant presence.

\(^10\)The TLC recently established a possibility to hail cabs via a mobile app, but this was not possible during our observation period.
requirements, and restrictions on the leasing rates (daily or weekly) that medallion owners can charge drivers.

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

The TLC imposes several restrictions on the terms of the leases between medallion owners and drivers. Leases can either be for a shift or an entire week. A rental for a shift has to last twelve consecutive hours and a weekly lease seven consecutive days. Minifleets must operate their cabs for a minimum of two nine hour shift per day every day of the week. 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. The fixed fare for transporting a passenger is $2.50 and the fare for an additional unit is $0.4.

### 3.2 Data

Our main data source is the TLC’s Taxicab Passenger Enhancements Project (TPEP) which creates an electronic record of every yellow cab trip. For each trip it records a unique identifier for the driver as well as the medallion; the length, distance, and duration of the trip, the fare and any surcharges, and the geo-spatial start and endpoint of the trip. The TPEP data can be obtained from the TLC. In this project we only use a subset of the data from 2011 and 2012, which ranges from

---

11 Some fleet owners operate more than one thousand taxis. Source: [http://www.nycitycab.com/Services/AgentsandFleets.aspx](http://www.nycitycab.com/Services/AgentsandFleets.aspx)

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


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

15 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 boroughs and trips from La Guardia are still governed by a variable fare.
the October 1st, 2011 to November 22nd, 2011; and August 1st, 2012 to September 30th, 2012. The data from 2012 encompasses the time at which the unit charge was increased from $0.4 to $0.5. During the time spanned by our data we see the universe of 13,520 medallions. We also observe all 37,406 licensed drivers that have been active in that period. We complement this data with information about the medallion type (minifleet or owner-operated), and the vehicle type.

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

4 Descriptive Evidence

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

4.1 Entry Restrictions

As we mentioned in the introduction, there are tight entry restrictions in most taxi markets. In NYC, the number of medallions during the time-period of our data is 13,520. This is an absolute limit on the number of possible taxis on the street at any moment in time. Prices of medallions are an indicator of the quantitative importance of the entry restriction.\(^{16}\) During the period we consider these prices exceeded half a million dollars. Of course, such medallions would not be valuable in a market with no entry restriction. We have also verified that the percentage of medallions that are driven at least once a day is close to 97% during weekdays, and 92% even on Sunday. Given that there is some natural failure rate of vehicles and other idiosyncratic reasons why taxis may fail to be utilized, this seems to be a very intense rate of utilization.\(^{17}\)

4.2 Daily Patterns of Activity

Figure 1 displays the fraction of taxis that are active at each hour of the day for a typical (Monday-Thursday) weekday, distinguishing whether it is a minifleet or

---

\(^{16}\)Medallions are often traded in auctions and prices are public data.

\(^{17}\)As we will see, this is quite different from the utilization rate for a typical hour, which is much lower and also depends on drivers’ hourly stopping behavior.
an owner-operated medallion.  

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

We wish to draw attention to several features of this figure. First, fleet medallions are more intensely utilized for every hour of the day.\textsuperscript{19} To place this in perspective the difference in utilization is larger than the difference between Uber and taxis reported by Cramer and Krueger (2016). Second, there is substantial intra-day variation in activity, but this variation does not seem to fully reflect expected patterns of intra-day variation of demand, despite the fact that activity is well below capacity for the entire day.\textsuperscript{20} Third, there is a large reduction in activity precisely during the evening rush hour. This is known as the witching hour. The drop in activity is more pronounced for fleet medallions.

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

4.3 Search Frictions

An important friction in this market relative to the ideal of a Walrasian market arises from the fact that drivers and passengers have to physically search for trading partners. We now provide some indirect evidence that matching frictions are important in this market. Figure 2 describes the fraction of time that an average taxi spends searching for passengers relative to the total time it is active, i.e., the unemployment rate for taxis.

\textsuperscript{18}One owner cannot manage multiple owner-operated medallions. However, he can lease the taxi to another driver for one of the shifts.

\textsuperscript{19}In subsection B.5 we discuss additional facts on the difference between fleet medallions and owner-operated medallions.

\textsuperscript{20}Below we provide evidence that activity is indeed less variable than demand.
Figure 2: Search-time Relative to Delivery Time During the Day

Notes: This plot shows the ratio of time a taxi spends searching for a passenger as a percentage of total time spent driving. For the plot we take the average over those ratios using data from Monday to 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.

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

Notes: The left panel shows search time for taxis in minutes, averaged for each hour of the day. The right panel shows waiting time for passengers as recovered by our simulation, again averaged for each hour of the day. Only data from Monday to Thursday is considered.
Two notable features are the following: First, the fraction of time taxis spend searching is almost never lower than thirty percent and display substantial variation during the day, going as high as 65% at its peak. It is important to note that, under the current system of fixed fares, most inter-temporal variation in driver profits and customer welfare are created by variation in delays in finding a partner. Indeed, a simple linear model reveals that variation in taxi search time explains about 60% of the variation in drivers’ hourly wages. The low point of search time is reached at 5PM shortly after the shift change when many medallions are still inactive. One interesting observation from Figure 3 in conjunction with Figure 2 is that search time increases during the night although the ratio of passengers to taxis is relatively stable. For passengers the wait time increases relative to the evening hours but is still lower than during the day time (the wait time for passengers is inferred from our simulation, which is described in detail later). This illustrates an “economy of density” that implies that both market sides benefit from the fact that density facilitates the matching process.

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

Figure 4: Matching function comparison

![Figure 4: Matching function comparison](image_url)

21Most of the remaining variation is explained by trip-length and the rate that is charged per minute of driving. This rate varies with the speed of traffic due to the mixture of time-based and distance-based metering.
from examining the degree of predictability of the order with which empty taxis find passengers in a given area: absent search frictions, one would expect the pick-up order of taxis in a specific area to be highly correlated to their drop-off order. In other words, focusing on a narrow enough area with uniform demand conditions, the $i$'th cab to drop off a passenger in this area should also be the $i$'th one to subsequently pick up the next passenger. We performed this exercise and, while we find that the drop-off sequence and the pick-up sequence are positively correlated, there is still a large amount of dispersion. We interpret this dispersion as evidence of search frictions. Results from this exercise are displayed in Figure 16 in Appendix A.

5 Model and Estimation

5.1 Demand Side Model

In the estimation we focus on trips that originate in Manhattan which accounts for most of the activity in the data (approximately 90% of trips). Since the NYC taxi market operates under a fixed fare system, the endogenous variables that adjust to clear the market are the wait time $w_t$ for passengers to find a taxi, and the search time $s_t$ for taxis to find a passenger.

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

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

Let $i \in \{1, ..., I\}$ be the index of an area in the city. We denote the total number of waiting passengers in the entire city by $d_t$. These are then divided up across

\[22\] Most of the remaining trips originate at one of the two NYC airports.
areas of the city according to proportions \( \{ p_i^d | i = 1, ..., I \} \). We denote the vector of waiting passengers \( d_t = (d_1 \cdot p_1^d, ..., d_I \cdot p_I^d) \). The matching process is captured by a function \( g \) that maps a vector of waiting passengers \( d_t \), and searching taxis \( c_t = (c_1 t, ..., c_I t) \) as well as other exogenous time varying variables \( \phi_t \) into an aggregate search time \( s_t \) and wait time \( w_t \):

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

If we knew \( g(\cdot) \), then, for given values of \( c_t \) and \( \phi_t \), inverting this function would allow us to infer \( d_t \), assuming, of course, that the function is invertible. We use our knowledge about the geographical nature of the matching process to infer the form of \( g(\cdot) \). In particular, we obtain an approximation of \( g \) by simulating the matching process of waiting passengers and searching taxis on a grid that represents an idealized version of the Manhattan street grid.

### 5.1.1 Implementing the Simulation

In the simulation, we assume that passengers wait at fixed locations on a two-dimensional grid. The map consists of nodes whose spacing is proportional to \( 1/20^{th} \) of a mile. Each of these nodes is a spot at which passengers potentially wait. Street blocks are assumed to be \( 4/20^{th} \) of a mile wide (east-west) by \( 1/20^{th} \) of a mile long (north-south), which corresponds to the approximate block size in Manhattan. Figure 5 shows the structure of the resulting grid. Gray nodes represent intersections between streets and avenues, at which cabs can change direction of travel. Turns at (gray) nodes are random with equal probability for each feasible travel direction.

We assume that the effect of time varying factors can be summarized by the speed \( mph_t \) at which the traffic flows and the average distance \( miles_t \) to deliver a passenger from his fixed position on the grid to the destination. Both factors are included in \( \phi_t \), 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. For each combination of variables that we feed into \( g \), we simulate the resulting average waiting time for passengers and search time for taxis over an hour long time interval. Every ten minutes \( d/6 \) potential passengers are born and placed on the map for a total of \( d \) passengers during the hour. We assume that the maximal waiting time is 20 minutes. After 20 minutes, a passenger stops waiting and counts as unmatched.

To account for the fact that the number of trips originating from different parts

---

23 These proportions are inferred from the data, see 5.1.1 below for more details.
24 Our results are robust to changing the maximal value of the waiting time. See the WebAppendix for more details.
of the city varies, we divide Manhattan in eight equally spaced areas (see Figure 17 in Appendix B for details). Passengers appear on the corresponding parts of the grid in proportion to the observed pick-up probabilities of those areas and cabs reappear according to the observed drop-off probabilities. In other words, for each of the passengers placed on the map, we first randomly determine an area according to a multinomial distribution with probabilities \( \{ \hat{p}_d^i | i = 1, ..., 8 \} \) and then a node within an area, where each node has equal probability. The probabilities \( \{ \hat{p}_d^i | i = 1, ..., 8 \} \) are estimated as the fractions of trips originating in these areas. Similarly, cabs re-appear on the map in area \( i \) according to multinomial probabilities \( \{ \hat{p}_c^i | i = 1, ..., 8 \} \) and with equal probability on each node within an area. The probabilities \( \{ \hat{p}_c^i | i = 1, ..., 8 \} \) are measured as the frequencies of drop-offs in those areas. In subsection B.1 we discuss the robustness of our inversion results to alternative (finer) divisions of the map. Note that, while our approach allows us to account for some of the heterogeneity in locations, we do not consider endogenous location choices as in Buchholz (2015).

Note that, with the exception of the division into areas according to multinomial probabilities, none of the steps so far required the use of observed data. Our simulation simply generates an approximation of \( g(\cdot) \) for any point in its do-

---

25To explore the robustness of the results with respect to the fineness of the division into areas, we have also performed the simulation for 16 areas and results are almost identical. We report these in Appendix B.

26We therefore do not consider conditional drop-off and pick-up probabilities. These would only be of interest if we wished to follow individual drivers. For the purpose of this simulation, however, driver identities are irrelevant and only aggregates matter.
As long as the simulation approximates the true matching process closely enough, we can use $\hat{g}$ to back out $d_t$ for each combination of hourly averages of search time $s_t$, number of taxis $c_t$, traffic speed, and trip distance observed in the data. Once the number of passengers is known we can insert it into $\hat{g}$ to determine wait-time $w_t$ as well. Additional details on the simulation, including the algorithm, are provided in appendix Appendix B.

5.1.2 Properties of the Matching Function

Figure 6: Graphical Illustration of Matching Function

Figure 6 graphically illustrates properties of the matching function. The figure on the left shows what happens to wait time and search time as the market becomes thicker. On the horizontal axis the number of taxis and passengers vary while keeping constant the ratio between the two. Both passenger wait-time and  

\footnote{Due to computational limitations, it is not feasible to repeat this simulation for each point in the domain of $g$. For example, if we assume that in an hour there are at most 70,000 passengers waiting, and multiply this by the maximal number of medallions, we would already obtain 945 million different points in the domain. This is prior to considering variation in $\phi$. We therefore simulate $g$ for a lower number of grid points and we interpolate linearly between those points to obtain the image for points in between. For each of the four independent variables of $g(\cdot)$, we pick eight different evenly spaced grid points. Because the outcome of the matching process is random, we have to repeat the simulation multiple times for each of those points. In practice we have found that the average of these simulations does not change much after ten iterations, which is therefore what we use to produce this average.}
search-time decrease as the market becomes thicker, but the improvements decrease relatively quickly. For the numbers observed during the late hours of the night, the returns to additional thickness are substantial. In contrast, at the scale corresponding to the average number of daily taxis in the data (approximately 7800 when averaging over the entire 24 hours), additional returns to market thickness are fairly small, especially regarding search time. The increasing returns nature of the matching function has important consequences for the welfare implications of a variety of policies. For instance, as we discuss in our counterfactuals, the segmentation of the market into multiple platforms such as Uber and Yellow taxis is harmful.

The figures on the right give a sense of how market tightness affects outcomes for both market sides. On the top right panel, the number of passengers is increased while holding fixed the number of taxis at the median observed in the data. In the bottom right panel, the number of taxis varies, holding fixed the number of passengers at the median. Both figures also display level changes due to differences in traffic speed, one of the exogenous inputs to the matching function. The dotted line is obtained under a traffic speed that is one standard deviation below the median and the solid line is for traffic speed that is one standard deviation above the median. While faster traffic speed is unequivocally good for passengers, there are two contrasting effects for cabs. On the one hand, faster traffic speed allows an individual cab to more quickly reach a waiting passenger, thereby reducing search time. On the other hand, faster traffic speed speeds up aggregate deliveries of passengers, and therefore leads to an increase in competition, effectively increasing the number of available cabs. At the median number of observed passengers, the latter effect outweighs the former at about 4000 cabs. We will make use of this effect in our demand instrument.

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

In order to provide additional validation for the results of our demand inversion, we considered the effect of weather shocks which are one of the important demand shifters in this market. subsection B.4 shows that our inferred wait-times and demand are both highly correlated to rainfall shocks.
Figure 7: Graphical Results of Demand and Wait Time

5.1.3 Estimating the Demand Function

We now proceed to estimate a demand function that relates the number of passengers to wait time. We assume a constant elasticity demand function of the following form:\(^{28}\)

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

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

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

\(^{28}\)Since we break up the data at the hourly levels and only use a subset of weekdays, we are left with slightly more than 700 observations for this estimation. The assumption of log-linearity is not unusual for this type of problem, see for example Kalouptsidi (2014).
A potential problem is that the wait time itself is a function of the number of passengers as well as the number of cab drivers. In particular, unobserved factors that shift demand will directly affect \( w_t \). Furthermore, drivers may condition their decisions on factors included in the error term \( \xi_t \), which would lead to a decrease in waiting time. For both of these reasons, the error term \( \xi_t \) and the wait time \( w_t \) may be correlated. This would, of course, introduce a bias in the estimation of \( \eta \). To address this concern we instrument for wait time. We need a variable that is correlated with wait time and that affects demand only through wait time. A supply shifter satisfies this requirement. We present results for three different instruments, all of which lead to similar elasticities with respect to wait time.

Table 1: Different Specifications for Demand Estimation.

<table>
<thead>
<tr>
<th></th>
<th>( \log(d_t) )</th>
<th>( \log(d_t) )</th>
<th>( \log(d_t) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \log(w_t) )</td>
<td>-0.949**</td>
<td>-1.078**</td>
<td>-0.670**</td>
</tr>
<tr>
<td></td>
<td>(0.174)</td>
<td>(0.366)</td>
<td>(0.0629)</td>
</tr>
<tr>
<td>Observations</td>
<td>714</td>
<td>714</td>
<td>714</td>
</tr>
<tr>
<td>Hour FE</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>2-Hour FE</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>( R^2 )</td>
<td>0.932</td>
<td>0.921</td>
<td>0.934</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. All regressions include log of traffic speed as a control. An observation is comprised of an hourly average over all trips in that hour. Standard errors clustered at the date level.

The first instrument uses the shift change as a supply shifter. Around 4 o’clock, there is a reduction in the number of active taxis by approximately one third. This is known as the “witching hour” in NYC: a time when it is very hard to find a taxi. In Appendix D we argue that the shift change is timed so that the day shift and the night shift have similar returns on average. This ensures that both shifts are attractive to potential drivers.\(^{29}\) Thus, the shift change is anticipated and is not chosen with regard to the details of intertemporal variation in demand.

\(^{29}\)Because fleets face caps on the lease rates that they can charge to drivers, that excess supply of drivers in one shift does not translate in higher ability to charge more for leasing a taxi during that shift.
The second specification makes use of shift indivisibilities. As we argue in detail in our supply side model, hourly supply is determined by drivers’ longer term decisions involving starting and ending a shift. Early starting and stopping decisions therefore respond to unobserved shocks that are different from later ones. Supply decisions that respond to these shocks will nevertheless shift supply in later hours. For example, drivers who have made decisions to start because early hours are profitable have sunk the entry cost and will tend to remain on the shift later even if subsequent hours are not particularly profitable.

Lastly, we instrument for the wait time with the requested trip distance outside of Manhattan. This instrument uses the fact that the requested trip distance determines the effective supply of taxis within Manhattan. To the extent that trip distance if uncorrelated to unobservables that shift the extensive margin of demand it serves as a valid instrument for the wait time.

In our view, among these three specifications, the shift change instrument presents the most compelling case for satisfying the exclusion restriction. However, it has the disadvantage of being relatively coarse, and that it does not allow us to include hourly fixed effects. We therefore use the second specification to compute counterfactuals (note that the supply side models is estimated against equilibrium earnings and does not depend on these demand specifications). The fact that all three specifications lead to very similar elasticity estimates is reassuring.

Our demand analysis does not incorporate an explicit model of the consumer choice problem among transportation options. Because this is not the main focus of the paper, we model substitution to other modes of transport and to the outside option via a reduced form demand approach in which these are captured by varying demand intercepts of the log linear demand function. We have, however, made an effort to validate our demand estimates with data on subway rides, which is the other major form of transportation in Manhattan. Because the data on subway rides is only measured at four hours intervals, we aggregate data at a daily level. For the validation we first compute the expected number of rides conditional on weekday, month, and year fixed effects. We then include residuals from this regressions as controls in our specifications. This has no discernable effect on our results. Alternatively, the left tail of the residual distribution from this regression can be interpreted as a short-fall in subway rides, possibly due to delays or break-downs. Conditioning our demand estimation on the lowest 50% of residuals we find that, as expected, demand becomes slightly less elastic, but that the average elasticity over the two events changes very little.30

30 Conditioning on the lowest 50% of residuals reduces elasticities from -0.67 to -0.62 for the shift instrument, -0.94 to -0.81 for the indivisibilities instrument, and -1.1 to -0.87 for the miles instrument. The average elasticity from both subsets (above the median residual value and below the median residual value) deviate only slightly from our estimates above, from -0.67 to -0.665 for the shift instrument, -0.94 to -0.85 for the indivisibilities instrument, and -1.1 to -1.266 for the miles instrument. For the purpose of our counterfactuals we are interested in the average elasticity in
Figure 8 shows what demand would be if the wait time was held fixed at the daily mean throughout the day. This highlights how the inelastic portion of demand is varying throughout the day. In other words, it is a representation of the hourly demand shocks. All else being equal, demand would be highest in the evening hours between 5PM and 7PM and lowest in the morning hours from 2AM to 7AM.

5.2 Supply Side Model

In modeling the driver’s problem, we aim to incorporate the regulatory and organizational constraints imposed by the medallion system that we discussed in section 3.

Agents make their decisions in discrete hourly intervals; the index $t$ denotes an hour from the set $H = \{1, \ldots, 24\}$. We interpret these as representative hours of a weekday. At each hour $t$ there are $N_t$ active and $M_t$ inactive medallions, which sum up to the total number (13,520) of medallions issued by the city. Each hour an empty medallion is probabilistically filled with a driver, depending on the outside option of drivers and their entry decisions. Active drivers make an hourly stopping decision. As mentioned above, we focus on Monday-Thursday.

In section 3 and section 4 we highlighted the fact that minifleet medallions are more heavily utilized than owner-operated medallions, and are also more likely to transition between shifts at 5PM, and therefore over-proportionally contribute to the supply shortage at that time of the day. In addition to medallion-type heterogeneity, we also allow for heterogeneity in the type of shift experience observed for different medallions. We capture this by allowing for different times of shift-transition. For a driver $i$ the index $z_i \in \{\text{Fleet, Own}\}$ denotes the ownership of the medallion and takes on those values depending on whether it is a minifleet or an owner-operated medallion. The model parameters are allowed to differ for minifleet and owner-operated medallions. The index $k_i \in K$ captures the time at a stationary market environment, which appears not too be affected much by the omission of other transportation options.
which the medallion transitions between day-shift and night shift. We allow for four possible transitions in the morning and four in the evening. Drivers have to pay fines if they return the car after the designated end of their shift. The transition type determines when those fines need to be paid. If, for example, the most common transition time in the morning (as measured by the mode) is 5PM, we assume that a driver has to pay a fine for handing in the medallion later than this.

We first discuss the optimal stopping decision for a driver who is already on a shift, then we discuss the decision to start a shift. The continuation value for a driver conditional on state \((x_{it}, \pi_t, \epsilon_{it})\) is given by:

\[
V(x_{it}, \epsilon_{it}) = \max\{\epsilon_{i0}, \pi_{it} - C_{z_i, h_t}(l_{it}) - f(h_t, k_i) + \epsilon_{i1} + \mathbb{E}_{\epsilon_{i(t+1)}}[V(x_{i(t+1)}, \epsilon_{i(t+1)})]\}
\]

At the beginning of each hour the driver decides whether she wants to collect the flow payoff from driving plus the continuation value of an active shift, or the random value of the outside option. This expression depends on an observable state vector \(x_{it} = (h_t, l_{it}, z_i, k_i)\), hourly earnings \(\pi_t\), as well as an idiosyncratic unobservable vector \(\epsilon_{it} = (\epsilon_{i0}, \epsilon_{i1})\), assumed to be distributed according to i.i.d. T1EV distributions with scale parameter \(\sigma_\epsilon\). The cost of driving \(C_{z_i, h_t}(l_{it})\) is a function of the total shift time \(l_{it}\) worked by driver \(i\) at time \(t\). It allows for the cost of driving to rise as the shift gets longer. The parameters of this function are indexed both by the medallion type \(z_i\) and by the hour of the day \(h_t\). We interpret this cost function as a combination of the hourly opportunity cost of driving, which may vary throughout the day, as well as the disutility of driving. We assume that the cost function takes the following form:

\[
C_{z_i, h_t}(l_{it}) = \lambda_{0, z_i, h_t} + \lambda_{1, z_i} \cdot l_{it} + \lambda_{2, z_i} \cdot l_{it}^2
\]

While the fixed cost components can depend on the hour, we only allow for two different values of \(\lambda_0\) for each 12-hour shift.

The term \(f(h_t, k_i)\) is a fine that has to be paid in the event that the driver

---

\[31\] Assuming such fixed times for fines partially endogenizes transition time and gives the model some flexibility in matching patterns where drivers drive longer than the typical pattern. However, transition times are in principle endogenous and one could imagine that drivers take into consideration daily variation in the value of day-shifts and night-shifts to time the transition in a negotiation process. Such a fully endogenous model is currently outside the scope of this paper. However, as we argue in Appendix D, the current regulatory arrangements limit the flexibility of transitions.

\[32\] We assume that there is no discounting since these are hourly decisions. In the estimation, the maximal shift length that we allow is 13 hours. This is longer than the regulatory maximum of 12 hours and it is surpassed only in a very small number of cases.

\[33\] We have also explored a specification with higher order polynomials but this does not make a difference in the results.
returns the taxi after hour $h_t$. Such fines are very common to insure that drivers do not operate the cab longer than contractually specified. Since fines are not directly observable we estimate them as parameters. We use the fact that we can observe medallions operate over a long period of time to classify them into categories $k_i$ of morning and evening transition times. For each medallion we compute the most common starting hour (the mode) in the morning as well as in the evening. We then assume that a morning driver has to pay a fine $f_{z_i}$ whenever he goes past the common night shift starting time and analogously for a night driver. The fine is again indexed by $z_i$ to account for the fact that owner-operated medallions seem to have less stringent transition times. We denote by $h_{k_i}$ the set of hours for which a driver of medallion $k_i$ is subject to the fine when driving. We therefore have: $f_{z_i}(h_t, k_i) = f_{z_i} \cdot 1 \{h_t \in h_{k_i}\}$.

We now discuss the decision of whether to start a shift. For each hour $t$ during which a medallion is inactive, a driver $i$ has an opportunity to be matched with this medallion. He decides to enter if the value of driving given optimal stopping behavior is higher than his outside option. If he decides not to enter, then the medallion will be available to another potential driver the following hour. We assume that the utility from the outside option is comprised of a fixed value $\mu_z$ that depends on the medallion type and on whether it is a day or night shift, as well as an idiosyncratic random component $v_{it0}$. The utility of driving depends on the entire expected value of a shift (described below) as well as an idiosyncratic component $v_{it1}$. 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 $\sigma_v$. Drivers also have to pay a daily rental fee $r$ that depends on whether they drive during the day shift or the night shift. If they own the medallion they have an opportunity cost of driving equal to $r$. We set $r_{h_t}$ equal to the rate caps, which, according to anecdotal evidence, were always binding during the data period. The deterministic part of the state vector is $x_{it} = (h_t, l_{it}, z_i, k_i)$. Note that $h_t$ as well as $l_{it}$ evolve deterministically and that $z_i$ and $k_i$ do not change over time.

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

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

and the utility of starting a shift is given by:

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

---

34Of course, all that matters for drivers’ choices is the difference between $v_{it0}$ and $v_{it1}$.

35The scale parameter $\sigma_v$ is identified because $\mathbb{E}_{e_i}V(x_{t+1})$ is a given value from the stopping problem and not pre-multiplied by any parameter.

36Recall that the sample period precedes the time when UBER entry became substantial in NYC. By 2015 it is likely that rental rate caps were no longer binding. In our discussion of counterfactuals we discuss how the endogenous determination of lease rates may change our results.
As is well known, a convenient feature of assuming a $T1EV$ distribution for the error terms is the closed form expression for the choice probabilities. Denoting by $q(x_{it})$ the probability that an inactive driver starts a shift at time $t$, conditional on state $x_{it}$, we obtain:

$$q(x_{it}) = \frac{\exp((\mathbb{E}_{it}^{e_{i(t+1)}}[V(x_{i+1}, e_{i(t+1)})] - r_{it})/\sigma_v)}{\exp((\mathbb{E}_{it}^{e_{i(t+1)}}[V(x_{i+1}, e_{i(t+1)})] - r_{it})/\sigma_v) + \exp(\mu_{it}/\sigma_v)}.$$  

### 5.2.1 Equilibrium Definition

Hourly earnings $\pi_t$ are determined by the equilibrium distributions of active medallions $F_c(.)|h_t)$ and searching passengers $F_d(.)|h_t)$. The distributions $F_c(.)|h_t)$ and $F_d(.)|h_t)$ determine a distribution of wait times $F_w(.)|h_t)$ and search time $F_s(.)|h_t)$ in a way that depends on the matching process.

**Definition 1** A competitive equilibrium in the taxi market is a set of distributions {$F_s(.)|h_t)$, $F_w(.)|h_t)$, $F_c(.)|h_t)$, $F_d(.)|h_t)$, $F_{\pi}(.|h_t) : h_t \in H)$, such that:

1. $F_d(|h_t)$ results from the demand function $d_t(w_t)$ under the distribution of waiting times $F_w(.)|h_t)$ and demand shocks $\xi_t$.
2. $F_s(.)|h_t)$ and $F_w(.)|h_t)$ result from $F_d(.)|h_t)$ and $F_c(.)|h_t)$ under the matching function $g(.)$.
3. $F_c(.)|h_t)$ results from optimal starting and stopping under $F_{\pi}(.|h_t)$.
4. $F_{\pi}(.|h_t)$ results from the distribution of search times $F_s(.)|h_t)$.

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

---

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

In this section we briefly discuss how the primitives of the model are identified.

There are six main objects to be identified. (1) There are cost function terms that vary with the duration of the shift, (2) an hourly fixed component of the cost function, (3) the fines, (4) the standard deviation of the hourly outside option, (5) the mean of the daily outside option, and (6) the standard deviation of the daily outside option. The identification of (1) can be best understood by using backward induction for the driver’s decision problem. At the maximum allowed shift length (drivers are not allowed to drive more than 13 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 earning in the last shift hour, which is a data object. Once the value of the cost function in the last hour is identified, it determines the continuation value from the perspective of the preceding hour. Hence, the second to last value of the cost function is identified: earnings in that hour and the continuation value are composed of data and identified objects. We can repeat the argument until we reach the first hour of the shift. However, (2) is also dependent on the hour of the day. This part of the cost function is identified by systematic inter-temporal variation in the stopping probabilities throughout the day even after conditioning on shift-length and earnings. For example, the stopping probability increases sharply after 12pm even though there is not a contemporaneous sharp decline in the earnings. This kind of variation in the data identifies the differences in the values of $\lambda_0$. (3) is identified by the increase in the stopping probabilities at shift transition times, again after conditioning on shift-length and expected earnings. (4) is identified by the variation in the earnings. This concludes the identification of the value function, which can be treated as a known object for the discussion of the primitives of the entry decision. The varying values throughout the day of $\mathbb{E}_{\pi_{t+1}, \epsilon_{i(t+1)}}[V(x_{i(t+1)}, \pi_{t+1}, \epsilon_{i(t+1)})] - r_{ht}$, which is composed of data and identified objects, identify the different values of $\mu_{ht}$ (5) and their dispersion, i.e., the value of $\sigma_{\pi}$ (6).

5.4 Estimation

5.4.1 Constructing the Data for Supply Side Estimation

To estimate the model, the trip based TPEP data has to be transformed into shift dataset where the unit of observation is a medallion-hour combination. For estimation we use data from 2011 as well as the August data from 2012. Shifts are defined following Farber (2008) as a consecutive sequence of trips where breaks between two trips cannot be longer than five hours. This definition might sometimes lead to long breaks within a shift if there is a long interval be-
tween two trips. This conflicts with our assumption that drivers’ make decisions given earnings for each hour of their shift. We do not model breaks: we assume them to be an exogenous process and estimate the likelihood of a break for each hour conditional on the state and compute hourly earnings as the expected wage that is earned while searching for passengers multiplied by the probability that the driver is not on a break.\footnote{In the WebAppendix we report that our results are robust to alternative definition of breaks.} 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 search time relates the aggregate market conditions to the hourly earnings potential of drivers. Recall from the discussion above that earnings are a result of a combination of time-based and mile-based metering. We first calculate the \textit{actual hourly based rate} $\pi^0_t$ 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.\footnote{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.} For each hour we also compute the average trip length as well as the average search time for a taxi to find a passenger. Before we compute these averages all variables are winsorized at the 1% level to avoid them being driven by large outliers in the data. Based on these hourly averages we can then compute a realization of the hourly wage rate as $\pi_t = \pi^0_t \cdot (e_t / (e_t + s_t))$, i.e. the actual hourly rate times the fraction of the time the driver is delivering a passenger as opposed to searching.

\subsection*{5.4.2 Estimation Procedure}

For the estimation of supply side parameters, that we denote by $\theta$, we make use of the fact that we can compute the supply side problem as if it were a single agent decision problem against given hourly equilibrium earnings $\pi_t$. In other words, since we observe equilibrium earnings directly in the data, there is no need to compute equilibria in the estimation. 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).\footnote{Su and Judd (2012) demonstrates the computational advantage of MPEC over a nested fixed point computations (NFXP) in the classical example of Rust’s bus engine replacement problem, Rust (1987). Applications of MPEC to demand models and dynamic oligopoly models can be found in, for example Conlon (2010) and Dubé et al. (2012). An intuitive explanation for the computational advantage of MPEC is that the constraints imposed by the economic model are not required to be satisfied at each evaluation of the objective function.}

In our case the equilibrium constraint comes from the assumption that the data is generated by a model of decentralized 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 \( \hat{\theta} \).\footnote{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.} MPEC allows the constraint imposed by the value function to be slack during the search but requires that it be satisfied for the final set of recovered parameters.\footnote{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.}

We specify a likelihood objective function. Following the model discussion above, we allow all parameters to be different for minifleet and owner-operated medallions. For a driver \( j \) we allow for two different daily outside options \( \mu_{z_j} \), one for the daytime shift 5AM to 5PM and one for the evening/night shift. We also allow the fine \( f_{h_1,z_j} \) to depend on whether the driver is on a night-shift \( f_{0,z_j} \) or a day-shift \( f_{1,z_j} \).

The constant part of the cost function is allowed to vary by hour in the following way: from 12AM to 5AM, from 5AM to 12PM, and from 5PM to 12AM. The other two parameters of the cost function \( \lambda_{1,z_j} \) and \( \lambda_{2,z_j} \) 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 \( \sigma_{e,z_j} \).

### 5.5 Parameter Estimates

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

<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>353.4 (15.79)</td>
<td>367.49 (20.45)</td>
</tr>
<tr>
<td>$\mu_{z_j,1}$</td>
<td>outside-option, 5am-5pm</td>
<td>369.18 (16.44)</td>
<td>369.82 (20.58)</td>
</tr>
<tr>
<td>$f_{0z_j}$</td>
<td>fine (nightshift)</td>
<td>99.42 (3.61)</td>
<td>102.95 (4.0)</td>
</tr>
<tr>
<td>$f_{1z_j}$</td>
<td>fine (dayshift)</td>
<td>103.87 (3.3)</td>
<td>81.87 (2.99)</td>
</tr>
<tr>
<td>$\lambda_{0z_j,0}$</td>
<td>fixed cost (1am-5am),</td>
<td>88.44 (0.51)</td>
<td>84.87 (1.14)</td>
</tr>
<tr>
<td>$\lambda_{0z_j,1}$</td>
<td>fixed cost (6am-12pm),</td>
<td>65.71 (0.69)</td>
<td>45.69 (0.67)</td>
</tr>
<tr>
<td>$\lambda_{0z_j,2}$</td>
<td>fixed cost (1pm-5pm),</td>
<td>49.23 (1.03)</td>
<td>43.44 (0.73)</td>
</tr>
<tr>
<td>$\lambda_{0z_j,3}$</td>
<td>fixed cost (6pm-12am),</td>
<td>60.22 (0.94)</td>
<td>41.03 (1.01)</td>
</tr>
<tr>
<td>$\lambda_{1z_j}$</td>
<td>linear cost coefficient</td>
<td>-16.01 (0.0)</td>
<td>-7.53 (0.0)</td>
</tr>
<tr>
<td>$\lambda_{2z_j}$</td>
<td>quadratic cost coefficient</td>
<td>1.66 (0.01)</td>
<td>1.0 (0.02)</td>
</tr>
<tr>
<td>$\sigma_\epsilon$</td>
<td>sd iid hourly outside option</td>
<td>67.06 (2.0)</td>
<td>67.06 (2.78)</td>
</tr>
<tr>
<td>$\sigma_\sigma$</td>
<td>sd iid daily outside option</td>
<td>61.34 (2.09)</td>
<td>61.34 (2.28)</td>
</tr>
</tbody>
</table>

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

5.5.1 Discussion of Parameter Estimates

A few observations about the estimates are worth highlighting. As shown in Figure 1, minifleet medallions follow the 5AM to 5PM shift pattern much more stringently than owner-operated taxis. This is reflected in the estimates. Owner-operated medallions have a higher standard deviation in the hourly error terms, which are the random parts that determine stopping behavior. This leads to stopping behavior that is “smoother”, than for minifleet medallions. This means that
owner operators have 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 6. To induce stopping of medallions near the end of the shift, the cost function and the fines therefore have to be higher for owner-operated medallions compared to minifleets, which is indeed the case. Lastly, we see that the daily outside option for owner-operated medallions is larger than for minifleets. This outside option captures all the surplus from driving, which is the wage plus the continuation value minus the cost of driving. This surplus is larger for owner-operated medallions because the cost function is steeper (Figure 14 in Appendix A), and the expected random term for continuing to drive is larger because of the larger standard deviation of the hourly outside option.

5.5.2 Model Fit

To evaluate the model fit it is useful to transform aggregate drivers’ decisions 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 given by the probability that the driver who utilizes stops and that no other driver decides to utilize it in the same hour: \( \hat{p}^M(t) = \hat{p}(h_t) \cdot (1 - \hat{q}(h_t)) \). The probability that an inactive medallion becomes active is the probability that a driver starts utilizing an inactive medallion \( \hat{q}^M = \hat{q}(h_t) \). The stopping probabilities unconditional on the shift length are obtained from the conditional stopping probabilities. Let \( N_t \) be the number of inactive medallions and let \( M_t \) be the number of active medallions. The law of motion for medallions discretized into hourly intervals is:

\[
N_{t+1} = (1 - \hat{p}^M_t) \cdot N_t + \hat{q}^M_t \cdot M_t \quad \text{and} \quad M_{t+1} = (1 - \hat{q}^M_t) \cdot M_t + \hat{p}^M_t \cdot N_t.
\]

The model fit for the aggregate law of motion is presented in Figure 9, which shows that we are able to replicate the daily pattern of supply activity quite well.

5.6 Labor Supply Elasticity

We now discuss the supply elasticities implied by our model. Given the shift indivisibilities and the dynamic nature of the labor supply problem, the supply responsiveness to earnings depends on the type of change (shock) to the earnings process that we consider. We contrast a uniform (anticipated) shock that increases earnings throughout the day to an hour specific (anticipated) shock for each hour that has the same normalized value, and we describe the response of total daily supply to each of these hourly shocks. Figure 10 displays the results. The uniform shock results in an elasticity estimate of 1.3 which, interestingly, is remarkably close to the number (1.2) reported in Angrist et al. (2017) in an experiment on Uber data.

The effect of the homogeneous daily shock results from a combination of ad-
ditional entry and longer continuation decisions, as continuation becomes more profitable at all hours relative to the (unchanged) outside options.

The response to hourly shocks varies from .7 to 1.5 depending on the hour, with beginning and end of shift elasticities lower than those for the middle of the shift. The reason for this heterogeneity in the responsiveness of total labor supply is the following. To a first order, the effect on daily entry decisions is the same for all hourly shocks and of the same order as the homogeneous shock. However, the effects on continuation decisions are different depending on the hour. An earnings shock in hour $t$ has an effect on all continuation decisions in hours prior to $t$ because of the expected future “bonus” at hour $t$, as well as a “mechanical” effect on continuation outcomes in hours following $t$ that comes from the fact that more drivers are active and may therefore take advantage of potential positive outside options shocks. This effect on balance penalizes earlier hours. It is less obvious why shocks in later hours in the shift generate less responsiveness. This is due to a combination of the fact that the cost is convex and the that the likelihood of reaching the later hours is lower since it requires a long sequence of positive continuation decisions.

6 Counterfactual Experiments

As we argued earlier, two important sources of inefficiency in the taxi market are regulatory entry barriers and matching frictions. Part of the success of startup

\[\text{Footnote 43: There is of course a difference due to the different driving costs implied by different continuation decisions. This is, of course, taken into account by the model but we ignore this for the purpose of the expositional discussion.}\]
ride hailing services can be attributed to the fact that they address both of these inefficiencies.\textsuperscript{44} Our counterfactuals attempt to distinguish the effects of additional entry from those of more efficient matching. We also investigate whether market segmentation between different operators may lead to inefficiencies: the introduction of a dispatch system à la Uber that only covers part of the market may reduce overall market thickness and lead to overall inferior outcomes in spite of an inherently superior dispatch platform. We then evaluate how matching frictions are affected by density by considering a city that is identical to Manhattan except that its passengers are spread over an area that is three times as large.

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

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

All counterfactuals are computed in two steps. We first compute an equilibrium in which we fix demand. This scenario is then compared to the full counterfactual in which demand is allowed to adjust to changes in wait time. This highlights the impact of demand responsiveness to wait time.

In the estimation there was no need to compute market equilibria since the supply side parameters were estimated using the observed process of hourly

\footnote{An additional effect is that supply is made more responsive to demand shocks through surge pricing. We do not address this issue here.}
earnings. For the counterfactuals we have to address the challenge of equilibrium computation, which is particularly demanding since all endogenous variables are allowed to vary by hour of day. We iteratively solve for the supply and demand side parameters, holding the variables of the respective other side of the market fixed until updates on both sides of the market reach a specified threshold. The exact algorithm is described in detail in item C.

Recall that in the estimation we set the daily lease rates that drivers pay to rent taxis equal to the binding rate cap. In the counterfactuals we do not allow lease rates to adjust endogenously to changes in market conditions. This is innocuous when we consider changes that improve drivers’ earning, such as the universal dispatch counterfactual. In such cases the lease cap would be even more binding. For the other counterfactuals, assuming that the lease cap is still binding amounts to an assumption that the change is not large enough to move out of the region in which the rate cap is binding. Unfortunately we do not have variation in the data that allows us to pin down where the lease cap would stop being binding. However, we acknowledge that for large changes, once the cap stops being binding, then lease rates would have to adjust to maintain drivers’ entry unchanged. This would moderate the negative effects on taxi aggregate supply, although there would still be effects on stopping decisions where lower lease rates would have no direct effect. (Lower lease rates would have an indirect offsetting effect through larger entry of reducing continuation earnings.)\(^{45}\)

\(^{45}\)The natural model of lease determination would be one in which the market is competitive on both sides (fleets and drivers), with rental rates equating the supply (daily entry) of drivers with the demand for drivers. The demand for drivers should be completely inelastic since the number of medallions is fixed and investment in taxis is sunk.
Table 3: Counterfactual Results

<table>
<thead>
<tr>
<th></th>
<th>hourly active cabs</th>
<th>hourly demand</th>
<th>passenger waittime per day</th>
<th>matches per day</th>
<th>taxi searchtime</th>
<th>hourly taxi revenues</th>
<th>consumer surplus (minutes)</th>
<th>medallion revenue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>6090.0</td>
<td>21973.0</td>
<td>2.7</td>
<td>19622.55</td>
<td>7.3</td>
<td>39.54</td>
<td>2.0</td>
<td>2.54</td>
</tr>
<tr>
<td>Entry</td>
<td>6624.0</td>
<td>23607.0</td>
<td>2.51</td>
<td>21133.85</td>
<td>7.47</td>
<td>39.19</td>
<td>2.09</td>
<td>2.51</td>
</tr>
<tr>
<td>Entry (perc. change)</td>
<td>8.78</td>
<td>7.44</td>
<td>-6.97</td>
<td>7.7</td>
<td>2.38</td>
<td>-0.89</td>
<td>4.74</td>
<td>-0.97</td>
</tr>
<tr>
<td>Entry(PE)</td>
<td>6404.0</td>
<td>21973.0</td>
<td>2.41</td>
<td>21706.27</td>
<td>8.0</td>
<td>38.13</td>
<td>2.37</td>
<td>2.44</td>
</tr>
<tr>
<td>Entry(PE) (perc. change)</td>
<td>5.16</td>
<td>0.0</td>
<td>-10.67</td>
<td>10.62</td>
<td>9.67</td>
<td>-3.58</td>
<td>18.37</td>
<td>-3.92</td>
</tr>
<tr>
<td>Dispatch</td>
<td>6358.0</td>
<td>22892.0</td>
<td>2.66</td>
<td>21768.85</td>
<td>6.49</td>
<td>34.63</td>
<td>2.12</td>
<td>2.68</td>
</tr>
<tr>
<td>Dispatch (perc. change)</td>
<td>7.69</td>
<td>4.19</td>
<td>-1.47</td>
<td>10.94</td>
<td>-11.11</td>
<td>5.27</td>
<td>6.11</td>
<td>5.42</td>
</tr>
<tr>
<td>Dispatch(PE)</td>
<td>6323.0</td>
<td>21973.0</td>
<td>2.75</td>
<td>21881.93</td>
<td>7.0</td>
<td>40.81</td>
<td>2.2</td>
<td>2.61</td>
</tr>
<tr>
<td>Dispatch(PE) (perc. change)</td>
<td>3.83</td>
<td>0.0</td>
<td>1.91</td>
<td>11.51</td>
<td>-4.08</td>
<td>3.2</td>
<td>9.91</td>
<td>2.68</td>
</tr>
<tr>
<td>Segmented(80)</td>
<td>5445.0</td>
<td>18101.0</td>
<td>3.32</td>
<td>16780.07</td>
<td>8.4</td>
<td>37.39</td>
<td>1.8</td>
<td>2.32</td>
</tr>
<tr>
<td>Segmented(80) (perc. change)</td>
<td>-10.58</td>
<td>-17.62</td>
<td>23.08</td>
<td>-14.49</td>
<td>15.04</td>
<td>-5.43</td>
<td>-9.88</td>
<td>-8.71</td>
</tr>
<tr>
<td>Segmented(2-Search)</td>
<td>4656.0</td>
<td>15966.0</td>
<td>3.78</td>
<td>14319.3</td>
<td>8.73</td>
<td>34.54</td>
<td>1.61</td>
<td>2.35</td>
</tr>
<tr>
<td>Segmented(2-Dispatch)</td>
<td>5985.0</td>
<td>20295.0</td>
<td>2.99</td>
<td>1924.83</td>
<td>7.42</td>
<td>39.07</td>
<td>1.99</td>
<td>2.49</td>
</tr>
<tr>
<td>Segmented(2-Dispatch) (perc. change)</td>
<td>-1.72</td>
<td>-7.63</td>
<td>10.87</td>
<td>-1.95</td>
<td>1.69</td>
<td>-1.2</td>
<td>-0.66</td>
<td>-1.98</td>
</tr>
<tr>
<td>SegmentedEnd</td>
<td>5359.0</td>
<td>20585.0</td>
<td>2.93</td>
<td>1933.72</td>
<td>9.1</td>
<td>37.29</td>
<td>2.01</td>
<td>2.27</td>
</tr>
<tr>
<td>SegmentedEnd (perc. change)</td>
<td>-12.0</td>
<td>-6.32</td>
<td>8.42</td>
<td>-1.45</td>
<td>24.67</td>
<td>-5.69</td>
<td>0.45</td>
<td>-10.75</td>
</tr>
<tr>
<td>Density</td>
<td>4581.0</td>
<td>13852.0</td>
<td>4.33</td>
<td>1474.71</td>
<td>11.1</td>
<td>32.61</td>
<td>1.43</td>
<td>1.97</td>
</tr>
<tr>
<td>Density (perc. change)</td>
<td>-24.76</td>
<td>-36.96</td>
<td>60.38</td>
<td>-36.41</td>
<td>52.14</td>
<td>-17.54</td>
<td>-28.27</td>
<td>-22.22</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 counterfactual results. The percentage changes ∆% are the changes of the means over all hours compared to the baseline. Consumer surplus is computed under the assumption that the demand function is truncated above the maximal waiting time observed in the data. The reason is that, for our parameter specifications, consumer surplus would be infinite if we integrated over all waiting time. This issue results from the assumption of constant elasticity, log-linear demand. A similar issue arises, for example, in Wolak (1994), who also truncates the demand distribution. Note, however, that except for the limit case, the absolute difference in consumer surplus will be well defined and the same, no matter how high we choose the truncation point.
6.1 Relaxing Entry Restrictions

To explore how additional entry affects the market, we consider an increase in the number of medallions of 10%, from 13,500 to 14,850. To place this magnitude in an applied context, note that Uber served 4% of the total number of trips in 2014 and 13% at the beginning of 2015.46

Table 3 describes counterfactual outcomes aggregated at the daily level. Changes for this counterfactual are of uniform sign at the hourly level, although there is some difference in the magnitudes for daytime versus nighttime responses. For instance, taxi activity changes less during the night. On average, taxi activity expands by 8.78%, less then proportionally to the medallion increase, because earnings drop, causing drivers to work less. This increase in available taxis reduces wait time for passengers and, as a result, demand expands moderately. On average, the expansion of demand is 7.44%, while the average reduction in wait-time for passengers over all hours is 6.97%.47 The demand expansion moderates the drop in taxi earnings caused by increased supply. This is a sizable effect: if demand were held fixed, supply would only expand by 5.16%. The demand expansion almost completely compensates drivers for the additional competition: taking the mean over all changes in hourly income, the hourly wage of drivers decreases by only 0.89%. Earnings would instead fall by 3.58% if demand were held constant. The increase in the number of matches (or total number of trips) is 7.7%, very close to the increase in demand. The present discounted revenue stream for a medallion decreases only by 0.97%, quite a bit less than if demand did not increase (in which case it would drop by 3.92%). Consumer surplus increases by roughly 2 percent and would be higher without demand response.48 In order to understand this, note first that wait time drops a lot more when demand does not adjust than when it does. Second, the additional gain in the number of serviced trips when demand adjusts is relatively small, partly because of the modest demand elasticity. Thus, the gain in the inframarginal trips that are already served in the scenario with no demand adjustment is larger.

Our model does not account for the traffic externality and adverse environmental effects due to additional taxis on the street. While environmental damage is hard to assess, the traffic externality has recently been investigated by the city. Their finding was that ride-hailing services are only a minor contributing factor to the recent decline in NYC traffic speed.49

---

46See http://fivethirtyeight.com/features/uber-is-taking-millions-of-manhattan-rides-away-from-taxis/. This is based on data that the city obtained from Uber for a traffic study.
47All averages across hours are weighted in proportion to the number of trips taken in each hour.
48This is also the case for other counterfactuals for the same reason discussed here.
49See http://www1.nyc.gov/assets/operations/downloads/pdf/For-Hire-Vehicle-Transportation-Study.pdf. However, Daniel Mangrum (2017) find that the introduction of outer borough green cabs has lead to a local decrease in traffic spec of about 9%.
6.2 Improved Matching

6.2.1 Universal dispatcher

We now consider a counterfactual that considers an alternative matching technology. We consider dispatch platforms that matches each empty (searching) cab with a waiting passenger. The dispatcher matches a passenger with the closest empty cab, if one is available within a one-mile radius.\textsuperscript{50} This matching process is a natural alternative to the street hailing system and approximates the one used, for instance, by Uber, especially in its first few years of operation. Ideally, a matching algorithm would not only search across empty taxis but would also take into account the possibility that a soon to be empty taxi may be closest to a passenger. Such an algorithm would be difficult to compute and is potentially difficult to implement so we focus on a simpler case. But, since our dispatcher does not optimize in a forward looking way, it can be very costly to allow for matches of passengers and cabs that are too far apart. The restriction to one-mile radius alleviates this problem.\textsuperscript{51}

![Figure 11: Matching function comparison](image)

\textsuperscript{50}Passengers who do not find an immediate match wait for taxis to become available. Taxis that are unmatched drive randomly until matched.

\textsuperscript{51}The following extreme-case illustrates why a pure spatial global search may not be optimal. Consider a scenario in which there is only one passenger left waiting, only one empty cab searching, and they are on opposite ends of the grid. The remaining cabs are delivering passengers. If the dispatcher were to commit them to a match, there is a high chance that a better outcome could be obtained by waiting. The passenger is likely to obtain a faster match by waiting for one of the busy cabs to finish its trip and become available. Analogously, for the empty cab a new passenger may appear on the map closer to its position. A dynamic algorithm that searches spatially as well as across time could account for these better match opportunities.
Before presenting the counterfactual results, it is useful to discuss how the dispatch matching function differs from the matching function that results from the search process that we have assumed so far and that was discussed in Section sub-subsection 5.1.2. We describe some key features of a simulation of the dispatch matching function over a range of outcomes. Results are displayed in figure Figure 11. This figure considers what happens to passenger wait time as the market becomes thicker, increasing the market size (or density) while keeping constant the ratio between taxis and passengers. (Recall that a taxi serves multiple passengers within an hour.) Several features of this figure are worth commenting on. First, both matching functions display increasing returns with respect to wait time. Second, the dispatch matching function has lower slope, implying that it is less sensitive to density. Third, for low density the search matching function is superior so that wait times can become longer under dispatch when density is low. The reason for this potentially surprising outcome is related to the issue detailed above: passengers may be better off waiting for a random cab that is currently delivering a passenger and therefore not available for dispatch. This effect is only relevant for numbers that are low relative to daytime traffic in NYC.

We now describe results under the different dispatch scenarios, starting with the extreme case of an entire market operating under the dispatcher. ?? gives an overview of the changes across different hours. The panel for each respective variable shows the difference between what happens in the baseline case minus what happens in the counterfactual. We see that both demand and supply expand at all hours of the day, with wait times and search times going down, particularly at night, demonstrating the higher returns to a dispatcher under a more sparsely populated map. In this counterfactual there is an increase in the number of active taxis (7.69%), and a substantial reduction in the search time for taxis (−11.11%). (If we didn’t allow for an expansion of demand in response to reduced wait times, the supply increase would in contrast only be 2.17%.) Passenger wait time is reduced by 1.47%, consumer surplus increases by 6.11%, and the number of trips increases by 10.94%, more than proportionally to demand.
(because frictions were also reduced and there are fewer unmatched passengers). The value of medallions increases by 5.4%. Therefore, all stakeholders appear to benefit from the introduction of the dispatch technology, under the assumption that this is universally adopted.

### 6.3 Partial Dispatching and Market Thickness

In reality, a new dispatch system is likely to only cover part of the market, at least initially. Thus, the introduction of a new matching technology may result in a segmentation of the market into multiple platforms. If consumers are divided between competing platforms, both segments of the market become thinner, which could potentially lead to losses in matching and wait time that could offset the improvements due to better matching within the dispatch platform. To explore whether this is a plausible outcome, consider a scenario where we divide the medallions equally across platforms: half operate on the search platform (baseline), half on the dispatch platform. In our first step we also divide up demand exogenously across platforms, by simply pre-multiplying the constant elasticity demand function $d(w_t)$ by the shares of 0.5. This ensures that, if the wait-time in the two platforms were the same, total demand would add up the baseline total demand.  

We first report the outcome when we do not allow passengers to choose among platforms. In particular, in the first segmented platform scenario, we assume that both drivers and passengers are committed to a match: neither can cancel should another match option become available sooner.  

Appendix A shows that the dispatch platform always serves more than 50% of the rides and that this share is highest during the night, where it can account for more than 60% of all rides. This is because the dispatch platform has lower wait times, and therefore, higher demand. Appendix A, however, illustrates that, in the aggregate, compared to the baseline, the reduction in market thickness has negative consequences that more than offset the improved matching in the segment of market served by the dispatcher. Regardless of the hour, in the aggregate there are substantially fewer active cabs (-10.58%) and lower demand (-17.62%). Wait time and search time are much larger (+23% and +14% respectively).

In order to separate the respective roles of the market thickness external-

---

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

53 E.g., we do not allow a waiting passenger who is matched to a dispatch cab to accept a ride from a searching cab.

54 This advantage of the dispatch platforms in hours with lower demand density is related to the findings in Cramer and Krueger (2016). They show that Uber’s advantage in capacity utilization (defined as the fraction of time delivering passengers) is relatively minor in NYC but large in other cities which are less dense.
ity and the of the dispatch platform, we perform the same exercise with two competing segmented dispatch platforms as well as with two competing street-hailing systems.\footnote{This is entirely a conceptual exercise: it is not clear what segmentation would mean for two street hailing platforms.} Lines Segmented 2-Search and Segmented 2-Dispatch in Table ?? present the results of these scenarios. The table shows that the two competing dispatch platforms perform substantially better, while the two competing search platforms deliver the worst outcomes. However, even the case with two competing dispatch platforms performs worse than the baseline, highlighting the importance of market thickness.

Finally, we allow passengers to sort themselves endogenously into either the dispatch or the search platforms. Because we do not consider additional dimensions of heterogeneity across the two platforms, it must be the case that passenger choice induces equal wait time across the two platforms. As more passengers move to the dispatch platform, this becomes more congested, while fewer passengers compete for the search taxis resulting in increased wait time for the dispatch platform and lower wait time for the search platform.\footnote{For given numbers of taxis in the two platforms existence of such indifference is guaranteed by this logic. In our setting, a complication arises because taxis supply is endogenous and therefore it responds to this process. However, in the simulation this is not sufficient to lead to a corner solution.} Lines SegmentedEnd in Table ?? presents the results for this scenario. Compared to the case in which demand is segmented exogenously, it turns substantially more market share is captured by the dispatch platform (especially during the night time). This turns out to moderate some of the negative effects of the market thickness externality. More?

### 6.4 Density Counterfactual

The last counter-factual explicitly addresses the spatial dimension of matching in this market. We now assume that matching takes place on an area, which is three times the area of our baseline simulation of Manhattan, but is otherwise identical. We first assume that there is decentralized search and compare outcomes in this less dense market to the baseline outcomes. We then introduce a universal dispatcher in order to evaluate how density affects the importance of the different matching technology. The welfare and driver income effects displayed in Table 4 highlight the important effects of density in this market.

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>Density</th>
<th>Δ%</th>
<th>Density Dispatch</th>
<th>Δ%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Consumer Surplus</strong></td>
<td>2.0</td>
<td>1.43</td>
<td>-28.3</td>
<td>1.97</td>
<td>-1.59</td>
</tr>
<tr>
<td>(million minutes / day)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Driver Revenue</strong></td>
<td>39.54</td>
<td>$32.61</td>
<td>-17.54</td>
<td>$39.12</td>
<td>-1.06</td>
</tr>
<tr>
<td>(hourly income)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Medallion Revenue</strong></td>
<td>2.54</td>
<td>$1.97</td>
<td>-22.2</td>
<td>$2.52</td>
<td>-1.59</td>
</tr>
<tr>
<td>(PV in millions)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Wait time</strong></td>
<td>2.7</td>
<td>4.33</td>
<td>60.4</td>
<td>3.02</td>
<td>11.87</td>
</tr>
<tr>
<td>(average in minutes)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4: Density
7 Conclusion

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

Counterfactual results from the model show that an improvement in the matching technology leads to substantial increases in consumers welfare as well as drivers’ earnings. However, our results also point to the fact that competition among dispatch platforms can lead to decreases in welfare because it leads to market segmentation and lower market thickness. Our analysis of segmented platforms is only suggestive as the model does not allow for any additional heterogeneity among the platforms and assumes exogenous assignments of passengers to platforms. Including such richness is not within the scope of the current paper but it would be interesting the extend the analysis to study this issue in more depth.\footnote{\cite{Cantillon2008}}

We have found that more efficient utilization of existing medallions due to the elimination of favored treatment for owner-operators can lead to comparable gains as a policy that allows for a substantial number of additional entrants. This points to the potential importance of regulations favoring small firms. Such regulations exists in many industry and countries. Few studies have explored and quantified the potential impact of such restrictions.

We have not considered the issue of surge pricing. In order to study this question, one would need to estimate a richer demand system that allows for dependence on both prices and wait time, while recognizing that the likely correlation

\cite{Cantillon2008} study a related question in their analysis of competition among financial exchanges that trade the same securities.
between consumers responsiveness to wait time and to prices. Uber data may be helpful for studying such a question.

References


A Additional Figures

**Figure 14:** Cost functions at different times of day

Medallion Type: — Fleet —– Owner-operated

<table>
<thead>
<tr>
<th>Shift Hour</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>(6am–12am)</td>
<td>$20, $40, $60, $80, $100, $120, $140, $160</td>
</tr>
<tr>
<td>(1pm–5pm)</td>
<td>$20, $40, $60, $80, $100, $120, $140, $160</td>
</tr>
<tr>
<td>(1am–5am)</td>
<td>$20, $40, $60, $80, $100, $120, $140, $160</td>
</tr>
<tr>
<td>(6pm–12pm)</td>
<td>$20, $40, $60, $80, $100, $120, $140, $160</td>
</tr>
</tbody>
</table>

B Details on Simulation

The goal of the simulation is to obtain a mapping of the number waiting passengers and searching cabs within an hour to the waiting-time and search-time of those passengers and cabs. The mapping is used to infer the number of waiting passengers from observed number of active cabs and their search time. Waiting and search-time are also influenced by other exogenous factors, which therefore need to be arguments of the matching function. These factors are the speed $mph_t$.
Figure 15: Time medallion is unutilized conditional on hour of drop off.

Figure 16: Evidence of search frictions
at which the traffic flows, the average trip length $\text{miles}_t$ requested by passengers. Table 5 provides an overview both over the taxi search time observed in the data as well as the observed inputs to the matching function.

Table 5: Summary Statistics for Simulation Variables.

<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.

The baseline simulation is performed under the assumption that cabs search randomly for passengers. The search is performed on an idealized map of Manhattan. Figure 5 provides a schematic of the grid that we use for the simulation. In line with the topography of Manhattan we require the area to be four times as long in north-south direction ($y_t$) than wide in east west-direction ($x_t$). Cabs are moving on nodes that are 1/20 mile segments apart from each other, which is based on the average block length in north-south direction. In the north-south direction they can turn at each node whereas in the east-west direction they can only turn at every fourth node. Figure 5 highlights nodes on which cabs can turn as gray. This corresponds to the block structure of Manhattan where a block is approximately 1/20 miles long in north-south and 4/20 miles wide in east-west direction. Under the random search assumption cabs take random turns at nodes with equal probability weight on each permissible direction. However, we assume that they never turn back to the direction from which they were coming (i.e. no U-turns).

Since we only model the Manhattan market (below 128th street), our grid corresponds to an area of 16 square miles. Figure 17 shows the modeled part on the map in its division in the eight equally sized different areas for which we separately compute the pick up and drop-off probabilities. Correspondingly, our grid is divided into eight equal parts, which correspond to those areas.

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

If a cab hits a node with a passenger a match occurs. There are no additional frictions on a node (which corresponds to a street corner) and the number of matches is the minimum of the number of passengers and the numbers of taxis on
Figure 17: Division of Manhattan
the node, which corresponds to the assumption of a Leontieff matching function on each node. Once the match takes place the cab is taken of the grid for 60 \( \cdot \frac{\text{miles}}{\text{mph}} \) minutes, i.e. the average measured delivery time from the data, after which is has delivered the passenger and is again placed randomly on the map with a random travel direction. Cabs reappear in locations on the grid in proportion to observed drop-off locations of the eight areas (Figure 17).

The full algorithm is described in pseudo-code below. It takes the following inputs: the number of cabs \( c \), the number of passengers \( d \), the trip length miles, and the trip speed mph. A unit of time in the algorithm is scaled so that it always represents the time it takes a cab to travel from one node to the next since there is no need for a smaller time unit. Passengers are added to the map for one hour in ten minute intervals. Since nodes are spaced 1/20 miles apart, the last time passengers are added to the map is at \( \bar{t} = 20 \cdot \text{mph} \). The set of times at which new passengers arrive is given by: \( \{\bar{t}/6 \cdot k|k = 1, ..., 6\} \). The following additional variables are used to describe the algorithm: \( \text{npick} \) refers to the number of matches that have already taken place, \( \text{deliverytime}_i \) to the remaining delivery time of taxi \( i \), and \( \text{searchtime}_i \) to the time that taxi \( i \) has spent searching since the last delivery, \( \text{total-searchtime} \) refers to the total time that taxis have spent searching for passengers, and \( \text{total-waittime} \) to the total wait time that passengers have been waiting.
while npick < d do
  \( t = t + 1 \) (time units represent travel time from one node to the next, scales with mph);
  if \( t \in \{ \frac{t}{6} \cdot k | k = 1, \ldots, 6 \} \) then
    add \( d/6 \) passengers to random nodes on map, stratified by eight areas;
  end
for \( i = 1 : c \) do
  if deliverytime \(_i\) = 0 (cab \( i \) is not occupied) then
    update the node of cab \( i \). Cabs only take turns on gray nodes (Figure 5) and do not make u-turns. All feasible travel directions are chosen with equal probability;
    if new node of cab \( i \) has a passenger then
      cab becomes occupied, set deliverytime \(_i\) to \( 20 \cdot \text{miles} \), and add searchtime \(_i\) to total_searchtime;
    else
      searchtime \(_i\) = searchtime \(_i\) + 1
    end
  else
    deliverytime \(_i\) = deliverytime \(_i\) - 1;
    if deliverytime \(_i\) == 0 then
      place cab in random area on map according to observed drop-off probabilities (all nodes within area equal probability), give cab random feasible travel direction;
    end
  end
end
Add one to total_waittime for each passenger that is on the map;
end

**Result:** Use total_waittime and total_searchtime to compute the average search time for taxis and average wait time for passengers.

In the dispatcher simulation we assume that each cab - as soon as the previous passenger has been delivered - is matched with the closest passenger available. We also assume that neither the driver nor the passenger has an option to cancel this match for another match option. It might for example happen that a waiting passenger, who is promised to a cab, encounters a cab that was not previously available before the promised cab arrives. The option to cancel might in some instances be beneficial to a market-side because our search for the optimal match
is only over the currently available cabs and passengers and does not take into account cabs and passengers that will soon appear somewhere close on the map.

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

### B.1 Details on Heterogeneity across areas.

This robustness check explores how our demand inversion results as well as the demand estimates change when we account for more granular heterogeneity in local demands. Note first that Table 6 shows that taking the eight areas that we are currently using, search time is remarkably stable across locations. Since this is the variable that informs us about the number of passengers and their wait-time, the data seems to suggest that more granular conditioning would not change results much. To explore, whether this is true, we subdivide the map into 16 areas (using conditional drop-offs and pick-ups for these smaller areas) instead and re-estimate demand. Our demand estimates barely change.

<table>
<thead>
<tr>
<th>Area</th>
<th>Median</th>
<th>Mean</th>
<th>Median Day</th>
<th>Mean Day</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.55</td>
<td>11.64</td>
<td>4.37</td>
<td>10.3</td>
</tr>
<tr>
<td>2</td>
<td>4.37</td>
<td>9.15</td>
<td>4.37</td>
<td>7.7</td>
</tr>
<tr>
<td>3</td>
<td>4.37</td>
<td>8.6</td>
<td>4.37</td>
<td>7.7</td>
</tr>
<tr>
<td>4</td>
<td>4.37</td>
<td>9.21</td>
<td>4.37</td>
<td>8.04</td>
</tr>
<tr>
<td>5</td>
<td>4.37</td>
<td>8.34</td>
<td>4.37</td>
<td>7.8</td>
</tr>
<tr>
<td>6</td>
<td>4.37</td>
<td>8.79</td>
<td>4.37</td>
<td>8.29</td>
</tr>
<tr>
<td>7</td>
<td>4.37</td>
<td>9.68</td>
<td>4.37</td>
<td>9.4</td>
</tr>
<tr>
<td>8</td>
<td>6.55</td>
<td>11.42</td>
<td>4.37</td>
<td>11.12</td>
</tr>
</tbody>
</table>

Next, in Table 7 we show results on search time wait time and matches four a matching simulation with 4 areas, 8 areas, and 16 areas. These results show that while going from 4 to eight areas leads to higher search, higher wait time, and less matches. However, allowing for more granularity leads to only very small additional changes in these variables. This is reflected in the inferred wait time and number of passengers, which changes very little from 8 areas to 16 areas. Under the less granular map with four areas, we back out 1.96 minutes of average wait time and 21246 passengers per hour. Under our baseline of 8 areas these numbers are 2.58 and 22450 and only increase to 2.73 and 22746 with 16 areas. Lastly, the
inferred demand elasticities are almost unchanged across all three specifications for the simulation.

Table 7: Robustness area division.

<table>
<thead>
<tr>
<th>Passengers</th>
<th>Cabs</th>
<th>s_4</th>
<th>s_8</th>
<th>s_16</th>
<th>w_4</th>
<th>w_8</th>
<th>w_16</th>
<th>m_4</th>
<th>m_8</th>
<th>m_16</th>
</tr>
</thead>
<tbody>
<tr>
<td>15000.0</td>
<td>4000.0</td>
<td>4.83</td>
<td>5.58</td>
<td>5.77</td>
<td>4.05</td>
<td>4.8</td>
<td>4.93</td>
<td>13750.0</td>
<td>13201.0</td>
<td>13056.0</td>
</tr>
<tr>
<td>15000.0</td>
<td>6000.0</td>
<td>11.53</td>
<td>11.84</td>
<td>11.96</td>
<td>1.32</td>
<td>1.74</td>
<td>1.83</td>
<td>14747.0</td>
<td>14599.0</td>
<td>14534.0</td>
</tr>
<tr>
<td>15000.0</td>
<td>9000.0</td>
<td>23.34</td>
<td>23.48</td>
<td>23.51</td>
<td>0.35</td>
<td>0.46</td>
<td>0.47</td>
<td>14876.0</td>
<td>14826.0</td>
<td>14813.0</td>
</tr>
<tr>
<td>25000.0</td>
<td>4000.0</td>
<td>1.41</td>
<td>2.18</td>
<td>2.3</td>
<td>8.24</td>
<td>8.68</td>
<td>8.72</td>
<td>16998.0</td>
<td>16145.0</td>
<td>16013.0</td>
</tr>
<tr>
<td>25000.0</td>
<td>6000.0</td>
<td>3.51</td>
<td>4.42</td>
<td>4.59</td>
<td>4.09</td>
<td>4.9</td>
<td>4.99</td>
<td>22190.0</td>
<td>21014.0</td>
<td>20791.0</td>
</tr>
<tr>
<td>25000.0</td>
<td>9000.0</td>
<td>9.29</td>
<td>9.73</td>
<td>9.86</td>
<td>1.32</td>
<td>1.7</td>
<td>1.8</td>
<td>24204.0</td>
<td>23798.0</td>
<td>23675.0</td>
</tr>
<tr>
<td>35000.0</td>
<td>4000.0</td>
<td>0.75</td>
<td>1.17</td>
<td>1.29</td>
<td>10.7</td>
<td>10.79</td>
<td>10.84</td>
<td>17670.0</td>
<td>17251.0</td>
<td>17111.0</td>
</tr>
<tr>
<td>35000.0</td>
<td>6000.0</td>
<td>1.46</td>
<td>2.35</td>
<td>2.48</td>
<td>6.83</td>
<td>7.41</td>
<td>7.47</td>
<td>25272.0</td>
<td>23818.0</td>
<td>23606.0</td>
</tr>
<tr>
<td>35000.0</td>
<td>9000.0</td>
<td>4.2</td>
<td>5.05</td>
<td>5.19</td>
<td>2.89</td>
<td>3.6</td>
<td>3.68</td>
<td>31672.0</td>
<td>30234.0</td>
<td>29981.0</td>
</tr>
</tbody>
</table>

4 areas:

- El(logcabd) = -0.88, El(logmiles) = -0.86, El(shift-instr) = -0.58

8 areas:

- El(logcabd) = -0.94, El(logmiles) = -0.94, El(shift-instr) = -0.59

16 areas:

- El(logcabd) = -0.98, El(logmiles) = -0.94, El(shift-instr) = -0.58
B.2 First Stage Results.

Table 8: First Stage Regressions.

<table>
<thead>
<tr>
<th></th>
<th>( \log(w_t) )</th>
<th>( \log(w_t) )</th>
<th>( \log(w_t) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>DELAYED SUPPLY INSTRUMENT</td>
<td>0.0864**</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0180)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MILES INSTRUMENT</td>
<td>0.415**</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.0558)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SHIFT INSTRUMENT</td>
<td></td>
<td>0.229**</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>(0.0128)</td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>714</td>
<td>714</td>
<td>714</td>
</tr>
<tr>
<td>( R^2 )</td>
<td>0.0132</td>
<td>0.152</td>
<td>0.0448</td>
</tr>
<tr>
<td>( F )</td>
<td>22.94</td>
<td>55.26</td>
<td>323.1</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.

B.3 Details on Estimation.

As defined in the text, let \( \pi_t \) be a realization of the earnings and \( x_{it} \) denote the other part of the observable state (the shift length, the hour of the day as well as the medallion invariant characteristics). Let \( p(x_{it}, \pi_t) \) be the theoretical probability that an active medallion/driver \( i \) stops at time point \( t \) and \( q(x_{it}) \) be the probability that a inactive medallion/driver \( i \) starts at \( t \). Correspondingly, let \( d^A \) be the indicator that is equal to one if an active driver stops and \( d^I \) be an indicator that an inactive driver starts. Using this notation we maximize a constrained log-likelihood that we formulate as an MPEC problem. MPEC does not perform any intermediate computations, such as value function iterations, to compute the objective function. It instead treats these objects as parameters. This means that the solver will be maximizing both over the parameters of interest \( \theta \) and an additional set of parameters \( \delta \). The parameter vector \( \delta \) consists of all \( p(x_{it}, \pi_t), q(x_{it}), E_{\pi, \epsilon}[V(x_{i(t+1)}, \pi_{t+1}, \epsilon_{i(t+1)}) | h_t] \) for \( x_{i(t+1)} \in X, \pi_t \in \text{supp}(F_{\pi}(\cdot | h_t)) \). In other words, \( \delta \) consists of expected values and choice probabilities for each point in the observable state space. Note, however that \( \pi_t \) follows a continuous distribution and it is therefore not possible to specify a constraints for each value in the support of its distribution. We instead approximate the distribution of \( \pi_t \) with a
B.4 More Details on Estimation.

One of the important demand shocks for taxi rides are weather patterns, most notably rainfall (cite farber?). To check whether our demand recovery picks up such patterns we have merged hourly rainfall data. We then regress both log-wait-times and log-recovered-demand on a dummy whether rainfall occurred in this hour. As the table below shows, our recovered demand and wait-times are highly correlated to rainfall. Estimates from these regressions suggest that in an hour with rain-fall our recovered demand is about 25% higher and waiting-times about 46% higher.

Table 9: Demand Validation with Weather Data.

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \log(d_t) )</td>
<td>0.261**</td>
<td>0.247**</td>
<td>0.247**</td>
<td>0.355**</td>
<td>0.462**</td>
<td>0.463**</td>
</tr>
<tr>
<td>( \log(w_t) )</td>
<td>(0.0141)</td>
<td>(0.00852)</td>
<td>(0.00828)</td>
<td>(0.0113)</td>
<td>(0.00856)</td>
<td>(0.00826)</td>
</tr>
</tbody>
</table>

Rainfall Dummy

<table>
<thead>
<tr>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Note: + \( p < 0.10 \), * \( p < 0.05 \), ** \( p < 0.01 \). The dependent variables are the log of wait time and demand. The rainfall dummy is an indicator whether it has rained in this hour.

B.5 Medallion Ownership and Utilization

Figure 18 shows the cross-sectional distribution of the fractions of time a medallion spends delivering a passenger out of the total time that we observe a medallion. The distribution for owner-operated cabs displays a much thicker left tail of low utilization rates and is overall more dispersed.\(^{59}\)

The left panel of Figure 15 in Appendix A shows the length of time a medallion is inactive conditional on the stopping time of the last shift. Since most day shifts start around 5AM and most night shifts around 5PM, the time of non-utilization is minimized for stops that happen right around these hours, while

\(^{58}\) We use six nodes.

\(^{59}\) The observed differences might be due to the fact that minifleets enables a more efficient utilization; but another plausible argument is a selection-effect.
Figure 18: Histogram of Utilization Separated by Medallions

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.

A stop at any other time causes the medallion to be stranded for a longer period. We see that minifleets typically return a medallion to activity faster after each drop-off. This difference is particularly large after the common night shift starting times (6PM and later), which suggests that minifleets have access to a larger pool of potential drivers, and this in turn makes it easier for them to find a replacement for someone who does not show up at the normal transition time. In the structural model we allow for a different set of parameters for minifleets and owner-operated to capture these differences. The right panel of Figure 15 in Appendix A shows the number of shifts that end conditional on the hour. We see that minifleet medallions have a more regular pattern with most day-shifts ending at 4PM. This is also reflected in Figure 1 which shows a stronger supply decrease for minifleets before the evening shift relative to owner-operated medallions.

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

\[
\min_{\theta, p(x_{it}, \pi_t), q(x_{it})} \mathbb{E}_{\pi, \epsilon} [V(x_{i(t+1)}, \pi_{t+1}, \epsilon_{i(t+1)}|h_{it})] \sum \sum d_{it}^A \cdot \log(p(x_{it}, \pi_t)) + (1 - d_{it}^A) \\
\cdot (\log(1 - p(x_{it}, \pi_t))) + d_{it}^I \cdot \log(q(x_{it})) + (1 - d_{it}^I) \cdot (\log(1 - q(x_{it})))
\]

subject to:
The constraint given by equation 5 ensures that the starting and stopping probabilities obey the intertemporal optimality conditions imposed by the value functions. The log-formula is the closed form expression for the expectation of the maximum over the two choices of stopping and continuing, which integrates out the T1EV unobserved valuations. Equation 6 and Equation 7 are again the closed form expressions for the choice probabilities under extreme value assumption. We also restrict the search for the cost-functions to the domain of increasing functions by requiring \( \lambda_{0,z_j,0}, \lambda_{1,z_j,0} \) and \( \lambda_{2,z_j,0} \) to be larger than zero.

C Details on the Computation of Counterfactuals

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

1. For each hour simulate from the observed empirical distributions of speed of traffic flow and the length of requested trips, under current guess \( c_h^i \) and \( d_h^i \) to determine
the search time for taxis $s^i_h$, $h \in \{0, \ldots, 23\}$ under $g(\cdot)$.

2. Simulate drivers earnings $\pi^i_h$, $h \in \{0, \ldots, 23\}$ from the ratios of passenger delivery time over delivery and search time (computed in step 2) and rate earned per minute of driving. Simulate new expected number of passengers $d^i_h$, $h \in \{0, \ldots, 23\}$ from the waiting times $w^i_h$ and the estimated demand function $d(\cdot)$.

3. Compute the optimal starting and stopping probabilities $p^i(x, \pi; \theta), q^i(x; \theta)$ under the new earnings (computed in step 3).

4. Use $p^i(x, \pi; \theta)$ and $q^i(x; \theta)$ to simulate a new law of motion for drivers $c^i_h$, $h \in \{0, \ldots, 23\}$. For each medallion type $(z,k)$ we simulate thirty medallions, where each medallion starts inactive at 12PM and iterate forward for 48 hours. Across these thirty medallions we then compute the fraction of times the medallion has been active in this hour (using only the last 24 hours) and multiply this by the total number of medallions.\footnote{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.}

5. Compute $\text{sumsq}_1 = \sum_h (c^i_h - c^{(i-1)}_h)^2$

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

1. For each hour simulate values from the observed empirical distributions of speed of traffic flow and the length of requested trips, under $c^i_h$, $h \in \{0, \ldots, 23\}$, to determine the waiting times $w^i_h$, $h \in \{0, \ldots, 23\}$ for passengers.

2. Simulate the new number of passengers $d^i_h$, $h \in \{0, \ldots, 23\}$ from the waiting times $w^i_h$, $h \in \{0, \ldots, 23\}$ and the estimated demand function $d(\cdot)$.

3. Compute $\text{sumsq}_2 = \sum_h (d^i_h - d^{i-1}_h)^2$
Using those definitions, the algorithm can be described as follows:

```plaintext
while STOP ≠ 1 do
  Compute sumsq₁ using Block1(i).
  if sumsq₁ < tol then
    STOP = 1
  else
    while sumsq₁ > tol do
      Block1(i)
    end
  Compute sumsq₂ using Block2(i).
  if sumsq₂ < tol then
    STOP = 1
  else
    while sumsq₂ > tol do
      Block2(i)
    end
  end
  i = i + 1
end
```

### D Shift Transition Instrument

We now argue that the timing of the shift transition is such a supply side driven shifter. The kink in the number of active taxis in the later afternoon hours is clearly visible in Figure 1 and ?? shows that this is due to the transitioning of shifts.\(^{61}\) 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.\(^{62}\) Second, there is a cap on the lease price for both day and night shifts. Anecdotal evidence from the TLC and individuals in the industry suggests that these lease caps are binding. Given those rules, minifleets may try to equate the earning potentials for the day and night shifts, as a way to ensure they will get similar number of drivers willing to drive each shifts. A similar argument applies for owner drivers that might want to ensure they always find a driver for the second shift, which they do not drive themselves. Figure 19 shows the earnings for night and day-shifts under different hypothetical shift divisions. The x-axis shows each potential division-point, i.e. each point

\(^{61}\)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.

\(^{62}\)See section 58-21(c) in TLC (2011).
Figure 19: Earnings of Day and Night Shift for Different Split Times

\[ \text{Value of Shift} \]

\[ \begin{array}{c|c|c}
\text{Shift} & \text{day shift} & \text{night shift} \\
\hline
\text{Value} & \text{day shift} & \text{night shift} \\
\hline
$320 & \bullet & \\
$340 & \bullet & \\
$360 & \bullet & \\
$380 & \bullet & \\
$400 & \bullet & \\
$420 & \bullet & \\
\end{array} \]

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.

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

\[ ^{63} \text{Clearly this comparison ignores any equilibrium effects of changing the shifts structure. The graph can therefore be understood as the earnings that one deviating medallions could have under the current system.} \]