Frictions in a Competitive, Regulated Market
Evidence from Taxis

Guillaume R. Fréchette (NYU) Alessandro Lizzeri (NYU)

Tobias Salz (MIT) *

March 14, 2019

Abstract

This paper presents a dynamic 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, matching frictions created by the need for both market sides to physically search for trading partners, and second, regulatory limitations to entry. To assess the importance of these features, we use the model to simulate the effect of changes in entry, alternative matching technologies, and different market density. We use the geographical features of the matching process to back out unobserved demand through a matching simulation. The matching function exhibits increasing returns to scale, which is important to understand the impact of changes in this market and has welfare implications. For instance, although alternative dispatch platforms can be more efficient than street-hailing, platform competition is harmful because it reduces effective density.

1 Introduction

This paper estimates a dynamic 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

*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 familiarize us with it. Jean-Francois Houde, Myrto Kalouptsidi, Nicola Persico, and Bernardo S. da Silveira, an editor and four referees offered very helpful feedback. Milena Almagro Garcia provided excellent research assistantship. We gratefully acknowledge financial support from the National Science Foundation via grant SES-1558857.

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 market between traditional taxis and a dispatch platform such as Uber could lead to a reduction in market thickness that can worsen matching frictions in the aggregate.

Taxi services offer limited room for product differentiation. 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 New York, taxi drivers make 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, inherent matching frictions are present that create barriers to trade. Furthermore, most taxi markets are subject to licensing restrictions that limit entry and to regulations of fares.[1]

We first document some important patterns in this market that will then motivate some of the key ingredients of our model. We provide evidence that entry restrictions are strongly binding in NYC. If no other frictions were present, one might therefore expect all taxis to be utilized at least during the daytime. 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 shift-indivisibilities in labor supply.

Because of regulations, this market does not feature any price flexibility. This, inflexibility, together with capacity limits, implies that regular (and predictable) patterns of variation in demand for rides during the day (e.g., rush hours) lead to large fluctuations 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 16 minutes depending on the time of day. To appreciate the magnitude of these numbers, note the average duration of a trip is 12 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

[1] As an important component of the urban transport infrastructure, the functioning of the taxi market also has potential consequences for congestion and pollution.

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.[2]

In our model, drivers make daily entry and hourly stopping decisions. Licenses to operate a taxi (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. Starting (entry) decisions result from comparing 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 2011 and 2012. These data include every trip of the yellow cab fleet in this time span. The data entry of a trip includes the fare, tip, distance, and duration, as well as the geo-spatial start and end points of the trip. We have a panel identifier for the medallion as well as the driver which allows us to account for an important source of heterogeneity in drivers' characteristics (i.e., owner-operators vs. fleet drivers) that affects the intensity of utilization.[3]

Prior papers have used the taxi market as a useful environment to study labor-supply decisions because drivers have more flexibility in deciding when to stop than workers employed in firms. However, we show that important rigidities are tied to the two-shift structure of operation. Our model delivers responses to earnings shocks that strongly depend on their timing during the day. 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 1.8, which, interestingly, is not too far from the number (1.2) reported by Angrist, Caldwell and Hall (2017) from an experiment on Uber data. The latter results in estimates that vary from 0.9 to 2.6 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. Although we observe the number of matches, neither the passengers' wait time nor the number of hailing passengers is directly observable in the data. However, we are able to recover these variables 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 the number of matches, as well as values for search time for taxis and

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

[3]These identifiers are no longer available for more recent data from NYC.

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 we observe (successful matches), how long taxis search for passengers, the number of cabs on the street, and the speed at which traffic flows; we observe all of these variables 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.[4] 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 inputs to the matching function, that is, the number of waiting passengers and their wait time. Interestingly, the matching process displays varying degrees of returns to scale depending on the level of activity. At low levels of activity, such as during the nighttime, returns to scale are substantial. However, for daytime levels of activity, returns to scale become essentially constant. This feature of the matching process has important welfare consequences.

With the recovered demand data in hand, we proceed to estimate a demand function in terms of the expected wait time for a cab (recall that fares are fixed). We find that the demand elasticity with respect to wait time is not large, but demand is sufficiently responsive that this plays a significant role in the counterfactuals.

Our first counterfactual evaluates the effects of additional entry. A 10% increase in the number of medallions leads to 8.9% increase in the number of active taxis.[5] 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 reduce 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 en-

[4]See, for instance, Petrongolo and Pissarides (2001)

[5]This number is an average across the day for a typical weekday. The extent of the increase in activity depends on the hour of the day.

tire 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% more medallions, despite the fact that the number of medallions is left unchanged. The number of matches increases by 12%, a larger amount than the increase in the number of taxis. The difference is due to the fact that the dispatch system reduces matching frictions.

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.[6] We show that market segmentation on different platforms creates an inefficiency that is due to a reduction in market thickness for both platforms.[7] Partial coverage by a dispatcher has two effects compared with a market with no dispatcher: on the one hand, at the market sizes we consider, the partial dispatcher is a more effective platform for taxis and passengers that are served by it; on the other hand, segmentation of the market makes both segments thinner, with the consequence of longer average distances between a random taxi and a random passenger. When we consider the case in which there is an equal number of potential taxis divided between dispatch and decentralized platforms, we find 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 nighttime 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 one third as dense; that is, the same number of potential passengers is spread over a larger territory covered by the same number of potential taxis. We find 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.

The insights that we gain from our counterfactuals are of broader relevance than the taxi market. Entry restrictions are related to the issue of occupational licensing, which affects a large number of workers in the United States. For instance, Kleiner and Krueger (2013) report that 29% of workers are subject to licensing regulations. Our results also speak to the effects of shift indivisibility and are therefore relevant to the literature studying the flexibility of work arrangements.[8] Our results are directly relevant for thinking about search and

[6] In order to isolate this effect we keep the total number of cabs the same and only change the dispatch process.

[7] Uber seems to be well aware of this effect and therefore subsidizes drivers, especially when they first enter a city. See, for instance, https://www.theguardian.com/technology/2016/apr/27/how-uber-conquered-london.

[8] See, for instance, Chen et al. (2017), who discuss the value of flexibility for Uber drivers. See

matching frictions, which have been greatly emphasized in labor markets and housing markets. In this respect, our paper is the first to offer an explicit simulation of the matching process as deriving from spatial matching frictions. This spatial simulation could be a useful metaphor for other markets in which matching frictions are important.[9] Finally, our results on competing dispatch platforms speak to the literature on network externalities and trading platforms.[10]

2 Related Literature

This project combines elements from the entry/exit literature, neoclassical labor-supply models, and search. Structural estimation of entry and exit models goes back to Bresnahan and Reiss (1991). This entry/exit perspective on the problem is motivated by the fact that drivers in the New York taxi industry, like in many other cities, are private independent contractors and decide freely when to work subject to the regulatory constraints. The labor-supply decisions of private contractors have, for example, been studied in Oettinger (1999), who uses data from stadium vendors. In the spirit of the entry/exit literature, we recover a sunk cost, which in our case is 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 prior work in Industrial Organization is that our market contains tens of thousands of entrants. Entry is therefore competitive and entrants 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, 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, Kalouptsidi and Papageorgiou (2017) consider the related application of exporters matching to ships. They develop and estimate a model that 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), who calibrates an equilibrium model (with frictions) of the taxicab market, and includes some heterogeneity among locations. However, he assumes

also Mas and Pallais (2017).

[9] See Eckstein and Eckstein and Van den Berg (2007) for a survey on empirical work on labor search, and Petrongolo and Pissarides (2001) for a survey on the matching function. On the housing market, see Wheaton (1990)

[10] Cantillon and Yin (2008) discuss competing trading platforms in financial markets; Hendel, Nevo and Ortalo-Magné (2009) discuss the role of competing platforms in the housing market.

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.[11] Using the model, he quantifies the impact of policies increasing fares and the number of medallions.

Buchholz (2018) 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 (2018) relies on spatial variation in fares due to the two components of fares, namely, a fixed fare and a fare that is variable in the travel distance, to estimate demand as a function of price. By contrast, absent aggregate intertemporal variation in fares, we allow demand to depend on the expected passenger wait time. Although 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, although Buchholz (2018) can explore counterfactuals with respect to fare structure, our paper can investigate the relaxation of entry restrictions and the extensive margin of supply more broadly.

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 argue that this finding is inconsistent with neoclassical labor-supply analysis. This interpretation has been challenged by Farber (2008). Crawford and Meng (2011) estimate a structural model of a taxi driver's stopping decision, allowing for a more sophisticated version of reference-dependent preferences. They do not consider a cab driver's entry decision, and do not analyze the industry equilibrium. We opted to stay within the neoclassical framework to study the overall 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 some drivers might not fit this assumption, the aggregate patterns are consistent with a standard model. Hence, a standard model of labor supply seems to be a reasonable starting place.[12] 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.[13]

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

[12] 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.

[13] Other papers that use these data are less directly related. For instance, Haggag and Paci (2014) study the impact of suggested tips on the NYC taxi-driver payment screen for clients on the realised tip. Haggag, McManus and Paci (2014) study how taxi drivers learn driving strategies based on their experiences. Buchholz, Xu and Shum (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 than we do: they develop a new closed-form estimator for the stopping problem. Thakral and Tô (2017) find reductions in cab

3 Industry Details and Data

3.1 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.[14] This regulation 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.[15] The cab market is regulated by New York's Taxi and Limousine Commission (TLC), which sets rules such as the fare drivers can charge, the qualifications for a taxi-driver license, the insurance and maintenance requirements, and restrictions on the leasing rates (daily or weekly) medallion owners can charge drivers.

Approximately 40% of the medallions, are owner-operated and require that the owner of the medallion drive the taxi for at least 210 shifts in a year. The remaining 60% of medallions, are called minifleets, and are operated by approximately 70 fleet companies that manage an average of 115 taxis each.[16] Fleet companies therefore manage many medallions and rent taxis out to drivers on a daily or weekly basis.[17] 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 allowing for heterogeneity among ownership types in our estimation is important.

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 12 consecutive hours, and a weekly lease must last seven consecutive days. Minifleets must operate their cabs for a minimum of two nine-hour shifts per day every day of the week.[18] The TLC also specifies a cap on the price medallion owners can charge that varies with the time of the lease and

driver 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. Hall, Horton and Knoepfle (2017) use fare changes on Uber to explore short run and long run labor supply responses. 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.

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

[15] The TLC recently established an ability to hail cabs via a a mobile app, but this option was not available during our observation period.

[16] Some fleet owners operate more than 1,000 taxis. Source: http://www.nycitycab.com/Services/AgentsandFleets.aspx

[17] 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.

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

the type of vehicle.[19]

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. It also records 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 which includes roughly four months: October 1, 2011, to November 22, 2011, and August 1, 2012, to September 30, 2012. 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 these data with information about the medallion type (minifleet or owner-operated), and the vehicle type.

We focus our analysis 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 doing so would make obtaining counterfactuals computationally prohibitive, and, as we will see, these days display very similar patterns. Furthermore, in the estimation, we focus on trips that originate in Manhattan, which account for most of the activity in the data (approximately 90% of trips).[20]

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, most taxi markets are subject to tight entry restrictions. In NYC, the number of medallions during the time-period of our data is 13,520. This number 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

[19] The leasing-rate caps are between $115 and $141 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.

[20] Most of the remaining trips originate at one of the two NYC airports.

quantitative importance of the entry restriction.[21] 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 some natural failure rate of vehicles and other idiosyncratic reasons explain why taxis may fail to be utilized, this seems to be a very intense rate of utilization. As we will see next, the daily utilization rate is quite different from the utilization rate for a typical hour, which is much lower and also depends on drivers' hourly stopping behavior.

4.2 Daily Patterns of Activity

subsection 4.2 displays the fraction of taxis that are active at each hour of the day for a typical weekday, distinguishing whether it is a minifleet or an owner-operated medallion.[22]

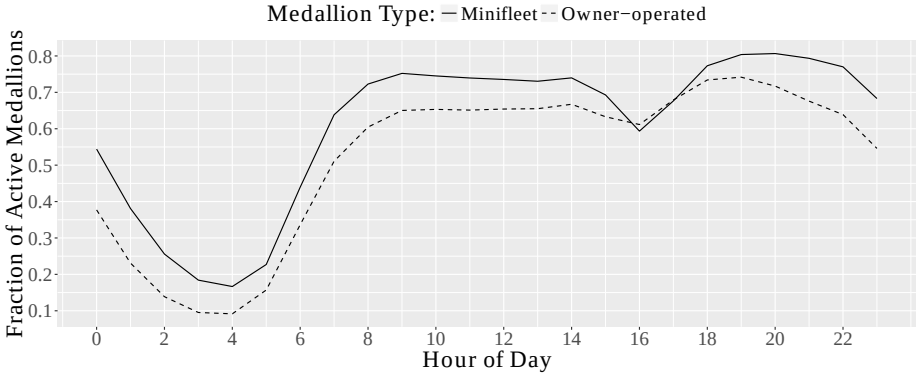


Figure 1: Comparing Activity of Owner-operated and Fleet Medallions

We wish to draw attention to several features of this figure. First, fleet medallions are more intensely utilized for every hour of the day.[23] To place the magnitude of the difference in utilization in perspective, this differnce is larger than the difference between Uber and taxis reported by Cramer and Krueger (2016). Second, substantial intra-day variation in activity exists, 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.[24] Third, a large

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

[22] At most one owner-operated medallion can be owned at any time. However, owners can lease the taxi to other drivers as long as the owner operates the taxi for a minimal number of shifts.

[23] In Appendix G we discuss additional facts on the difference in utilization between fleet medallions and owner-operated medallions.

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

reduction in activity occurs precisely during the evening rush hour. This time 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. We have also explored these in differences in counterfactual computations where we allow all medallions to be operated by mini-fleets. Such a change would lead to similar improvements in consumer welfare as a 10% increase (the same as in our entry counterfactual) in the number of medallions (see Table 13).

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. subsection 4.3 describes the fraction of time an average

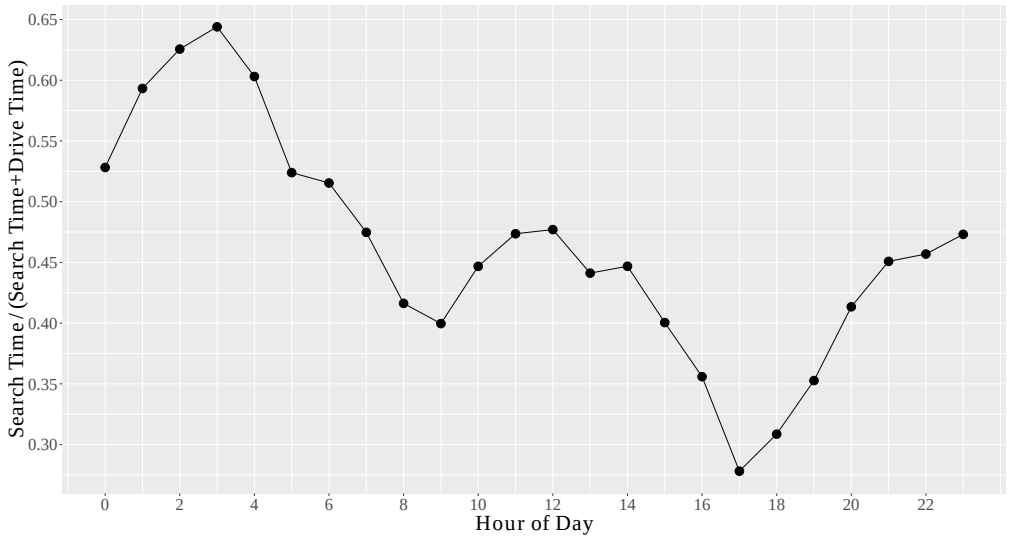
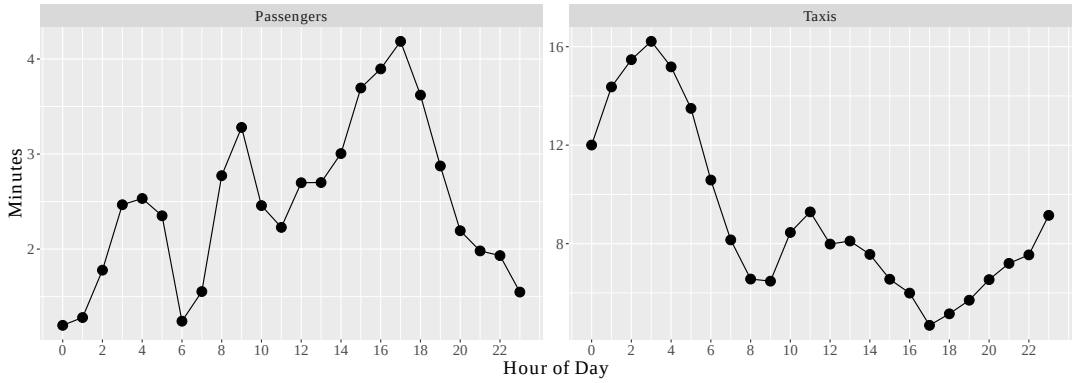


Figure 2: Search Time Relative to Delivery Time during the Day

Notes: This plot shows the ratio of the average time a taxi spends searching for a passenger as a percentage of total time spent driving. 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.

taxi spends searching for passengers relative to the total time it is active, that is, the unemployment rate for taxis.



Notes: The left panel shows search time for taxis in minutes, averaged for each hour of the day. The right panel shows waiting time, also in minutes, for passengers as recovered by our simulation, again averaged for each hour of the day.

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

Two notable features are the following. First, the fraction of time taxis spend searching is almost never lower than 30% and displays substantial variation during the day, going as high as 65% at its peak. 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.[25] The low point of search time is reached at 5PM, shortly after the shift change, when many medallions are still inactive. The high point of search time is during the late night hours. For passengers, the wait time increases at night relative to the late evening hours but is still lower than during the peak of daytime values (the wait time for passengers is inferred from our simulation, which is described in detail later). We will see that the negative relationship between search time and wait time during much of the day is mostly due to changes in the ratio of passengers and cabs. However, the fact that both search time and wait time increase during the night relative to the evening hours, 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 the observed drop-off location and the subsequent pick-up location of a cab (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 1,500 randomly selected drop-off and pick-up locations from our data. For each of these observa-

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

tions, 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 traveling between these points. The final piece of evidence for

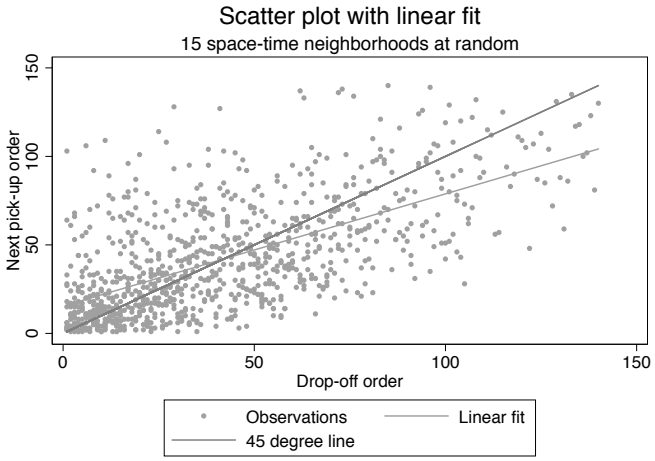


Figure 4: Drop-off Order Against Pick-up Order

search frictions emerges from examining the degree of predictability of the order in 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 with 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. Figure 4.3 displays the result of this exercise, along with the 45 degree line (steeper line and the linear fit (flatter line). Although we find that the drop-off sequence and the pick-up sequence are positively correlated (correlation coefficient is 0.51), a large amount of dispersion is still present. We interpret this dispersion as evidence of search frictions.[26]

5 Model and Estimation

Because 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

[26] As a placebo we also conduct the same analysis at JFK airport, where cabs are queuing for passengers. Under such a system one would expect drop-offs and pick-ups to be much more correlated. Here we find a correlation coefficient of 0.73, which statistically significantly different from the one above. The reason that the correlation at JFK is not even higher is that we must pool drop-offs at different terminals and that there are exemptions from queuing for drivers that were previously servicing short trips from the airport.

a taxi and the search time s_t for taxis to find a passenger. There are three key ingredients in our equilibrium model of the taxi market. First, hourly taxi supply will be obtained as a result of an optimal dynamic entry and stopping process for drivers and determined by outside options and hourly earnings, which themselves are largely determined by s_t. Second, hourly taxi demand is assumed to take a log-linear form as a function of hourly wait time w_t. Furthermore, we do not observe the number of passengers d_t or wait time w_t directly. We infer these variables from our observations of the number of matches, the number of searching cabs and the search time, given the third key ingredient: the matching process. This process is captured by a function $g(\cdot)$ that relates the number of taxis and passengers, as well as other observable market conditions to wait time and search time. We provide a simulation based approximation for this matching function that replicates the geographic nature of the matching process and features of the taxi market in NYC.

5.1 Demand-Side Model

The estimation of the demand function presents two distinct challenges. The first problem is that, although we observe the number of matches between passengers and taxis, we do not directly observe either the number of waiting passengers d_t (demand) or their wait time w_t(price). A fraction of this demand may not be fulfilled because some passengers may go unmatched. 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 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 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 times, $s_1 > s_2$. If all other relevant factors (e.g., 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 passengers are distributed across the areas of the city according to proportion $\{p_i^d | i = 1, ..., I\}$.[27] We denote the vector of waiting passengers $\mathbf{d}_t = (d_t \cdot p_1^d, ..., d_t \cdot p_I^d)$.

The matching process is captured by a function g that maps a vector of waiting passengers $\mathbf{d}_t$ and searching taxis $\mathbf{c}_t$=$(c_{1t}, ..., c_{It})$, as well as other exogenous time-

[27] These proportions are inferred from the data; see Section 5.1.1 below for more details.

varying variables ϕ_t, into an aggregate search time s_t and wait time w_t:

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

If we knew $g(\cdot)$, then, for given values of $\mathbf{c}_t$ and ϕ_t, inverting this function would allow us to infer d_t, assuming, of course, 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.[28]

5.1.1 Implementing the Simulation

We assume 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 $4/20th$ of a mile wide (east-west) by $1/20th$ of a mile long (north-south), which corresponds to the approximate block size in Manhattan. subsubsection 5.1.1 shows the structure of the resulting grid. Gray nodes represent intersections between streets and avenues, at which cabs can change their direction of travel. Turns at (gray) nodes are random with equal probability for each feasible travel direction.

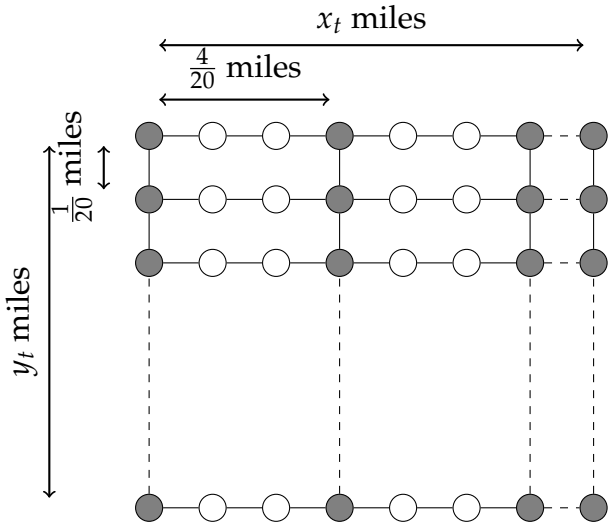
We assume the effect of time-varying factors can be summarized by the speed, mph_t, at which the traffic flows and the average distance, $miles_t$, to deliver a passenger from his fixed position on the grid to the destination.[29] Both factors are included in ϕ_t, and both are directly observed in our data by using the average hourly speed of the entire taxi fleet as well as the average distance of all trips on an hourly basis. For each combination of variables we feed into g, we simulate the resulting average wait 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.[30]

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

[28]Our matching function is effectively deterministic. While we could have allowed for randomness in the matching process, this feature does not seem important in the aggregate given the large number of matches at the hourly level. This, of course, does not mean that the components of the matching function, such as demand and supply, are deterministic. In fact, we will allow for error terms in the demand function, and it is important to do so.

[29]Note the trip distance is an important ingredient because it determines the effective supply. Longer trips mean fewer taxis are available for the delivery of passengers.

[30]Our results are robust to changing the maximal value of the waiting time. See subsection D.2 in the Appendix for more details.



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

Figure 5: Schematic of the Simulation Grid

parts of the city varies, we divide Manhattan into eight equally spaced areas (see Appendix D in Appendix B for details).[31] Passengers appear on the corresponding parts of the grid in proportion to the observed pick-up probabilities of those areas and cabs reappear according to the observed drop-off probabilities. In other words, for each of the passengers placed on the map, we first randomly determine an area according to a multinomial distribution with probabilities $\{\hat{p}_i^d | i = 1, ..., 8\}$ and then a node within an area, where each node has equal probability. The probabilities $\{\hat{p}_i^d | i = 1, ..., 8\}$ are estimated as the fractions of trips originating in these areas. Similarly, cabs re-appear on the map in area i according to multinomial probabilities $\{\hat{p}_i^c | i = 1, ..., 8\}$ and with equal probability on each node within an area. The probabilities $\{\hat{p}_i^c | i = 1, ..., 8\}$ are measured as the frequencies of drop-offs in those areas.[32] In subsection D.1, we discuss the robustness of our inversion results to alternative (finer) divisions of the map. Although our approach allows us to account for some of the heterogeneity in locations, we do not consider endogenous location choices as in Buchholz (2018).

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 $g()$ for any point in its do-

[31] To 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.

[32] We therefore do not consider conditional drop-off and pick-up probabilities. These probabilities 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.

main.[33] 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 D.

5.1.2 Properties of the Matching Function

Figure 6: Graphical Illustration of Matching Function

subsubsection 5.1.2 graphically illustrates properties of the matching function. The figure on the left shows what happens to wait time and search time as the market becomes thicker. On the horizontal axis, the number of taxis and passengers vary while keeping constant the ratio between the two. Both passenger wait-time and search-time decrease as the market becomes thicker, but the

[33] Due to computational limitations, performing this simulation for each point in the domain of g is not feasible. For example, if we assume that in an hour there are at most $70,000$ passengers waiting, and multiply this number by the maximal number of medallions, we would already obtain 945 million different points in the domain. Furthermore, there is variation in ϕ. We therefore simulate g for a lower number of grid points, and we interpolate linearly between those points to obtain the image for points in between. For each of the four independent variables of $g(.)$, we pick eight different evenly spaced grid points. Because the outcome of the matching process is random, we have to repeat the simulation multiple times for each of those points. In practice, we have found that the average of these simulations does not change much after 10 iterations, which is therefore what we use to produce this average.

improvements decrease relatively quickly. For the numbers observed during the late hours of the night, the returns to additional thickness are substantial. By contrast, at the scale corresponding to the average number of daily taxis in the data (approximately 7,800 when averaging over the entire 24 hours), additional returns to market thickness are fairly small, especially regarding search time. For a city with low density, however, increasing returns would be significant at all hours of the day.[34]

All of this richness, which arises naturally from the physical structure of the matching process, would be missed by assuming the standard matching function. In particular, the variation in returns to scale across different times of the day might have easily been missed under a parametric specification. Because density is a crucial feature of this market, and an important ingredient in evaluating the impact of various changes in the market as discussed in our counterfactuals, we view our departure from standard parametric assumptions as a strength of our approach.

The figures on the right give a sense of how market *tightness* affects outcomes for taxis and passengers. In 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 a traffic speed that is one standard deviation above the median. Although a faster traffic speed is unequivocally good for passengers, two contrasting effects exist 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 less than 3,000 cabs.

Figure 5.1.2 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 that is inferred from our demand recovery. The passenger figures confirm an expected pattern of strong rush-hour demand in the morning and evening hours on all weekdays. We find that demand is lower on Sundays than during weekdays, and no clear division exists between morning and evening rush hours. This,

[34]Regarding the number of matches, we consider a 10% increase for both taxis and passengers. If this increase is applied to the numbers we observe late at night, (2,000 cabs, 5,000 passengers), the resulting number of matches increases by 11%, so that the matching function displays modest increasing returns. When instead we scale it to something closer to the daytime numbers (10,000 cabs, 25,000 passengers), then we obtain a 9.8% increase in the number of matches, implying, if anything, a slight level of decreasing returns for matches at these higher values.

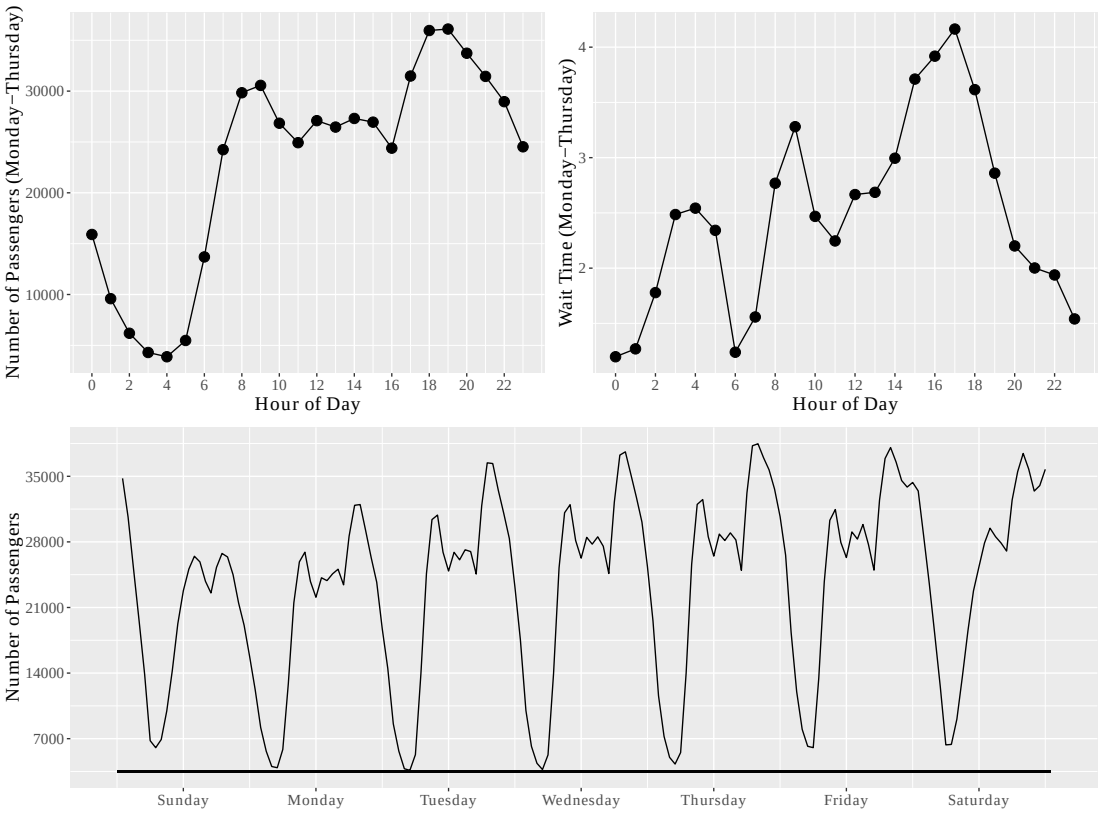


Figure 7: Graphical Results of Demand and Wait Time

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 and 10AM. This finding stands in contrast to the peak in wait time in the evening that occurs at 5PM, before the spike in demand occurring at 7PM. We believe that the reason for this pattern is the coordinated shift change (witching hour) that spills over beyond 5PM and leads to the more unfavorable ratio of active cabs to searching passengers.

To provide additional validation for the results of our demand recovery, we consider the effect of weather shocks, one of the important demand shifters in this market. Appendix F shows that our inferred wait-times and demand are both highly correlated with rainfall shocks, an external source of data that was not used in the initial demand recovery.

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 do not explicitly model the passenger choice problem among modes of transportation or whether to travel at all.[35] Our demand function has a similar interpretation to the standard case: some consumers will not travel by taxi because the realized price (wait time) is too high, whereas others do choose to travel. We assume a constant elasticity demand function of the following form.[36]

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

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

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

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

[35] This absence of choice model is not an issue for the counterfactuals that we consider. We do discuss how demand is related to determinants of outside options such as subway functioning and weather.

[36] Because we break up the data at the hourly levels and only use a subset of weekdays, we are left with slightly more than 1,500 observations. The assumption of log-linearity is not unusal for this type of problem; see, for example, Kalouptsidi (2014).

that shift demand will directly affect w_t. Furthermore, drivers may condition their decisions on factors included in the error term ξ_t, which would lead to a decrease in wait time. For both of these reasons, the error term ξ_t and the wait time w_t may be correlated. This correlation would, of course, introduce a bias in the estimation of η. 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.

Table 1: Demand Estimation

Dependent Variable:	$log(w_t)$	$log(d_t)$
Shift Instrument	0.179**	
	(0.00762)	
$log(w_t)$		-1.225**
		(0.0601)
Observations	1531	1531
2-Hour FE	Yes	Yes
R^2	0.572	0.637

Note: $+\ p < 0.10$, $*\ p < 0.05$, $**\ p < 0.01$. All regressions are based on our subset of 2011-12 TPEP data, excluding Fridays, Saturdays and Sundays. An observation consists of an hourly average over all trips in that hour. Standard errors clustered at the date level.

We make use of the shift change (or witching hour) as a supply shifter. In Appendix I, we argue that the shift change is timed so that the day shift and the night shift have similar returns on average. This equalization ensures both shifts are attractive to potential drivers.[37] Thus, the timing of the shift change is not chosen to accommodate hourly demand shocks.[38]

[37] Because fleets face caps on the lease rates they can charge to drivers, an excess supply of drivers in one shift does not lead to a higher ability to charge more for leasing a taxi during that shift. Additional explanation is provided in Appendix I.

[38] The shift changes occur from 5AM to 7AM and 3PM to 5PM. We have also experimented with another instrument: the lagged number of active taxis (every hour up to 6 hours before). 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

Although we do not consider an explicit model of consumer choice, we have made an effort to validate our demand estimates with data on subway rides, which is the other major form of transportation in Manhattan.[39] Because the data on subway rides are only measured at four-hour intervals, we aggregate data at a daily level. For the validation, we compute the expected number of subway rides conditional on weekday, month, and year fixed effects and we consider the residuals of these regressions as measure departures from routine subway travel conditions. We take two approaches to verify the robustness of our results. First, we include these residuals as controls in our specifications. This inclusion has no discernible effect on our results. Our second approach uses the idea that the left tail of the residual distribution from this regression can be interpreted as a shortfall in subway rides, possibly due to delays or breakdowns. Conditioning our demand estimation on the data corresponding to the lowest 20% of residuals, we find that, as expected, at these times, demand is slightly less elastic, but that the average elasticities (computing these regressions on both subsets and then taking the weighted average of the two elasticities) is very similar to the one we obtain without conditioning.[40]

Table 5.1.3 shows what demand would be if the wait time were held fixed at the daily mean throughout the day. This finding highlights how the inelastic portion of demand is varying throughout the day. In other words, it is a representation of predictable hourly variation in demand. All else being equal, demand would be highest in the evening hours between 5PM and 7PM and lowest in the morning hours from 12AM to 6AM.

5.2 Supply-Side Model

Agents make their decisions in discrete hourly intervals; for each time point t in our data, h_t denotes the hour of the day from the set $H = \{0, ..., 23\}$. We interpret h_t as representative hours of a weekday. At each time point t, N_t medallions are active and M_t are inactive summing up to the total number (13,520) of medal-

to remain on the shift later even if subsequent hours are not particularly profitable. From this instrument we get an elasticity of -0.7 ($p < 0.001$). Although our main instrument is coarser, we opted not to focus on this one as the shift change instrument seems more clearly exogenous.

[39] In related work in progress Buchholz et al. (2019) report hourly demand elasticities with respect to wait time. Their data allows them to observe trip level waiting times and therefore allows them to obtain clean estimates of hourly elasticities. For the daytime, the values range between -1.01 and -0.83 which is not a lot of intraday variation and are relatively close to the estimates we find with our shift-change instrument.

[40] Estimating the specification of Table 1 on the subset of days with the lowest 20% of residuals from the subway regression reduces elasticities from -1.225 to -1.19. For the purpose of our counterfactuals, we are interested in the average elasticity in a stationary market environment, which is almost unaffected (-1.224) by the omission of other transportation options. We have also considered other values for the cutoff of the residuals and results are not very sensitive to this definition.

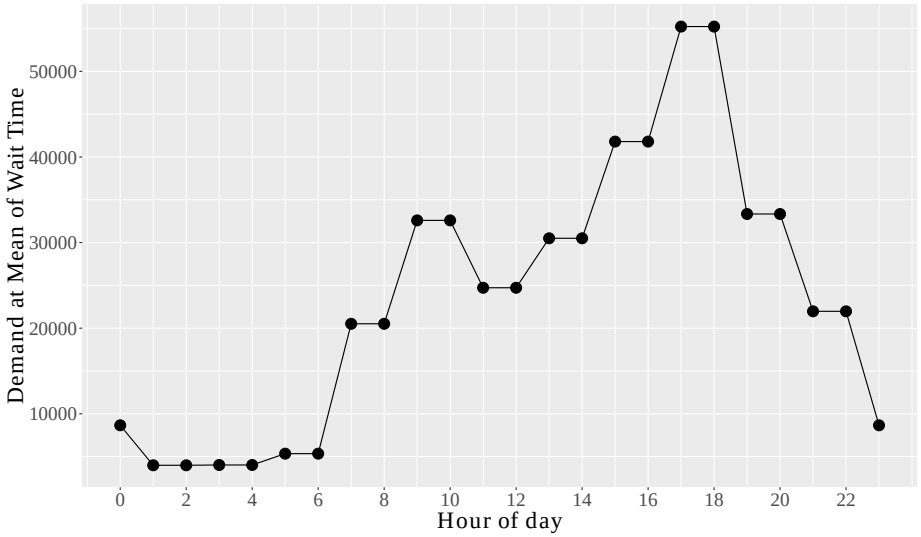


Figure 8: Demand Function Evaluated at the Mean of Wait-Time

lions 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. At each hour, active drivers choose whether to stop or continue driving. As mentioned above, we focus on Monday through Thursday.

5.2.1 Accounting for Observed Medallion Heterogeneity

The model incorporates the main regulatory and organizational constraints imposed by the medallion system that we discussed in section 3. In section 3 and section 4, we highlighted the fact that minifleet medallions are more heavily utilized than owner-operated medallions, and are also slightly more likely to transition between shifts at 5PM. For a driver i, the index $z_i \in \{Fleet, Own\}$ denotes the ownership type of the medallion. The model parameters are allowed to differ by ownership type.[41]

In addition to ownership-type heterogeneity, the model incorporates the observed heterogeneity in the shift transitions. We capture this heterogeneity in the model by first classifying medallions according to a coarser set of transition patterns than what is observed in the data. We allow each medallion to have one of four morning transition times (3AM-6AM) and one of four evening transition times (2PM-5PM). Thus, there are in total 16 possible shift-transition times. Each medallion is then assigned to the transition time that most closely resembles its

[41] We could alternatively impose that cost parameters are the same for minifleets and owner-operated medallions, but minifleets have a deeper pool of drivers to draw from. This is essentially an equivalent problem and we expect that our results would be robust to this alternative approach.

transition pattern. We index medallions according to their group ($k_i \in K$). The relevance of these transition times is the following. We know that drivers have to pay fines if they return the car after the designated end of their shift. These fines exist to ensure drivers do not operate the cab longer than contractually specified. Because fines are not directly observable to us, we estimate them as parameters. The transition types determine at what hours a driver of a particular medallion has to pay the fines, should it drive during the hours in which fines are active.

If, for example, the most common transition time in the morning from night to day-shift (as measured by the modal transition hour) is 5AM, we assume a driver has to pay a fine for handing in the medallion later than this.[42]

5.2.2 Optimal Stopping for Active Drivers

We first discuss the optimal stopping decision for a driver who is already on a shift, and then we discuss the decision to start a shift. At the beginning of each hour, the driver decides whether he wants to collect the flow payoff from driving plus the continuation value of an active shift, or the random value of an outside option.[43] The state vector is given by $\boldsymbol{x}_{it} = (h_t, l_{it}, z_i, k_i)$, which includes the hour of the day, h_t, the number of hours the driver has been on a shift, l_{it}, as well as an idiosyncratic unobservable vector, $\boldsymbol{\epsilon}_{it} = (\epsilon_{it0}, \epsilon_{it1})$, assumed to be distributed according to i.i.d. $T1EV$ distributions with scale parameter σ_ϵ.[44] Note h_t as well as l_{it} evolve deterministically and that z_i and k_i do not change over time. Therefore, the only random component is $\boldsymbol{\epsilon}_{it}$.

We specify a cost of driving, $C_{z_i,h_t}(l_{it})$, which is a function of l_{it} and therefore allows for a rising dis-amenity value as the shift progresses.[45] 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

[42] Transition times are in principle endogenous, and one could imagine these times 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 I, the current regulatory arrangements limit the flexibility of transitions. Furthermore, as we show in Figure A, the transition types we specify account for a large fraction of observed transition patterns.

[43] We assume no discounting because decisions occur hourly. In the estimation, the maximal shift length that we allow is 13 hours, which is longer than the regulatory maximum of 12 hours and is surpassed in the data only in very few cases.

[44] Thus, we do not include any persistence in these hourly shocks. The model could be modified to incorporate persistence but, as we will see below, this additional generality does not seem required in order to match the daily aggregate patterns of activity.

[45] Also note the cost function absorbs the fuel costs, but does not include time varying components. We do not incorporate data on fuel expenses because there is little variation in fuel costs in our sample. We have verified that in our sample there is no significant relation between taxi activity and fuel costs.

driving. We assume the cost function takes the following form:

$$C_{z_i,h_t}(l_{it}) = \lambda_{0,z_i,h_t} + \lambda_{1,z_i} \cdot l_{it} + \lambda_{2,z_i} \cdot l_{it}^2.^{46}$$

The fixed components, λ_{0,z_i,h_t}, take one of four possible values for each type z_i, depending on the time of day. The term $f(h_t, k_i)$ is a fine that has to be paid in the event that the driver returns the taxi after the transition time; this time is obtained from the classification into types that we describe above.[47] 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.[48] The fine is indexed by z_i to account for the fact that owner-operated medallions seem to have less stringent transition times. We denote by h_{k_i} the set of hours for which a driver of medallion k_i is subject to the fine when driving. We therefore have: $f_{z_i}(h_t, k_i) = f_{z_i} \cdot \mathbf{1}\{h_t \in h_{k_i}\}$. Putting all these ingredients together, the value function for a driver conditional on state $\boldsymbol{x}_{it} = (h_t, l_{it}, z_i, k_i)$ is given by:

$$V(\boldsymbol{x}_{it}, \boldsymbol{\epsilon}_{it}) = \max\{\epsilon_{it0}, \pi_{h_t} - C_{z_i,h_t}(l_{it}) - f(h_t, k_i) + \epsilon_{it1} + \mathbb{E}_{\boldsymbol{\epsilon}_{i(t+1)}}[V(\boldsymbol{x}_{i(t+1)}, \boldsymbol{\epsilon}_{i(t+1)})]\}.$$

As is well known, $T1EV$ distributions for the error terms lead to closed form expressions for the choice probabilities. We obtain the conditional stopping probabilities as:

$$p(\boldsymbol{x}_{it}) = \frac{exp\left(\frac{1}{\sigma_v}\right)}{exp\left(\frac{1}{\sigma_v}\right) + exp\left(\frac{\pi_{h_t} - C_{z_i,h_t}(l_{it}) - f(h_t,k_i) + \mathbb{E}_{\boldsymbol{\epsilon}_{i(t+1)}}[V(\boldsymbol{x}_{i(t+1)}, \boldsymbol{\epsilon}_{i(t+1)})]}{\sigma_v}\right)}.$$

5.2.3 Starting Decision for Idle Medallions

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, the

[46] We have also explored a specification with higher order polynomials, but it does not affect the results.

[47] Because shift transitions are approximately 2-hours long on average, the fine is imposed for those two hours.

[48] We view the fines as a reduced form of a potentially more complicated set of sanctions without explicitly modeling the many possible arrangements between drivers and owners. Unfortunately, we are not able to observe whether there are private arrangements to deviate from the regular transition patterns. We elaborate on the role of these fines when we discuss identification.

medallion will be available to another potential driver the following hour. We assume the utility from the outside option consists of a fixed value μ_z that depends on the medallion type and on whether it is a day or night shift, as well as an idiosyncratic random component v_{it0}. The utility of driving depends on the entire expected value of a shift (described below) as well as an idiosyncratic component v_{it1}. We assume v_{it0} and v_{it1} are i.i.d. random variables distributed according to a type 1 extreme value ($T1EV$) with scale parameter σ_v.[49] 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.[50] To summarize, the utility of the outside option is given by

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

and the utility of starting a shift is given by:

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

Denoting by $q(\boldsymbol{x}_{it})$ the probability that an inactive driver starts a shift at time t, conditional on state $\boldsymbol{x}_{it}$, we obtain:

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

5.2.4 Equilibrium Definition

Hourly earnings, π_{h_t}, are determined by the equilibrium distributions of active medallions, c_{h_t}, and searching passengers, d_{h_t}.[51] The number of actively searching cabs and passengers (c_{h_t} and d_{h_t}) determine wait times, w_{h_t}, and search time, s_{h_t} through $g(\cdot)$.

Definition 1 *A competitive equilibrium in the taxi market is a set: $\{s_{h_t}, w_{h_t}, c_{h_t}, d_{h_t}, \pi_{h_t} : h_t \in H_t\}$, such that:*

1. *d_{h_t} results from the demand function $d(\cdot)$ under the wait time w_{h_t}.*

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

[50] Recall that the sample period precedes the time when Uber entry became substantial in NYC. By 2015, rental rate caps were likely no longer binding. In our discussion of counterfactuals, we discuss how the endogenous determination of lease rates may change our results.

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

2. *s_{h_t} and w_{h_t} result from d_{h_t} and c_{h_t} under the matching function $g(\cdot)$.*

3. *c_{h_t} results from optimal starting and stopping under π_{h_t}.*

4. *π_{h_t} results from s_{h_t}.*

The idiosyncratic uncertainty on the supply side, such as the random outside options, averages out across the large number of taxis.[52] We do not allow for autocorrelation in the shocks or earnings. In fact, it turns out that most of the variation in demand and wait time is explained by hourly fixed effects. The same is true for search time.

5.3 Identification of Supply-Side Parameters

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

We need to identify six main objects: **(1)** 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 of 13 hours, the continuation value is zero; thus, the driver only compares expected income in that last hour against the cost of driving. The value of the cost function for l_{max} is therefore determined to match the expected 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 after conditioning on shift-length and earnings. For example, the stopping probability increases sharply after 12PM, even though no contemporaneous sharp decline in earnings occurs. This kind of variation in the data identifies the differences in the values of λ_0.

(3) is identified by the increase in the stopping probabilities at shift-transition times independently of the length of time a driver has been on a shift. While

[52] Since we are interested in aggregate market condition, over the entire city, we use the simple concept of a competitive equilibrium. This distinguishes our work from settings with many firms where firms still have some market power, in which case an equilibrium concept such as Weintraub, Benkard and Van Roy (2008) would be more appropriate.

drivers' shift lengths vary, the arrangements with other drivers are very consistent, and transitions occur mostly at the same hours. This is the data feature that we would have difficulty explaining without the fines. For instance, we frequently observe that some drivers start their day shift late, around 12:00, let's say, and then hand over the car to the next driver at four before the night shift starts. If stopping times were purely rationalized by shift length, this pervasive data pattern would not be replicated by our model. In aggregate, our model would predict that too few people stop at the hours 15:00, 16:00, etc. To illustrate this point further, Appendix J shows that stopping probabilities rise during 3AM-6AM and 2PM-5PM even for short shifts.

(4) is identified by the variation in 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 $\mathbb{E}_{\boldsymbol{\epsilon}_{i(t+1)}}[V(\boldsymbol{x}_{i(t+1)}, \boldsymbol{\epsilon}_{i(t+1)})] - r_{h_t}$, which is composed of data and identified objects, identify the different values of μ_{h_t} (**5**) and their dispersion, that is, the value of σ_v (**6**). In Appendix J we show the conditional choice probabilities.

5.4 Estimation

5.4.1 Constructing the Data for Supply-Side Estimation

To estimate the model, the trip based TPEP data have to be transformed into a shift dataset in which the unit of observation is a medallion-hour combination.

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 the interval between two trips is long but shorter than five hours. We do not model breaks: instead, we call a break a gap between a dropoff and a pickup that is longer than 45 minutes. We assume breaks to be an exogenous process, we estimate the likelihood of a break for each hour conditional on the state, and compute hourly earnings as the expected wage that is earned while searching for passengers, multiplied by the probability that the driver is not on a break.[53] 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 these data we drop shifts that are only one hour long, which make up less than 0.3% of the active shift data.[54]

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

[53] In subsection D.3, we report that our results are robust to alternative definition of breaks.

[54] These disparate hours might be part of an interrupted longer shift and may not be captured by the shift definition used here.

actual hourly based rate π_t^0 for each trip by dividing the total fare of each trip by the duration of a trip. These rates also include the tip that drivers earn. Because tips are only recorded for credit card transactions, we impute tips for trips that have been paid in cash.[55] 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. The length of the ride e_t determines together with search time the effective number of taxis searching on the street as the ratio of search versus delivery time. If the average ride length in an hour is e_t, the average search time s_t, and there are c_t active cabs on a shift then the number of searching taxis is determined as $c_t \cdot (s_t/(e_t + s_t))$. Based on these hourly averages we can then compute a realization of the hourly wage rate as $\pi_t = \pi_t^0 \cdot (e_t/(e_t + s_t))$, that is, the actual hourly rate times the fraction of the time the driver is delivering a passenger as opposed to searching. Note that s_t is the search time and an endogenous variable, which will adjust in our counterfactual computations according to demand, supply and the matching process.

5.4.2 Estimation Procedure

For the estimation of supply-side parameters, which we denote by θ, 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 π_t. In other words, because we observe equilibrium earnings directly in the data, we do not 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).[56]

In our case, constraints are derived from the assumption that the data are generated by a model of decentralized optimal starting and stopping decisions. 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.[57]

[55] We first run a regression with hourly dummy variables predicting the tip rate for each hour of the day. We then use predicted rates to impute the tips for trips for which the tip is not observed. About 47% of all transactions are paid by credit card.

[56] Su and Judd (2012) demonstrates the computational advantage of MPEC over a nested fixed point computations (NFXP) in the classical example of Rust (1987) bus-engine-replacement problem. Applications of MPEC to demand models and dynamic oligopoly models can be found in, for example, Conlon (2010) and Dubé, Fox and Su (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.

[57] Because 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 $\hat{\theta}$. Note the literature provides suggestions that would avoid the NFXP, such as Bajari, Benkard and Levin (2007), where value functions are forward simulated. A second advantage of MPEC is that it provides a conve-

We specify a likelihood objective function. For a driver j, we allow two different daily outside options μ_{z_j}, one for the daytime shift, 5AM to 5PM, and one for the evening/night shift. We also allow the fine, f_{h_t,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 take different values in four time intervals: from 12AM to 5AM, from 5AM to 12PM, and from 5PM to 12AM. The other two parameters of the cost function, λ_{1,z_j} and λ_{2,z_j}, are assumed to be time invariant. The remaining parameters are the standard deviation of the idiosyncratic shocks to the starting decision, σ_v, and to the stopping decision, σ_ϵ. We require those to be the same across the two medallion types.

5.5 Parameter Estimates

Table 12 gives an overview of the estimated parameters. Results are shown separately for minifleet and owner-operated medallions. Standard error calculations are bootstrapped: we drew 50 samples with replacement at the medallion level.

Table 2: Parameter Estimates (standard errors in parentheses)

parameter	**description**	**minifleet** $(z_j = F)$	**owner-operated** $(z_j = NF)$
$\mu_{z_j,0}$	outside-option, 6pm-4am	167.56 (16.31)	179.27 (16.61)
$\mu_{z_j,1}$	outside-option, 5am-5pm	174.98 (16.889)	175.44 (16.918)
f_{0,z_j}	fine (nightshift)	59.14 (3.613)	56.79 (3.911)
f_{1,z_j}	fine (dayshift)	63.8 (3.507)	50.3 (2.863)
$\lambda_{0,z_j,0}$	fixed cost (1am-5am),	66.6 (2.383)	59.79 (2.253)
$\lambda_{0,z_j,1}$	fixed cost (6am-12pm),	54.56 (1.355)	39.6 (0.934)
$\lambda_{0,z_j,2}$	fixed cost (1pm-5pm),	44.03 (1.083)	34.44 (0.947)
$\lambda_{0,z_j,3}$	fixed cost (6pm-12am),	53.89 (1.084)	38.34 (0.784)
λ_{1,z_j}	linear cost coefficient	-9.68 (0.649)	-3.92 (0.434)
λ_{2,z_j}	quadratic cost coefficient	1.0 (0.064)	0.56 (0.046)
σ_ϵ	sd iid hourly outside option	40.33 (2.448)	40.33 (2.448)
σ_v	sd iid daily outside option	37.3 (2.166)	37.3 (2.166)

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 \$167.56 and \$174.98 for minifleet drivers, and \$179.27 and \$175.44 for owner-operators. As we discussed in the identification section, these values are pinned down by the values of starting a shift. The values for minifleets are slightly lower, consistent with the descriptive evidence that fleet medallions are utilized more intensely. The estimated fines, f_{0,z_j} and f_{1,z_j}, are

nient 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.

$59.14 (nightshifts) and $63.8 (dayshift) for minifleet medallions. For owner-operated medallions, the night-shift fine is $56.79 and the day-shift fine is $50.3. The fixed part of the cost function parameters for minifleets are estimated as $66.6 from 1AM to 5AM, $54.56 from 6AM to 12PM, $44.03 from 1PM to 5PM, and $53.89 from 6PM to 12AM. For owner-operated medallions, the corresponding values are $59.79, $39.6, $34.44, and $38.34. The linear parameter of the hourly increase in cost is estimated to be $\$ - 9.68$ for minifleet and $\$ - 3.92$ for owner-operated cabs; the quadratic parameters are 1.0 for minifleets and 0.56 for owner-operated medallions. Because the cost parameter values by themselves are not very informative about the shape of the cost function we provide a graphical representation below (subsubsection 5.5.1). The standard deviation of the hourly outside option is 40.33. The standard deviation of the daily outside option is 37.3.

5.5.1 Discussion of Parameter Estimates

A few observations about the estimates are worth highlighting. As shown in Appendix A, minifleet medallions follow the 5AM to 5PM shift pattern more stringently than owner-operated taxis. This pattern is reflected in the estimates. subsubsection 5.5.1 displays the cost functions for both fleets and owner-operated taxis at different times of the day.[58] The graph shows the cost functions for a typical day and night shift (starting at 5AM and 5PM respectively). Owner-operated medallions have higher costs but less convex cost functions, which leads to stopping behavior that is "smoother", than for minifleet medallions. It also means owner-operators have a higher percentage of short shifts.

To evaluate the value of the outside option, comparing it to average daily earnings is useful. The hourly wage distribution is centered around $38.7 and a shift is on average 9.4 hours. Thus, expected earnings during a shift are approximately $364. If we subtract the rental rate of $112, we find that the outside option is not out of line relative the expected monetary value of the shift.

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 using it stops and that no other driver decides to use it in

[58] A potential anomaly is the slightly decreasing costs at the beginning of a shift. This reflects higher stopping probability after one hour of driving than after two or three hours. We interpret these extremely short shifts as due to drivers discovering problems with the car or being assigned the wrong vehicle. After this, costs are increasing, except for the small blip for Fleet drivers, which occurs at the time where the intercept of the cost functions changes (which varies by time of day) on the *typical* day shift.

Figure 9: Night and day shift cost function

Notes: The graph shows the cost functions for a typical day and night shift (starting at 5AM and 5PM respectively). Note that the model allows the intercept to take on two values depending on time of day. Since the intercept varies not by hours on a shift but by time of day, this discontinuity might happen at a different shift duration value for a drivers that starts at a different point in time than those displayed.

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 hours on a shift 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: $N_{t+1} = (1 - p_t^M) \cdot N_t + q_t^M \cdot M_t$ and $M_{t+1} = (1 - q_t^M) \cdot M_t + p_t^M \cdot N_t$.

The model fit for the aggregate law of motion is presented in subsubsection 5.5.2, which shows 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 we consider. We contrast a uniform (anticipated) increase in earnings throughout the day to an hour-specific (anticipated) increase that has the same normalized value, and we describe the response of total daily supply to each of these hourly shocks. subsection 5.6 displays the results. The uniform increase results in an elasticity estimate of 1.8, which, interestingly, is not too far from the number (1.2) reported in Angrist, Caldwell and Hall (2017) in an experiment on Uber data.

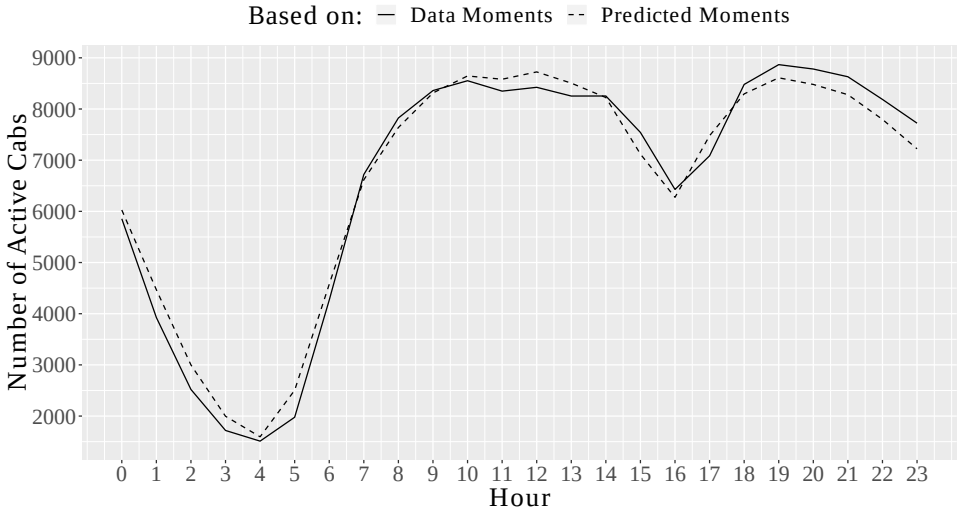


Figure 10: Model Fit

The effect of the homogeneous daily increase is due to a combination of additional entry and longer continuation decisions, as continuation becomes more profitable at all hours relative to the (unchanged) outside options.

The response to hourly increases varies from 0.9 to 2.6 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 complex, but we can highlight some of the forces at play. As we consider hourly increase later in the shift, two of the contrasting effects on labor supply are the following. The effect on daily entry decisions is highest for earnings increases in the first few hours of the shift. This is because the probability of actually receiving these earnings decreases as the shift becomes longer since the probability of stopping increases (due to both increasing costs as well as repeated exposure to hourly outside options). This effect works against later hours, which drivers are unlikely to reach. However, an earnings increase in hour t has an effect on all continuation decisions in hours prior to t, because of the expected future "bonus" at hour t. This effect penalizes earlier hours, although this is made more complicated by the fact that the effects depends on the likelihood of reaching those hours. In addition to those, there is a "mechanical" effect from a "bonus" at time t that comes from the fact that now that a driver is active at hours they would not have been active otherwise, they may receive an increase in those hours that induces them to continue working. Finally, an additional effect at play and working against hours that are very close to the end of the shift is that the fines are a strong deterrent to ending a shift late, so we find little responsiveness of supply in those hours. All of these countervailing forces come together to create the non-monotonic relation between hours in a shift and elasticities.

Comparing the pattern of hourly supply elasticities and of total taxi activity in subsubsection 5.5.2, it is apparent that there are some important differences. In particular, while supply in the early evening hours is high, the supply elasticity is relatively low. This is explained by the fact that earnings are especially high in the early evening hours.

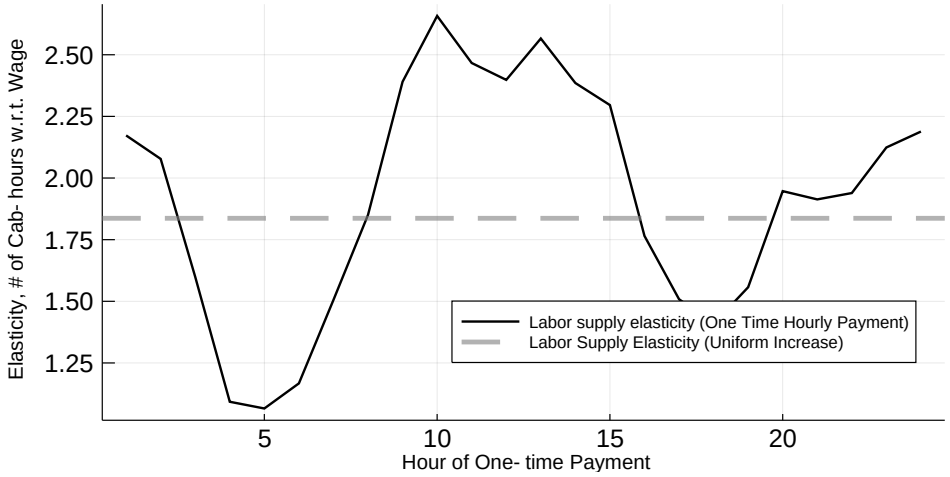


Figure 11: Labor-Supply Elasticity

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 ride hailing services can be attributed to the fact that they address both of these inefficiencies.[59] 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 Ã 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: we discuss how outcomes change if we consider a city that is identical to Manhattan except that its passengers are spread over an area that is three times as large.[60]

[59] An additional effect is that surge pricing makes supply more responsive to demand shocks. We do not address this issue here.

[60] We expect the main qualitative insights that we obtain in this section to be robust, but some caution should be used in evaluating the magnitude that we report in our counterfactuals, as our model is fairly elaborate and does rely on a number of simplifying assumptions.

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).[61]

Medallion revenue streams are obtained via simulation: we compute the number of times a medallion is rented out in a year in equilibrium and then use this number to compute the present discounted revenue under an annual interest rate of 5%. 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 computation highlights the impact of demand responsiveness to wait time.

The estimation did not require us to compute market equilibria because the supply side parameters were estimated using the observed process of hourly earnings. For the counterfactuals, we have to address the challenge of equilibrium computation. We iteratively solve for the supply- and demand-side parameters, holding the variables of the respective other side of the market fixed until updates on both sides of the market fall below some specified tolerance. The exact algorithm is described in detail in item H.

Recall that in the estimation we set the daily lease rate a driver pays to rent a taxi equal to the binding rate cap. In the counterfactuals, we do not allow lease rates to adjust endogenously to changes in market conditions. This assumption 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 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 is no longer binding, lease rates would have to adjust to maintain drivers' entry unchanged. Thus, if we consider a change in market conditions that is large enough to affect lease rates, our model would potentially overstate the supply response because the adjustment in lease rates would moderate the negative effects on taxi aggregate supply. Stopping decisions would still be affected because for these decisions, the rental rate is sunk. (Lower lease rates would have an indirect offsetting effect through larger entry of reducing continuation earnings.)[62]

[61] Please note that the medallion revenues should not be confused with medallion price since we have no direct measure regulatory uncertainty, and cost variables associated with owning medallions. If we had been able to compute the full medallion prices from our model one could use observed medallion prices as additional identifying restrictions, such as in Heckman and Navarro (2007) and Kalouptsidi, Scott and Souza-Rodrigues (2015).

[62] The natural model of lease determination would be one in which the market is competitive

Our model does not account for the traffic externality and potential adverse environmental effects due to additional taxis on the street. Note, however, that when we view search times and wait times as measures of frictions, these times indirectly incorporate concerns such as congestion: if moving to a dispatch model reduces search time and wait time, this move also improves congestion because empty taxis that do not deliver any passengers contribute to congestion without any social value. Analogously, a policy that results in increases in wait time and search time can be viewed as having a negative contribution to congestion. Although environmental damage is hard to assess, the city recently investigated this issue and found that ride-hailing services are only a minor contributing factor to the recent decline in NYC traffic speed.[63] These findings stand in contrast to Mangrum and Molnar (2017), who documents that the introduction of outer borough green cabs has lead to a local decrease in traffic speed of about 9%.

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 because the number of medallions is fixed and investment in taxis is sunk.

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

Table 3: Counterfactual Results

	Hourly Active Cabs	Hourly Demand	Passenger Wait Time	Matches per Day	Taxi Search Time	Hourly Taxi Revenues	Consumer Surplus (Minutes)	Medallion Revenue
Baseline	6785.0	23592.0	2.58	20898.63	7.62	40.37	1.66	1.53
Entry	7386.0	25530.0	2.43	22633.47	7.69	40.21	1.74	1.52
Entry (perc. change)	8.86	8.22	-6.01	8.3	0.95	-0.39	4.99	-0.61
Entry(PE)	7049.0	23592.0	2.35	22960.99	8.14	39.3	1.96	1.46
Entry(PE) (perc. change)	3.89	0.0	-8.87	9.87	6.84	-2.66	18.21	-4.59
Dispatch	7434.0	24739.0	2.53	23457.65	6.96	42.22	1.77	1.65
Dispatch (perc. change)	9.57	4.87	-1.94	12.24	-8.65	4.57	6.89	7.64
Dispatch(PE)	7121.0	23592.0	2.6	23525.06	7.45	41.41	1.84	1.59
Dispatch(PE) (perc. change)	4.95	0.0	0.58	12.57	-2.21	2.57	10.71	3.46
Segmented(50)	5591.0	17740.0	3.25	16350.99	8.82	38.2	1.43	1.29
Segmented(50) (perc. change)	-17.6	-24.8	25.97	-21.76	15.74	-5.37	-14.17	-16.23
SegmentedEnd	5377.0	21725.0	2.78	20319.2	9.5	37.8	1.66	1.26
SegmentedEnd (perc. change)	-20.75	-7.91	7.74	-2.77	24.59	-6.38	0.25	-17.93
Density	3662.0	10621.0	4.93	9228.03	12.45	31.88	0.91	0.86
Density (perc. change)	-46.03	-54.98	90.96	-55.84	63.3	-21.03	-44.98	-43.67
DensityDispatch	6534.0	20675.0	2.89	19613.32	8.04	39.7	1.61	1.5
DensityDispatch (perc. change)	-3.7	-12.36	12.1	-6.15	5.46	-1.67	-3.02	-2.29

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 $\Delta\%$ are the changes in 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 wait time observed in the data. The reason is that, for our parameter specifications, consumer surplus would be infinite if we integrated over all wait times. This issue results from the assumption of constant elasticity, log-linear demand. A similar issues 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 to be.

6.1 Relaxing Entry Restrictions

To explore how additional entry affects the market, we consider a 10% increase in the number of medallions, from 13,500 to 14,850. To place this magnitude in context, note that Uber served 4% of the total number of trips in 2014 and 13% at the beginning of 2015.[64]

Table 3 describes counterfactual outcomes aggregated at the daily level. Changes for this counterfactual are of uniform sign at the hourly level, although the magnitudes for daytime versus nighttime responses are somewhat different. For instance, taxi activity changes less during the night. On average, taxi activity expands by 8.86%, which is proportionally less than the medallion increase, because earnings drop, causing drivers to work less. This increase in available taxis reduces wait time for passengers, and demand expands. On average, the expansion of demand is 8.22%, whereas the average reduction in wait time for passengers over all hours is 6.01%.[65] The demand expansion moderates the drop in taxi earnings caused by increased supply. This effect is sizable: if demand were held fixed, supply would only expand by 3.89%. 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 0.39%. Earnings would instead fall by 2.66% if demand were held constant. The increase in the number of matches (or total number of trips) is 8.3%, which is very close to the increase in demand. The present discounted revenue stream for a medallion decreases only by 0.61%, which is a lot lower than if demand did not increase (in which case it would drop by 4.59%). Consumer surplus increases by roughly 5% and, perhaps surprisingly, would be higher without demand response. To understand this last result, 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. Thus, the gain in the infra-marginal trips that are already served in the scenario with no demand adjustment is larger.[66]

On a more speculative note, we have also explored what the socially optimal number of medallions would be. A social planner would ideally also adjust fares along with the number of medallions. However, this unconstrained planner's problem would require knowledge of the sensitivity of demand to price; and we do not have enough fare variation to gain information on this. However, we can conclude that if fares were fully flexible, then the optimal structure (absent considerations such as congestion), is to allow free entry. We can still compute the optimal number of medallions taking as given the current fare structure. Because

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

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

[66] The same logic applies to the other counterfactuals.

we have measured consumer surplus in minutes and producer (driver) surplus in dollar terms, we first need to find a translation in order to make these consistent. We convert consumer wait-time according to the median hourly wage in Manhattan. We also assume equal welfare weights on consumer and producer surplus. Under these assumptions, we find that the optimal number of medallions is approximately 21000, i.e., about 55% higher than under the cap in 2012.[67]

6.2 Improved Matching

6.2.1 Universal dispatcher

We now describe a counterfactual that considers an alternative matching technology: a dispatch platform 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.[68] This matching process is a natural alternative to the street-hailing system and approximates the one used, for instance, by Uber in its first 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, because our dispatcher does not optimize in a forward-looking way, allowing for matches of passengers and cabs that are too far apart can be very costly. The restriction to a one-mile radius alleviates this problem.[69]

Before presenting the results, we discuss how the dispatch matching function differs from the matching function that results from the street-hail process that we have assumed so far and modeled in subsubsection 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 subsubsection 6.2.1. This figure considers what

[67] Allowing for flexible fares is likely to increase this number substantially. In our view, the two biggest caveats for interpreting this number are the following. 1. A sizable increase in the number of medallions may lead to significant externalities, such as traffic congestion and pollution, that we do not consider; 2. The lease cap for rental rates for medallions may not be binding any longer for such a large increase in medallions. 3. Passengers have higher than median wages, which would also mean a higher optimal number of medallions.

[68] Passengers who do not find an immediate match wait for taxis to become available. Taxis that are unmatched drive randomly until matched. For the dispatcher we also count the time it takes the taxi to get to the passenger as wait and search-time respectively.

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

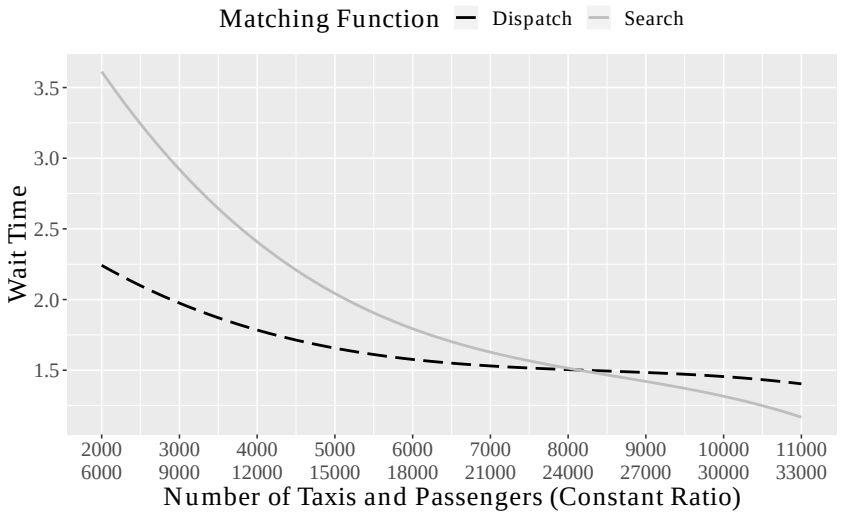


Figure 12: Matching Function Comparison

happens to passenger wait time as the market becomes thicker, that is, when we increase the market size (or density) while keeping constant the ratio of taxis to passengers. (Recall that a taxi serves multiple passengers within an hour.) Several features of this figure are worth noting. First, both matching functions display increasing returns, implying that as the market becomes thicker, wait time falls. Second, the dispatch matching function has a lower slope, implying that it is less sensitive to density. Third, for high values of density, the search matching function is superior so that wait times can become longer under dispatch when density is high. Point 3 may be surprising, but it is closely related to the non forward looking nature of the dispatch platform: passengers may be better off waiting for a random cab that is currently delivering a passenger and therefore not available for dispatch. The street-hailing system implicitly incorporates this option value and therefore can be superior to the dispatch matching function.

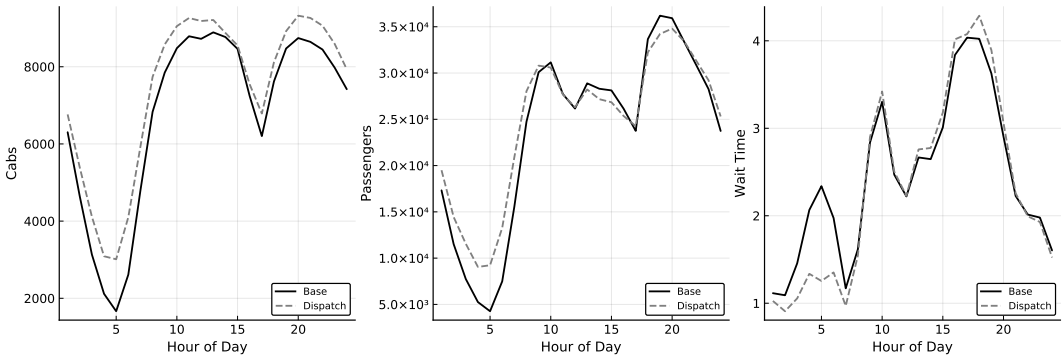


Figure 13: Dispatch Counterfactual

We now describe results under the different dispatch scenarios, starting with the extreme case of the entire market operating under the dispatcher. Figure 6.2.1 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. In this counterfactual, the number of active taxis increases 9.57%, and the search time for taxis decreases substantially (-8.65%). (If we did not allow for an expansion of demand in response to reduced wait times, the supply increase would, by contrast, only be 4.95%.) There is 4.87% more demand and the wait time for all passengers drops by 1.94%, consumer surplus increases by 6.89%, and the number of trips increases by 12.24%, more than proportionally to demand. The reason for the disproportionate change in consumer surplus and number of trips (successful matches) relative to demand is that fewer unmatched passengers are present. These passengers were previously unmatched because their wait time would have exceeded 20 minutes.[70] The value of medallions increases by 7.64%. To summarize, averaging across hours, all stakeholders benefit from the introduction of the dispatch technology, under the assumption that it 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.[71] To explore whether this outcome is plausible, we consider a scenario in which we divide the medallions equally across platforms: half operate on the search platform (baseline), half on the dispatch platform.[72] We assume both drivers and passengers are

[70] We record the wait time of unmatched passengers as 20 minutes.

[71] In addition to segmenting the market between different platforms, Uber entry also added capacity to the market. We also considered a counterfactual that combines entry and market segmentation: 10% additional capacity is offered with a dispatch technology, whereas the remaining stock of taxis remains on the street-hailing system. The results of this counterfactual are a hybrid of the two separate counterfactuals. Consumer surplus and wait times improved relative to the baseline but less than if entry had all occurred on the street-hailing system as in our first counterfactual.

[72] For this step, we also divide up potential demand exogenously across platforms, by simply pre-multiplying the constant elasticity demand function $d(w_t)$ by the shares of 0.5. This procedure ensures that if the wait time in the two platforms were the same, total demand would add up the baseline total demand. Realized demand, of course, depends on the wait time in the relevant platform. For the results we weigh wait times, driver earnings, and all the other variables by the respective hourly number of trips taken on both market sides and report those averages. The exception is medallion values, where we directly multiply the results by the respective share of

committed to a match: neither can cancel should another match option become available sooner.[73] We also first assume that passengers are exogenously assigned ex ante to one of the two platforms. We remove this assumption in what follows.

Figure A in Appendix C 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. The advantage of the dispatch platform is due to the fact that it has lower wait times, and therefore, higher demand.[74] However, 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 we see substantially fewer active cabs (-17.6%) and lower demand (-24.8%). Wait time and search time are much larger than baseline (+25.97% and +15.74% respectively).

To separate the respective roles of the market-thickness externality and of the dispatch platform, we perform the same exercise with two competing segmented dispatch platforms as well as with two competing street-hailing systems.[75] The table below present the results of these scenarios.

Table 4: Different Types of Segmentation

	Baseline	**Search/ Dispatch**	Δ%	**Search/ Search**	Δ%	**Dispatch/ Dispatch**	Δ%
Consumer Surplus (million minutes / day).	1.66	1.43	-14.17	1.2	-27.83	1.65	-0.86
Driver Revenue (hourly income)	\$40.37	\$38.2	-5.37	\$33.85	-16.14	39.8	-1.41
Medallion Revenue (PV in millions)	\$1.53	\$1.29	-16.23	\$1.09	-29.25	1.48	-3.22
Wait time (average in minutes)	2.58	3.25	25.97	3.92	52.03	2.84	10.02

The table shows that the two competing dispatch platforms perform substantially better, whereas 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.[76]

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, passenger choice must induce

issued medallions, in this case 50% of each type.

[73] For example, we do not allow a waiting passenger who is matched to a dispatch cab to accept a ride from a searching cab.

[74] The higher advantage of the dispatch platform in hours with lower demand density is related to the findings in Cramer and Krueger (2016) who show Uber's advantage in capacity utilization (defined as the fraction of time delivering passengers) is relatively minor in NYC but large in other cities that are less dense.

[75] This case is entirely hypothetical: what segmentation would mean for two street hailing platforms is not clear.

[76] One caveat of our analysis is that it does not take into account the potential difference in variance in wait times across platforms. It is likely that dispatch platforms deliver reduced uncertainty in wait times.

equal wait time across the two platforms. As more passengers move to the dispatch platform, it 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.[77] Line SegmentedEnd in Table 3 presents the results for this scenario. Compared to the case in which demand is segmented exogenously, substantially more market share is captured by the dispatch platform (especially during the night-time). This increased penetration of the dispatch platform moderates some of the negative effects of the market-thickness externality for passengers but makes drivers worse of on average. On balance, however, the number of matches per day only drops by 2.77% compared to the baseline, much better than with exogenous demand division across platforms. Allowing passengers to choose also increases drivers earnings disparity, roughly doubling the earnings advantage that drivers on the dispatch platform enjoy.

6.4 Density Counterfactual

The last counterfactual explicitly addresses the spatial dimension of matching in this market. We now assume matching takes place in an area that is three times the area of our baseline simulation of Manhattan but is otherwise identical. This counterfactual emphasizes the role of geography. It shows how the market may function in less dense areas, such as more sparsely populated cities, or suburban areas. Of course, the counterfactual retains some of the basic functioning of the market which are in fact specific to NYC, but we believe that it is in fact informative. We first assume decentralized search (a street-hailing system) 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. Table 5 reports the average welfare, driver income, medallion revenue and wait time in these two scenarios. This demonstrates the magnitude of the effects of density in our environment. For instance, wait time increases by 90% and driver revenue decreases by about 21%. However, the dispatcher has a much more sizable, positive effect in this low density scenario, leading to a large improvement in all dimensions. Table 6.4 displays the hourly number of active taxis, passengers, and wait time in these two scenarios compared to the baseline.

[77] For given numbers of taxis in the two platforms, this logic guarantees existence of such an indifference condition. In our setting, a complication arises because taxi supply is endogenous and therefore responds to this process. However, in the simulation, this endogeneity is not sufficient to lead to a corner solution.

Table 5: Density

	Baseline	Density	Δ%	Density Dispatch	Δ%
Consumer Surplus (million minutes / day)	1.65	0.91	-44.54	1.61	-3.02
Driver Revenue (hourly income)	40.37	$31.88	-21.59	$39.68	-1.71
Medallion Revenue (PV in millions)	1.53	$0.86	-43.28	$1.5	-2.33
Wait time (average in minutes)	2.6	4.93	90.1	2.89	12.07

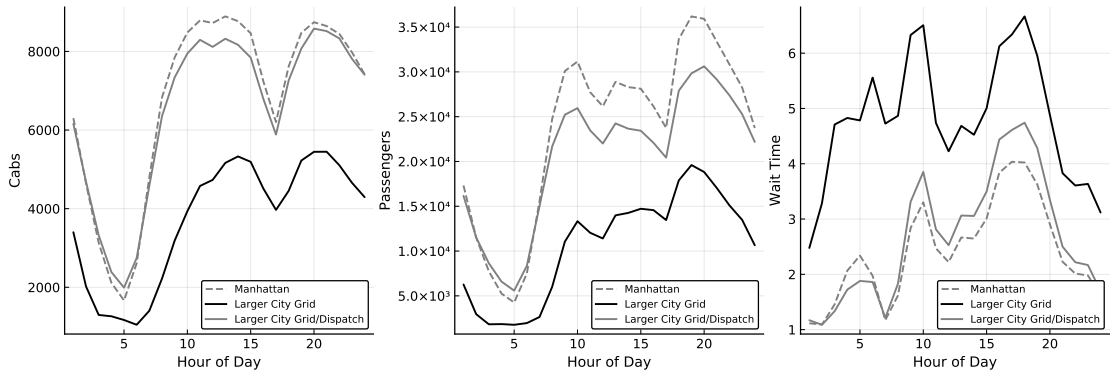


Figure 14: Density Intradaily

7 Conclusion

This paper develops and estimates a dynamic 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 wait time for a cab. To estimate the model, we first obtain a numerical representation of the matching function by explicitly modeling the geographic nature of the matching process. We then back out unobserved demand by inverting this matching function. The number of active taxis is determined by daily starting decisions and hourly stopping decisions given the anticipated level of earnings from the equilibrium level of cabs and passengers.

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, because the model does not incorporate any additional heterogeneity among the platforms and assumes exogenous assignments of drivers. Including such richness is not within the scope of the current paper but extending the analysis to study this issue in more depth would be interesting.

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

References

Angrist, Joshua D, Sydnee Caldwell, and Jonathan V Hall. 2017. "Uber vs. Taxi: A Driver's Eye View." National Bureau of Economic Research.

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

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

Brancaccio, Giulia, Myrto Kalouptsidi, and Theodore Papageorgiou. 2017. "Geography, Search Frictions and Endogenous Trade Costs."

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

Buchholz, Nicholas. 2018. "Spatial Equilibrium, Search Frictions and Efficient Regulation in the Taxi Industry."

Buchholz, Nicholas, Haiqing Xu, and Matthew Shum. 2016. "Semiparametric Estimation of Dynamic Discrete-Choice Models." *arXiv preprint arXiv:1605.08369*.

Buchholz, Nicholas, Laura Doval, Jakub Kastl, Filip Matejka, and Tobias Salz. 2019. "The Value of Time."

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

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

Chen, M Keith, Judith A Chevalier, Peter E Rossi, and Emily Oehlsen. 2017. "The value of flexible work: Evidence from uber drivers." National Bureau of Economic Research.

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

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

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

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

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

Eckstein, Zvi, and Gerard J Van den Berg. 2007. "Empirical labor search: A survey." *Journal of Econometrics*, 136(2): 531–564.

Farber, Henry S. 2008. "Reference-dependent preferences and labor supply: The case of New York City taxi drivers." *The American Economic Review*, 98(3): 1069–1082.

Farber, Henry S. 2014. "Why You Can't Find a Taxi in the Rain and Other Labor Supply Lessons from Cab Drivers."

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

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

Hall, Jonathan V, John J Horton, and Daniel T Knoepfle. 2017. "Labor market equilibration: Evidence from uber." *URL http://john-joseph-horton.com/papers/uber_price. pdf, working paper*.

Heckman, James J, and Salvador Navarro. 2007. "Dynamic discrete choice and dynamic treatment effects." *Journal of Econometrics*, 136(2): 341–396.

Hendel, Igal, Aviv Nevo, and François Ortalo-Magné. 2009. "The relative performance of real estate marketing platforms: MLS versus FSBOMadison. com." *American Economic Review*, 99(5): 1878–98.

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

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

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

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

Kalouptsidi, Myrto, Paul T Scott, and Eduardo Souza-Rodrigues. 2015. "Identification of counterfactuals in dynamic discrete choice models." National Bureau of Economic Research.

Kleiner, Morris M, and Alan B Krueger. 2013. "Analyzing the extent and influence of occupational licensing on the labor market." *Journal of Labor Economics*, 31(S1): S173–S202.

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

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

Mangrum, Daniel, and Alejandro Molnar. 2017. "The marginal congestion of a taxi in New York City."

Mas, Alexandre, and Amanda Pallais. 2017. "Valuing alternative work arrangements." *American Economic Review*, 107(12): 3722–59.

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

Petrongolo, Barbara, and Christopher A Pissarides. 2001. "Looking into the black box: A survey of the matching function." *Journal of Economic literature*, 39(2): 390–431.

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

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

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

Thakral, Neil, and Linh T Tô. 2017. "Daily Labor Supply and Adaptive Reference Points."

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

Weintraub, Gabriel Y, C Lanier Benkard, and Benjamin Van Roy. 2008. "Markov perfect industry dynamics with many firms." *Econometrica*, 76(6): 1375–1411.

Wheaton, William C. 1990. "Vacancy, search, and prices in a housing market matching model." *Journal of Political Economy*, 98(6): 1270–1292.

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

Frictions in a Competitive, Regulated Market: Evidence from Taxis.

Guillaume R. Fréchette, Alessandro Lizzeri, and Tobias Salz

Online Appendix

A Additional Figures

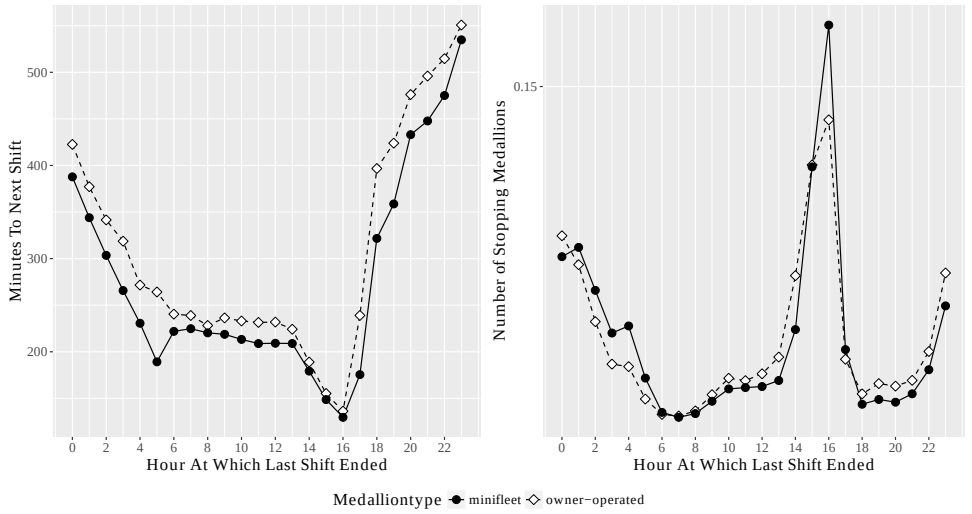


Figure 15: Time Medallion is Unutilized Conditional on Hour of Drop-Off.

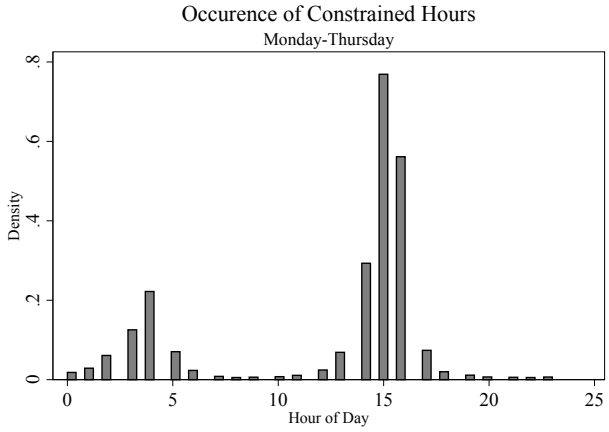


Figure 16: Prevalence of Constrained Hours Throughout the Day

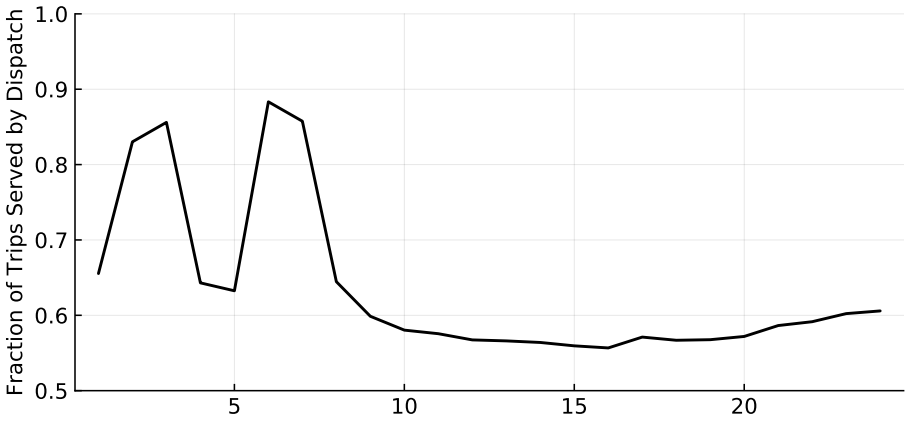


Figure 17: Fraction of Rides Served by Dispatch

B Multiplicity

A look at the inter-temporal patterns of the data shows that it follows a remarkably stable pattern. This clearly suggests that equilibrium multiplicity is not an issue in the time dimension. To illustrate this we plot the same weekdays (for those weekdays that we use in the data) on one plot (Figure 19), which allows a comparison of intra-daily demand and supply patterns. We therefore believe that a single equilibrium in the data is a reasonable assumption. Even across weekdays, the numbers at different hours are remarkably similar (one of the Mondays is a holiday).

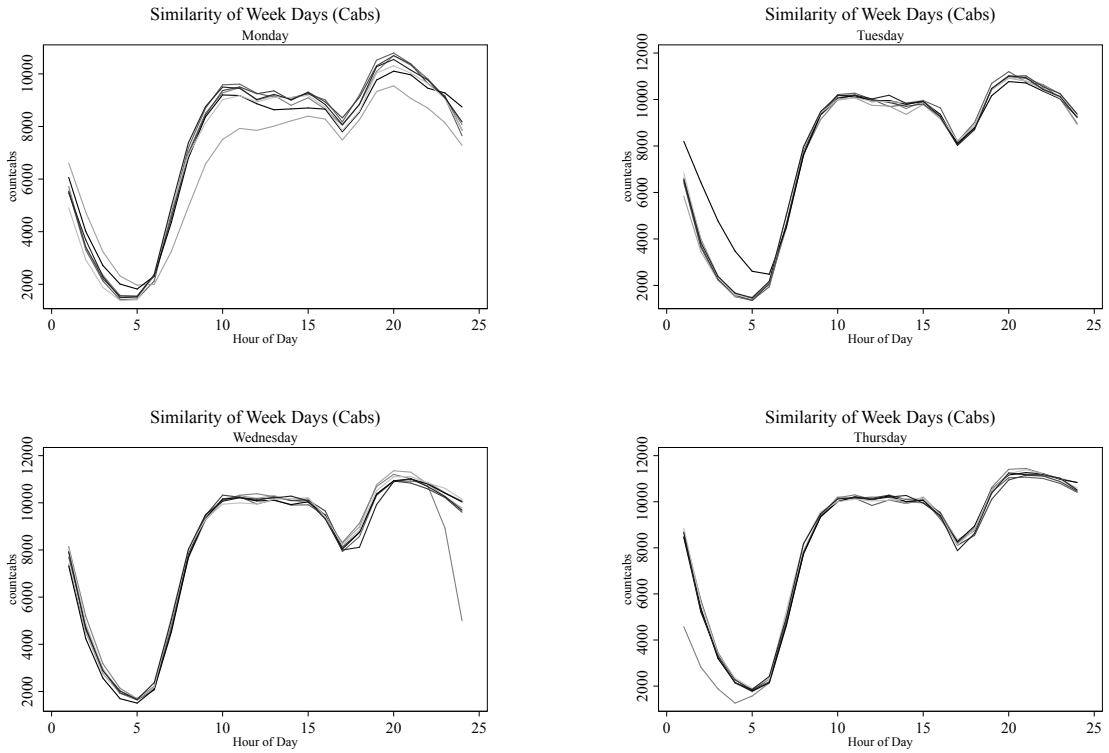


Figure 18: Activity per Hour on Different Mondays and Tuesdays in our Sample, Cabs

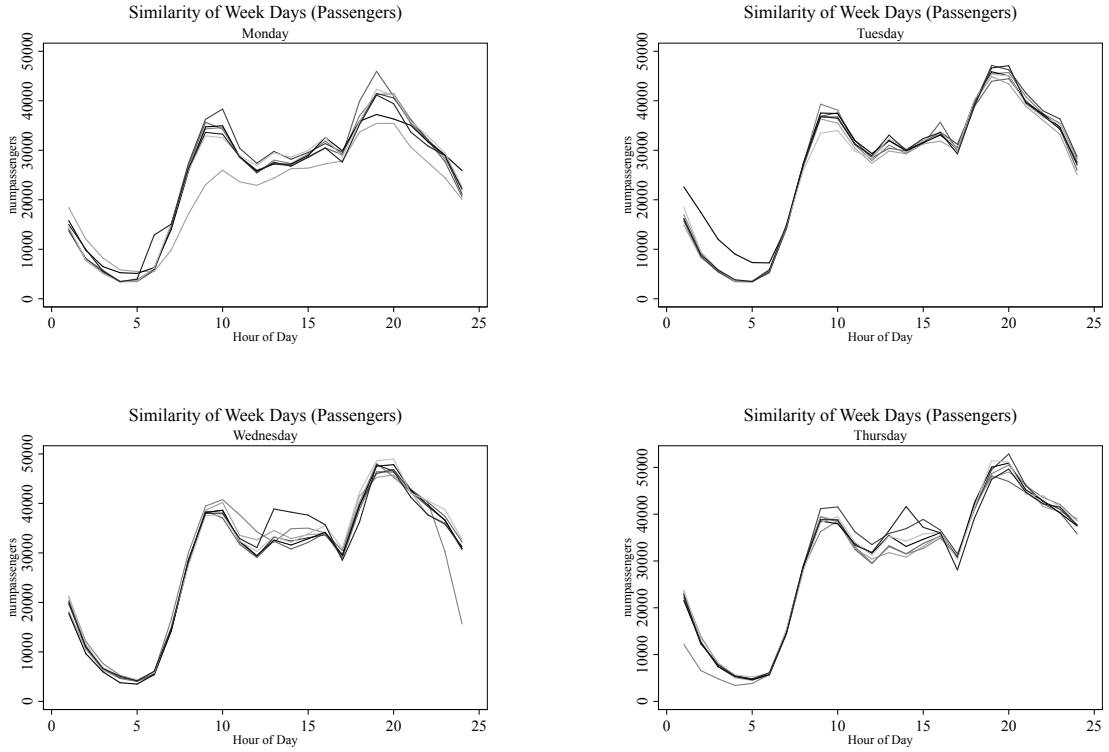


Figure 19: Activity per Hour on Different Weekdays in our Sample, Passengers

C Full Table of Summary Statistics

Table 6: Summary Statistics by Hour, Mean and (Standard Deviation)

Hours	Cabs	Passenger	Search time (minutes)	Wait time (minutes)	MPH (Manhattan)	MPH (other)
0	9188.2 (1311.6)	28849.74 (6536.2)	9.3 (1.7)	1.24 (.31)	15.13 (.907)	17.62 (1.05)
1	6997.4 (1273.6)	18201.71 (3401.0)	12.3 (1.0)	1.03 (.20)	16.6 (1.09)	18.44 (1.55)
2	4351.2 (904.2)	10481.33 (2307.3)	14.4 (1.1)	1.14 (.19)	17.58 (.77)	18.05 (1.07)
3	2744.3 (608.9)	6611.52 (1448.7)	15.6 (.86)	1.5 (.26)	18.24 (.59)	17.47 (1.03)
4	1875.7 (435.4)	4533.09 (1208.4)	16.4 (1.05)	1.94 (.24)	18.62 (.48)	17.21 (.75)
5	1652.1 (248.6)	4224.4 (831.1)	15.1 (.99)	2.00 (.14)	20.49 (.65)	18.28 (.63)
6	2208.0 (132.4)	6056.1 (1344.2)	13.2 (2.0)	1.95 (.24)	21.08 (.74)	21.19 (1.25)
7	4699.1 (348.2)	14592.1 (1015.5)	10.1 (.74)	1.18 (.17)	17.51 (.60)	19.22 (1.04)
8	7541.5 (606.6)	27263.6 (2187.9)	7.5 (.60)	1.53 (.18)	13.86 (.77)	15.05 (1.07)
9	9081.8 (618.8)	36598.5 (3289.6)	5.8 (.55)	2.29 (.22)	10.89 (.79)	12.70 (1.28)
10	9824.1 (561.8)	36818.3 (3028.7)	6.2 (.53)	2.48 (.23)	10.03 (.76)	12.65 (1.30)
11	9903.3 (528.3)	31471.9 (2649.3)	8.2 (.69)	2.00 (.24)	10.17 (.82)	13.06 (1.58)
12	9722.8 (570.9)	28986.3 (2548.5)	9.1 (.77)	1.83 (.23)	10.25 (.80)	13.14 (1.23)
13	9775.4 (579.1)	31866.6 (3153.1)	7.9 (.81)	2.08 (.30)	10.03 (.82)	12.99 (1.26)
14	9650.0 (548.6)	31281.1 (3677.4)	8.0 (1.0)	2.05 (.36)	10.27 (.88)	13.34 (1.2)
15	9745.5 (475.2)	32456.2 (2929.8)	7.4 (.79)	2.23 (.35)	10.31 (.83)	13.55 (1.16)
16	9204.7 (345.0)	33786.9 (2233.8)	6.2 (.55)	2.71 (.35)	10.16 (.85)	12.99 (1.14)
17	8072.0 (207.8)	29813.4 (1011.3)	5.9 (.29)	2.88 (.34)	10.86 (.87)	13.24 (.83)
18	8726.0 (296.2)	39035.8 (2479.2)	4.6 (.4)	2.94 (.39)	10.73 (.76)	13.01 (.81)
19	10320.8 (315.0)	45458.0 (3697.0)	4.9 (.63)	2.65 (.39)	10.28 (.73)	12.74 (.84)
20	10872.3 (403.6)	45170.5 (4402.3)	5.6 (.7)	2.12 (.40)	11.10 (.81)	13.90 (.97)
21	10767.6 (526.9)	40399.2 (4159.9)	6.7 (.92)	1.64 (.27)	12.68 (.90)	15.80 (1.04)
22	10457.3 (643.0)	37413.6 (4455.7)	7.3 (1.0)	1.49 (.25)	13.84 (.81)	16.89 (.83)
23	10027 (787.3)	34763.1 (4989.8)	7.7 (1.1)	1.44 (.26)	14.47 (.75)	17.36 (.77)

D Details on Simulation

The goal of the simulation is to obtain a mapping of the number of waiting passengers and searching cabs within an hour to the wait 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. Wait and search time are also influenced by other exogenous factors, which therefore need to be arguments of the matching function. These factors are the speed, mph_t, at which the traffic flows, and the average trip length, $miles_t$, requested by passengers. Note that the average trip length determines how long a taxi is utilized for each unit of demand, and therefore shifts effective supply. The longer the trip length the higher the utilization. Table 6 provides an overview of taxi search time, the number of taxis, as well as the recovered number of passengers and their wait time.

The baseline simulation is performed under the assumption that cabs search randomly for passengers. The search is performed on an idealized map of Manhattan. subsubsection 5.1.1 provides a schematic of the grid we use for the simu-

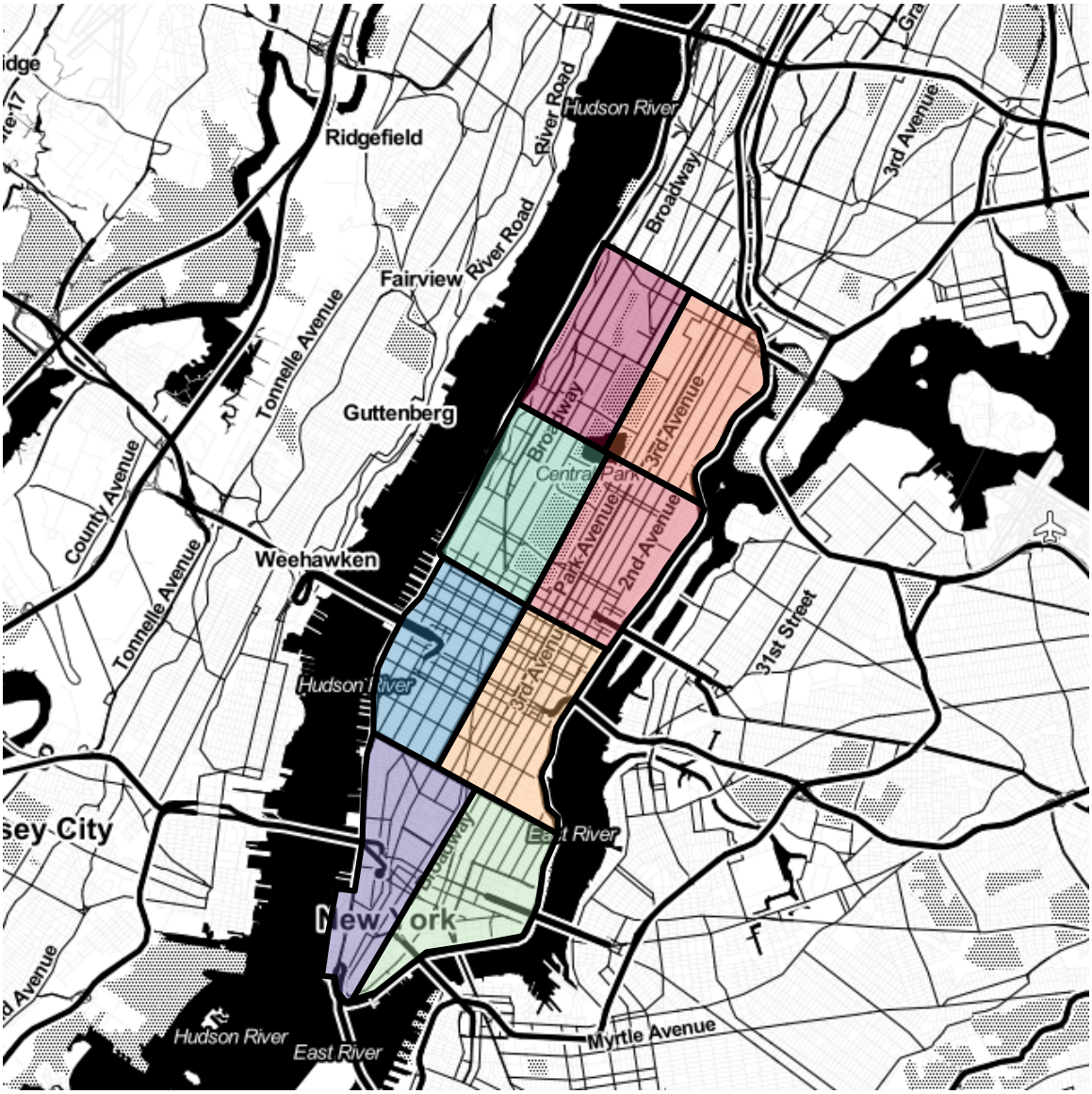


Figure 20: Division of Manhattan

lation. In line with the topography of Manhattan, we require the area to be four times longer in the north-south direction (y_t) than wide in the east-west-direction (x_t). Cabs move on nodes that are $1/20$ mile segments apart from each other, which is based on the average block length in the 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. subsubsection 5.1.1 highlights nodes on which cabs can turn as gray. This map corresponds to the block structure of Manhattan, where a block is approximately $1/20$ of miles long in the north-south and $4/20$ of a mile wide in the 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 they never turn back in the direction from which they were coming (i.e., no U-turns).

Because we only model the Manhattan market (below 128th Street), our grid corresponds to an area of 16 square miles. Appendix D shows the modeled part on the map and its division into the eight equally sized different areas for which we separately compute the pickup and drop-off probabilities. Correspondingly, our grid is divided into eight equal parts, which correspond to those areas. Note however, that we map those areas onto our rectangular grid and do, therefore, not precisely model the actual street grid and shape of Manhattan.

Each node on the grid is a possible passenger location. For each hourly simulation, $\frac{d_t}{6}$ passengers are placed in 10-minute intervals randomly on the map. Those $\frac{d_t}{6}$ are divided up and placed in proportion to the corresponding (observed) pickup probabilities on the eight areas on the grid. Within those areas passengers appear with equal probability on each node.

If a cab hits a node with a passenger a match occurs, implying no additional frictions on a node, and so the number of matches is the minimum of the number of passengers and the numbers of taxis on the node, which corresponds to the assumption of a Leontieff matching function on each node. Once the match takes place the cab is taken of the grid for $60 \cdot \frac{miles_t}{mph_t}$ minutes, that is, 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 (Appendix D).

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 because a smaller time unit is unnecessary. Passengers are added to the map for one hour in 10-minute intervals. Because nodes are spaced $1/20$ of a mile 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 ad-

ditional variables are used to describe the algorithm: *npick* refers to the number of matches that have already taken place; $deliverytime_i$, to the remaining delivery time of taxi i; $searchtime_i$ to the time that taxi i has spent searching since the last delivery, *total_searchtime* refers to the total time taxis have spent searching for passengers; and *total_waittime* to the total wait time that passengers have been waiting.

while $npick < d$ **do**
  $t = t + 1$ (time units represent travel time from one node to the next, scales with mph);
  if $t \in \{\bar{t}/6 \cdot k | k = 1, ..., 6\}$ **then**
    add $d/6$ passengers to random nodes on map, stratified by eight areas;
  end
  for $i = 1 : c$ **do**
    if $deliverytime_i = 0$ *(cab i is not occupied)* **then**
      update the node of cab i. Cabs only take turns on gray nodes (subsubsection 5.1.1) and do not make u-turns. All feasible travel directions are chosen with equal probability;
      if *new node of cab i has a passenger* **then**
        cab becomes occupied, set $deliverytime_i$ to $20 \cdot miles$, and add $searchtime_i$ to *total_searchtime*;
      else
        $searchtime_i = searchtime_i + 1$
      end
    else
      $deliverytime_i$ = $deliverytime_i$ - 1;
      if $deliverytime_i$ == 0 **then**
        place cab in random area on map according to observed drop-off probabilities (all nodes within area equal probability), give cab random feasible travel direction;
      end
    end
  end
  Add one to *total_waittime* for each passenger that is on the map;
  for $j = 1 : p$ **do**
    Increase counter for all passengers on the map; remove passengers that have been waiting for 20 minutes.
  end
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, as soon as a cab has delivered a passenger, it 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 an-

other match option. For example, a waiting passenger, who is promised to a cab, may encounter another (previously unavailable) cab earlier than the promised cab. 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.

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

D.1 Details on Heterogeneity across Areas.

In the current baseline simulation we allow for heterogeneity by dividing the city map into eight different areas and we match pickup and dropoff probabilities by area. To test how sensitive our results are to this subdivision, we recompute the simulation, including inferred passenger waiting times and the demand function, for two alternative subdivisions of the map. In one case, we double the number of areas to 16; in the other, we only divide the map into four areas. We perform this simulation in the same way in which we do it for the baseline, but instead using drop-off and pick-up probabilities that are matched at different (higher and lower respectively) levels of aggregation. We find aggregate numbers that look almost identical when we move to 16 areas, whereas there are some differences when we reduce the number of areas to 4. This can be seen in table Table 7. Since the natural concern is about robustness with respect to finer subdivisions, we feel comforted by this, and also feel justified in keeping our 8 area baseline. We also re-estimate demand both for the 4-area division and for the 16-area division. The demand estimates for all specifications do not change notably. For 4 areas, the numbers are as follows. Mean wait-time: 1.88; mean number of passengers: 22234; elasticity: -1.08. For eight areas (current specification) we have the following. Mean wait-time: 2.47; mean number of passengers: 23262; elasticity: -1.225. For 16 areas the numbers are as follows. Mean wait-time: 2.61; mean number of passengers: 23551; elasticity: -1.21. Lastly, taking the eight areas that we currently use, search time is remarkably stable across locations (see Table Table 8).

Table 7: Robustness to Number of Areas (from our simulated matching function)

		Search Time			Wait Time			Matches		
Passengers	Cabs	s^4	s^8	s^{16}	w^4	w^8	w^{16}	m^4	m^8	m^{16}
15000.0	4000.0	4.83	5.58	5.77	4.05	4.8	4.93	13750.0	13201.0	13056.0
15000.0	6000.0	11.53	11.84	11.96	1.32	1.74	1.83	14747.0	14599.0	14534.0
15000.0	9000.0	23.34	23.48	23.51	0.35	0.46	0.47	14876.0	14826.0	14813.0
25000.0	4000.0	1.41	2.18	2.3	8.24	8.68	8.72	16998.0	16145.0	16013.0
25000.0	6000.0	3.51	4.42	4.59	4.09	4.9	4.99	22190.0	21014.0	20791.0
25000.0	9000.0	9.29	9.73	9.86	1.32	1.7	1.8	24204.0	23798.0	23675.0
35000.0	4000.0	0.75	1.17	1.29	10.7	10.79	10.84	17670.0	17251.0	17111.0
35000.0	6000.0	1.46	2.35	2.48	6.83	7.41	7.47	25272.0	23818.0	23606.0
35000.0	9000.0	4.2	5.05	5.19	2.89	3.6	3.68	31672.0	30234.0	29981.0

Note: superscript indicates number of areas.

Table 8: Taxi Search Time Across Areas (data)

Area	Median	Mean	Median Day	Mean Day
1	6.55	11.64	4.37	10.3
2	4.37	9.15	4.37	7.7
3	4.37	8.6	4.37	7.7
4	4.37	9.21	4.37	8.04
5	4.37	8.34	4.37	7.8
6	4.37	8.79	4.37	8.29
7	4.37	9.68	4.37	9.4
8	6.55	11.42	4.37	11.12

D.2 Robustness Check on Maximal Wait Time

Table 9: Robustness to Different Maximal Wait Time (from our simulated matching function)

		Search Time			Wait Time			Matches		
Passengers	Cabs	s^{15}	s^{20}	s^{25}	w^{15}	w^{20}	w^{25}	m^{15}	m^{20}	m^{25}
15000	4000	8.48	10.01	11.5	2.07	2.02	1.52	9809.04	10046.88	10072.36
25000	4000	10.97	13.5	15.97	1.03	1.0	0.99	10281.26	10386.78	10377.66
35000	4000	11.97	14.96	17.5	0.53	0.5	0.5	10420.64	10470.94	10453.68
15000	6000	4.99	5.12	5.68	5.25	5.25	4.95	13397.34	13579.86	13706.74
25000	6000	8.97	10.5	12.15	1.29	1.03	1.0	15174.32	15412.9	15457.64
35000	6000	10.46	12.52	15.01	1.03	0.63	0.5	15514.7	15650.06	15643.38
15000	9000	2.76	2.67	2.89	34.52	35.04	34.61	14529.56	14463.64	14666.5
25000	9000	5.47	5.55	6.29	2.9	2.97	2.65	21524.88	21686.48	22057.66
35000	9000	7.98	9.02	11.0	1.07	1.03	1.0	22894.04	23181.42	23315.9

Note: superscript indicates maximal wait time.

D.3 Robustness to Definition of Breaks

Our definition of a shift is a sequence of trips that have not been interrupted for longer than 300 minutes (we follow Farber's shift definition). However, within a shift, cabs sometimes take breaks. In the paper, define a break every gap between a drop-off and the next pickup that is longer than 45 minutes (note that we restrict the sample to trips within Manhattan). Of course, the value of mean search-time does depend to some extent on this choice, but it turns out not to be very sensitive to it. The average search time under our current break definition is 7.5 minutes. If instead we chose a 50-minute cutoff, it would be 7.73, whereas with a 40-minute cutoff, it would be 7.22. An important question is how our demand recovery is affected by the break-time cutoff. In this matter, a countervailing effect arises. Note first that a change in the break cutoff changes both the search time that we use to recover demand, but also the effective number of searching taxis. As an example, if we lower the cutoff, search time is lower, but so is effective supply since more cabs are on a break according to the new cutoff. To understand how these two changes affect the inferred number of passengers, it is helpful to consider how each change separately affects the inversion of the matching function. First, decreasing supply, ceteris paribus, leads us to infer a smaller number of passengers. Lower search time, on the other hand, implies a larger number of passengers, all else equal. These two changes counteract each other. We now report our computation of how different the inferred number of passengers is as we change the cutoff. Moving the cutoff from 45 to 50 minutes, an 11% change, leads to an inferred number of passengers that is 0.77% higher. A reduction in the cutoff to 40 minutes, leads us to infer that the number of passengers is 0.89% lower. Lastly, even if we consider a more drastic change, a cut-off of 30 minutes,

a reduction of one third, we would infer 3.8% fewer passengers. Hence, the total effect on the implied number of passengers from varying the definition of a break within reasonable values is quite small. Furthermore, the changes in our counter-factuals should be even smaller because all scenarios would be similarly affected by any such change.

E Details on Supply Side Estimation.

As defined in the text, let $\boldsymbol{x}_{it}$ denote the observable part of the state (the hours on a shift, the hour of the day as well as the medallion-invariant characteristics). Let $p(\boldsymbol{x}_{it})$ be the theoretical probability that an active medallion/driver i stops at time point t, and let $q(\boldsymbol{x}_{it})$ be the probability that an 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. Therefore, the solver maximizes both over the parameters of interest θ and an additional set of parameters δ. The parameter vector δ consists of all $p(\mathbf{x}_{it})$, $q(\mathbf{x}_{it})$, $\mathbb{E}_\epsilon[V(\mathbf{x}_{i(t+1)}, \epsilon_{i(t+1)})]$ for $\mathbf{x}_{i(t+1)} \in \mathbf{X}$. In other words, δ consists of expected values and choice probabilities for each point in the observable state space. With this notation in place we can express the maximization problem as follows:

$$\begin{aligned}\min_{\theta, p(\mathbf{x}_{it}), q(\mathbf{x}_{it}), \mathbb{E}_\epsilon[V(\mathbf{x}_{i(t+1)}, \epsilon_{i(t+1)})]} & \sum_{j \in J} \sum_{t \in T_j} d_{it}^A \cdot log(p(\mathbf{x}_{it})) + (1 - d_{it}^A) \\ \cdot (log(1 - p(\mathbf{x}_{it}))) + d_{it}^I \cdot log(q(\mathbf{x}_{it})) + (1 - d_{it}^I) \cdot (log(1 - q(\mathbf{x}_{it}))))\end{aligned} \tag{4}$$

subject to

$$\begin{aligned}\mathbb{E}_\epsilon[V(\mathbf{x}_{i(t+1)}, \epsilon_{i(t+1)})] = \sigma_\epsilon \cdot log\Bigg(exp\left(\frac{1}{\sigma_\epsilon}\right) \\ + exp\left(\frac{\pi_{h_t} - C_{z_i}(l_{it}) - f(h_t, k_i) + \mathbb{E}_{\boldsymbol{\epsilon}_{i(t+1)}}[V(\boldsymbol{x}_{i(t+1)}, \boldsymbol{\epsilon}_{i(t+1)})]}{\sigma_\epsilon}\right)\Bigg) \\ + \gamma * \sigma_\epsilon \quad \forall \boldsymbol{x}_{it} \in \boldsymbol{X}\end{aligned} \tag{5}$$

$$p(\boldsymbol{x}_{it}) = \frac{exp\left(\frac{1}{\sigma_v}\right)}{exp\left(\frac{1}{\sigma_v}\right) + exp\left(\frac{\pi_t - C_{z_i,h_t}(l_{it}) - f(h_t,k_i) + \mathbb{E}_{\boldsymbol{\epsilon}_{i(t+1)}}[V(\boldsymbol{x}_{i(t+1)},\boldsymbol{\epsilon}_{i(t+1)})]}{\sigma_v}\right)} \forall \ \boldsymbol{x}_{it} \in \boldsymbol{X} \tag{6}$$

$$q(\boldsymbol{x}_{it}) = \frac{exp\left(\frac{\mathbb{E}_{\boldsymbol{\epsilon}_{i(t+1)}}[V(\boldsymbol{x}_{i(t+1)},\boldsymbol{\epsilon}_{i(t+1)})] - r_{h_t}}{\sigma_v}\right)}{exp\left(\frac{\mathbb{E}_{\boldsymbol{\epsilon}_{i(t+1)}}[V(\boldsymbol{x}_{i(t+1)},\boldsymbol{\epsilon}_{i(t+1)})] - r_{h_t}}{\sigma_v}\right) + exp\left(\frac{\mu_{h_{t+1}}}{\sigma_v}\right)} \ \forall \boldsymbol{x}_{it} \in \boldsymbol{X} \tag{7}$$

The constraint given by Equation 5 ensures 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.

F Impact of weather on Demand.

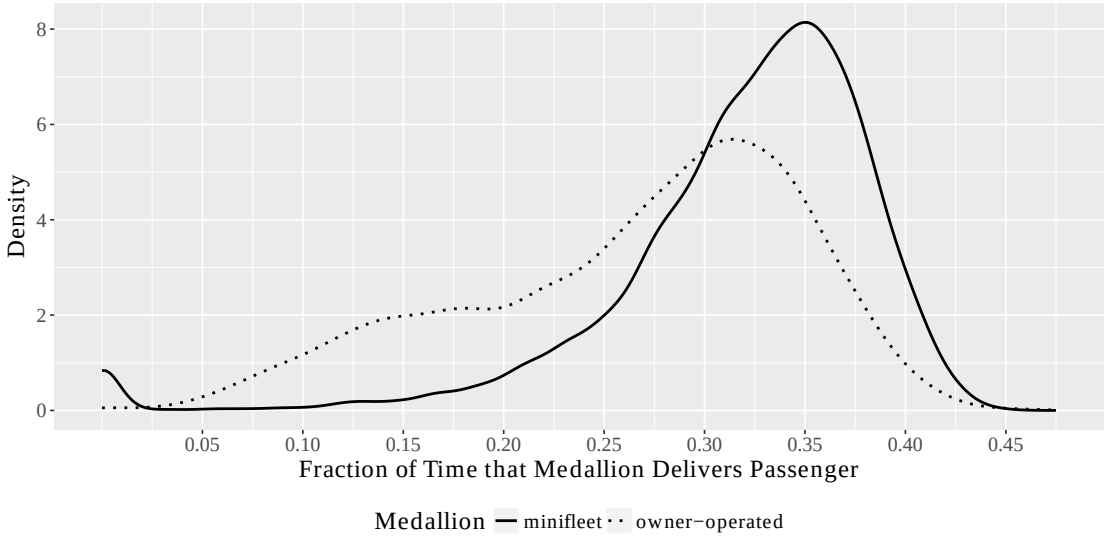
One of the important demand shocks for taxi rides is weather patterns, most notably, rainfall. To check whether our demand recovery picks up such patterns, we have merged hourly rainfall data. We then regress both log-wait-times and log-recovered-demand on a dummy for whether rainfall occurred in this hour. As the table below shows, our recovered demand and wait-times are highly correlated with rainfall. Estimates from these regressions suggest that in an hour with rainfall, our recovered demand is about 28% higher and wait times about 37% higher.

Table 10: Demand Validation with Weather Data.

	(1)	(2)	(3)	(4)	(5)	(6)
	$log(d_t)$	$log(d_t)$	$log(d_t)$	$log(w_t)$	$log(w_t)$	$log(w_t)$
Rainfall Dummy	0.292**	0.281**	0.281**	0.275**	0.371**	0.372**
	(0.0140)	(0.00839)	(0.00817)	(0.0108)	(0.00829)	(0.00800)
Hour FE	No	Yes	Yes	No	Yes	Yes
Day of Week FE	No	No	Yes	No	No	Yes

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 is has rained in this hour.

G Medallion Ownership and Utilization



Notes: These densities are based on medallion-level observations, where each observation is the fraction of time 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 densities show a stark differences between owner-operated and minifleet medallions. The lower tail of low utilization is much thicker for owner-operated medallions.

Figure 21: Density of Utilization Separated by Medallions

Appendix G 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.[78]

The left panel of Appendix A in Appendix C shows the length of time a medallion is *inactive* conditional on the stopping time of the last shift. Because 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, whereas a stop at any other time causes the medallion to be *stranded* for a longer time period. We see that minifleets typically return a medallion to activity faster after each drop-off. This difference is particularly large after the common night shift starting times (6PM and later). This in turn suggests that minifleets have access to a larger pool of potential drivers, making 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 medallions to capture these differences. The right panel of Appendix A in Appendix C 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 pattern is also reflected in subsection 4.2, which shows a stronger supply decrease for minifleets before the evening shift relative to owner-operated medallions.

H 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_h^i, $h \in \{0, \ldots, 23\}$ *under* $g(.)$.

2. *Simulate drivers earnings* π_h^i, $h \in \{0, \ldots, 23\}$ *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_h^i, $h \in \{0, \ldots, 23\}$ *from the wait times* w_h^i *and the estimated demand function* $d(.)$.

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

4. *Use* $p^i(\mathbf{x}, \pi; \theta)$ *and* $q^i(\mathbf{x}; \theta)$ *to simulate a new law of motion for drivers* c_h^i, $h \in \{0, \ldots, 23\}$. *For each medallion type (z,k), we simulate 30 medallions, where each medallion starts inactive at 12PM, and iterate forward for 48 hours. Across these 30 medallions, we then compute the fraction of times the medallion has been active*

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

in this hour (using only the last 24 hours) and multiply this amount by the total number of medallions.[79]

5. *Compute* $sumsq_1 = \sum_h (c_h^i - c_h^{(i-1)})^2$

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_h^i , $h \in \{0, \ldots, 23\}$, *to determine the wait times* w_h^i, $h \in \{0, \ldots, 23\}$ *for passengers.*

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

3. *Compute* $sumsq_2 = \sum_h (d_i^h - d_h^{i-1})^2$

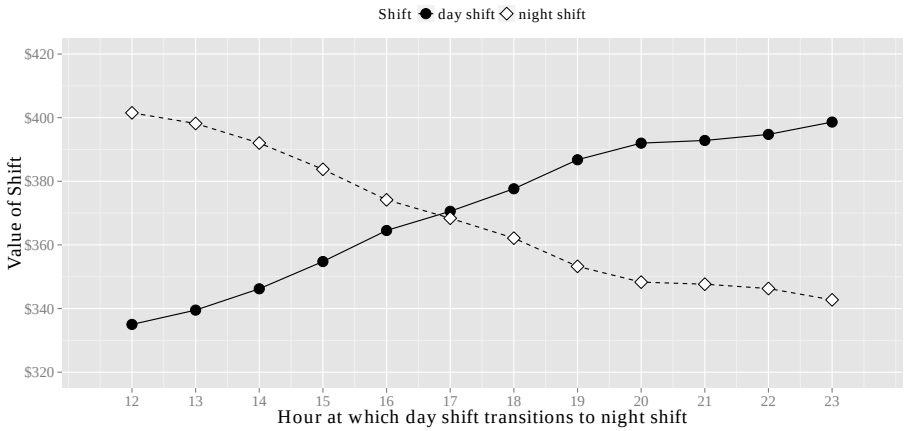
Using these definitions, the algorithm can be described as follows:

while $STOP \neq 1$ **do**
  Compute $sumsq_1$ using **Block1(i)**.
  if $sumsq_1 < tol$ **then**
    STOP=1
  else
    while $sumsq_1 > tol$ **do**
      Block1(i) (in each iteration guess new vector of cabs with some non-linear solver)
    end
  end
  Compute $sumsq_2$ using **Block2(i)**.
  if $sumsq_2 < tol$ **then**
    STOP=1
  else
    while $sumsq_2 > tol$ **do**
      Block2(i) (in each iteration guess new vector of passengers with some non-linear solver)
    end
  end
  i = i + 1
end

[79] We have also experimented with different numbers in this step, for example simulating each medallion for more than 48 hours or increasing the number of simulations. For the final counterfactual results, these alternatives do not make a large difference.

I Shift-Transition and the Witching Hour

We now argue that the timing of the shift transition is such a supply side driven shifter. The dip in the number of active taxis in the later afternoon hours is clearly visible in subsection 4.2. The right-hand panel of Appendix A shows that this dip is mostly due to taxis ending their shift. The left-hand panel shows that the transition to the next driver is quite long. Together these suggest that this common shift transition is responsible for a prolonged reduction in the number of cabs between 4 and 6 o'clock. Multiple reasons could explain why most shifts are from 5AM to 5PM and 5PM to 5AM, but the data (and the rules) suggests some key factors.



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. Because 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-shift and night-shift drivers. The graph shows that earnings are almost equal at 5PM, the prevailing division for most medallions.

Figure 22: Earnings of Day and Night Shift for Different Split Times

First, the rules are such that minifleets can only lease a medallion for exactly two shifts per day: they must operate a medallion for at least two shifts of nine hours and the lease must be on a per-day or per-shift basis.[80] Second, a cap is placed 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 were binding during our sample period. Given these rules, minifleets may try to equate the earning potentials for the day and night shifts, as a way to ensure they will get a similar number of drivers willing to drive each shifts. A similar argument applies for owner-drivers that want to ensure they always find a driver for the

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

second shift, which they do not drive themselves. Appendix I shows the earnings for night and day-shifts under different hypothetical shift divisions. The x-axis shows each potential division-point, that is, each point at which a day shift could end and a night shift start. The y-axis reports the earnings for the day-shift (black dots) and night-shift (white diamonds).[81] 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. Because 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.

J Table of Choice Probabilities

Table 11: Stopping probabilities by hour and hours on a shift.

	$l=1$	$l=2$	$l=3$	$l=4$	$l=5$	$l=6$	$l=7$	$l=8$	$l=9$	$l=10$	$l=11$	$l=12$	$l=13$	$l=14$
$h=0$	.18	.11	.13	.14	.14	.14	.15	.21	.32	.37	.43	.43	.41	.4
$h=1$	.15	.18	.13	.15	.15	.16	.16	.19	.24	.35	.4	.42	.42	.37
$h=2$	.07	.11	.22	.15	.17	.18	.17	.19	.23	.29	.44	.46	.39	.36
$h=3$	.03	.04	.15	.32	.19	.22	.23	.24	.28	.34	.44	.58	.5	.42
$h=4$	.02	.03	.08	.23	.53	.36	.4	.45	.51	.59	.73	.81	.78	.65
$h=5$	.01	.01	.03	.07	.18	.42	.39	.46	.54	.61	.7	.79	.79	.64
$h=6$	.01	0	.01	.02	.05	.12	.27	.32	.4	.53	.62	.71	.63	.52
$h=7$	.01	0	0	.01	.01	.03	.06	.15	.21	.32	.42	.46	.41	.29
$h=8$	.01	0	0	.01	.01	.01	.03	.07	.12	.2	.29	.27	.24	.19
$h=9$	.02	.01	.01	.01	.01	.02	.03	.04	.11	.17	.21	.23	.14	.13
$h=10$	.03	.01	.01	.01	.01	.02	.02	.04	.09	.16	.25	.25	.12	.1
$h=11$	.04	.01	.02	.02	.02	.02	.02	.04	.09	.15	.22	.2	.16	.08
$h=12$	.04	.02	.02	.02	.02	.02	.03	.04	.07	.17	.29	.28	.16	.1
$h=13$	.05	.02	.03	.04	.04	.04	.04	.06	.08	.15	.29	.34	.22	.14
$h=14$	.04	.03	.05	.07	.09	.1	.1	.13	.17	.22	.34	.38	.27	.14
$h=15$	.03	.03	.06	.1	.15	.19	.24	.29	.37	.45	.5	.58	.46	.3
$h=16$	.01	.02	.05	.09	.16	.24	.33	.44	.58	.68	.74	.73	.55	.4
$h=17$	.01	.01	.02	.04	.06	.09	.12	.16	.22	.27	.3	.27	.17	.1
$h=18$	.01	0	.01	.02	.04	.05	.08	.09	.12	.14	.15	.16	.14	.1
$h=19$	.01	.01	.01	.01	.03	.05	.07	.09	.1	.14	.15	.16	.17	.15
$h=20$	.02	.01	.01	.01	.02	.04	.07	.09	.11	.13	.16	.18	.18	.17
$h=21$	.04	.02	.02	.01	.02	.03	.06	.09	.12	.14	.17	.19	.19	.19
$h=22$	.06	.03	.03	.03	.03	.03	.06	.1	.14	.18	.2	.23	.22	.23
$h=23$	.1	.05	.05	.05	.05	.05	.06	.11	.16	.2	.22	.23	.23	.23

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

Table 12: Starting probabilities by hour.

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
.01	.01	.02	.05	.11	.22	.29	.26	.2	.17	.14	.12	.12	.14	.2	.43	.73	.47	.28	.15	.09	.05	.03	.03

K Other Counterfactual Numbers

Table 13: Counterfactual Results for other Segmentations

	hourly active cabs	hourly demand	passenger waittime	matches per day	taxi searchtime	hourly taxi revenues	consumer surplus (minutes)	medallion revenue
Segmented(2-Search)	4580.0	14163.0	3.92	12543.18	10.44	33.85	1.2	1.09
Segmented(2-Search) (perc. change)	-32.49	-39.96	52.03	-39.98	37.0	-16.14	-27.83	-29.25
Segmented(2-Dispatch)	6602.0	21316.0	2.84	20158.8	7.84	39.8	1.65	1.48
Segmented(2-Dispatch) (perc. change)	-2.7	-9.65	10.2	-3.54	2.87	-1.41	-0.86	-3.22
Fleet	7217.0	24926.0	2.46	22115.01	7.72	40.26	1.71	1.53
Fleet (perc. change)	6.37	5.65	-4.62	5.82	1.29	-0.26	2.9	-0.1

Note: The changes are the mean over all 24 hours of the day. The wait time and search time averages 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 $\Delta\%$ are the changes in 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 wait times. This issue results from the assumption of constant elasticity, log-linear demand. A similar issues 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 to be.