Spatial Pricing: An Empirical Analysis of Taxi Rides in New York City

Baris Ata[†], Nasser Barjesteh[†], and Sunil Kumar[⋆]

[†]Booth School of Business, The University of Chicago

[⋆]Johns Hopkins University

Abstract

This paper studies how spatial pricing and search friction can impact the taxi market in New York City. We use a mean field model, in which the taxi drivers strategically search for customers in different neighborhoods across the city, taking into account the spatial and temporal distribution of the supply and demand as well as the prices across the city. Our model captures the interplay between spatial pricing, where prices depend on either the origin of the ride alone or both its origin and destination, and search friction, due to empty taxis and customers within the same neighborhood failing to pair efficiently. Spatial pricing can incentivize relocation of empty taxis to a neighborhood while the use of mobile applications can alleviate search friction within that neighborhood. We fit our model to a dataset of New York City taxi rides over four years and conduct a series of counterfactual studies to explore how spatial pricing impacts demand for and supply of rides, consumer welfare, and drivers' profit. Our analysis reveals that spatial prices that only use origin information can increase consumer surplus by 7.0% of the average fare and serve 2.6% more customers without hurting the drivers' profit. Moreover, we find that eliminating the (local) search inefficiency alone can increase consumer surplus by 13.9% of the average fare and serve 4.3% more customers while simultaneously increasing drivers' profit by 2.5% of the average fare. We also observe that improving search efficiency primarily impacts under-served neighborhoods such as upper Manhattan, Brooklyn and Queens, while pricing primarily impacts well-served neighborhoods, for example, the airports, midtown, and downtown Manhattan. This underscores the value of a hybrid mechanism. We propose a mechanism in which (local) search is eliminated in all neighborhoods while spatial pricing is only used in well-served neighborhoods. This mechanism increases consumer surplus by 21.5% of the average fare and serves 8.7% more customers, while avoiding price discrimination in less affluent neighborhoods of the city. The proposed mechanism achieves 96.3% of the benefits of a citywide spatial pricing and friction removal mechanism.

Keywords— taxi, spatial pricing, search friction, mean field equilibrium

1 Introduction

Taxi industry is an essential part of the transportation sector.[1] For example, in New York City (NYC), taxis offer over 150 million rides per year (TLC 2014). In this market, taxis and customers search for each other. This search friction results in a substantial welfare loss. Prices set by the taxi industry affect the inter-temporal and spatial distribution of the supply and demand and the interaction between them, impacting the consumer surplus and drivers' profit. This paper explores spatial pricing, a mechanism that prices rides based on their origin and destination. In this context, we study the following questions: How does spatial pricing impact consumer surplus and drivers' profit? What is the pattern of the optimal spatial prices? How does pricing based solely on the origin of the ride compare with pricing based on the origin-destination pair? How does spatial pricing compare with removing the local search friction using mobile applications?[2]

We consider two inefficiencies: First, the mismatch between supply and demand. Second, the (local) search friction. To address these, we consider spatial pricing and friction removal (through a better matching technology) as levers. Although spatial pricing[3] is not widely used in the taxi industry or ride-sharing platforms, it can help match supply and demand. Supply can be redistributed since the profitability of different neighborhoods is closely tied to prices and taxi drivers make relocation decisions to maximize their profit. The distribution of demand can also be adjusted since customers are price sensitive. By adjusting both supply and demand, spatial pricing provides the policy maker with a unique and powerful tool to intervene in the details of the market. Moreover, spatial pricing can be implemented with the existing equipment[4] and technological solutions such an mobile applications.

Spatial pricing does not resolve all the inefficiencies. The search by customers for taxis is spatially localized. A customer looks for a taxi only within a relatively small geographical area at any given time. Although empty taxis have more freedom and may relocate to different areas to seek customers, within any single area their ability to connect with customers is imperfect. As the spatial densities of customers and taxis increase, a higher fraction of customers are matched with taxis. However, it is not uncommon to find both empty taxis and unfulfilled customers in the same area. On the one hand, spatial pricing can impact the relocation decisions of the empty taxis. This in turn, impacts the spatial density of taxis, and consequently, the efficacy of the search. On the other hand, mobile (search) applications can improve the search efficacy as well, enabling better connections between empty taxis and customers even in locations

[1] With an annual revenue of $18.9 billion and annual profit of $1.5 billion, taxi and limousine industry is one of the major segments of the US economy; see e.g., Sayler (2017).

[2] Internet-based mobile applications such as Arro and Curb can provide a better match between customers and taxis; see https://www.ridearro.com and https://main.gocurb.com.

[3] Spatial pricing should not be confused with dynamic pricing.

[4] Smart meters installed on all NYC yellow taxis (since 2009) report all the information (longitude and latitude of the pick-up and drop-off locations) required for the implementation of spatial prices. Mobile applications such as Arro and Curb can inform customers of their fare before they hail a taxi.

with low density of customers and empty taxis, as measured by the number of customers and the number of empty taxis per street mile, respectively. Relocation of empty taxis induced by spatial pricing and the reduction of search friction complement each other. Understanding the interplay between spatial pricing and the removal of (local) search friction is the primary focus of this paper. Although removing the (local) search friction resolves the issue of having unfulfilled customers and unutilized taxis simultaneously at the same location, it is merely a local solution. It has little to no impact on neighborhoods with high density of supply and demand, where the majority of rides initiate (due to the already high efficacy of search in these neighborhoods). In other words, removing the (local) search friction does not address the global mismatch between supply and demand; see e.g., Lagos (2000).

The contributions of this paper are threefold: First, we propose a suitably refined spatial model and study spatial pricing in a setting where spatial redistribution of empty taxis is crucial and truly complementary to search. We also propose a matching model that captures the spatial aspect of the search friction and the interplay between the density of supply/demand and the efficiency of matching. Putting these together, we study the resulting mean field equilibrium. Second, we fit the model to the data of yellow taxi trips in NYC from January 2010 to December 2013 to estimate the primitives of the model. Certain primitives, such as demand, can be useful beyond this paper. Third, we quantify the impact of origin-only and origin-destination spatial pricing on consumer surplus and drivers' profits, study the pattern of the optimal spatial prices, and show that spatial pricing is complementary to removing the (local) search friction. This underscores the value of a "hybrid" mechanism, where pricing is used as a tool to shape supply of taxis in well-served areas, and removal of search friction induces appropriate supply in less-served ones. It is worth noting that a matching model that captures the spatial aspect of the search friction and a granular model for prices are crucial in capturing the complementarity of removing friction and spatial pricing.

Our model captures the interaction of supply and demand both locally and globally. To capture the global aspect, we use a mean field model that approaches the problem from a macroscopic perspective. In this model, an individual taxi driver is irrelevant and the focus is on the distribution of the taxis. To capture the local aspect, we use an aggregate matching function that captures the spatial and microscopic aspect of the search.

We take the view of a social planner and optimize total consumer surplus. We observe that a spatial pricing scheme that prices rides based on their origin and destination and allows prices to change between 50% and 150% of the current prices[5] in NYC can increase consumer surplus by $\$168,000$ in every day shift ($\$0.79$ per ride or 8.5% of the average fare), serve 3.2% more customers, and increase customer miles (total number of miles traveled by customers) by 7.4%, without hurting the drivers' profit. A similar spatial pricing

[5]In this paper, we do not change the ratio of the price per mile to the fixed portion of the fare. Instead, we use price multipliers to change the fares; see Section 6.

scheme that only uses origin information can increase consumer surplus by $\$135,000$ in every day shift ($0.63 per ride or 7.0% of the average fare), serve 2.6% more customers, and increase customer miles by 3.9%.

Prices in under-served areas (such as upper Manhattan, Brooklyn, and Queens)[6] increase to attract more taxi drivers. Due to higher prices, each served customer in under-served areas is worse off. However, this effect is offset by the increase in the number of served customers. In well-served areas, prices increase in nodes with low demand or shorter trips and decrease in nodes with high demand and longer trips.

Table 1: Improvement in various parameters of interest under spatial pricing and removing friction (prices deviation is limited to 50%).

	Origin-Only Pricing	Origin-Destination Pricing	Removing Local Search Friction	Hybrid Mechanism	Proposed Mechanism	Citywide Origin-Only Pricing & Friction Removal
Consumer Surplus						
Total increase	\$135K	\$168K	\$268K	\$322K	\$417K	\$433K
Per ride	\$0.63	\$0.79	\$1.26	\$1.52	\$1.96	\$2.04
In terms of average fare	7.0%	8.5%	13.9%	16.7%	21.5%	22.4%
Number of served customers	2.6%	3.2%	4.3%	5.7%	8.7%	8.9%
Miles traveled by customers	3.9%	7.4%	6.2%	8.1%	11.7%	12.0%
Drivers' Profit	\$0	\$0	\$48K	\$0	\$0	\$0

Considerably higher prices in the less affluent neighborhoods of the city (upper Manhattan, Brooklyn, and Queens) is an undesirable outcome of using spatial pricing alone. This effect can be mitigated by removing the (local) search friction. Table 1 compares the impact of origin-only spatial pricing (with price deviation smaller than 50%) and removing friction on various metrics of interest. Removing the (local) search friction alone (with no spatial pricing) can increase consumer surplus by $1.26 per ride, serve 4.3% more customers, and increase drivers' profit by $0.22 per ride. The majority of the benefits of removing the (local) search friction come from under-served nodes while the majority of the benefits of spatial pricing are from well-served nodes. This highlights the value of a hybrid mechanism. A hybrid mechanism, that uses spatial pricing in well-served neighborhoods and removes friction in under-served neighborhoods (keeping the price pattern), can increase consumer surplus by $1.52 per ride. Under the hybrid mechanism, we need only a little price variation to achieve the majority of the benefits. Table 1 also presents the impact of citywide spatial pricing and friction removal. This mechanism (with only 50% price variation) increases consumer surplus by $2.04 per ride, which is considerably higher than spatial pricing or removing friction alone. Since policy makers prefer avoiding price discrimination in less affluent neighborhoods of the city, we propose a mechanism in which friction is removed in the entire city while spatial pricing is used only in well-served neighborhoods. The mechanism captures almost all the benefits of citywide spatial pricing and friction removal.

[6] See Figure 14 for the definition of these geographical areas.

Spatial prices can increase drivers' profits, as well. Spatial prices, that only use origin information and do not deviate from the current prices in NYC by more than 50%, can increase drivers' profits by $\$100,000$ in every day shift ($0.47 per ride or 5.3% of the average base-fare), without lowering consumer surplus.

An important antecedent of this paper is the work of Buchholz (2018) that also studies removing the (local) search friction and origin-only spatial pricing (with four pricing neighborhoods). We use a different matching model to capture the spatial aspect of the search friction and study a more refined pricing scheme. This allows us to unveil the complementarity between removing friction and spatial pricing. Consequently, we obtain different results; see Sections 2-3 for a detailed comparison.

The rest of the paper is organized as follows. Section 2 reviews the literature. Section 3 introduces the mean field model. Section 4 describes the data. Section 5 describes the estimation procedure and results. Section 6 describes the counterfactual analysis and Section 7 concludes. Appendix A discusses the procedure used to define the nodes/neighborhoods. Appendix B provides complementary discussions on the matching model. Appendices C-D provides supplementary material for the Data and Estimations sections, respectively. Appendix E describes a Monte Carlo simulation study to illustrate the identification of our model. Appendix F uses five-fold cross-validation to examine the ability of our model in predicting the relocation decisions of the drivers. Appendix G provides supplementary material for the counterfactual analysis and Appendix H studies spatial pricing for maximizing drivers' profit. Appendix I provides the proofs and derivations.

2 Literature Review

This paper is related to four streams of literature. The first stream focuses on mechanisms and regulations used for improving the performance of the taxi market. The second stream studies spatial models of search with an emphasis on the ride-hailing industry. The third stream focuses on dynamic discrete choice models. The last stream studies mean field games and their applications.

Mechanisms and regulations used for improving the performance of the taxi market have been studied extensively. Entry restrictions and price controls are the most studied mechanisms in the literature; see e.g., Coffman and Shreiber (1977), Foerster and Gilbert (1979), Schroeter (1983), and Häckner and Nyberg (1995). The common theme in many of these papers is that price regulations and entry restrictions are helpful since they increase the availability of taxis in times and locations with low demand.[7] Frechette et al. (2016) follows this literature by showing that search frictions and entry restrictions are important inefficiencies in the taxi market, and one reason for the success of ride-sharing platforms is the fact that they can address these issues. Similar to Frechette et al. (2016), we observe that search friction is a major issue and removing

[7]This is achieved by reducing the friction of price negotiation and ensuring a suitable minimum profit for the drivers. For example, in exchange for serving all neighborhoods of a city, a firm could be granted a monopoly position.

it increases consumer surplus and drivers' profit by \$1.26 and \$0.22 per ride, respectively.

Following the success of ride-sharing platforms, taxi industry has received increasing scrutiny. Cramer and Krueger (2016) shows that the utilization of Uber drivers is higher than the utilization of taxi drivers. They argue that this is in part due to Uber's efficient matching technology and inefficient regulations in the taxi industry. Buchholz (2018) proposes a model of spatial search of taxi rides and computes the gain in consumer surplus from perfect matching of customers and taxis in each neighborhood. Buchholz concludes that although the elimination of the (local) search friction in NYC results in a 7.1% increase in the number of served customers, it has a negative impact on consumer surplus. In contrast, we observe that the elimination of the (local) search friction results in a 4.3% increase in the number of served customers while increasing consumer surplus by \$1.26 per ride. The majority of these benefits come from neighborhoods with low density of supply and demand. Our observation is consistent with the conclusions made in Lam and Liu (2017) and Shapiro (2018) that a platform that posses a superior matching technology outperforms NYC yellow taxis primarily in neighborhoods with low density of supply and demand.

Another class of mechanisms closely related to this paper is dynamic and spatial pricing in ride-hailing networks. This literature can be divided into spatial and non-spatial treatments of the problem. Banerjee et al. (2015), Bai et al. (2016), Ozkan and Ward (2017), and Cachon et al. (2017) use a non-spatial analytical approach. Banerjee et al. (2015) models the problem of dynamic pricing of rides in a single region as a queueing system. It shows that the throughput and revenue of no dynamic pricing strategy can exceed that of the optimal static pricing policy. However, dynamic pricing strategies are more robust to fluctuations in the system parameters. Relevant empirical non-spatial studies include Hall et al. (2015), Cohen et al. (2016), Ming et al. (2017), Lam and Liu (2017), and Shapiro (2018). Cohen et al. (2016) uses data on Uber rides in four major cities in the United States and estimates that for each dollar spent by customers, 1.6 dollars of consumer surplus is generated. Lam and Liu (2017) uses a dynamic choice model in which customers in NYC choose between Uber, Lyft, and Taxi rides. It finds that customers who use ride-sharing platforms gain 0.72 dollars for every dollar spent on rides and 64% of this gain is due to dynamic pricing. Although our paper focuses on spatial pricing and we observe more modest increases in consumer surplus from pricing and using a better matching technology (e.g., an increase in consumer surplus of \$2.04 per ride, 22.4% of the average ride, from city-wide spatial pricing and friction removal), important qualitative insights of Lam and Liu (2017) agree with ours. For example, Lam and Liu (2017) observes that pricing is more welfare-enhancing in thick markets (such as midtown Manhattan) while the matching technology is most beneficial in the outer boroughs.

The second stream of literature studies spatial search models. Lagos (2000) is one of the first papers that studied the taxi market with an emphasis on its spatial aspect and the strategic behavior of its drivers.

Lagos (2000) proposed a model in which taxi drivers search for customers on a graph, and highlighted the friction resulting from this search. A number of papers followed this work, by generalizing it, or using it to address other issues in the ride-hailing industry.

Bimpikis et al. (2016) builds on Lagos (2000) to study spatial pricing in ride-sharing platforms. It shows that if the demand pattern is not balanced, spatial pricing is beneficial and optimal prices can be written in terms of the optimal dual variables corresponding to the flow balance equations. Under the assumption that prices and compensations can be decoupled, the authors are able to solve for the optimal prices by focusing on the mass balance equations only. Since in the taxi market in NYC the entire fare is collected by the drivers and the Taxi and Limousine Commission (TLC) does not collect a fee, prices and compensations cannot be decomposed in our setting. Hence, no such simplification occurs in our analysis. Our paper allows for a non-complete graph with different distances/travel times between nodes, a non-uniform demand pattern on the graph, and inter-temporal variations in the system.

Buchholz (2018) builds upon Lagos (2000) to study the pricing of taxi rides based on pick-up location, time of day, and the distance of the ride. It concludes that distance-based pricing outperforms pricing based on the pick-up location and time of day. Buchholz finds that distance-based pricing outperforms origin-only pricing by 766% with respect to the increase in consumer surplus and by 356% with respect to the increase in the number of served customers. Although our paper does not analyze distance-based pricing, the origin-destination pricing subsumes distance-based pricing. However, we see a more modest increase with respect to consumer surplus (24%) and the number of served customers (23%) in origin-destination compared to origin-only pricing.

Other related papers in the spatial literature are Braverman et al. (2016), Banerjee et al. (2016), Yang et al. (2017), Afèche et al. (2018), and Besbes et al. (2018). Braverman et al. (2016) studies centralized empty car routing in ride-shairng platforms. It shows that a fluid-based optimization can be used to solve for the optimal network utility, which is an upper bound on the utility of all static and dynamic routing policies in the finite-car system. Banerjee et al. (2016) studies pricing in shared vehicle systems where taxis do not relocate when they are empty. It proposes a static spatial pricing algorithm with an approximation ratio that improves as the average number of vehicles per location grows. Afèche et al. (2018) takes prices as fixed and studies admission control of customers and centralized re-positioning of drivers and shows that the value of these control policies are largest at moderate capacity and they increase with demand imbalances. Besbes et al. (2018) uses a stylized model on a line, where taxis make myopic decisions based on the next ride and can relocate instantaneously. They show that the pricing problem can be spatially decomposed based on the attraction regions. The platform can use prices to create regions in which driver congestion is artificially high in order to motivate drivers to relocate to more profitable regions.

The third stream of literature studies discrete choice theory and structural estimation; see Ben-Akiva et al. (1985) for an introduction and Anderson et al. (1992) for examples. The most relevant papers in this area are Rust (1987), Nair (2007), and Su and Judd (2012). Rust (1987) studies a structural model in which a manager has to decide in each period whether to replace a bus engine or postpone the decision. Nair (2007) studies an equilibrium model in which consumers decide when to purchase a durable product based on their expectation of future prices and the seller decides on the optimal temporal pricing scheme. In a setting similar to Nair (2007), in this paper, taxi drivers make relocation decisions based on their expectation of future supply and demand, and the central planner sets the prices hoping to impact the relocation decisions of the drivers. Su and Judd (2012) shows that structural estimation problems can be viewed as constrained optimization problems. We use an approach similar to Su and Judd (2012) in Sections 5-6. Other relevant papers in this area include Akşin et al. (2013, 2016)[8], Li et al. (2014), Zheng (2016), and Zheng et al. (2018).

The last stream of literature studies mean field games. The mean field games theory was developed to study systems with an infinite number of rational agents in competition. The theory was developed independently by the mathematics community in Lasry and Lions (2007) and by the engineering community in Huang et al. (2003). This framework has been used to model various economic and engineering systems. Examples of such papers in the operations community include Adlakha and Johari (2013), Xu and Hajek (2013), Iyer et al. (2014), Adlakha et al. (2015), Gummadi et al. (2013), and Balseiro et al. (2015).

3 Model

This section introduces the empirical model used to estimate the effect of spatial pricing of taxi rides in NYC. We index the months between January 2010 and December 2013 by $k \in \{1, \ldots, K\}$ ($K = 48$). In what follows, we fix k and focus on non-holiday weekdays of month k. We assume that the primitives of the model are the same in all weekdays of the same month. In other words, although there is variation across months, we assume that the non-holiday weekdays in a month are i.i.d. copies of each other. Consider a connected directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with n nodes, i.e. $|\mathcal{V}| = n$. Each node represents an area in Manhattan, Brooklyn, Queens or one of the airports (La Guardia and JFK). Any pair of distinct nodes are connected with at most two directed edges, denoted by (i, j) and (j, i), respectively. Furthermore, each node has a loop, which represents travel within the area/node.

Let $t \in \{1, \ldots, T\}$ index time. Each t represents a five-minute interval between the hours of 6 AM and 4 PM (a typical NYC day-time shift), corresponding to $T = 120$. For all $i, j \in \mathcal{V}$, the average time and distance to travel from node i to node j are denoted by τ_{ij} and d_{ij}, respectively.[9] We allow (τ_{ij}, d_{ij}) to be different

[8] Ata et al. (2017) and Ata and Peng (2017) rigorously establish the existence and uniqueness of the equilibria in a similar setting.

[9] If $(i, j) \in \mathcal{E}$, τ_{ij} and d_{ij} are the travel time/distance on the edge connecting i to j. Otherwise, τ_{ij} and d_{ij}

from (τ_{ji}, d_{ji})[10] and do not impose a relationship between τ_{ij} and d_{ij}. We assume that τ_{ij} is integer-valued for all $i, j \in \mathcal{V}$ and $\tau_{ii} = 1$ for all i. To facilitate the analysis to follow, define

$$S_{ij} = \begin{cases} 1 & \text{if } (i,j) \in \mathcal{E}, \\ 0 & \text{otherwise.} \end{cases}$$

Let P_{ij} and F_{ij} denote the price per mile and the fixed portion of the fare, respectively, for a ride from node i to node j. The fare paid by a customer to his driver for a ride from node i to node j is $F_{ij} + P_{ij}d_{ij}$. The fixed portion is used to model the flag-drop and the JFK flat fare.[11] Taxes and tolls are not included in the calculations as they are fixed (over time) and taxis do not keep them. The price per mile and the fixed portion of the fare are allowed to depend on the origin and the destination of the ride in order to capture the flat fare structure of rides between JFK and Manhattan.[12] This also gives us sufficient flexibility to explore the origin-destination pricing in our counterfactual study.

Given prices $F = [F_{ij}]_{ij=1}^n$ and $P = [P_{ij}]_{ij=1}^n$, the arrival rate (per period) of potential customers who wish to go to node j from node i in period t, denoted by $\Lambda_{ij}^t(F,P)$, is given as follows:

$$\Lambda_{ij}^t(F,P) = A_{ij}^t \left[F_{ij} + P_{ij}d_{ij}\right]^\alpha \exp(\beta k) \qquad \text{for all } i,j,t. \tag{1}$$

The parameter A_{ij}^t is the fixed effect of demand and $\alpha < 0$ is the price-elasticity. As will be shown in Section 4, our data exhibits a trend over time (see Figure 3) and β captures the monthly trend in demand. Since α does not depend on the fare, the demand model in (1) is a constant-elasticity demand model.[13] The (potential) demand at node i in period t, denoted by $\Lambda_i^t(F,P)$, is given by

$$\Lambda_i^t(F,P) = \sum_{j=1}^n \Lambda_{ij}^t(F,P) \qquad \text{for all } i,t.$$

The probability that a customer at node i is headed to node j, denoted by $\pi_{ij}^t(F,P)$, is equal to

$$\pi_{ij}^t(F,P) = \frac{\Lambda_{ij}^t(F,P)}{\Lambda_i^t(F,P)} \qquad \text{for all } i,j,t. \tag{2}$$

correspond to the travel time/distance of full taxis, that are calculated from the data.

[10] This allows us to capture the spatial variation in the traffic speed as well as the impact of one-way streets and streets with different traffic speeds in each direction.

[11] The fare of a ride from JFK to Manhattan (and vice versa) after September 2012 is \$52 plus the \$4.5 rush hour surcharge (4PM to 8PM on weekdays, excluding legal holidays). Since this paper focuses on the day shift (6AM to 4 PM) on weekdays, the rush hour surcharge is not included in the calculations.

[12] Consider a customer at JFK airport at a non-rush hour time in September of 2012. If he is headed to Manhattan, his fare is \$52. This corresponds to $F_{ij} = 52$ and $P_{ij} = 0$. However, if he is headed to another location, for example to La Guardia airport for a connecting flight, his fare is calculated based on the distance he intends to travel. In this case, $F_{ij} = 2.5$ and $P_{ij} = 2.5$ per mile.

[13] See Van Zandt (2012, Page 152) for a discussion on constant-elasticity demand models.

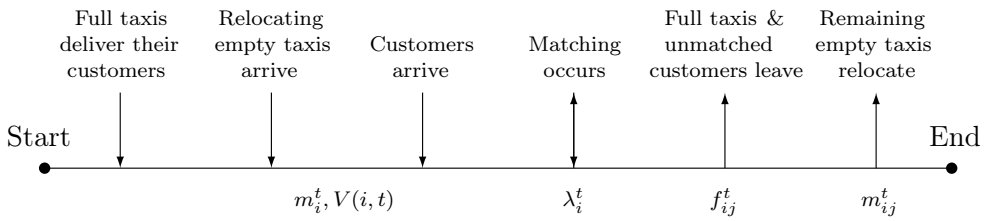


Figure 1: Timing of events and definition of variables in each period.

In what follows, we denote satisfied demand at node i in period t by λ_i^t. Note that $\lambda_i^t \leq \Lambda_i^t(F,P)$.

The timing of events in each period is depicted in Figure 1. We denote the number of active taxis during the day shift by M and the average of M across the 48 month by $\overline{M}$. We normalize the number of active taxis to an $M/\overline{M}$ mass (as done usually in mean field models).[14] The mass of empty cars at node i at the begining of period t is denoted by m_i^t. Also, let m_{ij}^t denote the mass of empty cars that decide to relocate from node i to node j at the end of period t, and f_{ij}^t denote the mass of full cars that picked up a customer at node i, who wishes to go to node j, in period t.

An empty taxi at node i in period t that could not pick up a customer relocates to node j with probability q_{ij}^t. In particular, it stays idle at node i until the next period with probability q_{ii}^t. These probabilities arise endogenously as drivers make their relocation decisions. Letting $\mathcal{A}(i) = \{j : S_{ij} = 1\}$ denote the set of nodes that can be reached from node i, an empty taxi at node i can only relocate to nodes in $\mathcal{A}(i)$. Moreover, since $q_{i\cdot}^t = \{q_{ij}^t : j \in \mathcal{A}(i)\}$ is a probability distribution for all i, it follows that

$$\begin{aligned} \sum_{j \in \mathcal{A}(i)} q_{ij}^t &= 1 \qquad \text{for all } i,t, \\ q_{ij}^t &= 0 \qquad \text{for all } j \notin \mathcal{A}(i), \\ q_{ij}^t &\geq 0 \qquad \text{for all } i,j,t. \end{aligned}$$

Let c denote the mean travel cost per mile (incurred by the drivers). We assume that the cost of traveling from node i to node j is equal to $c\, d_{ij} + \epsilon_{ij}$, where ϵ_{ij} denotes a Gumbel Min[15] (minimum extreme value type I) distributed idiosyncratic shock to the travel cost of the driver. The idiosyncratic cost shocks correspond to unobservable variables in the empirical industrial organization literature (see, e.g., Rust (1987)), which are observed by the driver but not recorded in the data. The cost shocks are assumed to be i.i.d. with mean zero and scale parameter σ. Furthermore, we let $V(i,t)$ denote the value function of an empty taxi's driver at node i in period t (before customers arrive); and N_i denotes the number of regions in node i. What

[14] Using $M/\overline{M}$ as opposed to the unit mass in all months allows us to capture the variation in the number of active taxis while ensuring that all quantities of interest are comparable across months.

[15] A Gumbel Min distributed random variable has the same distribution as the negative of a Gumbel Max (commonly referred to as Gumbel) distributed random variable.

constitutes a region will be discussed below in detail.

For the remainder of this section, we assume that the problem primitives $\overline{M}$, M, N_i, c, F_{ij}, P_{ij}, τ_{ij}, d_{ij}, S_{ij}, A^t_{ij}, α, β, k, σ, and the initial distribution of empty taxis, m^1_i, are given for all i, j, t. We then proceed to characterize λ^t_i, m^t_i, m^t_{ij}, f^t_{ij}, q^t_{ij}, and $V(i,t)$ by deriving a set of equations they must satisfy. To characterize the satisfied demand λ^t_i, we focus on the following question: *If there are m taxis and Λ customers at the beginning of a period in an area (node), what is the expected number of served customers (matches between taxis and customers)?* The answer depends on whether the matching of customers and taxis is perfect (frictionless) or imperfect (with friction). Matching can be perfect in nodes where customers and taxis are matched through a mobile application. In most other cases, the matching is imperfect and taxis need to search for customers block by block.

Perfect Matching. When matching is perfect, the number of matches is the minimum of the number of taxis and customers in the node in that period, i.e.

$$\lambda^t_i = \min\big(m^t_i, \Lambda^t_i(F,P)\big). \tag{3a}$$

When matching is imperfect, there can be unfulfilled customers and unutilized taxis simultaneously. We propose the following matching model in this case. Consider a node partitioned into $N \geq 2$ regions, which can be done so that it takes one period (i.e. five minutes) for a taxi to explore any of the regions.[16] We assume that customers and taxis randomly choose one of N regions (with equal probabilities) to explore and the number of matches in each region is the minimum of the number of customers and taxis in that region. The total number of matches in the node is equal to the sum of the matches made in the N regions.

Since customers and taxis choose regions with equal probabilities, all regions have the same expected number of matches. Consider an arbitrarily chosen region. Let X and Y denote the number of taxis and customers that chose that region. Since there are m taxis and each taxi chooses this region with probability $1/N$, we have $X \sim \text{Binomial}(m, 1/N)$. Similarly, $Y \sim \text{Binomial}(\Lambda, 1/N)$. Therefore, the expected number of matches in the region is equal to $\mathbb{E}\big[\min(X,Y)\big]$, and that in the node is equal to $N \times \mathbb{E}\big[\min(X,Y)\big]$. We would like to find a tractable approximation for $N \times \mathbb{E}\big[\min(X,Y)\big]$ that is monotone in Λ and equals zero when $\Lambda = 0$. To do so, we use a normal approximation to the Binomial distribution that is further refined by a linear approximation for small values of demand to ensure monotonicity.[17]

[16] Regions are the smallest geographic units in our model and they are obtained by dividing the street miles in a node such that each region takes exactly one period to explore. One way to compute the number of regions N is to use the data on the average speed of taxis and the total street-miles in that node.

[17] Our numerical results are robust to the approximation used in small values of demand. This approximation is required at demand values much smaller than one customer per period, which are observed in less than 1% of the (i,t) pairs in the numerical experiments of Section 6.

Imperfect Matching. When matching is imperfect, we propose the following model:

$$\lambda_i^t = \begin{cases} G\big(\Lambda_i^t; m_i^t\big) & \text{if } \Lambda_i^t \geq \hat{\Lambda}_{m_i^t}, \\ G\big(\hat{\Lambda}_{m_i^t}; m_i^t\big)\, \frac{\Lambda_i^t}{\hat{\Lambda}_{m_i^t}} & \text{otherwise,} \end{cases} \tag{3b}$$

where $G(\Lambda; m) = \Lambda\, \Phi(\nu) + m\, \Phi(-\nu) - (m - \Lambda)\, \phi(\nu)/\nu$ with $\Phi(\cdot)$ and $\phi(\cdot)$ denoting the cdf and pdf of the standard normal distribution, respectively,

$$\nu \triangleq \sqrt{\frac{\overline{M}}{N_i - 1}}\, \frac{m - \Lambda}{\sqrt{m + \Lambda}},$$

and $\hat{\Lambda}_m = \min\big\{\Lambda \geq 0 : G(\Lambda; m) \text{ is strictly increasing on } (\Lambda, \infty)\big\}$; see Appendix I.1 for the derivation of (3b).

The term $G(\Lambda; m)$ in (3b) is derived using a normal approximation for the Binomial distributions. Although this approximation elegantly captures the key features of search friction that we seek to model, it is not monotone in a neighborhood of the origin and fails to satisfy $\lambda_i^t = 0$ when $\Lambda_i^t = 0$. The linear approximation at $\Lambda_i^t < \hat{\Lambda}_{m_i^t}$ in (3b) remedies these two issues and retains the other desirable features of the normal approximation.

Buchholz (2018) tailors the imperfect matching model proposed in Burdett et al. (2001) to the taxi market. In Buchholz (2018)'s model, first the number of drivers and their locations are revealed to the customers. Then, each customer chooses a taxi. When more than one customer chooses a taxi, only one of the customers is served and the others leave the system unfulfilled. In our matching model, taxis and customers in each region observe each other (and not the taxis and customers in other regions) and matching in each region is frictionless. Due to this frictionless matching in each region, our matching model results in a higher expected number of matches when the number of taxis and customers are sufficiently higher than the number of regions (nodes with high density of supply and demand). However, when the number of taxis or customers is low (nodes with low density of supply or demand), due to the fact that taxis can be matched only with customers in their region, our matching model results in a lower number of matches. For a detailed comparison of our matching model with the matching model of Buchholz (2018), see Appendix B.

Next, we describe the flow balance equations and the Bellman equation, governing the system dynamics and the relocation decisions of empty taxis, respectively. These equations hold under both the perfect and imperfect matching.

Flow Balance of Full Cars. The mass of empty cars that picked up a customer at node i headed to node j in period t, f_{ij}^t, must be equal to λ_i^t, the satisfied demand at node i in period t, multiplied by the

probability that the customers are headed to node j, $\pi_{ij}^t(F,P)$. Therefore, we must have

$$f_{ij}^t \;=\; \lambda_i^t \, \pi_{ij}^t(F,P) \qquad \text{for all } i,j,t. \tag{4}$$

Flow Balance of Empty Cars. The mass of empty cars that decided to relocate from i to j at the end of period t (after they could not obtain a customer), m_{ij}^t, must be equal to the mass of empty cars after the customers are served, $\left(m_i^t - \lambda_i^t\right)$, multiplied by the probability that a car will relocate to node j, q_{ij}^t; see Figure 1 for the timing of events in a period. Therefore, we must have

$$m_{ij}^t \;=\; \left(m_i^t - \lambda_i^t\right) q_{ij}^t \qquad \text{for all } i,j,t. \tag{5}$$

Flow Balance at Nodes. The mass of empty cars in node i at the begining of period t is equal to the sum of the mass of empty cars that arrived at node i in period t and the mass of full cars that dropped their customer at node i in period t. Therefore, we must have

$$m_i^t \;=\; \sum_{j \in \mathcal{A}(i)} m_{ji}^{t-\tau_{ji}} \;+\; \sum_{j=1}^{n} f_{ji}^{t-\tau_{ji}} \qquad \text{for all } i,t. \tag{6}$$

Mass Balance. Since taxis are either empty or full and the mass of (relocating) empty taxis and full taxis at any period t must sum to $M/\overline{M}$, the total mass of taxis on the graph, we must have

$$\sum_{i=1}^{n} \sum_{j \in \mathcal{A}(i)} \sum_{s=1}^{\tau_{ij}} m_{ij}^{t-s} \;+\; \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{s=1}^{\tau_{ij}} f_{ij}^{t-s} \;=\; M/\overline{M} \qquad \text{for all } t. \tag{7}$$

Next, we describe how taxi drivers make their relocation decisions.

Bellman Equation. Consider an infinitesimal driver and assume that the mass of the empty and full taxis, $[m_{ij}^t]_{i,j=1}^n$ and $[f_{ij}^t]_{i,j=1}^n$, and satisfied demands, $[\lambda_i^t]_{i=1}^n$, are given for all $t \in \{1,\ldots,T\}$. The driver moves around on the graph until $t = T$, picks up customers, delivers them to their destination and makes relocation decisions when he can not pick up a customer. The objective of the driver is to maximize his total profit. Recall that the travel cost of the infinitesimal driver from node i to node j is equal to $c\,d_{ij} + \epsilon_{ij}$, where ϵ_{ij} denotes a Gumbel Min (minimum extreme value type I) distributed idiosyncratic shock to the travel cost of the driver. The cost shocks are assumed to be i.i.d. with mean zero and scale parameter σ.

The state of the empty infinitesimal taxi is $s = (i, t, \epsilon^{(i)})$, where $i \in \{1,\ldots,n\}$ denotes the location (node) of the taxi, $t \in \{1,\ldots,T\}$ denotes the period, and $\epsilon^{(i)} = \left(\epsilon_{ij}\,;\, j \in \mathcal{A}(i)\right)$ is the vector of iid cost shocks. Each element of $\epsilon^{(i)}$ has a Gumbel Min (minimum extreme value type I) distribution with mean zero and scale parameter σ. We denote the observable (to the researcher) state of the empty taxi by $x = (i, t)$. Therefore,

$s = (i, t, \epsilon^{(i)}) = (x, \epsilon^{(i)})$. The driver of the empty taxi at state $s = (i, t, \epsilon^{(i)})$ can choose to relocate to node $j \in \mathcal{A}(i)$. Therefore, the action set of the driver at state $s = (i, t, \epsilon^{(i)})$ is $\mathcal{A}(i)$. The following proposition characterizes the driver's value function and the relocation probabilities; see Appendix I.2 for its proof.

Proposition 1. *The value function of the infinitesimal empty taxi is given by*

$$V(i,t) = \frac{\lambda_i^t}{m_i^t}\left(\sum_{j=1}^{n} \pi_{ij}^t(F,P)\left(F_{ij} + [P_{ij} - c]\, d_{ij}\right) + \sum_{j=1}^{n} \pi_{ij}^t(F,P)\, V(j, t+\tau_{ij})\right) + \sigma\left(1 - \frac{\lambda_i^t}{m_i^t}\right)\log\left[\sum_{j\in\mathcal{A}(i)} \exp\left(\frac{V(j,t+\tau_{ij}) - c\, d_{ij}}{\sigma}\right)\right] \qquad \text{for all } i \text{ and } t \le T, \tag{8}$$

and $V(i,t) = 0$ for all i and $t > T$. The relocation probabilities are given by

$$q_{ij}^t = \begin{cases} \dfrac{\exp\left(\left[V(j,t+\tau_{ij}) - c\, d_{ij}\right]/\sigma\right)}{\sum_{l\in\mathcal{A}(i)} \exp\left(\left[V(l,t+\tau_{il}) - c\, d_{il}\right]/\sigma\right)} & \text{for } j \in \mathcal{A}(i), \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

Proposition 1 applies to any matching model. The first term on the right-hand side of Equation (8) captures the expected value of picking up a customer and the second term captures the expected value of the best relocation decision, where λ_i^t/m_i^t is the probability of picking up a customer. The value of picking up a customer is the expected fare of the ride plus the expected value function at the destination of the ride. Note that the arguments in the exponential function are the scaled values of the value function at the destination of the ride minus the travel cost. For an analysis of strategic relocation of the drivers that is similar to ours, see Lagos (2000) and Buchholz (2018).

Next, we define the mean field equilibrium, and Theorem 1 establishes its existence; see Appendix I.3 for its proof.

Definition 1. *Given the problem primitives $\overline{M}$, M, N_i, c, F_{ij}, P_{ij}, τ_{ij}, d_{ij}, S_{ij}, A_{ij}^t, α, β, k, σ, and m_i^1, the solution $\left(\lambda_i^t, m_i^t, m_{ij}^t, f_{ij}^t, q_{ij}^t, V(i,t)\right)$ to Equations (3)-(9) is called the mean field equilibrium.*

Theorem 1. *Given the problem primitives $\overline{M}$, M, N_i, c, F_{ij}, P_{ij}, τ_{ij}, d_{ij}, S_{ij}, A_{ij}^t, α, β, k, σ, and m_i^1, there exists a mean field equilibrium.*

4 Data

Our data set is the NYC yellow taxi trip record data spanning four years from January 2010 to December 2013.[18] In this time-span, NYC yellow taxis offered an average of 477,497 rides per day (174.3 Million per

[18]NYC taxi trip record data set is released by the NYC Taxi and Limousine Commission (TLC).

year). NYC yellow taxis are only available through street hails.[19] For each ride, the data set specifies the time-stamp (date and time up to the second) and location (longitude and latitude) of the pick-up/drop-off and itemized fares (fares, taxes, tolls, and tips) paid by the customers. The data set also includes identifiers for the taxi drivers that offered each ride. Therefore, we observe the status (full/empty) of the taxis at all times, the times at which taxis pick up or drop off customers, and the destinations of the customers. Note that we only observe the location of the taxis when they pick up or drop off a customer.

A preliminary look at the data indicates that weekdays and weekends as well as day and night shifts exhibit significant spatial and inter-temporal variations in the ride traffic. Figure 2 depicts the number of pick-ups per minute in the entire city on an average day. The number of pick-ups fluctuates between a minimum of 66 customers per minute (around 5:00AM) and a maximum of 507 customers per minute (around 7:30PM).[20] To focus on relatively busy hours of the day when the ride traffic is relatively stable, in the remainder of this paper, we focus on the day shift on weekdays (i.e., 6AM to 4PM). Furthermore, we disregard trips with a duration more than three hours or a distance longer than a hundred miles. These instances are a combination of out-of-town trips and inaccurate data entries and account for less than 0.5% of the trips. Figure 3 depicts the average number of pick-ups during the day shift on a weekday in each month between January 2010 and August 2012. Prices did not change in this time-span. The dashed line in Figure 3, which depicts the regression of the average number of pick-ups on the month index k as in (1), highlights the trend in the average number pick-ups.

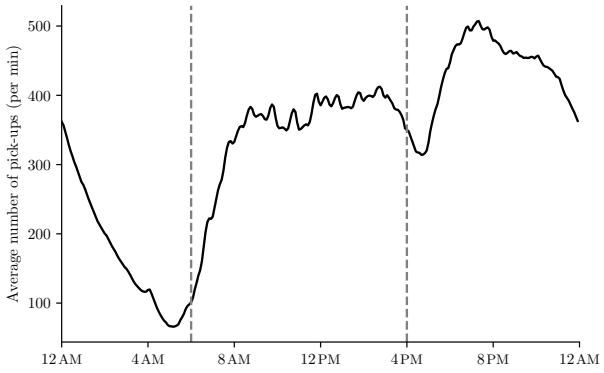


Figure 2: Average number of pick-ups per minute in NYC. The dashed vertical lines depict the begining and the end of the day shift (6AM and 4PM, respectively).

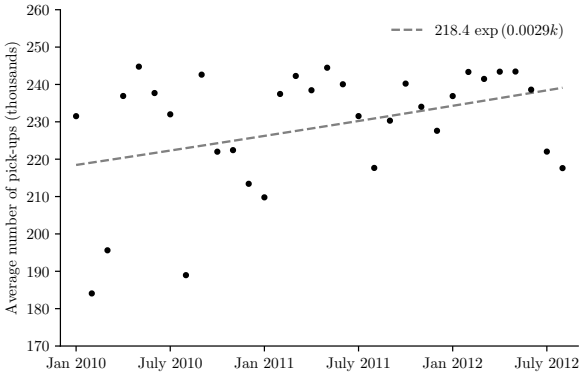


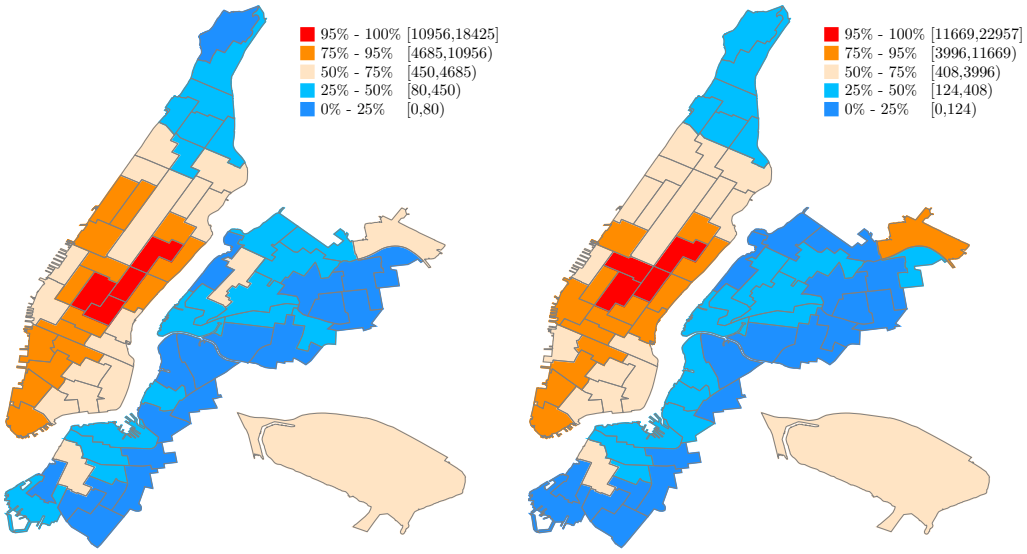
Figure 3: Average number of pick-ups in NYC during the day shift on a weekday of each month. The dashed line highlights the monthly trend in the number of pick-ups.

There is significant spatial variation in the number of pick-ups across the city, with 93.8% of the rides (during the day shift) originating at Manhattan, 3.5% in the airports, 1.2% in Brooklyn, 1.4% in Queens,

[19] Prearranged services are offered by For-Hire-Vehicles (FHV).

[20] A similar analysis can be carried out for the night shift.

and less than 0.1% of the rides originating in Staten Island and Bronx. We focus our analysis on the rides taking place between the two airports (JFK and La Guardia) and parts of Manhattan, Brooklyn, and Queens that have sufficiently high traffic; see Appendix A. This area constitutes 99.4% of the pick-ups, 98.5% of the drop-offs, and 97.9% of the pick-up, drop-off pairs in NYC.[21] We divide this area into seventy five nodes and study trip statistics in each of them; see Figure 16 for the nodes. This node definition provides sufficient separation between the major hubs of the city. Note that in the definition of the nodes, both size and density matter. This gives rise to a trade-off between size and density; for further details on the node definitions, see Appendix A. The average number of pick-ups and drop-offs in each of the seventy five nodes are depicted in Figure 4. Half of the pick-ups belong to the top ten nodes, with the top four nodes constituting a quarter of all pick-ups.



(a) Average number of pick-ups. (b) Average number of drop-offs.

Figure 4: Average number of pick-ups and drop-offs during the day shift on weekdays in each node.

A trip on average is 2.7 miles and 13.2 minutes long. There is considerable spatial variation in the average trip duration and distance across the city; see Figures 19-20 in Appendix C. Trips originating in Manhattan are the shortest with an average of less than three miles in most nodes and trips originating in the airports are the longest with an average of nine miles in La Guardia and fifteen miles in JFK. Trips originating in Brooklyn and Queens fall in between with an average of three to five miles.

In summary, our analysis focuses on 217,051,494 taxi rides taken place between 6AM and 4PM on weekdays

[21] Note that only 97.5% of the trips in the data set have pick-up coordinates that fall within the boundaries of NYC. Inaccurate pick-up coordinates (zero entry for longitude or latitude) account for 2.0% of the trips and out-of-town trips account for the remaining 0.5% of the trips.

(excluding federal holidays) from January 2010 to December 2013, with pick-up and drop-off coordinates in the shaded area in Figure 16.

5 Estimation

This section uses the nodes depicted in Figure 16 to estimate the primitives of the model; see Appendix D for a summary of the primitives. The procedure used to construct the nodes is outlined in Appendix A. The estimation is carried out in three steps. First, we directly estimate the problem primitives τ_{ij}, d_{ij}, S_{ij}, c, N_i, M, and m_i^1. Second, given τ_{ij}, d_{ij}, S_{ij}, c, N_i, M, and m_i^1, for each σ, we characterize the mean field equilibrium $\left(\lambda_i^t, m_i^t, m_{ij}^t, f_{ij}^t, q_{ij}^t, V(i,t)\right)$ at each month. The data set provides the sequence of pick-ups and drop-offs of the taxi drivers. For each σ, we use the equilibrium relocation probabilities to calculate the likelihood of the observed sequence of pick-ups and drop-offs. Then, we use maximum likelihood estimation to estimate σ. Third, we use the estimated (potential) demand at the estimated σ to estimate the demand curve parameters, A_{ij}^t, α, and β. Appendix F uses five-fold cross validation to examine the performance of our model in capturing the strategic relocations of the taxi drivers.

5.1 Offline Estimation of the Primitives τ_{ij}, d_{ij}, S_{ij}, c, N_i, M, and m_i^1

We use the trip duration and distances in the data set to estimate τ_{ij} and d_{ij}. We set $S_{ij} = 1$ if nodes i and j share a border or they are connected through a tunnel or bridge. Since JFK is geographically isolated and it does not share a border with any node, we add an arc between JFK and all nodes on the eastern border of Queens and Brooklyn. We use the fuel cost per mile and the depreciation cost per mile to estimate the cost of travel c, in each month. The fuel cost can be estimated from the historical gas prices[22] and the depreciation cost can be estimated from the salvage value of the taxis and the miles driven at the time of salvage.[23] Following Farber (2008) and Frechette et al. (2016), we define a shift of a taxi driver as a consecutive sequence of trips, where breaks between two trips cannot be longer than five hours. We use the shifts of the taxis drivers to calculate the number of active taxi drivers during the day. From 8AM to 4PM, the number of active taxi drivers shows little variation. Although the number of active taxi drivers between 6AM and 8AM is considerably lower than the rest of the day shift, for simplicity, we assume that the number of active taxis is fixed throughout the shift and let M equal to the average (across weekdays) of the maximum number of active taxi drivers during the day shift. Moreover, we use the distribution of the first pick-up of

[22] https://www.nyserda.ny.gov/Researchers-and-Policymakers/Energy-Prices/Motor-Gasoline/Monthly-Average-Motor-Gasoline-Prices.

[23] In this calculation, we use the price of the dominant NYC Taxi model (Camry), \$24,000 including taxes, the average retired taxi salvage value of \$3,000, and the average mileage of the NYC taxis at the time of retirement, 350,000; See https://nycitycab.com/Business/cabsforsalelist.aspx.

the taxis in the first fifteen minutes (average search time of the taxis until their next pick-up) of the day shift as the initial distribution of cars.

We estimate the number of regions in non-airport nodes as follows: Consider node i. Regions in node i are defined such that it takes a taxi exactly one period (five minutes) to explore each region. Therefore, it takes a taxi N_i periods ($5 \times N_i$ minutes) to search all the streets in node i. Thus, N_i can be estimated by dividing the average time it takes a taxi drivers to search all the streets of node i by the length of a period (five minutes). We use the average traffic speed in each zone and the shape file of NYC streets[24] to compute the average time it takes to explore all the streets in a node; see Table 5 and Figure 14 in Appendix A for the average traffic speeds. At the airports, we set N_i equal to the number of taxi stands in that airport.[25]

5.2 Estimation of the Standard Deviation of Cost Shocks

Identification. We assume that taxi drivers in our dataset relocated rationally based on the model in Proposition 1. As proved in Proposition 1, taxi drivers relocate to those nodes for which the difference of their value function and the travel cost is higher with higher probabilities. However, the impact of the value function at the destination diminishes with σ. For small values of σ, taxi drivers almost always relocate to the destination with the highest value function minus travel cost. As σ increases, drivers relocate to destinations with lower value function minus travel cost with a higher probability, i.e., relocation probabilities become more uniform. Intuitively, the impact of σ on the relocation probabilities helps us identify σ.

Estimation. We estimate σ in two steps. First, for each σ, we characterize the equilibrium. We use the equilibrium relocation probabilities to calculate the likelihood of the sequence of pick-ups and drop-offs observed in the data set for each feasible σ. We then use maximum likelihood estimation to estimate σ.

Recall that we index the months with $k \in \{1, \ldots, K\}$ and our data covers the months January 2010 through December 2013, i.e., $K = 48$. To emphasize the month dependence, we attach superscript k to various quantities of interest in the remainder of this section. We estimate λ_i^{tk}, the satisfied demand at node i in period t of month k, and $\pi_{ij}^{tk}(F^k, P^k)$, the distribution of the destinations of the customers at node i in period t of month k at fare structure (F^k, P^k)[26], directly from the data. Given $\overline{M}$, M^k, N_i, m_i^{1k}, c^k, F_{ij}^k, P_{ij}^k, $\pi_{ij}^{tk}(F^k, P^k)$, and λ_i^{tk}, for each σ, we solve for the mean field equilibrium that satisfies the Equations (4)-(9)

[24] https://data.cityofnewyork.us/City-Government/NYC-Street-Centerline-CSCL-/exjm-f27b/data.

[25] Taxi operation in airports is accurately captured by (3b). In each stand, taxis and customers see each other and matching is perfect. However, taxis and customers in different stands cannot be matched.

[26] Between January 2010 and August 2012, during the day shift, the fare of a ride between JFK and Manhattan was \$45 and the fixed portion of the fare and price per mile in the rest of the city were \$2.5 and \$2.0, respectively. On September 4th 2012, the fare structure of NYC taxi rides changed. The fare of a ride between JFK and Manhattan increased to \$52 and the price per mile increased to \$2.5. No price changes occurred between September 2012 and December 2013. We do not include Monday September 3rd 2012 in our analysis.

and $\lambda_i^{tk} \leq m_i^{tk}$.[27]

Next, we discuss how to use the equilibrium relocation probabilities and the sequence of pick-ups and drop-offs observed in the dataset to estimate σ. Let $m \in \{1, \ldots, M^k\}$ and $d \in \{1, \ldots, D^k\}$ index taxis and days, respectively, where D^k denotes the number of non-holiday weekdays in month k. Let $\mathcal{L}_{m,d,k}(\sigma)$ denote the likelihood of taxi m having the set of drop-off and subsequent pick-ups observed in the data on day d of month k that are not longer than two hours apart; see Appendix D for the calculation of $\mathcal{L}_{m,d,k}(\sigma)$. In the dataset, a taxi driver on average searches for 2.5 periods after a drop-off until it picks up his next customer. Therefore, we omit drop-off and subsequent pick-up tuples that are longer than two hours apart from the likelihood function because they are likely due to breaks the drivers may be taking in their shifts. Indeed, the great majority of these occur around noon and likely correspond to lunch breaks. Such drop-off and pick-up tuples constitute less than 3% of all such tuples.

The likelihood of observing the set of drop-off and subsequent pick-ups (that are not longer than two hours apart) observed in the data set given σ is $L(\sigma) = \prod_{k=1}^{K} \prod_{m=1}^{M} \prod_{d=1}^{D^k} \mathcal{L}_{m,d,k}(\sigma)$, where $\mathcal{L}_{m,d,k}(\sigma)$ is given in Equation (13) of Appendix D. Therefore, the maximum likelihood estimator of σ is the solution to

$$\text{maximize } \log\big(L(\sigma)\big) \text{ subject to (4) - (9)}, \tag{P1}$$

given $\overline{M}$, M^k, N_i, m_i^{1k}, c^k, F^k, P^k, $\pi_{ij}^{tk}(F^k, P^k)$, and λ_i^{tk} (estimated directly from data) for all i, j, t, k. We solve Equations (4)-(9) for each month $k \in \{0, \ldots, K\}$ and σ using the nonlinear optimization solver Knitro (Byrd et al. (2006)). Then, we compute the likelihood of each σ and search $\Re^+$ with a precision of 0.0001 for the solution to Problem (P1). We obtain the maximum likelihood estimate of $\sigma = 1.4048$; see Figure 21 for a graph of log-likelihood values. To compute its standard error, we use the parametric bootstrap method (see e.g., Horowitz (2001)). We generate 100 simulated datasets with the same size as our dataset using the estimated σ. We then estimate parameters of the simulated datasets and compute the standard error. This procedure results in the standard error of 0.001, a 95% confidence interval of $(1.4015, 1.4056)$. We also conduct a Monte Carlo experiment to show that Problem (P1) can recover the true σ; see Appendix E for details.

5.3 Estimation of Demand Curve Parameters

In this section, we use the mean field equilibrium at the estimated σ to estimate the demand curve parameters A_{ij}^t, α, and β. Recall that demand model parameters are assumed to be the same in all non-holiday weekdays of the same month. Equation (3b) is monotone in Λ_i^t for all m_i^t. Therefore, given the

[27] This approach and Definition 1 both characterize the mean field equilibrium. In Definition 1, λ_i^t satisfies equation (3) while in this characterization, λ_i^t is estimated from the data.

satisfied demand λ_i^t and the mass of empty taxis m_i^t, there exists a unique estimate for (potential) demand $\Lambda_{ij}^{tk}(F^k, P^k)$ that satisfies (3b). Using this estimate, we set[28] $\Lambda_{ij}^{tk}(F^k, P^k) = \pi_{ij}^{tk}(F^k, P^k)\ \Lambda_i^{tk}(F^k, P^k)$ for all i, j, t, k. This gives the demand for rides from node i to node j in period t of a day in month k. Recall that $\pi_{ij}^{tk}(F^k, P^k)$ has been estimated directly from the data. Since there are D^k non-holiday weekdays in month k, the (realized) total demand for rides from node i to node j in period t across all non-holiday weekdays in month k is $Y_{ij}^{tk} = \overline{M} D^k \Lambda_{ij}^{tk}(F^k, P^k)$ customers.[29] Similarly, given the demand curve parameters A_{ij}^t, α, and β, by equation (1), the expected total demand for the same rides is $\overline{M} D^k A_{ij}^t \left[F_{ij}^k + P_{ij}^k d_{ij}\right]^\alpha \exp(\beta k)$ customers.

We use a negative binomial (NB-2) model[30] to estimate the primitives of the demand model. To be more specific, we assume

$$Y_{ij}^{tk} \sim \text{Negative Binomial}\Big(\overline{M} D^k A_{ij}^t \left[F_{ij}^k + P_{ij}^k d_{ij}\right]^\alpha \exp(\beta k),\ \delta\Big) \tag{10}$$

and use maximum likelihood estimation to estimate its parameters. The parameter δ captures the over-dispersion of the estimated demands. As δ goes to zero, the negative binomial model approaches the Poisson model. For similar treatments and further details on negative binomial regression, see Hilbe (2011), Hardin et al. (2007, Chapter 13), and Lawless (1987). We obtain a price-elasticity estimate of $\alpha = -0.4735$ with standard error 0.0033, a monthly trend estimate of $\beta = 0.0024$ with standard error 1.7E-5 (equivalent to an annual trend of 2.92% with standard error 0.02%)[31], and a dispersion parameter estimate of $\delta = 0.1082$ with standard error 1.5E-4. The estimated demand across the forty eight months do not show drastic changes. For brevity, we focus on September 2012 and provide an overview of the demand estimates in this month. The estimated potential and satisfied demand during the day in the entire city and Brooklyn are depicted in Figure 5. As discussed in Section 3, the efficiency of matching is positively correlated with the density of supply and demand. Therefore, in Manhattan, where the density of supply and demand is high (see Figure 6), the majority of customers are served. In Brooklyn, however, due to high friction as a result of low density of supply and demand, only a small fraction of customers are served. Increasing matching efficiency could

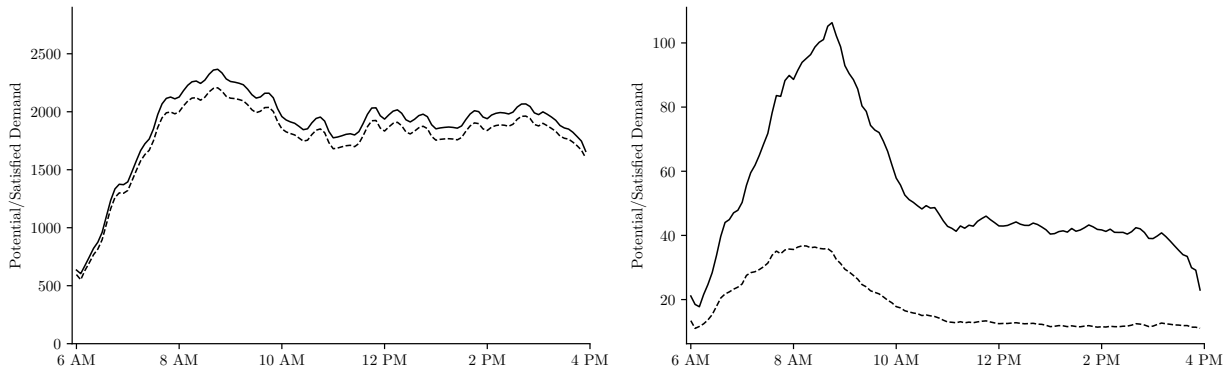
[28] Since customers who hailed a taxi are randomly chosen from all customers who needed a ride, the distribution of the destination of the served customers coincides with the distribution of the destinations of all customers.

[29] The term $\Lambda_{ij}^{tk}(F^k, P^k)$ is the (normalized) mean field demand. By multiplying $\Lambda_{ij}^{tk}(F^k, P^k)$ by the average number of active taxis $\overline{M}$ and the number of weekdays D^k in month k, we obtain the expected total number of customers who wanted a ride from node i to node j in period t on a weekday in month k.

[30] A random variable Y is said to have a negative binomial distribution with parameters (μ, δ) if $Y \sim \text{Poisson}(\lambda u)$, where $u \sim \Gamma(\delta, 1/\delta)$. In other words, a negative binomial model can be thought of as a Poisson model with gamma heteregeneity, where the gamma noise has a mean of one. As a result, negative binomial is commonly referred to as the Poisson-Gamma mixture. In a negative binomial model, the ratio of variance to mean is $1 + \delta\mu$. As δ goes to zero, the gamma noise vanishes and the negative binomial model approaches the underlying Poisson model. We use the negative binomial model as opposed to the commonly used Poisson model because it can capture the over-dispersion (variance being greater than the mean) observed in the estimated demand.

[31] Note that the trend of potential demand is not equal to the trend of satisfied demand; see Figure 3.

add substantial value, particularly in locations with low density of supply/demand, by increasing the number of served customers. Furthermore, better matching could impact the power of spatial prices; see Section 6.2 for further details.



(a) Entire city. (b) Brooklyn.

Figure 5: Estimated potential and satisfied demand (per period) during the day shift. The solid lines correspond to potential demand and the dashed lines correspond to satisfied demand.

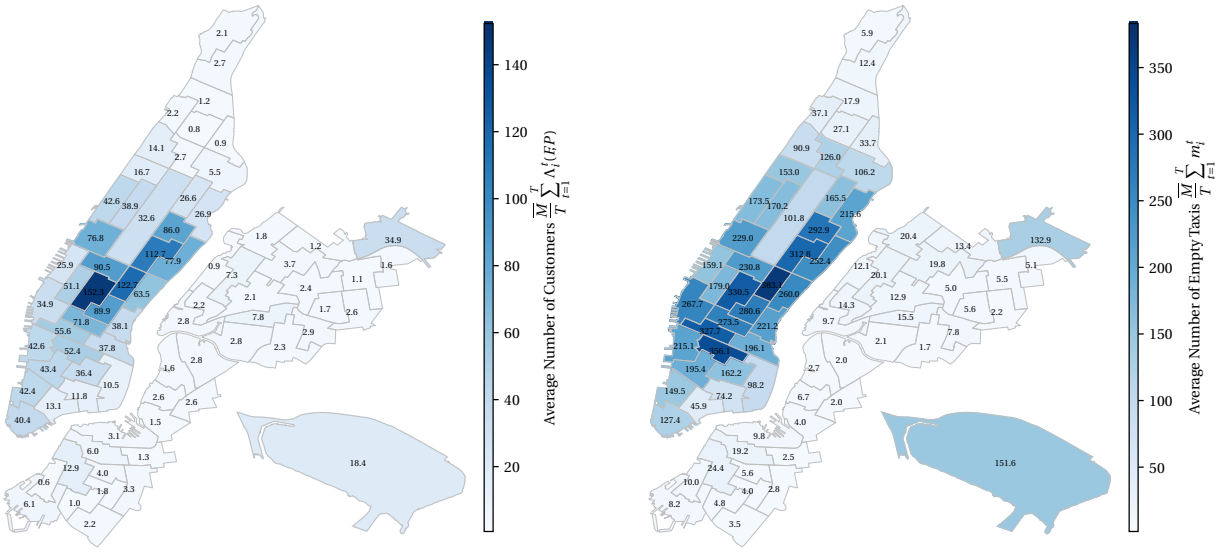
6 Counterfactual Analysis

Using the model primitives estimated in Section 5, this section studies the impact of spatial pricing of taxi rides and removing the (local) search friction on the taxi market in NYC. Following Section 5, we focus our analysis on September 2012 and study spatial prices that maximize the consumer surplus; see Appendix H for a similar analysis that focuses on maximizing drivers' profit. In what follows, we denote the fixed portion of the fare and the price per mile that were used in September 2012 by $\overline{F}_{ij}$ and $\overline{P}_{ij}$, respectively, and refer to them as the base prices. We assume that $F_{ij} = \eta_{ij}\overline{F}_{ij}$ and $P_{ij} = \eta_{ij}\overline{P}_{ij}$, where η_{ij} denotes the ratio of the price/fare to the base price/fare for rides from node i to node j. We refer to η_{ij} as the price multiplier for rides from node i to node j. We assume this particular form because of computational simplicity. Moreover, for the case of origin-destination pricing, this form is without loss of generality.

6.1 Spatial Pricing

Since the demand curve introduced in Equation (1) has a constant price-elasticity and demand is (locally) inelastic,[32] consumer surplus under this demand model is infinite. We use a truncation approach to tackle this issue (for a similar treatment, see Cohen et al. (2016) and Buchholz (2018)). We truncate the demand

[32]The absolute value of the price-elasticity of rides around the base prices is less than one.



(a) Average number of customers.

(b) Average number of empty taxis.

Figure 6: Average number of customers and empty taxis (in a five-minute period) during the day shift on weekdays in September 2012 (these are obtained through estimation; see Problem (P1) in Section 5).

at some price multiplier $\hat{\eta} > 1$. This is equivalent to assuming that no customer is willing to pay more than $\hat{\eta}$ times the base fare for a ride. We refer to $\hat{\eta}$ as the maximum price multiplier. The next proposition derives the consumer surplus; see Appendix I.4 for its proof.

Proposition 2. *Consumer surplus resulting from the rides from node i to node j in period t under price multiplier η_{ij} is equal to*

$$CS_{ij}^t(\eta_{ij}) \;=\; \frac{\lambda_i^t}{\Lambda_i^t(F,P)}\,\Lambda_{ij}^t(\overline{F},\overline{P})\,\left[\overline{F}_{ij}+\overline{P}_{ij}d_{ij}\right]\,\frac{\left[\hat{\eta}^{(\alpha+1)}-{\eta_{ij}}^{(\alpha+1)}\right]}{1+\alpha}. \tag{11}$$

We start by studying pricing based on the origin of the rides. Let $\overline{V}(i,t)$ denote the value function of

the taxi drivers under the base prices $(\overline{P}_{ij}, \overline{F}_{ij})$ and consider the following pricing formulation.

$$\underset{\eta_i}{\text{maximize}} \quad \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{t=1}^{T} \mathrm{CS}_{ij}^{t}(\eta_i) \tag{P2}$$

subject to (3) - (9) and

$$|\eta_i - 1| \leq \overline{\eta} \tag{P2a}$$

$$(F_{ij}, P_{ij}) \;=\; \eta_i\,(\overline{P}_{ij}, \overline{P}_{ij}) \tag{P2b}$$

$$\sum_{i=1}^{n} m_i^1 V(i,0) \geq \sum_{i=1}^{n} m_i^1 \overline{V}(i,0). \tag{P2c}$$

We call the solution $\{\eta_i\}_{i\in\mathcal{V}}$ to Problem (P2) the optimal origin-only price multipliers (or equivalently the optimal origin-only prices) with a maximum price variation of $\overline{\eta}$. Note that in Problem (P2), price multipliers only depend on the origin of the ride. Constraints (3)-(9) ensure that equilibrium conditions are satisfied. Constraints (P2a)-(P2b) ensure that prices do not deviate from the base prices by more than a factor of $\overline{\eta}$. Constraint (P2c) ensures that the drivers' profit is not hurt by the new pricing scheme (i.e., the average value function of the drivers in the begining of the day shift under the new pricing scheme is greater than or equal to the average value function of the taxi drivers in the begining of the day shift under the base prices.).

The increase in consumer surplus for four plausible values of maximum price multiplier $\hat{\eta}$ is presented in Table 2. As shown in Table 2, using an origin-only pricing scheme with a maximum price variation of 20% results in an increase in consumer surplus of $\$94,000$ - $\$325,000$ on every day shift on weekdays. This is equivalent to an increase in consumer surplus of \$0.44 - \$1.53 per ride (4.9% - 17.0% of the average fare paid by the customers). Although the increase in consumer surplus is sensitive to maximum price multiplier $\hat{\eta}$ (as shown in Table 2), the pattern of prices is not sensitive to $\hat{\eta}$; see Figure 24 in Section G.3. Therefore, in the remainder of this section we use maximum price multiplier $\hat{\eta} = 5$, which is equivalent to assuming that for a ride of \$9 (average fare before tax and tips in September 2012), no customer is willing to pay more than \$45. This is a conservative choice for $\hat{\eta}$ considering its impact on consumer surplus (see Table 2).[33]

Table 2: Consumer surplus under origin-only optimal spatial prices, with maximum price variation $\overline{\eta} = 0.2$.

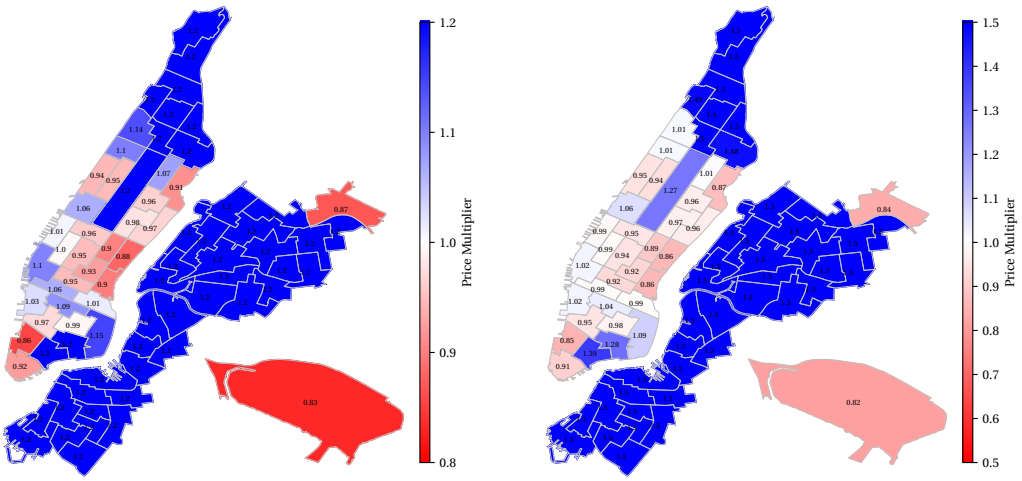
Maximum price multiplier ($\hat{\eta}$)	5	10	20	50
Total CS under the base prices	\$4.8M	\$8.4M	\$13.7M	\$24.4M
Total increase in CS	\$94K	\$130K	\$196K	\$325K
Increase in CS per ride	\$0.44	\$0.61	\$0.90	\$1.53
Increase in CS in terms of average fare	4.9%	6.8%	10.0%	17.0%

[33] In a similar setting, Cohen et al. (2016, Section 4) makes the conservative assumption that no customer is willing to pay more than 4.9 times the base prices for an Uber ride.

Table 3 presents the increase in consumer surplus, number of served customers, and the miles traveled by customers for different maximum price variations, $\overline{\eta}$. We observe that larger values of $\overline{\eta}$ results in larger increases in consumer surplus, as one would expect. However, the returns to the change in $\overline{\eta}$ are diminishing. The reason for the diminishing returns is illustrated in Figure 7. As the maximum price variation $\overline{\eta}$ increases, fewer nodes require a price variation higher than $\overline{\eta}$. Table 3 indicates that the optimal origin-only pricing scheme (for maximizing consumer surplus) results in a 2.6% increase in the number of served customers and a 3.9% increase in the miles traveled by customers for a maximum price variation of $\overline{\eta} = 0.5$.[34]

Table 3: Impact of maximum price variation on various performance metrics (maximum price multiplier $\hat{\eta} = 5$).

Maximum price variation ($\overline{\eta}$)	10%	20%	30%	40%	50%
Consumer surplus					
Total increase	\$62K	\$94K	\$113K	\$126K	\$135K
Per ride	\$0.29	\$0.44	\$0.53	\$0.59	\$0.63
In terms of average fare	3.3%	4.9%	5.9%	6.6%	7.0%
Number of served customers	1.1%	1.7%	2.1%	2.4%	2.6%
Miles traveled by customers	1.8%	2.7%	3.2%	3.6%	3.9%



(a) Maximum price variation $\overline{\eta} = 0.2$ (b) Maximum price variation $\overline{\eta} = 0.5$

Figure 7: Optimal origin-only price multipliers for different maximum price variations $\overline{\eta}$ (maximum price multiplier $\hat{\eta} = 5$).

Given the decomposition of the consumer surplus provided in Proposition 2, we expect the average fraction of customers served, demand, and the fare to impact the optimal price multiplier of a node. In what

[34] By using the total number of served customers or miles traveled by customers as objectives in Problem (P2), we can find spatial prices that result in larger increases in these metrics. However, such spatial prices result in a lower consumer surplus.

follows, we study the impact of these parameters on the optimal origin-only price multipliers.

The price multiplier of rides from node i to node j impacts consumer surplus only through the fraction of customers served, $\lambda_i^t/\Lambda_i^t(F,P)$, and the term $\big(\hat{\eta}^{(\alpha+1)}-\eta^{(\alpha+1)}\big)/\big(1+\alpha\big)$. As higher prices attract more drivers, they increase the fraction of customers served. The term $\big(\hat{\eta}^{(\alpha+1)}-\eta^{(\alpha+1)}\big)/\big(1+\alpha\big)$, however, decreases in price. Depending on which term dominates, the consumer surplus of a node could be increasing or decreasing in its price multiplier.

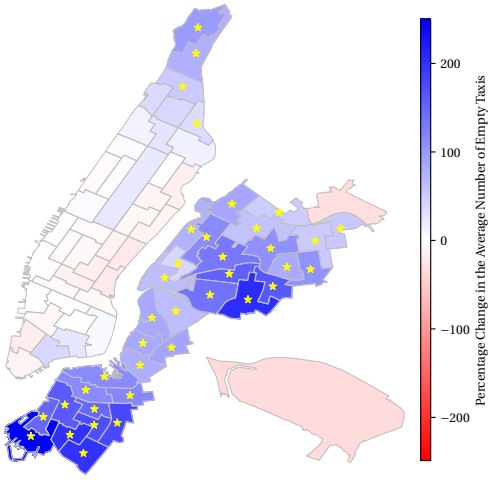
In well-served nodes, a high fraction of customers are served and the impact of price multiplier on the fraction of customers served is minimal. Therefore, in such nodes, the last term dominates and the consumer surplus is decreasing in the price multiplier. In contrast, in under-served nodes, a lower fraction of customers are served and the impact of price multiplier on the fraction of customers served is substantial. In under-served nodes, the first term dominates and the consumer surplus is locally increasing in the price multiplier. In such nodes, both customers and drivers benefit from higher prices. Therefore, the optimal price multiplier of these nodes is expected to be greater than one.

Nodes in which less than 80% of the customers are served (under the base prices) are marked with a star in Figure 8. We refer to these as under-served nodes. These nodes belong to upper Manhattan, Brooklyn and Queens, in which the supply of taxis is low; see Figure 6 for the map of the average supply and demand under the base prices. These nodes have a price multiplier greater than one. In response to a increase in prices in these nodes, more taxis relocate to them, which increases the supply of rides at these nodes; see Figure 8a. As a result of this increase in supply, a higher fraction of customers are served; see Figure 8b for the change in the fraction of customers served. This indicates that in these nodes the first term on the right hand-side of Equation (11) is dominant.

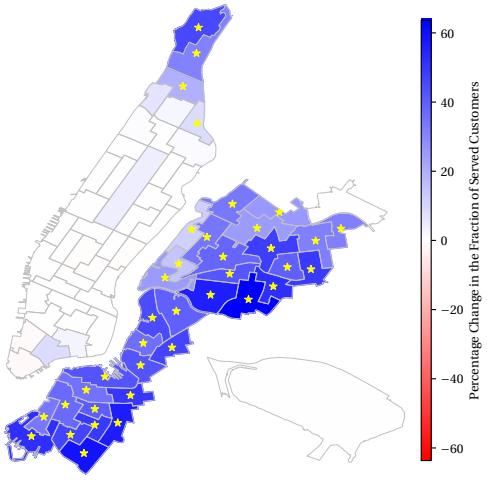
The fact that prices increase in less affluent neighborhoods in the city (upper Manhattan, Brooklyn, and Queens) highlights the necessity to use other mechanisms in addition to spatial prices (such as facilitating matching between customers and taxis through mobile applications) or subsidizing these neighborhoods. This issue is further explored in Section 6.2.

We refer to nodes in which at least 80% of the customers are served under the base prices as well-served nodes. For such nodes, the terms $\Lambda_{ij}^t(\overline{F},\overline{P})$ and $\overline{F}_{ij}+\overline{P}_{ij}d_{ij}$ are the critical drivers of consumer surplus. To see this, consider the potential revenue that can be collected at a node, $\sum_{t=1}^{T}\sum_{j=1}^{n}\Lambda_{ij}^t(\overline{F},\overline{P})\left[\overline{F}_{ij}+\overline{P}_{ij}d_{ij}\right]$. Note that both demand and fare impact the potential revenue of a node. In the majority of the well-served nodes with high potential revenue prices decrease while in the majority of well-served nodes with low potential revenue prices increase; see Figure 9.

Next, we study origin-destination pricing; see Appendix G.1 for its mathematical formulation. Similar to origin-only pricing, we use maximum price variation $\overline{\eta}=0.5$ and maximum price multiplier $\hat{\eta}=5$; see



(a) Average number of empty taxis.



(b) Fraction of served customers.

Figure 8: Percentage change in the average number of empty taxis and the fraction of served customers under the optimal origin-only prices compared to the base prices (maximum price variation $\overline{\eta} = 0.5$ and maximum price multiplier $\hat{\eta} = 5$).

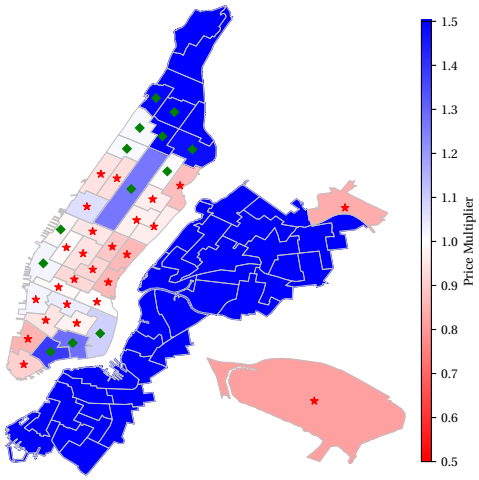


Figure 9: Well-served nodes with high (star) and low (diamond) potential revenue. Prices decrease in 88% of the nodes with high potential revenue and increase in 92% of the nodes with low potential revenue.

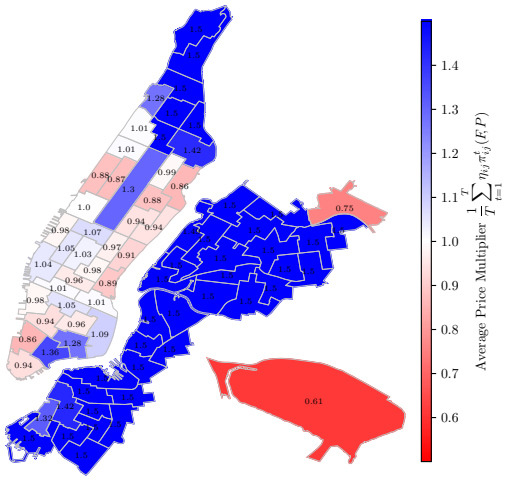


Figure 10: Average price multipliers for rides originating at each node in the city under origin-destination spatial pricing (the average is calculated with respect to the demand from the node to each destination).

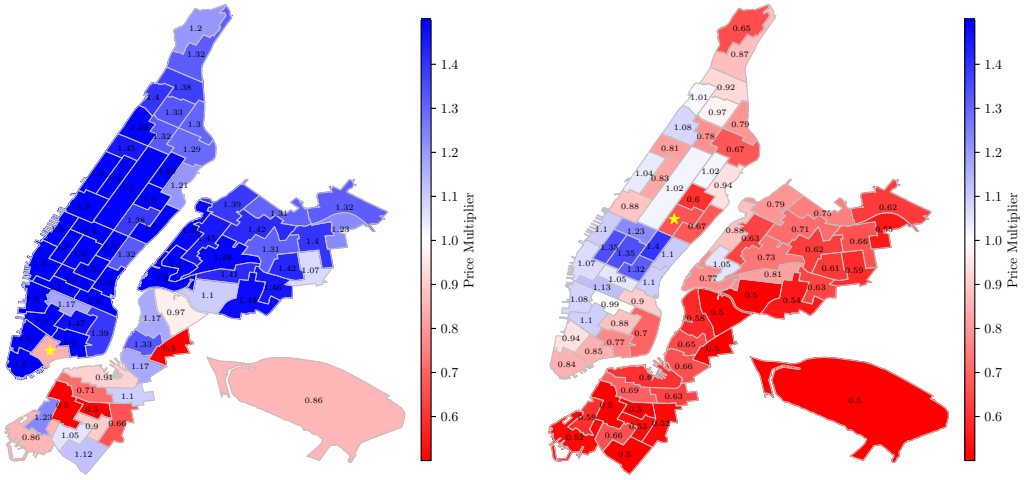
Appendix G.2 for further details. Origin-destination prices result in a \$0.79 increase in consumer surplus per ride, which is 24.4% higher than the increase in consumer surplus from origin-only pricing. They also results in a 3.2% increase in the number of served customers and an 7.4% increase in the miles traveled by customers.

The average optimal origin-destination price multipliers for the maximum price variation of 50% are depicted in Figure 10. Comparing Figures 7b and 10, we observe that the pattern of the average optimal origin-destination prices looks similar to the pattern of optimal origin-only prices. In the (majority of) under-served nodes, the same price is used for all destinations. In these nodes, the benefit of attracting more drivers by using higher average prices outweighs the benefit of price discrimination based on destinations. In the well-served nodes, however, price discrimination based on destinations is valuable. Figure 11 provides two examples of how destinations could impact the prices. Let us start with the node marked by a star in Figure 11a. Under origin-only pricing (see Figure 7b), the price of all rides from this node increase by 39%. However, under origin-destination pricing, the price of rides from this node to (parts of) Brooklyn are reduced. This reduction in prices increases the demand for rides to Brooklyn, which in turn, helps further increase the average number of the empty taxis in this region. That is, this increases the supply of taxis in Brooklyn for future rides. Since only 15% of the customers are headed to these destinations, this pricing pattern maintains (almost) the same average price as the origin-only pricing scheme. The pattern of reducing the price of rides headed to (some) under-served neighborhoods while increasing the price of rides headed to (some) well-served nodes is observed frequently. Figure 11b depicts another example of this pattern. Under origin-only pricing (see Figure 7b), the price of all rides from this node decrease by 3% (almost no change in prices). Under origin-destination pricing, the price of rides to midtown and parts of downtown increase while the price of all other rides decrease. This decreases the future supply in midtown and parts of downtown and increases the future supply in the rest of the city, particularly in Brooklyn and Queens. The price multiplier of the rides originating at this node range between 0.5 and 1.35 (a substantial change in prices). These examples exhibit how using destinations alongside origins in the pricing scheme allows us to maintain an average price similar to the origin-only pricing scheme while adjusting supply and demand in a more refined manner.

6.2 Removing Local Search Friction

Mobile application (such as Arro and Curb) can remove the local search friction by matching customers and taxi drivers in each neighborhood. The impact of such applications can be modeled by replacing the matching model (3b) with that in (3a).

Table 4 shows removing the local search friction alone results in a $\$268,000$ increase in consumer surplus in every day shift, a 4.3% increase in the number of served customers, and an 6.2% increase in customer miles. It also compares the improvement with those from spatial pricing. Removing the (local) search friction is on par with spatial pricing in terms of the number of served customers and customer miles. However, its impact on consumer surplus is considerably higher. Furthermore, although spatial prices introduced in



(a) A node in downtown Manhattan.

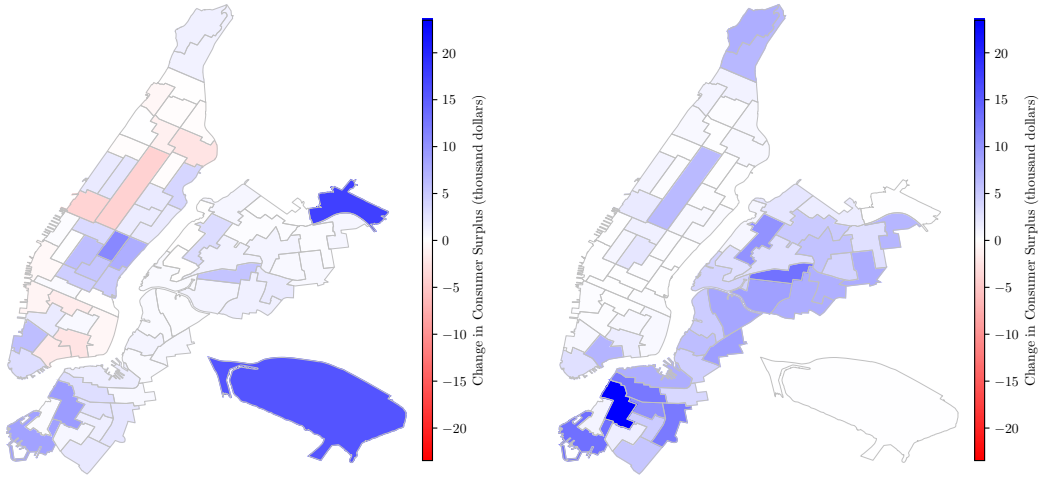
(b) A node in upper east side.

Figure 11: Optimal price multipliers under origin-destination pricing for the rides originating at the node denoted by a star (maximum price variation $\overline{\eta} = 0.5$ and maximum price multiplier $\hat{\eta} = 5$).

Section 6.1 have no impact on drivers' profit by design, since removing the local search friction increases the total number of served customers without changing the prices, drivers' profit increase by $\$48,000$ in every day shift when search friction is removed.

For brevity, we focus on origin-only pricing in the remainder of this section. Figure 12 depicts the change in consumer surplus from origin-only spatial pricing and removing the (local) search friction alone. The majority of the change in consumer surplus from spatial pricing is due to nodes with high demand and average fare, such as midtown Manhattan and the airports. Contrary to spatial pricing, when the local search friction is removed, well-served nodes (the majority of the nodes in Manhattan and the airports) do not benefit much, whereas under-served nodes enjoy the highest increase in consumer surplus. Under-served nodes benefit more from removing the (local) search friction than form spatial prices. In such nodes, matching is the key. In contrast, well-served nodes benefit more from spatial pricing than from removing the local search friction. In such nodes, spatial pricing is the key. This suggest that the impact of spatial prices and removing local search friction are complementary to each other. Therefore, we study a hybrid mechanism.

To be specific, the hybrid mechanism removes (local) search friction in under-served nodes and uses spatial prices in well-served nodes; see Appendix G.1 for its mathematical formulation. Table 4 provides a comparison of the impact of spatial prices, removing the (local) search friction, and the hybrid mechanism on consumer surplus and drivers' profits. The hybrid mechanism with maximum price variation $\overline{\eta} = 0.2$ captures 99.4% of the increase in consumer surplus from a hybrid mechanism with $\overline{\eta} = 0.5$. This mechanism



(a) Origin spatial pricing ($\overline{\eta} = 0.5$). (b) Removing the (local) search friction.

Figure 12: Change in consumer surplus from spatial pricing and removing the search friction (maximum price multiplier $\hat{\eta} = 5$).

needs only a little price variation, because the optimal price multipliers of the well-served nodes are close to one (see Figure 7).

Table 4: Comparison between spatial prices, removing local search friction, and the hybrid mechanism ($\overline{\eta} = 0.5$).

	Origin-Only Pricing	Origin-Destination Pricing	Removing Local Search Friction	Hybrid Mechanism	Proposed Mechanism	Citywide Origin-Only Pricing & Friction Removal
Consumer Surplus						
Total increase	\$135K	\$168K	\$268K	\$322K	\$417K	\$433K
well-served nodes	\$79K	\$109K	\$33K	\$82K	\$182K	\$186
under-served nodes	\$56K	\$59K	\$235K	\$240K	\$235K	\$247K
Per ride	\$0.63	\$0.79	\$1.26	\$1.52	\$1.96	\$2.04
In terms of average fare	7.0%	8.5%	13.9%	16.7%	21.5%	22.4%
Number of served customers	2.6%	3.2%	4.3%	5.7%	8.7%	8.9%
Miles traveled by customers	3.9%	7.4%	6.2%	8.1%	11.7%	12.0%
Drivers' Profit						
Total increase	\$0	\$0	\$48K	\$0	\$0	\$0
Per ride	\$0	\$0	\$0.22	\$0	\$0	\$0
In terms of average fare	0%	0%	2.5%	0%	0%	0%

Next, we study citywide spatial pricing and friction removal. The optimal price multipliers of this mechanism are the solution to Problem (P2), where (3b) is replaced with (3a). As shown in Table 4, using citywide spatial pricing and friction removal generates \$433,000 of consumer surplus. This is equivalent to a 34.4% improvement over the hybrid mechanism. The majority of this improvement is in well-served nodes. The optimal price multipliers under this mechanism are depicted in Figure 13. Although the pattern

of optimal prices in well-served nodes (under the base prices) is similar to the pattern of optimal prices in presence of friction (see Figure 7a), once (local) search friction is removed in well-served nodes, it is optimal to use more aggressive price variations in these nodes (i.e., we observe a larger deviation from the base-prices, both in nodes that have a price multiplier greater than one and those that have a price multiplier less than one). This implies that in the absence of friction, spatial prices are more powerful.

Policy makers may prefer to avoid price discrimination in less affluent neighborhoods of the city. As such, we propose a mechanism in which friction is removed in the entire city while spatial pricing is used only in well-served neighborhoods; see Appendix G.1 for its mathematical formulation. The proposed mechanism increases consumer surplus by \$1.96 per ride (96.3% of the consumer surplus generated by citywide spatial pricing and friction removal), serves 8.7% more customers, and increases customer miles by 11.7%.

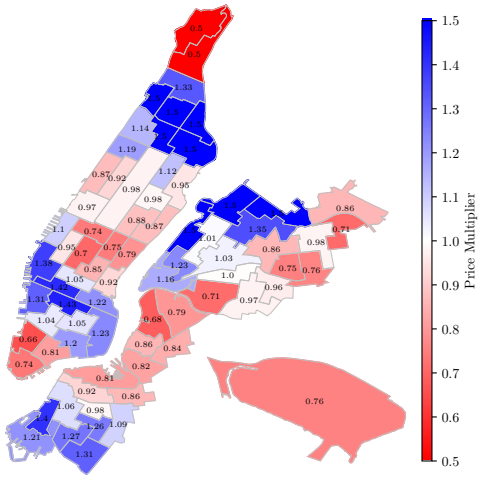


Figure 13: Optimal price multipliers under citywide origin-only pricing and friction removal (maximum price variation $\overline{\eta} = 0.5$ and maximum price multiplier $\hat{\eta} = 5$).

7 Concluding Remarks

We study the impact of spatial pricing on the taxi market in New York City. We use a mean field model, in which taxi drivers strategically search for customers in different neighborhoods across the city taking into account the spatial and temporal distribution of the supply and demand as well as the prices across the city. Our analysis reveals that spatial prices that only use origin information can increase consumer surplus by \$0.63 per ride, 7.0% of the average fare, and serve 2.6% more customers without hurting the drivers' profit. Similarly, spatial prices that utilize both origin and destination information can increase consumer surplus by \$0.79 per ride, 8.5% of the average fare, and serve 3.2% more customers. The optimal spatial prices increase in the under-served areas (e.g., Brooklyn, Queens, and upper Manhattan) whereas they decrease in some

well-served areas (e.g., midtown), encouraging the taxis to relocate accordingly, and hence, redistributing supply so as to increase consumer welfare.

We find that removing the local spatial search friction alone can increase consumer surplus by \$1.26 per ride, 13.9% of the average fare, and serve 4.3% more customers while simultaneously increasing drivers' profit by \$0.22 per ride. Removing the local search friction primarily impacts the under-served neighborhoods, whereas spatial prices primarily impact the well-served neighborhoods. This underscores the value of a hybrid mechanism. We propose a mechanism in which (local) search is eliminated in all neighborhoods while spatial pricing is only used in well-served neighborhoods. This mechanism increases consumer surplus by 21.5% of the average fare and serves 8.7% more customers, while avoiding price discrimination in less affluent neighborhoods of the city. The proposed mechanism achieves 96.3% of the benefits of a citywide spatial pricing and friction removal mechanism.

Our analysis has some limitations. First, a social planner might be most concerned with the impact of spatial pricing on the customer that is hurt by it the most. We limit the worst case impact of spatial pricing by introducing a maximum price constraint. Moreover, although we show that both customers and taxi drivers benefit from higher prices in the under-served areas in the presence of search friction, we do not study the implications of this phenomena fully. This phenomena leads to interesting work on subsidies and voucher mechanisms that is beyond the scope of our paper. However, our study of the hybrid mechanisms is motivated by this concern. Lastly, this paper uses the NYC taxi dataset from an era, 2010-2013, in which taxi substitutes were limited.[35] Future work in this area is needed to study the impact of spatial pricing on customers and taxi drivers in the presence of competition.

References

TLC factbook. NYC Taxi & Limousine Commission, 2014. URL nyc.gov/tlcfactbook.

TLC factbook. NYC Taxi & Limousine Commission, 2016. URL nyc.gov/tlcfactbook.

S. Adlakha and R. Johari. Mean field equilibrium in dynamic games with strategic complementarities. *Operations Research*, 61(4):971–989, 2013.

Sachin Adlakha, Ramesh Johari, and Gabriel Y Weintraub. Equilibria of dynamic games with many players: Existence, approximation, and market structure. *Journal of Economic Theory*, 156:269–316, 2015.

Philipp Afèche, Zhe Liu, and Costis Maglaras. Ride-hailing networks with strategic drivers: The impact of platform control capabilities on performance. Working paper, 2018.

[35] Ride-sharing platforms and street hail liveries (green taxis) had less than 5% market share in May 2014, five months after the period this paper studies ends; see TLC (2016) and Schneider (2018/11/22).

Frank J Aherne, Neil A Thacker, and Peter I Rockett. The bhattacharyya metric as an absolute similarity measure for frequency coded data. *Kybernetika*, 34(4):363–368, 1998.

Zeynep Akşin, Barış Ata, Seyed Morteza Emadi, and Che-Lin Su. Structural estimation of callers' delay sensitivity in call centers. *Management Science*, 59(12):2727–2746, 2013.

Zeynep Akşin, Baris Ata, Seyed Morteza Emadi, and Che-Lin Su. Impact of delay announcements in call centers: An empirical approach. *Operations Research*, 65(1):242–265, 2016.

Simon P Anderson, Andre De Palma, and Jacques-Francois Thisse. *Discrete choice theory of product differentiation*. MIT press, 1992.

Baris Ata and Xiaoshan Peng. An equilibrium analysis of a multiclass queue with endogenous abandonments in heavy traffic. *Operations Research*, 66(1):163–183, 2017.

Baris Ata, Peter W Glynn, and Xiaoshan Peng. An equilibrium analysis of a discrete-time markovian queue with endogenous abandonments. *Queueing Systems*, 86(1-2):141–212, 2017.

J Bai, K So, C Tang, X Chen, and H Wang. Coordinating supply and demand on an on-demand service platform: Price, wage, and payout ratio. Working paper, 2016.

Santiago R Balseiro, Omar Besbes, and Gabriel Y Weintraub. Repeated auctions with budgets in ad exchanges: Approximations and design. *Management Science*, 61(4):864–884, 2015.

Siddhartha Banerjee, Carlos Riquelme, and Ramesh Johari. Pricing in ride-share platforms: A queueing-theoretic approach. Working paper, 2015.

Siddhartha Banerjee, Daniel Freund, and Thodoris Lykouris. Pricing and optimization in shared vehicle systems: An approximation framework. Working paper, 2016.

Michele Basseville. Distance measures for signal processing and pattern recognition. *Signal processing*, 18 (4):349–369, 1989.

Moshe E Ben-Akiva, Steven R Lerman, and Steven R Lerman. *Discrete choice analysis: theory and application to travel demand*, volume 9. MIT press, 1985.

Omar Besbes, Francisco Castro, and Ilan Lobel. Surge pricing and its spatial supply response. Working paper, 2018.

Anil Bhattacharyya. On a measure of divergence between two multinomial populations. *Sankhyā: the indian journal of statistics*, pages 401–406, 1946.

Kostas Bimpikis, Ozan Candogan, and Saban Daniela. Spatial pricing in ride-sharing networks. Working paper, 2016.

Anton Braverman, JG Dai, Xin Liu, and Lei Ying. Empty-car routing in ridesharing systems. Working paper, 2016.

Nicholas Buchholz. Spatial equilibrium, search frictions and efficient regulation in the taxi industry. Working paper, July 2018.

Kenneth Burdett, Shouyong Shi, and Randall Wright. Pricing and matching with frictions. *J Polit Econ*, 109(5):1060–1085, 2001.

Richard H Byrd, Jorge Nocedal, and Richard A Waltz. Knitro: An integrated package for nonlinear optimization. In *Large-scale Nonlinear Optimization*, pages 35–59. Springer, 2006.

Gerard P Cachon, Kaitlin M Daniels, and Ruben Lobel. The role of surge pricing on a service platform with self-scheduling capacity. *Manufacturing & Service Operations Management*, 19(3):368–384, 2017.

John Gunnar Carlsson, Mehdi Behroozi, and Kresimir Mihic. Wasserstein distance and the distributionally robust tsp. *Operations Research*, 66(6):1603–1624, 2018.

Charles E Clark. The greatest of a finite set of random variables. *Operations Research*, 9(2):145–162, 1961.

Richard B Coffman and Chanoch Shreiber. The economic reasons for price and entry regulation of taxicabs (comment and rejoinder). *Journal of Transport Economics and Policy*, pages 288–304, 1977.

Peter Cohen, Robert Hahn, Jonathan Hall, Steven Levitt, and Robert Metcalfe. Using big data to estimate consumer surplus: The case of uber. Working paper, 2016.

Judd Cramer and Alan B Krueger. Disruptive change in the taxi business: The case of uber. *AER*, 106(5): 177–182, 2016.

Juan C Duque, Luc Anselin, and Sergio J Rey. The max-p-regions problem. *Journal of Regional Science*, 52 (3):397–419, 2012.

Henry S Farber. Reference-dependent preferences and labor supply: The case of new york city taxi drivers. *American Economic Review*, 98(3):1069–82, 2008.

James F Foerster and Gorman Gilbert. Taxicab deregulation: economic consequences and regulatory choices. *Transportation*, 8(4):371–387, 1979.

Guillaume R Frechette, Alessandro Lizzeri, and Tobias Salz. Frictions in a competitive, regulated market evidence from taxis. Working paper, 2016.

Sommer Gentry, Eric Chow, Allan Massie, and Dorry Segev. Gerrymandering for justice: redistricting us liver allocation. *Interfaces*, 45(5):462–480, 2015.

Ramki Gummadi, Ramesh Johari, Sven Schmit, and Jia Yuan Yu. Mean field analysis of multi-armed bandit games. Working paper, 2013.

Jonas Häckner and Sten Nyberg. Deregulating taxi services: a word of caution. *J Transp Econ Policy*, pages 195–207, 1995.

Jonathan Hall, Cory Kendrick, and Chris Nosko. The effects of uber's surge pricing: A case study. Working paper, 2015.

Robert E Hall. A theory of the natural unemployment rate and the duration of employment. *Journal of monetary economics*, 5(2):153–169, 1979.

James William Hardin, Joseph M Hilbe, and Joseph Hilbe. *Generalized linear models and extensions*. Stata press, 2007.

Joseph M Hilbe. *Negative binomial regression*. Cambridge University Press, 2011.

Joel L Horowitz. The bootstrap. In *Handbook of econometrics*, volume 5, pages 3159–3228. Elsevier, 2001.

Minyi Huang, Peter E Caines, and Roland P Malhamé. Individual and mass behaviour in large population stochastic wireless power control problems: centralized and nash equilibrium solutions. In *Dec & Con*, pages Vol 1. 98–103, 2003.

Krishnamurthy Iyer, Ramesh Johari, and Mukund Sundararajan. Mean field equilibria of dynamic auctions with learning. *Management Science*, 60(12):2949–2970, 2014.

Thomas Kailath. The divergence and bhattacharyya distance measures in signal selection. *IEEE transactions on communication technology*, 15(1):52–60, 1967.

Ron Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada, 1995.

Ricardo Lagos. An alternative approach to search frictions. *Journal of Political Economy*, 108(5):851–873, 2000.

Chungsang Tom Lam and Meng Liu. Demand and consumer surplus in the on-demand economy: The case of ride sharing. Working paper, 2017.

Jean-Michel Lasry and Pierre-Louis Lions. Mean field games. *Japanese Journal of Mathematics*, 2(1): 229–260, 2007.

Jerald F Lawless. Negative binomial and mixed poisson regression. *Canadian J. of Statistics*, 15(3):209–225, 1987.

Jun Li, Nelson Granados, and Serguei Netessine. Are consumers strategic? structural estimation from the air-travel industry. *Management Science*, 60(9):2114–2137, 2014.

Liu Ming, Tunay Tunca, Yi Xu, and Weiming Zhu. An empirical analysis of price formation, utilization, and value generation in ride sharing services. Working paper, 2017.

Harikesh Nair. Intertemporal price discrimination with forward-looking consumers: Application to the US market for console video-games. *Quantitative Marketing and Economics*, 5(3):239–292, 2007.

Erhun Ozkan and Amy Ward. Dynamic matching for real-time ridesharing. 2017.

Ofir Pele and Michael Werman. Fast and robust earth mover's distances. In *2009 IEEE 12th International Conference on Computer Vision*, pages 460–467. IEEE, September 2009.

Halsey Lawrence Royden and Patrick Fitzpatrick. *Real analysis*. Macmillan New York, 4 edition, 1968.

John Rust. Optimal replacement of GMC bus engines: An empirical model of harold zurcher. *Econometrica*, pages 999–1033, 1987.

Brian Sayler. IBISWorld Industry Report 48533 Taxi & Limousine Services in the US. Technical report, 2017.

Todd W Schneider. Analyzing 1.1 billion nyc taxi and uber trips, with a vengeance, 2018/11/22. URL http://toddwschneider.com/posts/analyzing-1-1-billion-nyc-taxi-and-uber-trips-with-a-vengeance.

John R Schroeter. A model of taxi service under fare structure and fleet size regulation. *The Bell Journal of Economics*, pages 81–96, 1983.

Matthew H Shapiro. Density of demand and the benefit of uber. Working paper, 2018.

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

Timothy Van Zandt. *Firms, Prices, and Markets*. 2012.

Jiaming Xu and Bruce Hajek. The supermarket game. *Stochastic Systems*, 3(2):405–441, 2013.

Pu Yang, Krishnamurthy Iyer, and Peter Frazier. Mean field equilibria for resource competition in spatial settings. Working paper, 2017.

Fanyin Zheng. Spatial competition and preemptive entry in the discount retail industry. Working paper, 2016.

Fanyin Zheng, Pu He, Elena Belavina, and Karan Girotra. Customer preference and station network in the london bike share system. Working paper, 2018.

A Node Definitions

This section discusses the procedure used to draw the boundaries of the nodes in Figure 16. We use the approach commonly referred to as regionalization/districting (e.g., see Duque et al. (2012)). In this approach, one starts with a set of small districts and aggregates them to create nodes of appropriate size. We use 2010 census tracts[36] as districts in our analysis and use a MIP (Mixed Integer Program) to aggregate the census tracts into nodes. Census tracts are permanent statistical subdivisions of a county defined by the United States Census Bureau and used in many statistical spatial analyses that require this level of granularity. NYC consists of 2166 census tracts of various shapes and sizes.[37] We limit our analysis to the 488 census tracts that have a minimum of one customer per street mile per day and are not isolated from the rest of the cluster (except for JFK airport). These census tracts cover 99.42% of the pick-ups in NYC between January 2010 and December 2013.

Objective. One of the assumptions used in the matching model introduced in Proposition **??** is that customers are uniformly distributed within the node. Therefore, we would like the census tracts that are aggregated to create a node to share similar demand densities (we use the number of pick-ups as a proxy for demand). We would also like the nodes to satisfy certain spatial contiguity, shape, and demand constraints. Specifically, we would like to avoid isolated census tracts in a node, avoid long and narrow nodes, and ensure a minimum number of pick-ups in each node. Moreover, we would like the number of regions (the only notion of node size in the matching model; see Section 3) in the nodes to show little variation across the city. This prevents the size of nodes from being a factor in the relocation decisions of the taxi drivers. We aim to achieve this goal by ensuring that the street miles in a node are such that it takes a taxi between $\underline{t}$ and $\bar{t}$ hours to search all the streets in the node.[38] Due to the significant spatial variation in the average traffic speed and density of pick-ups in NYC, we divide the city into six zones; see Figure 14 for the boundaries of the zones. These zones correspond to i) downtown/lower Manhattan (south of 14th street), ii) midtown Manhattan (14th street to 59th street), iii) central park area (59th street to about 110th street excluding central park), iv) upper Manhattan (north of 110th street to Inwood),[39] v) Brooklyn, and vi) Queens (excluding La Guardia and JFK airports). The average traffic speed and the density of pick-ups in each zone are presented in Table

[36] Census tracts should not be confused with the regions discussed in Section 3. Regions are defined in the matching section of this paper to model the search of taxi drivers for customers within a node. Census tracts, however, are permanent subdivisions of a county defined by the US Census Bureau that provide a stable set of geographic units for the presentation of statistical data.

[37] See https://data.cityofnewyork.us/City-Government/2010-Census-Tracts/fxpq-c8ku/data.

[38] The lower bound ensures that nodes are sufficiently large for computational tractability of the equilibrium. The upper bound helps avoid long travel times between adjacent nodes. Long travel times on adjacent nodes create an unintended incentive for taxi drivers to stay in their current node and avoid relocation due to the opportunity cost of long travels.

[39] For further information on Manhattan zones, see https://www.nybits.com/manhattan.

5.

Formulation. Next, we discuss the formulation we use to aggregate the census tracts in a fixed zone into nodes whilst satisfying the aforementioned conditions. We start by introducing the notation. Then, we proceed to the formulation and discuss each constraint in detail. This formulation is conceptually similar to the formulation used in Gentry et al. (2015). Let $(\mathcal{V}^c, \mathcal{E}^c)$ denote the graph of the census tracts that we would like to aggregate into nodes, where $(i,i) \notin \mathcal{E}^c$ and for $i \neq j$, $(i,j) \in \mathcal{E}^c$ if census tracts i and j share a border (of positive length). Let A^c denote the adjacency matrix of the graph, i.e, $A^c_{ij} = 1$ if $(i,j) \in \mathcal{E}^c$ and zero otherwise.

We would like the census tracts that will be aggregated into a node to be physically close to each other, i.e., we would like to avoid long and narrow nodes if possible. Therefore, we use an approach similar to Gentry et al. (2015) and choose a census tract as the center of each node and assign each census tract to the node whose center is physically closer to it. To that end, let z_i be a binary decision variable such that z_i is equal to one if census tract i is chosen as the center of a node and zero otherwise. Let x_{ij} be a binary variable that is equal to one if census tract i is assigned to the node whose center is census tract j and let y_j be a non-negative decision variable that denotes the maximum variation in the density of pick-ups (pick-ups per mile) across the census tracts assigned to the node whose center is census tract j. If census tract j is not the center of a node, $y_j = 0$.

Denote the average number of pick-ups per period in census tract i by D_i and the street miles of this census tract by S_i. Similarly, denote the lower bound on the number of pick-ups per period in a node by $\underline{D}$ and the lower and upper bounds on the street miles in a node by $\underline{S}$ and $\overline{S}$, respectively. Let $\hat{\delta}_{ij}$ denote the distance between the geometric centers of census tracts i and j and $\tilde{\delta}_{ijk} = 1$ if $\hat{\delta}_{ij} < \hat{\delta}_{ik}$ and zero otherwise. Also, let $\overline{\delta}$ denote the maximum allowable distance between the census tracts assigned to a node. We use a $\overline{\delta}$ that is equivalent to 10 minutes of travel in the zone.

The optimal node definition is the solution to

$$\underset{x,y,z}{\text{minimize}} \quad \sum_{j\in\mathcal{V}^c} y_j \tag{P0}$$

$$\text{subject to} \sum_{i\in\mathcal{V}^c} D_i\, x_{ij} \geq \underline{D} z_j \qquad \text{for all } j, \tag{P0a}$$

$$\sum_{i\in\mathcal{V}^c} S_i\, x_{ij} \leq \overline{S} z_j \qquad \text{for all } j, \tag{P0b}$$

$$\sum_{i\in\mathcal{V}^c} S_i\, x_{ij} \geq \underline{S} z_j \qquad \text{for all } j, \tag{P0c}$$

$$\overline{\delta} \geq \hat{\delta}_{ik}\Big(x_{ij} + x_{kj} - 1\Big) \qquad \text{for all } i,j,k, \tag{P0d}$$

$$\sum_{k\in\mathcal{V}^c} A^c_{ki}\, x_{kj} \geq x_{ij} \qquad \text{for all } i \neq j, \tag{P0e}$$

$$\sum_{k\in\mathcal{V}^c} \tilde{\delta}_{ijk}\, x_{ik} \leq 1 - z_j \qquad \text{for all } i,j, \tag{P0f}$$

$$\sum_{j\in\mathcal{V}^c} x_{ij} = 1 \qquad \text{for all } i, \tag{P0g}$$

$$x_{ij} \leq z_j \qquad \text{for all } i,j, \tag{P0h}$$

$$y_j \geq \left|\frac{D_i}{S_i} - \frac{D_k}{S_k}\right|\Big(x_{ij} + x_{kj} - 1\Big) \qquad \text{for all } i,j,k \tag{P0i}$$

$$x_{ij}, z_i \in \big\{0,1\big\},\ y_j \geq 0 \qquad \text{for all } i,j. \tag{P0j}$$

Problem (P0) minimizes the sum of the maximum variations in the density of pick-ups in each node (a measure of the heterogeneity in the density of demand within the nodes). We will show that in the optimal solution, if census tract j is not the center of a node, $y_j = 0$.

The constraints of Problem (P0) can be categorized into three groups: constraints concerning the demand and size of the nodes, constraints concerning the assignment of the census tracts, and constraints concerning the definition of the decision variables. We start with Constraints (P0a)-(P0d) that focus on the demand and size of the nodes. Constraint (P0a) ensures that all nodes have a minimum of $\underline{D}$ pick-ups per period. Similarly, Constraints (P0b)-(P0c) ensure that the street miles in each nodes fall between the lower bound of $\underline{S}$ and the upper bound of $\overline{S}$. Constraint (P0d) ensures that the census tracts in a node are not more that $\overline{\delta}$ miles apart. This constraint helps avoid long and narrow nodes. Note that the right hand-side of (P0d) is less than or equal to zero unless both census tracts i and k are assigned to the node whose center is census tract j, in which case the right-hand side of (P0d) is equal to $\hat{\delta}_{ik}$.

Constraints (P0e)-(P0g) focus on the assignment of the census tracts. Constraint (P0e) ensures that a census tract that is not the center of its node is connected to at least one other census tract in that node. Because, if census tract i is assigned to the node whose center is census tract j, the right-hand side of

(P0e) is equal to one. Therefore, there must be at least one census tract connected to census tract i who is also assigned to the node whose center is census tract j.[40] Constraint (P0f) ensures that census tracts are assigned to the node whose center is closest to them. Because, if census tract j is the center of a node, the right hand side of (P0f) is equal to zero. This ensures that $x_{kj} = 0$ for all nodes k in which $\tilde{\delta}_{ijk} = 1$. In other words, if census tract j is the center of a node, census tract i can not be assigned to a node whose center is further from i than census tract j; see Gentry et al. (2015) for a similar treatment. Constraints (P0g) ensures that all census tracts are assigned to a node.

Constraints (P0h)-(P0j) focus on the consistency in the definition of the decision variables. Constraint (P0h) ensures the consistency in the definitions of x_{ij} and z_i, i.e., if census tract i is assigned to the node whose center is census tract j, then $z_j = 1$. Constraint (P0i), alongside the objective function, ensures that y_j (at the optimal solution) is equal to the maximum variation in the density of pick-ups among the census tracts assigned to the node whose center is census tract j. Note that if census tract j is not the center of a node, the right-hand side of (P0i) is less than or equal to zero and $y_j = 0$ at the optimal solution. Finally, Constraint (P0j) enforces that x_{ij} and z_i are binary decision variables and y_j are non-negative.

Inputs. In what follows, we attach a subscript l to various quantities to indicate they correspond to zone l. Note that $\underline{S}_l$ = Average Traffic Speed in Zone $l \times \underline{t}_l$, where $\underline{S}_l$ denotes the lower bound on the total number of miles in nodes in zone l and $\underline{t}_l$ is the minimum time a taxi has to spend to travel all streets in a node in zone l. Similarly, $\overline{S}_l$ = Average Traffic Speed in zone $l \times \overline{t}_l$, where $\overline{S}_l$ and $\overline{t}_l$ are the analogous upper bounds. Recall that D_i/S_i denotes the pick-up density of census tract i and $\mathcal{V}_l^c$ denotes the set of census tracts in zone l for $l = 1, \ldots, L$, where L denotes the number of zones. We observe that $\{D_i/S_i : i \in \mathcal{V}_l^c\}$ exhibits significant variation for all zones; see the third column of Table 5 for their coefficient of variation. Define

$$\text{PD}_l = \frac{\sum_{i \in \mathcal{V}_l^c} D_i}{\sum_{i \in \mathcal{V}_l^c} S_i} \qquad \text{for } l = 1, \ldots, L$$

as the average pick-up density for zone l. There is also substantial variation in $\{\text{PD}_l : l = 1, \ldots, L\}$; see the second column of Table 5. For example, the pick-up density in midtown Manhattan is two orders of magnitude higher than pick-up density in Brooklyn. Therefore, to avoid vacuous constraints in nodes with high pick-up density, we use a lower bound $\underline{D}_l$ (for zone l) on demand that is proportional to the average

[40] Note that Constraint (P0e) does not enforce spatial contiguity, rather it ensures that no census tract is isolated (unless it is the center of a node). One can enforce spatial contiguity by using the notion of contiguity order (e.g., see Duque et al. (2012)). Such constraints increase the size of the MIP significantly, which has motivated Duque et al. (2012) to propose an approximation algorithm to their problem. Motivated by this observation, we do not introduce such constraints in Problem (P0).

pick-up density of zone l. That is, defining the overall pick-up density $\mathrm{PD} = \sum_i D_i / \sum_i S_i$, we set

$$\underline{D}_l = \frac{\mathrm{PD}_l}{\mathrm{PD}}\, \hat{D},$$

where $\hat{D}$ is a tuning parameter that relates the lower bounds across zones. We use $\underline{t}_l = \big(1 - 0.25\,\mathrm{CV}_l\big)\,\hat{t}$ and $\bar{t}_l = \big(1 + 0.25\,\mathrm{CV}_l\big)\,\hat{t}$, where $\hat{t}$ is a tuning parameter common across zones and CV_l is the coefficient of variation of pick-up density in zone l.[41] Note that the term $0.25\,\mathrm{CV}_l$ is used to allow variation in the pick-up densities of various tracts in a node in zone l in Problem (P0); and it is equivalent to controlling the variation in pick-up densities of the various tracts in zone l.

Let $\mathrm{LT}(\hat{t}, \hat{D})$ denote the fraction of arcs in $\mathcal{E}$ (see Section 3) which take longer than one period to traverse. That is,

$$\mathrm{LT}(\hat{t}, \hat{D}) \;=\; \frac{\big|\{(i,j) \in \mathcal{E} : \tau_{ij} > 1\}\big|}{|\mathcal{E}|}.$$

We would like to choose $\hat{t}$ and $\hat{D}$ such that this fraction is sufficiently small and that $\hat{D}$ is sufficiently large to ensure considerable demand in all nodes. Next, for each fixed $\hat{D}$, we consider $\min_{\hat{t}} \mathrm{LT}(\hat{t}, \hat{D})$. As depicted in Figure 15, this shows a steady increase for $\hat{D} \leq 5.9$ and a sharp increase beyond $\hat{D} = 5.9$. Therefore, to accommodate both goals, we choose $\hat{D} = 5.9$ and $\hat{t} = \operatorname{argmin}_{\hat{t}} \mathrm{LT}(\hat{t}, 5.9) = 1.2$.

Results. The solution to Problem (P0) for $(\hat{t}, \hat{D}) = (1.2, 5.9)$ is depicted in Figure 16. In particular, this node definition provides sufficient separation between the major hubs of the city. In Figure 16, the nodes that accommodate the major hubs of the city are highlighted. JFK and La Guardia airports are colored in red and blue, respectively. The nodes accommodating the major railroad terminals, Penn Station and Grand Central Terminal, are colored in yellow and purple, respectively. The node accommodating the major bus terminal, Port Authority Bus Terminal, is colored in green.

Table 5: Parameters used in Problem (P0).

	Traffic speed[1]	Average density of pick-ups[2]	CV of density of pick-ups	Maximum distance within a node ($\bar{\delta}$)[3]
Lower Manhattan	11.8	2.36	0.8	2.0
Midtown Manhattan	8.6	6.92	0.8	1.4
Central Park Area	10.2	4.75	0.7	1.7
Upper Manhattan	11.8	0.16	1.7	2.0
Brooklyn	11.4	0.08	1.8	1.9
Queens	12.8	0.08	1.6	2.1

[1] Miles per hour. [2] Pick-ups per period per mile. [3] Miles.

[41] Sensitivity analysis displays that node definitions show little sensitivity to the choice of the coefficient of CV_l.

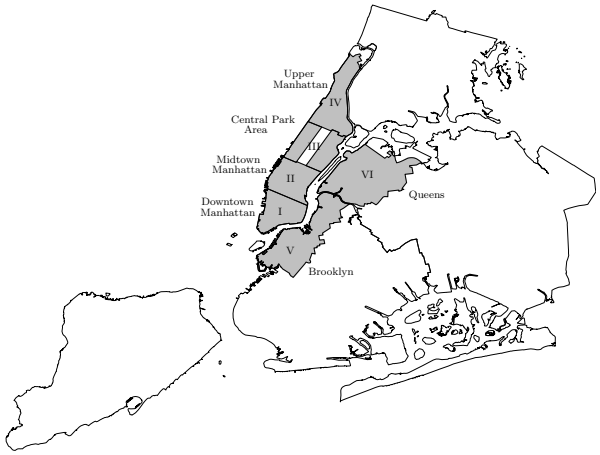


Figure 14: Boundaries of the zones used in node generation.

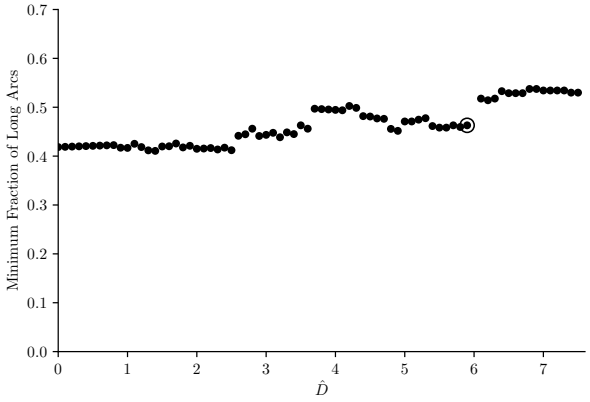


Figure 15: Minimum achievable fraction of arcs between adjacent nodes with travel time longer than one period for given $\hat{D}$.

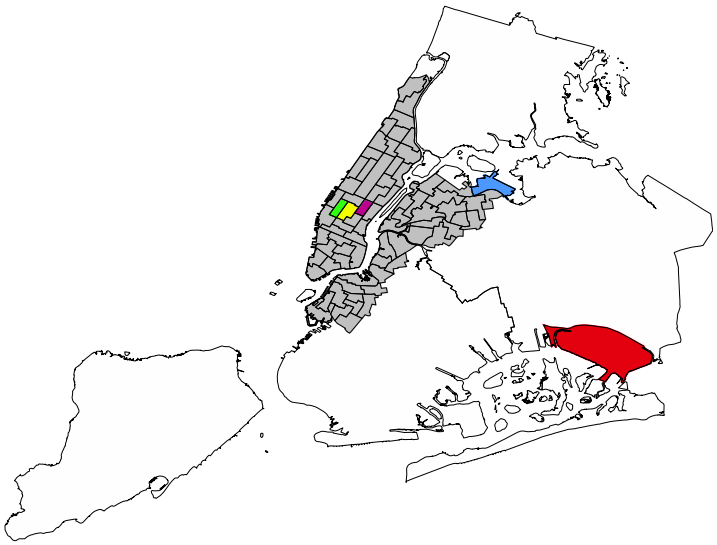


Figure 16: The boundaries of the seventy five nodes used in the analysis and the area under consideration.

B Further Discussion on the Matching Model

This section compares our matching model with that of Buchholz (2018). Buchholz uses the friction model introduced in Burdett et al. (2001).[42] Burdett et al. (2001) provides a formula for friction in the following setting: Suppose that n buyers each want one unit of a good and m sellers each have one unit.

[42]This was initially formulated as an urn-ball problem in Hall (1979), where n balls are randomly placed in m urns. In this problem, a match (only) occurs for the first ball placed in any urn.

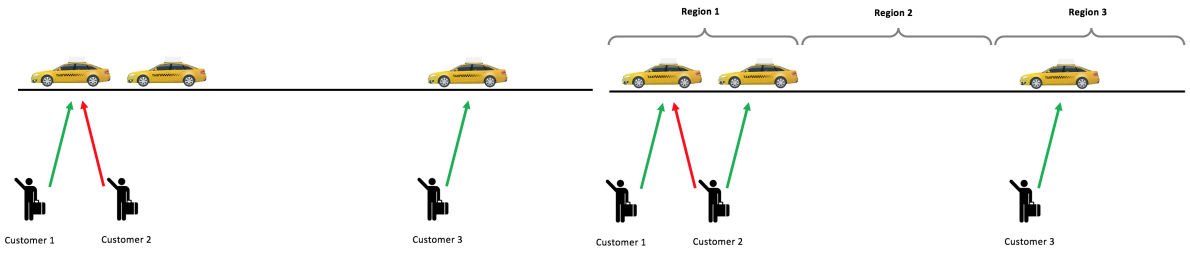
First, sellers are revealed to buyers, and then each buyer chooses a seller. If more than one buyer chooses a seller, the seller can serve only one of the buyers. Burdett et al. (2001) refers to this phenomena as friction and calculates the expected number of sales (matches between buyers and sellers) given n and m. Burdett et al. (2001) does not allow the buyers to search for sellers if they fail in their first attempt. It also assumes that buyers know the number of sellers and their exact location (sellers are revealed to buyers).

Buchholz (2018) tailors the imperfect matching model proposed in Burdett et al. (2001) to the taxi market. This imperfect matching model is equivalent to the following setting: First the number of drivers and their locations are revealed to the customers. Then, each customer chooses a taxi. When more than one customer chooses a taxi, only one of the customers is served and the others leave the system unfulfilled. Therefore, given m taxis and Λ customers in a node, the expected number of matches (served customers) is $m\big(1-[1-1/m]^{\Lambda}\big)$; see Burdett et al. (2001) for the derivation. Buchholz introduces an efficiency parameter γ and models the expected number of matches as follows:

$$\text{Expected Number of Matches} = m\Big(1-\big[1-\frac{1}{\gamma m}\big]^{\Lambda}\Big).$$

Lower values of γ correspond to more efficient matching.

Figure 17a depicts an example of matching under the model of Buchholz (2018). In this example, Customers 1 and 2 choose Taxi 1 and Customer 3 chooses Taxi 3. Taxi 1 chooses Customer 1 and Customer 2 is left without a taxi. He can not choose Taxi 2 although they are very close to each other. Figure 17b depicts the matches in the same example under our matching model with three regions. In this case, after failure to obtain a taxi in the first attempt, Customer 2 is allowed to re-choose among the remaining empty taxis in its region. Customer 2 chooses Taxi 2 and there are no unfulfilled customers left.

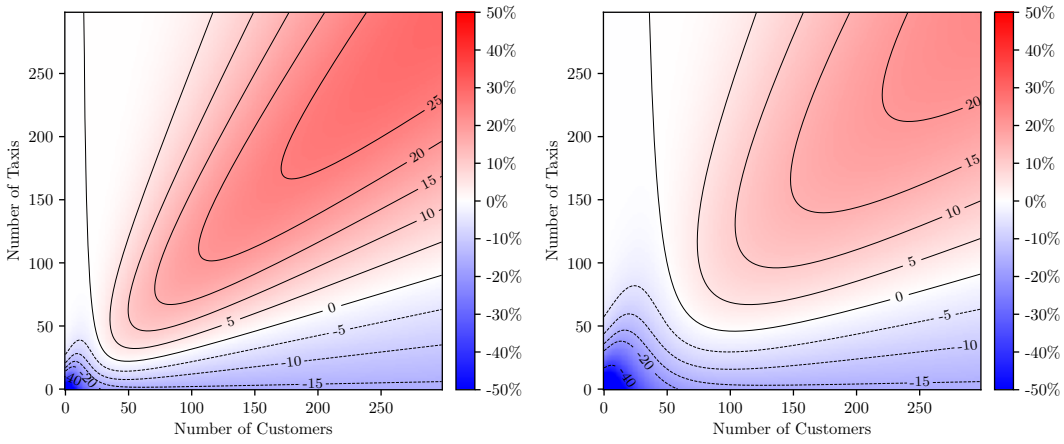


(a) Matching in the model of Buchholz (2018). (b) Matching in our model with three regions.

Figure 17: An illustration of matching in our imperfect matching model and the model of Buchholz (2018).

Figure 18 depicts the percentage increase in the number of matches under our matching model with $N = 10$ and $N = 20$, respectively, compared to the matching model of Buchholz (2018) with $\gamma = 0.84$

(as estimated by Buchholz for midtown Manhattan).[43] We observe that for small number of taxis and customers, which is common in areas such as Brooklyn and Queens, our matching model results in fewer matches. However, for large number of taxis and customers, which is common in areas such as midtown Manhattan, our matching model results in more matches.



(a) $N = 10$ (b) $N = 20$

Figure 18: Percentage change in the number of matches in our imperfect matching model compared to the imperfect matching model of Buchholz (2018).

C Supplemental Material for Section 4

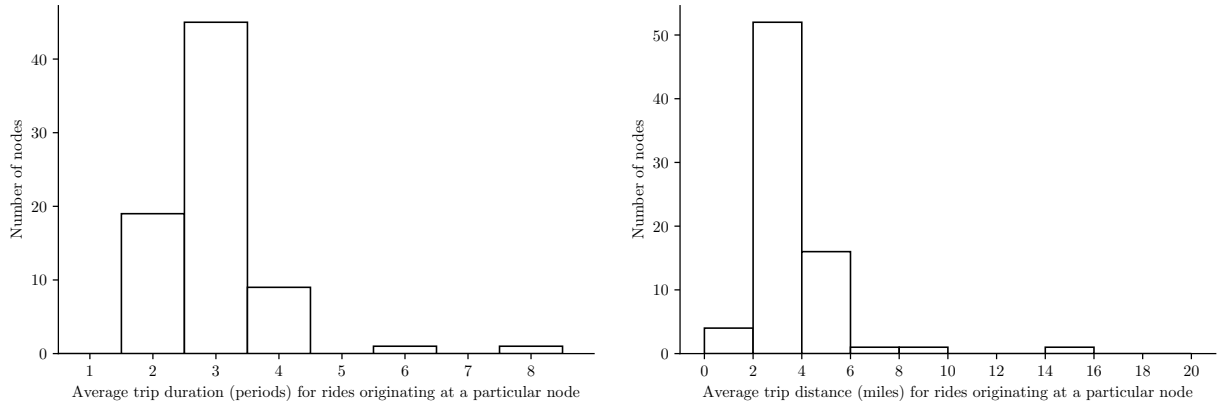
The histograms of the average trip duration and distance (for rides originating at each node) are depicted in Figure 19. Recall that a trip on average is 2.7 miles and 13.2 minutes long. There is considerable spatial variation in the average trip duration and distance across the city. Figure 20 depicts the average trip duration and distance for rides originating at each node.

D Supplemental Material for Section 5

Table 6 provides a summary of the model primitives. The estimation of these primitives is carried out in three steps; see Section 5 for further detail.

Next, we discusses the calculation of the likelihood of taxi m observing the sequence of drop-off and subsequent pick-ups on day d of month k that are no longer than two hours apart. We start by introducing some notation. Then, we calculate the likelihood of taxi m observing the entire sequence of drop-off and

[43] We use $N = 10$ and 20 as the number of regions in the overwhelming majority of the nodes in Figure 16 fall between these numbers.



(a) Average trip durations (in five-min periods).

(b) Average trip distances (in miles).

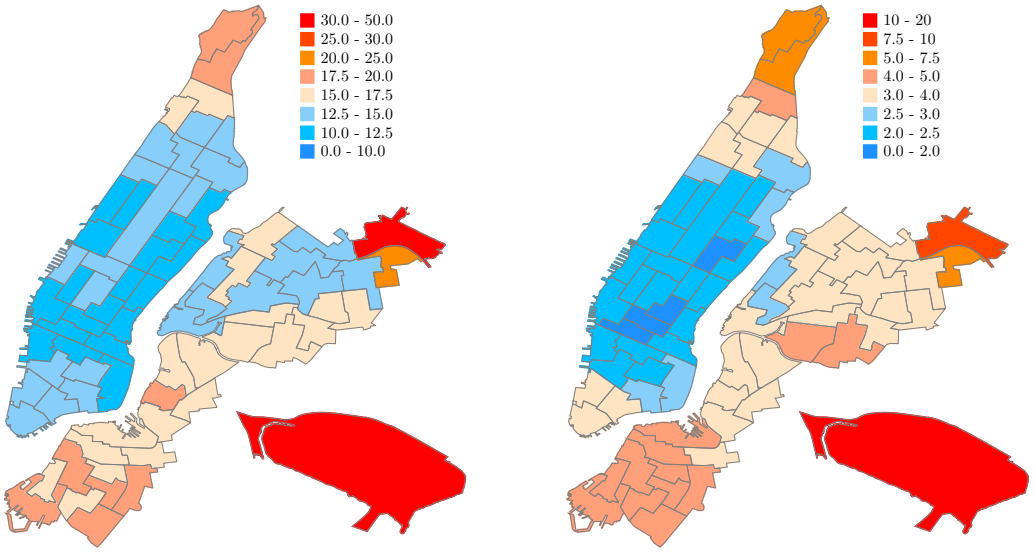
Figure 19: Histograms of the average trip duration and distances across nodes.

Table 6: Summary of the primitives.

Parameter	Number of elements	Estimation type
α	1	Endogenous
β	1	Endogenous
σ	1	Endogenous
A_{ij}^t	$120 \times 75 \times 75$	Endogenous
S_{ij}	75×75	Exogenous
τ_{ij}	75×75	Exogenous
d_{ij}	75×75	Exogenous
m_i^1	75×48	Exogenous
N_i	75	Exogenous
M	48	Exogenous
c	48	Exogenous

subsequent pick-ups on day d of month k. We conclude this section by removing the drop-off and subsequent pick-up tuples that are longer than two hours apart.

Let $N_{m,d,k}$ denote the number of riders that taxi m served on day d of month k. We denote the pick-up periods of taxi m on day d of month k by $t_{m,d,k}^{(1)}, \ldots, t_{m,d,k}^{(l)}, \ldots, t_{m,d,k}^{(N_{m,d,k})}$, where superscript l denotes the number of the pick-up. Similarly, we denote the pick-up and drop-off nodes of taxi m on day d of month k by $i_{m,d,k}^{(1)}, \ldots, i_{m,d,k}^{(l)}, \ldots, i_{m,d,k}^{(N_{m,d,k})}$ and $j_{m,d,k}^{(1)}, \ldots, j_{m,d,k}^{(l)}, \ldots, j_{m,d,k}^{(N_{m,d,k})}$, respectively. Then, the likelihood of



(a) Average trip duration (min). (b) Average trip distances (in miles).

Figure 20: Average trip duration and distance for rides originating at each node.

taxi m observing the entire sequence of pick-ups on day d of month k is

$$\mathbb{P}\left\{\left(t^{(1)}_{m,d,k},i^{(1)}_{m,d,k},j^{(1)}_{m,d,k}\right),\ldots,\left(t^{(N_{m,d,k})}_{m,d,k},i^{(N_{m,d,k})}_{m,d,k},j^{(N_{m,d,k})}_{m,d,k}\right)\right\}$$

$$=\left[\prod_{l=2}^{N_{m,d,k}}\mathbb{P}\left\{t^{(l)}_{m,d,k},i^{(l)}_{m,d,k},j^{(l)}_{m,d,k}\middle|t^{(l-1)}_{m,d,k},i^{(l-1)}_{m,d,k},j^{(l-1)}_{m,d,k}\right\}\right] \tag{12}$$

$$\times\,\mathbb{P}_I\{t^{(1)}_{m,d,k},i^{(1)}_{m,d,k},j^{(1)}_{m,d,k}\}\times\mathbb{P}_F\{t^{(l)}_{m,d,k},i^{(l)}_{m,d,k},j^{(l)}_{m,d,k}\}, \tag{13}$$

where $\mathbb{P}_I\{t^{(1)}_{m,d,k},i^{(1)}_{m,d,k},j^{(1)}_{m,d,k}\}$ is the probability that the customer picked up at period $t^{(1)}_{m,d,k}$ at node $i^{(1)}_{m,d,k}$ for destination $j^{(1)}_{m,d,k}$ is the first pick-up of the driver, $\mathbb{P}_F\{t^{(l)}_{m,d,k},i^{(l)}_{m,d,k},j^{(l)}_{m,d,k}\}$ is the probability that the customer picked up at period $t^{(l)}_{m,d,k}$ at node $i^{(l)}_{m,d,k}$ for destination $j^{(l)}_{m,d,k}$ is the final pick-up of the driver, and $\mathbb{P}\left\{t^{(l)}_{m,d,k},i^{(l)}_{m,d,k},j^{(l)}_{m,d,k}\middle|t^{(l-1)}_{m,d,k},i^{(l-1)}_{m,d,k},j^{(l-1)}_{m,d,k}\right\}$ is the probability that the driver picks up his l-th customer in period $t^{(l)}_{m,d,k}$ at node $i^{(l)}_{m,d,k}$ for destination $j^{(l)}_{m,d,k}$ given that his previous customer was picked up in period $t^{(l-1)}_{m,d,k}$ at node $i^{(l-1)}_{m,d,k}$ for destination $j^{(l-1)}_{m,d,k}$. Note that the distribution of $\left(t^{(l)}_{m,d,k},i^{(l)}_{m,d,k},j^{(l)}_{m,d,k}\right)$ depends on all elements of $\left(t^{(l-1)}_{m,d,k},i^{(l-1)}_{m,d,k},j^{(l-1)}_{m,d,k}\right)$. Because, all three elements of $\left(t^{(l-1)}_{m,d,k},i^{(l-1)}_{m,d,k},j^{(l-1)}_{m,d,k}\right)$ impact the period and node at which the taxi drops its $(l-1)$-th customer and starts the search for the l-th customer. Since it takes $\tau_{i,j}$ periods to transport a customer from node i to node j, the term

$$\mathbb{P}\left\{t^{(l)}_{m,d,k},i^{(l)}_{m,d,k},j^{(l)}_{m,d,k}\middle|t^{(l-1)}_{m,d,k},i^{(l-1)}_{m,d,k},\ j^{(l-1)}_{m,d,k}\right\} \tag{14}$$

is equal to the probability that driver m picks up his next customer in period $t^{(l)}_{m,d,k}$ at node $i^{(l)}_{m,d,k}$ for destination $j^{(l)}_{m,d,k}$ if he starts empty in period $t^{(l-1)}_{m,d,k} + \tau_{i^{(l-1)}_{m,d,k}, j^{(l-1)}_{m,d,k}}$ at node $j^{(l-1)}_{m,d,k}$.

We are interested in the likelihood of drop-off and subsequent pick-ups that are no longer than two hours apart. Therefore, we must omit drop-off and subsequent pick-up tuples that take longer than two hours. To do so, let $J\big(t^2, i^2, j^2 \big| t^1, i^1, j^1\big) = \mathbb{P}\big\{t^2, i^2, j^2 \big| t^1, i^1, j^1\big\}$ if $t^2 - t^1 - \tau_{j^1, i^2} > 24$ and one otherwise. Define J_I and J_F in a similar manner for the initial and final pick-up. Then, the likelihood of taxi m observing the sequence of drop-off and subsequent pick-ups on day d of month k that are no longer than two hours apart is

$$\begin{aligned}\mathcal{L}_{m,d,k}(\sigma) = &\left[\prod_{l=2}^{N_{m,d,k}} J\Big(t^{(l)}_{m,d,k}, i^{(l)}_{m,d,k}, j^{(l)}_{m,d,k} \Big| t^{(l-1)}_{m,d,k}, i^{(l-1)}_{m,d,k}, j^{(l-1)}_{m,d,k}\Big)\right] \\ &\times J_I\Big(t^{(1)}_{m,d,k}, i^{(1)}_{m,d,k}, j^{(1)}_{m,d,k}\Big) \times J_F\Big(t^{(l)}_{m,d,k}, i^{(l)}_{m,d,k}, j^{(l)}_{m,d,k}\Big).\end{aligned} \tag{15}$$

To compute (14), we transform the graph of the problem to an augmented graph, in which all arcs have length one. This can be achieved by adding $n-1$ auxiliary equi-spaced nodes on each arc of length n (for all n). Movements of an empty taxi on the augmented graph follow a non-homogeneous discrete time Markov Chain and (14) can be computed from the transition matrix of this Markov chain. As discussed in Section 5, solving (P1) using (15) gives an estimate of $\sigma = 1.4048$; see Figure 21 for log likelihood values in the vicinity of the maximum likelihood estimate $\sigma = 1.4048$.

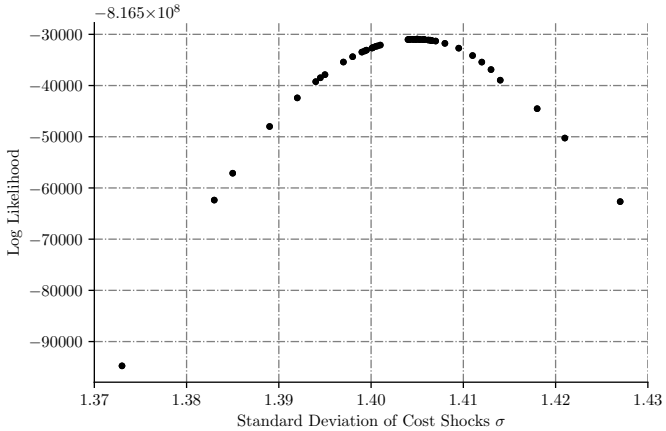


Figure 21: Log likelihood values in the vicinity of the maximum likelihood estimate $\sigma = 1.4048$.

Next, let us illustrates through an example how an augmented graph is created and used to compute (14). Consider the graph in Figure 22, where $\tau_{12} = \tau_{21} = 1$ and $\tau_{23} = \tau_{32} = 2$. The augmented graph is generated by placing node 4 on arc $(2, 3)$ and node 5 on arc $(3, 2)$, such that $\tau_{24} = \tau_{43} = \tau_{35} = \tau_{52} = 1$ and nodes 4 and 5 do not have a loop. The augmented graph is depicted in Figure 23. The relocation decisions of

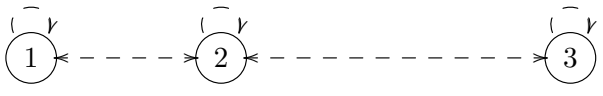


Figure 22: Original graph.

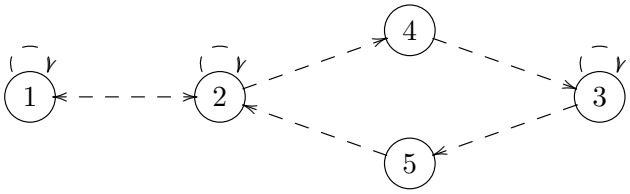


Figure 23: Augmented graph.

an empty taxi on the augmented graph correspond to the relocation decisions of the same taxi on the original graph. Consequently, the (path-wise) relocations of an empty taxi on the augmented graph are defined as follows:

1. If the taxi decides to relocate on an arc of unit length in the original graph, it relocates on the same arc in the augmented graph. For example, a taxi relocating from node 1 to node 2 in the original graph relocates from node 1 to node 2 in the augmented graph.

2. If the taxi decides to relocate on a long arc (arc of length longer than one period) in the original graph, it relocates to the first node created as a replacement of the long arc in the augmented graph. For example, a taxi relocating from node 2 to node 3 in the original graph relocates from node 2 to node 4 in the augmented graph.

3. If the taxi is currently on an auxiliary node in the augmented graph (which corresponds to a taxi relocating on a long arc in the original graph), it relocates on the only arc originating from the auxiliary node. For example, a taxi on node 4 of the augmented graph relocates to node 3 and a taxi on node 5 of the augmented graph relocates to node 2.

Given the aforementioned relocation policy on the augmented graph, the transition matrix of the augmented graph in period t is

$$\tilde{Q}^t = \begin{bmatrix} q_{11}^t & q_{12}^t & 0 & 0 & 0 \\ q_{21}^t & q_{22}^t & 0 & q_{23}^t & 0 \\ 0 & 0 & q_{33}^t & 0 & q_{32}^t \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix}, \tag{16}$$

where q_{ij}^t are relocation probabilities on the original graph. Let B^t be a matrix whose (i,j)-th element is

given by

$$B_{ij}^t = \begin{cases} (1 - \frac{\lambda_i^t}{m_i^t})\tilde{Q}_{ij}^t & \text{for} \quad i \leq n, \\ \tilde{Q}_{ij}^t & \text{for} \quad i > n, \end{cases}$$

and denotes the probability that an empty taxi at node i does not pick up a customer in period t and relocates to node j. Following the numbering in Figure 23, since n is the number of nodes on the original graph, $i < n$ corresponds to a node common to the original and the augmented graph and $i > n$ corresponds to an auxiliary node on the augmented graph.

Next, let us compute the probability that the taxi driver picks up his next customer in period t_1 at node i_1 for destination j_1 if he starts empty in period t_0 at node j_0, denoted by $\mathbb{P}\{t_1, i_1, j_1 | t_0, j_0\}$. Computing this probability is equivalent to computing (14).[44] By the definition of B^t, we have

$$\mathbb{P}\{t_1, i_1, j_1 | t_0, j_0\} \;=\; \left[\Pi_{t=t_0}^{t_1-1} B^t\right]\Big|_{i=j_0, j=i_1} \frac{\lambda_{i_1}^{t_1}}{m_{i_1}^{t_1}} \pi_{i_1,j_1}^{t_1}(F,P). \tag{17}$$

The first term on the right hand side is the probability that a taxi driver who starts empty at period t_0 in node j_0 unsuccessfully searches for customers until he arrives in period t_1 at node i_1. The second term on the right hand side is the probability that the driver picks up a customer in period t_1 at node i_1, and the third term is the probability that the destination of the customer is node j_1.

We conclude this section by discussing the details of the equilibrium calculations. Given $\overline{M}$, M^k, N_i, c^k, F_{ij}^k, P_{ij}^k, $\pi_{ij}^{tk}(F^k,P^k)$, and λ_i^{tk}, for each σ, we would like to solve for the mean field equilibrium that satisfies Equations (4)-(9) and $\lambda_i^{tk} \leq m_i^{tk}$, with m_i^1 having the same distribution as the first pick-up of the taxis in the first fifteen minutes of the day shift. However, due to the potential misspecification of the initial distribution of taxis m_i^1, such an equilibrium may not exist (in some months). Consider the following formulation:

$$\underset{m_i^1}{\text{minimize}} \;\; \sum_{i=1}^{n} \left|m_i^1 - \hat{m}_i^1\right|^2 \tag{18}$$

$$\text{subject to (4) - (9) and} \tag{19}$$

$$\lambda_i^t \leq m_i^t \tag{20}$$

In this formulation, m_i^1 denotes the initial distribution of empty taxis used in equilibrium calculations and $\hat{m}_i^1$ represents the desired (empirical) distribution (distribution of the first pick-up of the taxis in the first fifteen minutes of the day shift). Problem (18) searches among all mean-field equilibria (that satisfy

[44] Note that (14) is equal to the probability that driver m picks up his next customer in period $t_{m,d,k}^{(l)}$ at node $i_{m,d,k}^{(l)}$ for destination $j_{m,d,k}^{(l)}$ if he starts empty in period $t_{m,d,k}^{(l-1)} + \tau_{i_{m,d,k}^{(l-1)}, j_{m,d,k}^{(l-1)}}$ at node $j_{m,d,k}^{(l-1)}$.

constraints (19)-(20) given the estimated primitives) for the equilibrium whose initial distribution is closest (in the L^2 sense) to the desired (empirical) initial distribution. Numerical experiments show that the optimal objective function value of Problem (18) is non-zero in less than a handful of months. In other words, in the overwhelming majority of the months, there exists an equilibrium whose m_i^1 has the desired distribution.

E Monte Carlo Experiments

This section uses Monte Carlo experiments to evaluate the capability of the maximum likelihood estimator described in Section 5 in identifying the true σ. We generate 100 simulated datasets (of pick-ups and drop-offs) assuming a true σ and estimate new σ's from the simulated datasets. Then, we construct the 95% confidence interval and check whether the assumed true value falls within the confidence interval.

Assuming the true value of $\sigma = 0.7$ and using the estimated values of $\overline{M}$, M^k, N_i, m_i^{1k}, c^k, F_{ij}^k, P_{ij}^k, $\pi_{ij}^{tk}(F^k, P^k)$, and λ_i^{tk} from Section 5, we calculate the mean field equilibrium for each month. For simplicity, the month index k is suppressed in the remainder of this section. Using the mean field equilibria (for all months), we generate 100 simulated datasets with the same size as the original dataset (same number of months/days/active taxi drivers). To simulate the movements of the drivers, we use the following procedure for each active taxi driver on each weekday of each month.

1. Set $t = 1$ and generate a discrete r.v. with distribution $m_\cdot^1 = \{m_j^1 : 1 \leq j \leq n\}$. This r.v. corresponds to the initial location of the taxi. Set i equal to this location.
2. For an empty taxi in period t at node i, generate a Bernoulli r.v. with success probability λ_i^t / m_i^t, to simulate the search of the taxi for customers.
   (a) If the Bernoulli r.v. is equal to one (driver picking up a customer), generate r.v. X with distribution $\pi_{i,\cdot}^t = \{\pi_{i,j}^t : 1 \leq j \leq n\}$.
   (b) If the Bernoulli r.v. is equal to zero (driver failing to pick up a customer), generate r.v. X with distribution $q_{i,\cdot}^t = \{q_{i,j}^t : j \in \mathcal{A}(i)\}$.
3. Update $i \leftarrow X$ and add $t \leftarrow t + \tau_{ij}$.
4. If $t < T$, return to Step 2.

Estimating σ from the simulated datasets results in the 95% confidence interval of $(0.699, 0.702)$. We observe that the (assumed) true σ falls within the confidence interval. This shows that our estimator can recover the true structural parameter σ from the data.

F Cross-Validation

This section uses five-fold cross-validation to examine the ability of our model in predicting the relocation decisions of the empty taxi drivers (Kohavi et al. (1995)). We randomly split the non-holiday weekdays in the dataset into five groups of roughly equal sizes. Keeping the node definitions in Figure 16, we hold out one of the groups and use the other groups to estimate the primitives of the model (see Appendix D for a summary of the primitives and Section 5 for the estimation procedure). The goal is to use the estimated primitives from the other groups to make predictions about the hold-out group and assess the accuracy of the prediction. The relocation probabilities would be an ideal performance metric. However, in the data, we do not observe the location of the empty taxis at all times. Therefore, we use the distribution of the next pick-up location as a proxy.

Let μ_{it}^{kl} denote the ex ante distribution of the next pick-up location for a taxi driver who dropped off his customer at node i in period t of a weekday in month k and group l. Let ν_{it}^{kl} denote the corresponding empirical quantity and w_{it}^{kl} denote the number of drop-offs at node i in period t of a weekday in month k and group l. Furthermore, let $D(\cdot,\cdot)$ be a measure of similarity between two (spatial) distributions. We take

$$\frac{\sum_{l=1}^{5}\sum_{k=1}^{K}\sum_{i=1}^{n}\sum_{t=1}^{T} w_{it}^{kl}\, D(\mu_{it}^{kl},\nu_{it}^{kl})}{\sum_{l=1}^{5}\sum_{k=1}^{K}\sum_{i=1}^{n}\sum_{t=1}^{T} w_{it}^{kl}}, \tag{21}$$

the weighted average[45] of the similarity between the ex ante and empirical distributions of next pick-up location, as the performance metric for cross-validation. We consider three different similarity measures. The first similarity measure is the total variation distance

$$D_T(\mu,\nu) \;=\; \max_{E\subseteq\mathcal{E}} \big|\mu(E)-\nu(E)\big| \;=\; \frac{1}{2}\sum_{i=1}^{n} \big|\mu(i)-\nu(i)\big| \;\in\; [0,1].$$

Total variation distance is the most well-known tool for quantifying the similarity between two probability distributions. It captures the largest difference in the probability assigned to sets $E \subseteq \mathcal{E}$. We also consider Bhattacharyya distance (Bhattacharyya (1946))

$$D_B(\mu,\nu) \;=\; -\log\Big(\sum_{i=1}^{n}\sqrt{\mu(i)\nu(i)}\Big) \;\in\; [0,\infty]$$

that is commonly used in the information theory and pattern recognition literature; e.g., see Kailath (1967), Basseville (1989), and Aherne et al. (1998). Since the total variation and Bhattacharyya distances do not take advantage of the spatial aspect of the problem, we also consider earth mover's distance (also referred

[45] We use the number of drop-offs in a period at a node as weights in the performance metric since μ_{it}^{k} and ν_{it}^{k} describe the relocation behavior of a taxi driver who just dropped off his customer. Recall that we only observe the sequence of pick-ups and drop-offs of the taxi drivers.

to as Wasserstein distance). Earth mover's distance, that is commonly used in the transportation literature, is defined as the minimum cost of turning one distribution (or pile of dirt) into the other, where the cost is assumed to be the mass moved from one node/location to the other times the distance between them (Pele and Werman (2009); Carlsson et al. (2018)). In other words,

$$D_E(\mu,\nu) \;=\; \inf_{\gamma\in\mathcal{P}(\mu,\nu)} \sum_{i=1}^{n}\sum_{j=1}^{n} \Delta_{ij}\gamma_{ij} \;\in\; [0, \max_{ij} \Delta_{ij}],$$

where $\mathcal{P}(\mu,\nu)$ denotes the collection of all probability measures on $\mathcal{E}\times\mathcal{E}$ with marginals μ and ν, and $[\Delta_{ij}]_{i,j\in\mathcal{E}}$ is a metric distance matrix. The distance matrix Δ must have zero diagonal entries, positive off-diagonal entries, be symmetric, and satisfy the triangle inequality. We use

$$\Delta_{ij} = \begin{cases} (d_{ij}+d_{ji})/2 & \text{for} \quad i\neq j, \\ 0 & \text{for} \quad i=j, \end{cases}$$

that satisfies the necessary conditions of a metric distance and it is derived from the travel distances (d_{ij}, d_{ji}) estimated from the data.

Table 7 provides a summary of the cross-validation results. A performance metric of 0.22 under total variation distance can be interpreted as a maximum deviation of 22% between the ex ante and empirical distributions (in estimating the number of pick-ups in any set of nodes/locations). A performance metric of 0.52 under earth mover's distance can be interpreted as an average deviation of 0.52 miles between the next pick-up location under the ex ante and empirical distributions. This performance metric is particularly meaningful when we compare it to the 2.26 miles average distance between (distinct) adjacent nodes or the 1.41 miles average distance between (distinct) non-airport adjacent nodes.

Table 7: Cross-validation results.

	Performance Metric
Total variation distance	0.22
Bhattacharyya distance	0.08
Earth mover's distance	0.52

G Supplemental Material for Section 6

This section provides the mathematical formulation of the mechanisms discussed in the paper and a discussion on origin-destination spatial pricing.

G.1 Mathematical Formulations of the Mechanisms

This section discusses the mathematical formulation of the mechanisms introduced in the paper. The mathematical formulations are introduced in the order they appear in the paper. For numerical results and discussions on each mechanism, see Section 6.

G.1.1 Origin-Destination Pricing.

Consider the following pricing formulation.

$$\begin{aligned}
\underset{\eta_{ij}}{\text{maximize}} \quad & \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{t=1}^{T} \mathrm{CS}_{ij}^{t}(\eta_{ij}) && \text{(P3)}\\
\text{subject to } & (3)\text{-}(9) \text{ and} \\
& |\eta_{ij} - 1| \le \overline{\eta} \\
& (F_{ij}, P_{ij}) \;=\; \eta_{ij}\,(\overline{P}_{ij}, \overline{P}_{ij}) \\
& \sum_{i=1}^{n} m_i^1 V(i,0) \ge \sum_{i=1}^{n} m_i^1 \overline{V}(i,0).
\end{aligned}$$

We call the solution $\{\eta_{ij}\}_{i,j\in\mathcal{V}}$ to Problem (P3) the optimal origin-destination price multipliers (or equivalently the optimal origin-destination prices) with a maximum price variation of $\overline{\eta}$. Note that in Problem (P2), price-multipliers only depend on the origin of the ride while in Problem (P3), price multipliers depend on the origin-destination pair.

G.1.2 Hybrid Mechanism.

Let $\mathcal{R}$ denote the set of nodes in which less than 80% of customers are served under the base prices and refer to them as under-served nodes. Consider the following pricing formulation.

$$\begin{aligned}
\underset{\eta_{i}}{\text{maximize}} \quad & \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{t=1}^{T} \mathrm{CS}_{ij}^{t}(\eta_{i}) && \text{(P4)}\\
\text{subject to } & (4)\text{-}(9) \text{ and} \\
& (3b) \text{ and } |\eta_i - 1| \le \overline{\eta} \qquad \forall\, i \notin \mathcal{R} && \text{(P4a)}\\
& (3a) \text{ and } \eta_i = 1 \qquad \forall\ i \in \mathcal{R} && \text{(P4b)}\\
& (F_{ij}, P_{ij}) \;=\; \eta_{i}\,(\overline{P}_{ij}, \overline{P}_{ij}) \\
& \sum_{i=1}^{n} m_i^1 V(i,0) \ge \sum_{i=1}^{n} m_i^1 \overline{V}(i,0).
\end{aligned}$$

We call the solution $\{\eta_i\}_{i\in\mathcal{V}}$ to Problem (P4) the optimal price multipliers for the hybrid mechanism (with origin-only pricing). In Problem (P4), local search friction is removed in under-served nodes and spatial pricing is used in well-served nodes. Constraint (P4a) ensures that in well-served nodes, friction is present and spatial pricing is used. In contrast, Constraint (P4b) ensures that in under-served nodes, friction is removed and spatial pricing is not used. The remaining constraints are similar to the ones in Problem (P2) in Section 6.1.

G.1.3 Proposed Mechanism.

Similar to Section G.1.2, let $\mathcal{R}$ denote the under-served nodes. Consider the following pricing formulation.

$$\underset{\eta_i}{\text{maximize}} \quad \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{t=1}^{T} \mathrm{CS}_{ij}^{t}(\eta_i) \tag{P5}$$

$$\text{subject to (4) - (9) and}$$

$$(3a) \text{ and } |\eta_i - 1| \le \overline{\eta} \qquad \forall\, i \notin \mathcal{R} \tag{P5a}$$

$$(3a) \text{ and } \eta_i = 1 \qquad \forall\ i \in \mathcal{R} \tag{P5b}$$

$$(F_{ij}, P_{ij}) \;=\; \eta_i\,(\overline{P}_{ij}, \overline{P}_{ij})$$

$$\sum_{i=1}^{n} m_i^1 V(i,0) \ge \sum_{i=1}^{n} m_i^1 \overline{V}(i,0).$$

We call the solution $\{\eta_i\}_{i\in\mathcal{V}}$ to Problem (P5) the optimal price multipliers for the proposed mechanism (with origin-only pricing). In Problem (P5), local search friction is removed in all nodes while spatial pricing is limited to well-served nodes. Constraint (P5a) ensures that in well-served nodes, both friction removal and spatial pricing are used. In contrast, Constraint (P5b) ensures that in under-served nodes, friction is removed while spatial pricing is not used. The remaining constraints are similar to the ones in Problem (P2) in Section 6.1.

G.2 Further Discussion on Origin-Destination Pricing

Table 8 presents the increase in consumer surplus for different maximum price variations $\overline{\eta}$ when using the optimal origin-destination pricing scheme. Similar to origin-only pricing, when using origin-destination pricing, larger values of $\overline{\eta}$ results in larger increases in consumer surplus. However, there are diminishing returns to change in $\overline{\eta}$. Origin-destination pricing with a maximum price variation of 50% results in a \$0.79 increase in consumer surplus per ride, which is 24.4% higher than the increase in consumer surplus from origin-only pricing. The gap between origin-destination pricing and origin-only pricing is increasing in the maximum price variation $\overline{\eta}$. Origin-destination pricing (with a maximum price variation of 50%) also results

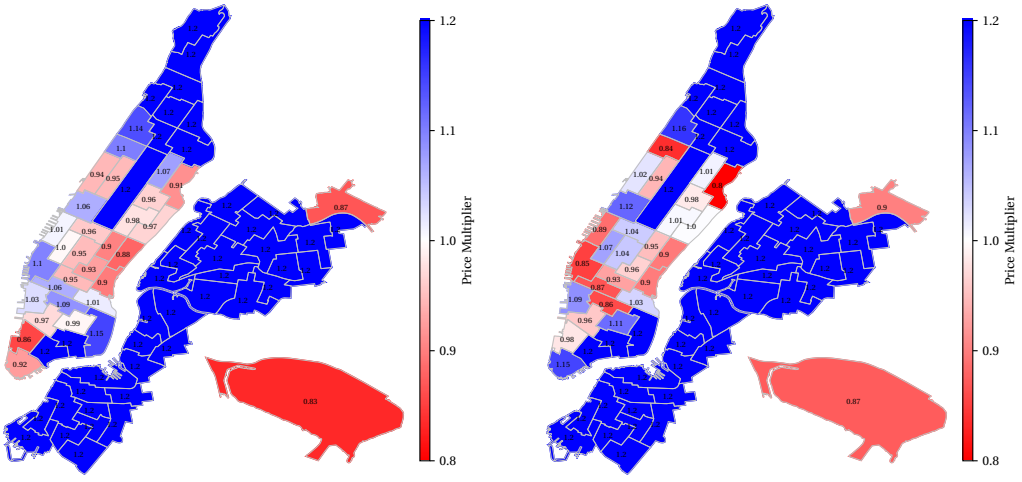
in a 3.2% increase in the number of served customers and an 7.4% increase in the miles traveled by customers.

Table 8: Impact of maximum price variation when using origin-destination pricing on various performance metrics (maximum price multiplier $\hat{\eta} = 5$).

Maximum price variation ($\overline{\eta}$)	10%	20%	50%
Consumer surplus			
Total increase	\$69K	\$112K	\$168
Per ride	\$0.33	\$0.53	\$0.79
In terms of average fare	3.6%	5.8%	8.5%
Over origin-only pricing	11.3%	19.1%	24.4%
Number of served customers	1.3%	1.8%	3.2%
Miles traveled by customers	2.0%	3.9%	7.4%

G.3 Further Discussion on the Impact of Maximum price multiplier on optimal spatial prices

The optimal origin-only price multipliers for the maximum price multipliers of $\hat{\eta} = 5$ and $\hat{\eta} = 50$ are depicted in Figure 24. The pattern of prices is not sensitive to $\hat{\eta}$. Therefore, in this paper, we use maximum price multiplier $\hat{\eta} = 5$, which is equivalent to assuming that for a ride of \$9, no customer is willing to pay more than \$45. This is a conservative choice for $\hat{\eta}$ considering its impact on consumer surplus (see Table 2).



(a) Maximum price multiplier $\hat{\eta} = 5$ (b) Maximum price multiplier $\hat{\eta} = 50$

Figure 24: Optimal origin-only price multipliers for different maximum price multipliers $\hat{\eta}$ (maximum price variation $\overline{\eta} = 0.2$).

H Maximizing Drivers' Profit

This section focuses on the spatial prices that maximize drivers' profit. Consider the following pricing formulation.

$$\underset{\eta_i}{\text{maximize}} \quad \sum_{i=1}^{n} m_i^1 V(i,0) \tag{P6}$$

$$\text{subject to (3) - (9) and}$$

$$|\eta_i - 1| \leq \overline{\eta}$$

$$(F_{ij}, P_{ij}) \;=\; \eta_i \,(\overline{P}_{ij}, \overline{P}_{ij})$$

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{t=1}^{T} \mathrm{CS}_{ij}^t(\eta_i) \geq \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{t=1}^{T} \mathrm{CS}_{ij}^t(1). \tag{P6a}$$

We call the solution $\{\eta_i\}_{i\in\mathcal{V}}$ to Problem (P6) the optimal origin-only price multipliers (or equivalently the optimal origin-only prices) subject to minimum consumer surplus with a maximum price variation of $\overline{\eta}$. Note that in Problem (P6), price multipliers only depend on the origin of the ride. Constraint (P6a) ensures that the consumers surplus is not hurt by the new pricing scheme. The remaining constraints are similar to the ones in Problem (P2) in Section 6.1.

The increase in drivers' profit for four plausible values of maximum price multiplier $\hat{\eta}$ is presented in Table 9. As shown in Table 9, using a origin-only pricing scheme with a maximum variation of 20% results in an increase in driver's profit of $\$64,000$ - $\$144,000$ on every day shift on weekdays. This is equivalent to an increase in drivers' profit of \$0.30 - \$0.71 per ride (3.4% - 7.9% of the average fare paid by the customers). Since similar to Section 6.1, the pattern of the optimal origin-only prices is insensitive to the maximum price multiplier $\hat{\eta}$, in the remainder of this section we use $\hat{\eta} = 5$. This is a conservative choice for $\hat{\eta}$ considering its impact on drivers' profit (see Table 9).[46]

Table 9: Impact of origin-only prices on drivers' profit (maximum price variation $\overline{\eta} = 0.2$).

Maximum price multiplier ($\hat{\eta}$)	5	10	20	50
Total increase	\$64K	\$81K	\$105K	\$144K
Increase per ride	\$0.30	\$0.39	\$0.51	\$0.71
Increase in terms of average fare	3.4%	4.3%	5.6%	7.9%

Table 10 presents the increase in drivers' profit for different maximum price variations $\overline{\eta}$. We observe that larger values of $\overline{\eta}$ results in larger increases in consumer surplus. Table 10 indicates that the optimal origin-only pricing scheme (for maximizing drivers' profit) has minimal impact on the the number of served

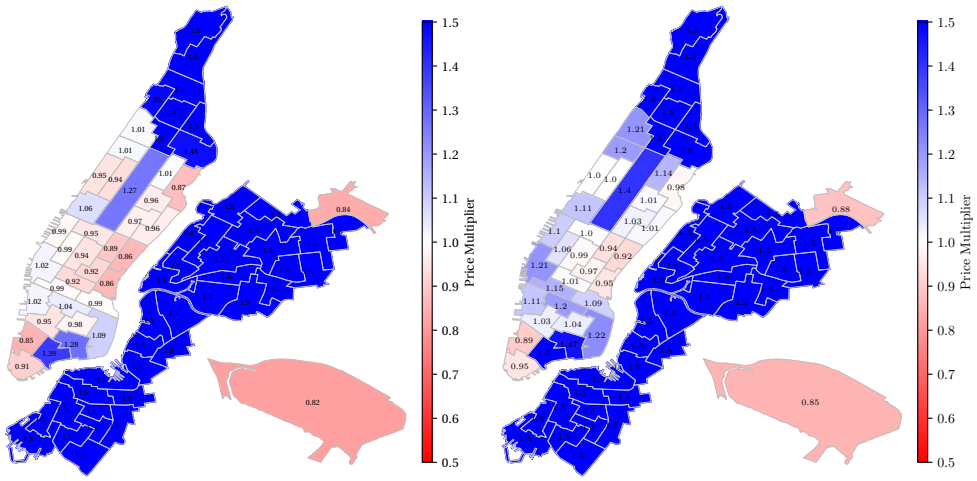
[46] In a similar setting, Cohen et al. (2016, Section 4) makes the conservative assumption that no customer is willing to pay more than 4.9 times the base prices for an Uber ride.

customers and the miles traveled by customers. This is not surprising since the objective function of Problem (P6) is drivers' profit as opposed to a measure of consumer welfare.

Table 10: Impact of maximum price variation on drivers' profit (maximum price multiplier $\hat{\eta} = 5$).

Maximum price variation ($\overline{\eta}$)	10%	20%	30%	40%	50%
Drivers' profit					
Total increase	\$43K	\$64K	\$79K	\$91K	\$100K
Per ride	\$0.20	\$0.30	\$0.38	\$0.43	\$0.47
In terms of average fare	2.3%	3.4%	4.2%	4.8%	5.3%
Number of served customers	-0.5%	-0.6%	-0.7%	-0.8%	-0.8%
Miles traveled by customers	0.3%	0.4%	0.4%	0.5%	0.5%

The optimal origin-only price multipliers (with a maximum price variation of 50%) for maximizing consumer surplus and drivers' profit are depicted in Figure 25. The pattern of the optimal prices are similar. In upper Manhattan, Brooklyn, and Queens, prices increase since both customers and drivers benefit from the increase in prices. Higher prices in the high demand/fare areas in downtown and midtown Manhattan as well as the airports generate consumer surplus. This consumer surplus is then transformed into drivers' profits through higher prices in low demand/fare neighborhoods. In Figure 25b, the increase in consumer surplus from better serving the under-served neighborhoods is transformed into drivers' profit through (slightly) higher prices compared to Figure 25a in the well-served neighborhoods.



(a) Maximizing consumer surplus. (b) Maximizing drivers' profit.

Figure 25: Comparison of the optimal origin-only price multipliers for maximizing consumer surplus and drivers' profit (maximum price variation $\overline{\eta} = 0.5$ and maximum price multiplier $\hat{\eta} = 5$).

I Miscellaneous Proofs and Derivations

This section provides a collection of proofs and derivations for the propositions, theorems, and equations discussed in the paper. The proofs/derivations are discussed in the order they appear in the paper.

I.1 Derivation of (3b)

We start by deriving $G(\Lambda, m)$. Then, we discuss the non-monotonicity of $G(\cdot, m)$ at small values of Λ and use a linear approximation to resolve this issue. Consider node i in period t. Let us denote the de-normalized mass of empty taxis by m, the de-normalized potential demand by Λ, and the de-normalized satisfied demand by λ. Then,

$$m \,=\, \overline{M}\, m_i^t, \qquad \Lambda \,=\, \overline{M}\, \Lambda_i^t(F,P), \qquad \lambda \,=\, \overline{M}\, \lambda_i^t. \tag{22}$$

Recall that the expected number of matches in node i is $\lambda \,=\, N_i \times \mathbb{E}\big[\min(X,Y)\big]$, where $X \sim \text{Binomial}(m, \frac{1}{N_i})$ and $Y \sim \text{Binomial}(\Lambda, \frac{1}{N_i})$. By approximating the Binomial distributions with Normal distributions, we obtain

$$\lambda \,\simeq\, N_i \times \mathbb{E}\big[\min(X,Y)\big], \tag{23}$$

where $X \sim \mathcal{N}(\mu_m, \sigma_m)$ and $Y \sim \mathcal{N}(\mu_\Lambda, \sigma_\Lambda)$ are independent random variables with

$$\begin{aligned} \mu_m \,&=\, \frac{m}{N_i}, \qquad \text{and} \qquad \sigma_m^2 \,=\, \frac{m}{N_i}\,(1-\frac{1}{N_i}),\\ \mu_\Lambda \,&=\, \frac{\Lambda}{N_i}, \qquad \text{and} \qquad \sigma_\Lambda^2 \,=\, \frac{\Lambda}{N_i}\,(1-\frac{1}{N_i}). \end{aligned}$$

By substituting (22) into (23), we obtain

$$\lambda_i^t \,=\, \frac{1}{\overline{M}} \times N_i \times \mathbb{E}\big[\min(X,Y)\big], \tag{24}$$

where $X \sim \mathcal{N}(\mu_m, \sigma_m)$ and $Y \sim \mathcal{N}(\mu_\Lambda, \sigma_\Lambda)$ are independent random variables with

$$\mu_m \,=\, \frac{\overline{M} m_i^t}{N_i} \qquad \text{and} \qquad \sigma_m^2 \,=\, \mu_m\,(1-\frac{1}{N_i}), \tag{25}$$

$$\mu_\Lambda \,=\, \frac{\overline{M} \Lambda_i^t}{N_i} \qquad \text{and} \qquad \sigma_\Lambda^2 \,=\, \mu_\Lambda\,(1-\frac{1}{N_i}). \tag{26}$$

Using *Equation (2)* of Clark (1961), (24) simplifies to

$$\begin{aligned}\lambda_i^t &= -\frac{N_i}{\overline{M}}\Bigg[-\mu_m\Phi\Big(\frac{\mu_\Lambda-\mu_m}{\sqrt{\sigma_m^2+\sigma_\Lambda^2}}\Big) - \mu_\Lambda\Phi\Big(\frac{\mu_m-\mu_\Lambda}{\sqrt{\sigma_m^2+\sigma_\Lambda^2}}\Big) + \sqrt{\sigma_m^2+\sigma_\Lambda^2}\phi\Big(\frac{\mu_\Lambda-\mu_m}{\sqrt{\sigma_m^2+\sigma_\Lambda^2}}\Big)\Bigg] \\ &= \frac{N_i}{\overline{M}}\Bigg[\mu_m\Phi\Big(\frac{\mu_\Lambda-\mu_m}{\sqrt{\sigma_m^2+\sigma_\Lambda^2}}\Big) + \mu_\Lambda\Phi\Big(\frac{\mu_m-\mu_\Lambda}{\sqrt{\sigma_m^2+\sigma_\Lambda^2}}\Big) - \sqrt{\sigma_m^2+\sigma_\Lambda^2}\phi\Big(\frac{\mu_m-\mu_\Lambda}{\sqrt{\sigma_m^2+\sigma_\Lambda^2}}\Big)\Bigg], \qquad (27)\end{aligned}$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ are the cumulative distribution function and probability density function of the standard normal distribution. Then, by plugging $\nu \triangleq (\mu_m-\mu_\Lambda)/\sqrt{\sigma_m^2+\sigma_\Lambda^2}$ into (27), we obtain

$$\begin{aligned}\lambda_i^t &= \frac{N_i}{\overline{M}}\Big(\mu_m\Phi(-\nu) + \mu_\Lambda\Phi(\nu) - \sqrt{\sigma_m^2+\sigma_\Lambda^2}\,\phi(\nu)\Big), \\ &= \frac{N_i}{\overline{M}}\Big(\mu_m\Phi(-\nu) + \mu_\Lambda\Phi(\nu) - \frac{\mu_m-\mu_\Lambda}{\nu}\,\phi(\nu)\Big), \qquad (28) \\ &= \frac{N_i}{\overline{M}}\Bigg(\frac{\overline{M}m_i^t}{N_i}\Phi(-\nu) + \frac{\overline{M}\Lambda_i^t}{N_i}\Phi(\nu) - \frac{\frac{\overline{M}m_i^t}{N_i}-\frac{\overline{M}\Lambda_i^t}{N_i}}{\nu}\phi(\nu)\Bigg), \qquad (29) \\ &= m_i^t\,\Phi(-\nu) + \Lambda_i^t\,\Phi(\nu) - \frac{m_i^t-\Lambda_i^t}{\nu}\,\phi(\nu), \qquad (30)\end{aligned}$$

where (28) follows from the definition of ν and (29) follows from (25)-(26). Next, we rewrite ν in terms of m_i^t and Λ_i^t as follows:

$$\begin{aligned}\nu &= \frac{(\mu_m-\mu_\Lambda)}{\sqrt{\sigma_m^2+\sigma_\Lambda^2}} = \frac{(\mu_m-\mu_\Lambda)}{\sqrt{(1-\frac{1}{N_i})(\mu_m+\mu_\Lambda)}}, \qquad (31) \\ &= \frac{(\frac{\overline{M}m_i^t}{N_i}-\frac{\overline{M}\Lambda_i^t}{N_i})}{\sqrt{1-\frac{1}{N_i}}\sqrt{\frac{\overline{M}m_i^t}{N_i}+\frac{\overline{M}\Lambda_i^t}{N_i}}}, \qquad (32) \\ &= \frac{\frac{\overline{M}}{N_i}(m_i^t-\Lambda_i^t)}{\sqrt{1-\frac{1}{N_i}}\sqrt{\frac{\overline{M}}{N_i}}\sqrt{m_i^t+\Lambda_i^t}}, \\ &= \frac{\sqrt{\frac{\overline{M}}{N_i}}}{\sqrt{1-\frac{1}{N_i}}}\frac{m_i^t-\Lambda_i^t}{\sqrt{m_i^t+\Lambda_i^t}}, \\ &= \sqrt{\frac{\overline{M}}{N_i-1}}\frac{m_i^t-\Lambda_i^t}{\sqrt{m_i^t+\Lambda_i^t}},\end{aligned}$$

where (31)-(32) follow from (25)-(26).

For small value of m_i^t, (30) is not monotone in Λ_i^t for small values of Λ_i^t. This issue follows from the inaccuracy of the normal approximation for the binomial distributions of X and Y at small values of m_i^t and

Λ_i^t. We use a linear approximation to resolve the issue. In other words, we use

$$\lambda_i^t \;=\; \begin{cases} G\big(\Lambda_i^t, m_i^t\big) & \text{if } \Lambda_i^t \geq \hat{\Lambda}_{m_i^t}, \\ G\big(\hat{\Lambda}_{m_i^t}, m_i^t\big)\, \frac{\Lambda_i^t}{\hat{\Lambda}_{m_i^t}} & \text{otherwise,} \end{cases}$$

where $G(\Lambda, m) \;=\; \Lambda\, \Phi(\nu) + m\, \Phi(-\nu) - \frac{m-\Lambda}{\nu}\, \phi(\nu)$ follows from (30) and $\hat{\Lambda}_m = \min\big\{\Lambda \geq 0 : G(\Lambda, m) \text{ is strictly increasing on } (\Lambda$

I.2 Proof of Proposition 1

First, we derive the expected payoff of the driver in each state from each action. Then, we derive the state transition matrix of the driver. We conclude the proof by writing out the Bellman equation and simplifying it. Consider the empty infinitesimal driver at state $s = (i, t, \epsilon)$. For readability and ease of notation, in the remainder of this proof, we write $\epsilon^{(i)}$ as ϵ and ϵ_{ij} as ϵ_j. Assume that the driver chooses action $j \in \mathcal{A}(s) = \{j : S_{ij} > 0\}$. The driver picks up a customer with probability $p(s) = \frac{\lambda_i^t}{m_i^t}$ and relocates to node j with probability $1 - p(s)$. If the driver picks up a customer, his expected payoff will be $\sum_{l=1}^n \pi_{il}^t(F,P)\,\big(F_{il} + P_{il} d_{il} - c_{il}\big)$, where $c_{il} = c\, d_{il} + \epsilon_l$ is the cost of traveling from node i to node l at state $s = (i, t, \epsilon)$. If the driver does not pick up a customer, he will receive the payoff $-c_{ij} = -c d_{ij} - \epsilon_j$. Therefore, the expected (immediate) payoff of the driver at state $s = (i, t, \epsilon)$ from action $j \in \mathcal{A}(s)$ is

$$\begin{aligned} u_j(s) \;&\triangleq\; p(s)\Big[\sum_{l=1}^n \pi_{il}^t(F,P)\,\big(F_{il} + P_{il} d_{il} - c_{il}\big)\Big] \;-\; \big[1 - p(s)\big] c_{ij} \\ &=\; p(s) \sum_{l=1}^n \pi_{il}^t(F,P)\,\Big(F_{il} + P_{il}\, d_{il} - c\, d_{il} - \epsilon_l\Big) \;-\; \big[1 - p(s)\big]\big(c\, d_{ij} + \epsilon_j\big) \end{aligned}$$

and the state transition probabilities are

$$\mathbb{P}\Big(s' = (i', t', \epsilon')\Big| s = (i, t, \epsilon),\, a = j\Big) \;=\; \mathbb{P}\Big(x' = (i', t')\Big| x = (i, t),\, a = j\Big) \times g(\epsilon'),$$

where

$$\mathbb{P}\bigg(x' = (i', t')\Big| x = (i, t),\, a = j\bigg) \;=\; \begin{cases} p(s)\, \pi_{ii'}^t(F,P) & \text{if } i' \neq j \text{ and } t = t + d_{ii'}, \\ p(s)\, \pi_{ii'}^t(F,P) + \big(1 - p(s)\big) & \text{if } i' = j \text{ and } t = t + d_{ii'}, \\ 0 & \text{otherwise.} \end{cases}$$

Let $v(i,t,\epsilon)$ denote the value function of the driver at state $s=(i,t,\epsilon)$. Then, by Bellman's equation, the value function must satisfy

$$\begin{aligned} v(s) \;&=\; \max_{j\in\mathcal{A}(s)} \Big\{ u_j(s) \;+\; \mathbb{E}_{s'}\Big[v(s'|s,a=j)\Big] \Big\} \\ &=\; \max_{j\in\mathcal{A}(s)} \Big\{ p(s)\sum_{l=1}^{n} \pi_{il}^{t}(F,P)\,\big(F_{il}+P_{il}d_{il}-cd_{il}-\epsilon_l\big) \;-\; \big[1-p(s)\big]\big(c\,d_{ij}+\epsilon_j\big) \\ &\qquad\qquad +\int \Big(p(s)\sum_{l=1}^{n}\pi_{il}^{t}(F,P)\,v(l,t+\tau_{il},\epsilon') \;-\; \big[1-p(s)\big]v(t+\tau_{ij},j,\xi')\Big) g(\epsilon')d\epsilon' \Big\}. \end{aligned}$$

Let $V(i,t) = \int v(i,t,\epsilon)d\epsilon$ denote the integrated value function of the driver at (the observable state) $x=(i,t)$. Then,

$$\begin{aligned} V(i,t) \;&=\; \int v(i,t,\epsilon)g(\epsilon)d\epsilon \\ &=\; \int \max_{j\in\mathcal{A}(s)} \Big\{ u_j(s) \;+\; \mathbb{E}_{s'}\Big[v(s'|s,a=j)\Big] \Big\} g(\epsilon)d\epsilon \\ &=\; \int \max_{j\in\mathcal{A}(s)} \Big\{ p(s)\sum_{l=1}^{n} \pi_{il}^{t}(F,P)\,\big(F_{il}+P_{il}d_{il}-cd_{il}-\epsilon_l\big) \;-\; \big[1-p(s)\big]\big(c\,d_{ij}+\epsilon_j\big) \\ &\qquad\qquad +\int \Big(p(s)\sum_{l=1}^{n}\pi_{il}^{t}(F,P)\,v(l,t+\tau_{il},\epsilon') \;+\; \big[1-p(s)\big]v(j,t+\tau_{ij},\epsilon')\Big) \\ &\qquad\qquad\qquad g(\epsilon')d\epsilon' \Big\} g(\epsilon)d\epsilon \\ &=\; \int \max_{j\in\mathcal{A}(s)} \Big\{ p(s)\sum_{l=1}^{n} \pi_{il}^{t}(F,P)\,\big(F_{il}+P_{il}d_{il}-cd_{il}-\epsilon_l\big) \;-\; \big[1-p(s)\big]\big(c\,d_{ij}+\epsilon_j\big) \\ &\qquad\qquad +\int p(s)\sum_{l=1}^{n}\pi_{il}^{t}(F,P)\,v(l,t+\tau_{il},\epsilon')g(\epsilon')d\epsilon' \\ &\qquad\qquad +\int \big[1-p(s)\big]v(j,t+\tau_{ij},\epsilon')g(\epsilon')d\epsilon' \Big\} g(\epsilon)d\epsilon, \end{aligned}$$

where in the last equality we have separated the integral over ϵ' into two parts. Then, by linearity of integration, we have

$$\begin{aligned} V(i,t) \;=\; \int \max_{j\in\mathcal{A}(s)} \Big\{ & p(s)\sum_{l=1}^{n} \pi_{il}^{t}(F,P)\,\big(F_{il}+P_{il}d_{il}-cd_{il}-\epsilon_l\big) \;-\; \big[1-p(s)\big]\big(c\,d_{ij}+\epsilon_j\big) \\ & +\; p(s)\sum_{l=1}^{n}\pi_{il}^{t}(F,P) \int v(l,t+\tau_{il},\epsilon')g(\epsilon')d\epsilon' \\ & +\; \big[1-p(s)\big]\int v(j,t+\tau_{ij},\epsilon')g(\epsilon')d\epsilon' \Big\} g(\epsilon)d\epsilon. \end{aligned}$$

Next, using the definition of the integrated value function, we obtain

$$V(i,t) \;=\; \int \max_{j\in\mathcal{A}(s)} \Bigg\{ p(s) \sum_{l=1}^{n} \pi_{il}^{t}(F,P) \left(F_{il} + P_{il} d_{il} - c d_{il} - \epsilon_l\right) \;-\; \big[1-p(s)\big]\big(c\, d_{ij} + \epsilon_j\big) \\ +\; p(s) \sum_{l=1}^{n} \pi_{il}^{t}(F,P)\, V(l, t+\tau_{il}) \;+\; \big[1-p(s)\big] V(j, t+\tau_{ij}) \Bigg\} g(\epsilon) d\epsilon.$$

The first term in the maximum does not depend on j. Therefore, we can take it out. Thus,

$$\begin{aligned} V(i,t) \;=\; & \int \Bigg(p(s) \sum_{l=1}^{n} \pi_{il}^{t}(F,P) \left(F_{il} + P_{il} d_{il} - c d_{il} - \epsilon_l\right) \;+\; \max_{j\in\mathcal{A}(s)} \Bigg\{ -\big[1-p(s)\big]\big(c\, d_{ij} + \epsilon_j\big) \\ & +\; p(s) \sum_{l=1}^{n} \pi_{il}^{t}(F,P)\, V(l, t+\tau_{il}) \;+\; \big[1-p(s)\big] V(j, t+\tau_{ij}) \Bigg\} \Bigg) g(\epsilon) d\epsilon \\ =\; & \int p(s) \sum_{l=1}^{n} \pi_{il}^{t}(F,P) \left(F_{il} + P_{il} d_{il} - c d_{il} - \epsilon_l\right) g(\epsilon) d\epsilon \\ & +\; \int \max_{j\in\mathcal{A}(s)} \Bigg\{ \;-\; \big[1-p(s)\big]\big(c\, d_{ij} + \epsilon_j\big) \\ & +\; p(s) \sum_{l=1}^{n} \pi_{il}^{t}(F,P)\, V(l, t+\tau_{il}) \;+\; \big[1-p(s)\big] V(j, t+\tau_{ij}) \Bigg\} g(\epsilon) d\epsilon. \end{aligned}$$

The expected value of the cost shocks is zero. Thus, we can simplify the first term of the equation. Furthermore, the second term in the maximum does not depend on j and ϵ. Therefore, we can take it out of the

maximization and integration. Combining these two steps gives

$$
\begin{aligned}
V(i,t) \;=& p(s)\sum_{l=1}^{n}\pi_{il}^{t}(F,P)\left(F_{il}+P_{il}d_{il}-cd_{il}\right)\;+\;p(s)\sum_{l=1}^{n}\pi_{il}^{t}(F,P)\,V(l,t+\tau_{il})\\
&+\int\max_{j\in\mathcal{A}(s)}\left\{\;-\;\big[1-p(s)\big]\big(c\,d_{ij}+\epsilon_j\big)\;-\;\big[1-p(s)\big]V(j,t+\tau_{ij})\right\}g(\epsilon)d\epsilon\\
=& p(s)\sum_{l=1}^{n}\pi_{il}^{t}(F,P)\left(F_{il}+P_{il}d_{il}-cd_{il}\right)\;+\;p(s)\sum_{l=1}^{n}\pi_{il}^{t}(F,P)\,V(l,t+\tau_{il})\\
&+\;\big[1-p(s)\big]\int\max_{j\in\mathcal{A}(s)}\left\{V(j,t+\tau_{ij})\;-\;c\,d_{ij}\;-\;\epsilon_j^{t}\right\}g(\epsilon)d\epsilon\\
=& p(s)\sum_{l=1}^{n}\pi_{il}^{t}(F,P)\left(F_{il}+P_{il}d_{il}-cd_{il}\right)\;+\;p(s)\sum_{l=1}^{n}\pi_{il}^{t}(F,P)\,V(l,t+\tau_{il})\\
&+\;\sigma\big[1-p(s)\big]\log\left[\sum_{j\in\mathcal{A}(s)}\exp\Big(\frac{V(j,t+\tau_{ij})\;-\;c\,d_{ij}}{\sigma}\Big)\right]\\
=& p(s)\left(\sum_{l=1}^{n}\pi_{il}^{t}(F,P)\left(F_{il}+[P_{il}-c]d_{il}\right)\;+\;\sum_{l=1}^{n}\pi_{il}^{t}(F,P)\,V(l,t+\tau_{il})\right)\\
&+\;\sigma\big[1-p(s)\big]\log\left[\sum_{j\in\mathcal{A}(s)}\exp\Big(\frac{V(j,t+\tau_{ij})\;-\;c\,d_{ij}}{\sigma}\Big)\right],
\end{aligned}
$$

where $p(s)=\lambda_i^t/m_i^t$, and

$$
q_{ij}^{t}\;=\;\begin{cases}\dfrac{\exp\Big(\big[V(j,t+\tau_{ij})-c\,d_{ij}\big]/\sigma\Big)}{\sum_{l\in\mathcal{A}(s)}\exp\Big(\big[V(l,t+\tau_{il})-c\,d_{il}\big]/\sigma\Big)} & \text{for } j\in\mathcal{A}(s),\\[2ex] 0 & \text{otherwise.}\end{cases}
$$

I.3 Proof of Theorem 1

First, we show that a mean field equilibrium is a solution $\big(m_i^t, V(i,t)\big)$ to Equations (6) and (8), where λ_i^t, m_{ij}^t, f_{ij}^t, and q_{ij}^t are characterized by Equations (3), (4), (5), and (9). Next, we use Brouwer's Fixed Point Theorem to show that there exists a solution $\big(m_i^t, V(i,t)\big)$ to Equations (6) and (8).

Assume that the problem primitives $\overline{M}$, M, N_i, c, F_{ij}, P_{ij}, τ_{ij}, d_{ij}, S_{ij}, A_{ij}^t, α_i, β_i, k, σ, and the initial distribution of (empty) cars m_i^1 are given. By Definition 1, a mean field equilibrium is a solution $\big(\lambda_i^t, m_i^t, m_{ij}^t, f_{ij}^t, q_{ij}^t, V(i,t)\big)$ to Equations (3)-(9). Given $\big(m_i^t, V(i,t)\big)$, the equilibrium values λ_i^t, m_{ij}^t, f_{ij}^t, and q_{ij}^t are uniquely determined by Equations (3), (4), (5), and (9). In fact, Equations (3), (4), (5), and (9), provide no further information besides determining the values of λ_i^t, m_{ij}^t, f_{ij}^t, and q_{ij}^t given $\big(m_i^t, V(i,t)\big)$. Therefore, we can think of a mean field equilibrium as a solution $\big(m_i^t, V(i,t)\big)$ to Equations (6) and (8), where λ_i^t, m_{ij}^t, f_{ij}^t, and q_{ij}^t are characterized by Equations (3), (4), (5), and (9). Note that Equation (7) is satisfied by any solution to Equations (4)-(6) and we need not worry about it.

Let

$$V \;=\; \Big(V(i,t) :\, i \in \{1,\ldots,n\}, t \in \{1,\ldots,T\}\Big)$$
$$m \;=\; \Big(m_i^t :\, i \in \{1,\ldots,n\}, t \in \{1,\ldots,T\}\Big)$$

denote the $n \times T$ dimensional vectors of the value function and the distribution of the empty taxis. Furthermore, let

$$S_V \;=\; \big\{V :\, |V(i,t)| \le \big(\max_{ij}\{F_{ij} + P_{ij} d_{ij}\} + \log(n)\big) \times (T-i)\big\}$$
$$S_m \;=\; \big\{m :\, \|m\|_1 \le T, m \ge 0\big\}$$

Let $\mathcal{G} : S_m \times S_V \to S_V$ be the continuous function on the right-hand side of (8), i.e.,

$$\begin{aligned}\mathcal{G}(m,V)_{it} \;=& \frac{\lambda_i^t}{m_i^t}\left(\sum_{j=1}^n \pi_{ij}^t(F,P)\left(F_{ij} + [P_{ij} - c]\, d_{ij}\right) \;+\; \sum_{j=1}^n \pi_{ij}^t(F,P)\, V(j, t+\tau_{ij})\right) \\ &+\; \sigma\Big(1 - \frac{\lambda_i^t}{m_i^t}\Big) \log\Big[\sum_{j \in \mathcal{A}(i)} \exp\Big(\frac{V(j,t+\tau_{ij}) - c\, d_{ij}}{\sigma}\Big)\Big].\end{aligned}$$

Similarly, let $\mathcal{F} : S_m \times S_V \to S_m$ be (a continuous function) such that the (i,t)-th element of $\mathcal{F}$ for $t = 1$ is the initial condition, m_i^1, and (i,t)-th element of $\mathcal{F}$ for $t > 1$ is given by

$$\mathcal{F}_{it}(m,V) \;=\; \sum_{j \in \mathcal{A}(i)} q_{ji}^{t-\tau_{ji}} \left(m_j^{t-\tau_{ji}} - \lambda_j^{t-\tau_{ji}}\right) \;+\; \sum_{j=1}^n \lambda_j^{t-\tau_{ji}}\, \pi_{ji}^{t-\tau_{ji}}(F,P),$$

where λ_i^t and q_{ij}^t are given by (3b) and (9). A mean field equilibrium is a solution $\big(m_i^t, V(i,t)\big)$ to Equations (6) and (8), which is equivalent to a solution (m,V) to the fixed point equation

$$(m,V) \;=\; \big(\mathcal{F}(m,V), \mathcal{G}(m,V)\big) \tag{33}$$

given the initial distribution m_i^1 and terminal value $V(i,t) = 0$ for $t > T$. Since $\mathcal{F}$ and $\mathcal{G}$ are continuous mappings, $(m,V) \to \big(\mathcal{F}(m,V), \mathcal{G}(m,V)\big)$ is a continuous mapping from the compact convex set $S_m \times S_V$ onto itself. Therefore, by Brouwer's Fixed Point Theorem (see Royden and Fitzpatrick (1968, Section 10.3)), there exists a solution to (33). Hence, there exists a mean field equilibrium.

I.4 Proof of Proposition 2

Let $H_{ij} = F_{ij} + P_{ij} d_{ij}$ denote the total fare paid for a ride from node i to node j. Following the notation introduced in Section 6, $\overline{H}_{ij}$ denotes the fare paid under the base prices $(\overline{P}_{ij}, \overline{F}_{ij})$, and H_{ij} denotes the fare

paid under (P_{ij}, F_{ij}). By the definition of η_{ij}, it follows that $H_{ij} = \eta_{ij}\overline{H}_{ij}$ for all i, j. By Equation (1), the demand for rides from i to j in period t at fare level H_{ij} is equal to $\Lambda_{ij}^t(H_{ij}) = A_{ij}^t H_{ij}^{\alpha}$. Therefore, the consumer surplus[47] generated by all customers who want a ride from node i to node j in period t at fare level $H_{ij} < \hat{\eta}\,\overline{H}_{ij}$ is equal to

$$\begin{aligned}
\tilde{\mathrm{CS}}_{ij}^t(\eta_{ij}) &= \int_{H_{ij}}^{\hat{\eta}\overline{H}_{ij}} \Lambda_{ij}^t(h)dh && (34)\\
&= \int_{H_{ij}}^{\hat{\eta}\overline{H}_{ij}} A_{ij}^t h^{\alpha} dh \\
&= \int_{H_{ij}}^{\hat{\eta}\overline{H}_{ij}} A_{ij}^t [\eta\overline{H}_{ij}]^{\alpha} d(\eta\overline{H}_{ij}) && (35)\\
&= \int_{1}^{\hat{\eta}} A_{ij}^t \overline{H}_{ij}^{(\alpha+1)} \eta^{\alpha} d\eta && (36)\\
&= A_{ij}^t \overline{H}_{ij}^{(\alpha+1)} \int_{1}^{\hat{\eta}} \eta^{\alpha} d\eta \\
&= A_{ij}^t \overline{H}_{ij}^{(\alpha+1)} \frac{\left[\hat{\eta}^{(\alpha+1)} - \eta^{(\alpha+1)}\right]}{1+\alpha} \\
&= \Lambda_{ij}^t(\overline{H}_{ij})\,\overline{H}_{ij}\,\frac{\left[\hat{\eta}^{(\alpha+1)} - \eta^{(\alpha+1)}\right]}{1+\alpha} \\
&= \Lambda_{ij}^t(\overline{F}, \overline{P})\left[\overline{F}_{ij} + \overline{P}_{ij} d_{ij}\right] \frac{\left[\hat{\eta}^{(\alpha+1)} - \eta^{(\alpha+1)}\right]}{1+\alpha}. && (37)
\end{aligned}$$

Equation (34) follows from the definition of consumer surplus,[48] Equations (35)-(36) use the change of variables $h = \eta\overline{H}_{ij}$, and Equation (37) uses the definition of demand curve and H_{ij}.

Since $\lambda_i^t/\Lambda_i^t(F, P)$ fraction of the customers who want a ride originating at node i in period t are served, the consumer surplus generated by the served customers for rides from node i to node j in period t is equal to

$$\begin{aligned}
\mathrm{CS}_{ij}^t(\eta_{ij}) &= \frac{\lambda_i^t}{\Lambda_i^t(F,P)}\,\tilde{\mathrm{CS}}_{ij}^t(\eta_{ij}) \\
&= \frac{\lambda_i^t}{\Lambda_i^t(F,P)}\Lambda_{ij}^t(\overline{F}, \overline{P})\left[\overline{F}_{ij} + \overline{P}_{ij} d_{ij}\right] \frac{\left[\hat{\eta}^{(\alpha+1)} - \eta^{(\alpha+1)}\right]}{1+\alpha}.
\end{aligned}$$

[47] See Van Zandt (2012, Page 59) for an introduction to consumer surplus.

[48] Consumer surplus is defined as the difference between the total amount that consumers are willing to pay for a ride (indicated by the demand curve) and the total amount that they pay (fare).