

Who Benefits from Surge Pricing?*

Juan Camilo Castillo[†]

December 28, 2019

[Click here to download the most recent version](#)

Abstract

In the last decade, new technologies have led to a boom in dynamic pricing. I analyze the most salient example, surge pricing in ride hailing. Using data from Uber in Houston, I develop an empirical model of spatial equilibrium to measure the welfare effects of surge pricing. My model is composed of demand, supply, and a matching technology. It allows for temporal and spatial heterogeneity as well as randomness in supply and demand. I find that, relative to a counterfactual with uniform pricing, surge pricing increases total welfare by 3.53% of gross revenue. The gains mainly go to riders: rider surplus increases by 6.97% of gross revenue, whereas driver surplus and platform profits decrease by 1.97% and 1.42% of gross revenue, respectively. Disparities in driver surplus are magnified. Riders, on the other hand, are overwhelmingly better off.

Keywords: Surge Pricing, Dynamic Pricing, Ride Hailing

JEL Codes: L11, R41, D47

*I would especially like to thank Matthew Gentzkow, Liran Einav, Susan Athey, and Lanier Benkart for their invaluable advice and support. I am also grateful to Tim Bresnahan, Emma Harrington, Caroline Hoxby, Brad Larsen, Jonathan Levin, Jesse Shapiro, Paulo Somaini, Glen Weyl, Heidi Williams, and Ali Yurukoglu, as well as participants at the Stanford IO workshop and lunch for their valuable comments. I would also like to thank Tiago Caruso, Jonathan Hall, Dan Knoepfle, Chenfei Lu, Helin Zhu, and several other people at Uber whose support and feedback made this project possible. This research is supported by the Kapnick Foundation Fellowship through a grant to the Stanford Institute for Economic Policy Research.

[†]Economics Department, Stanford University. E-mail: jccast@stanford.edu

1 Introduction

Until about ten years ago, dynamic pricing was mostly limited to a few industries, such as airlines and hotels. New technologies, however, have led to rapid changes. Companies can now use the internet and smartphones to communicate prices instantly, and they can use big data to create better pricing algorithms. Consequently, more and more companies are using dynamic pricing, especially in two-sided markets and e-commerce. This shift has most likely increased welfare: flexible prices allow markets to clear, bringing efficiency gains. However, the adoption of dynamic pricing might not always be desirable because it often hurts some market participants.

Ride-hailing platforms like Uber and Lyft have become the most salient adopters of dynamic pricing—or surge pricing, as Uber calls it. To ensure that the market runs smoothly, these platforms adjust prices in response to demand and supply in real time. This flexibility strongly suggests that surge pricing increases welfare. However, the magnitude and distribution of the welfare gains are far from clear. Many critics suggest that surge pricing can hurt riders, calling it a form of price discrimination, or even price gouging (Dholakia, 2015; Crilly, 2016). Others have suggested that it could hurt drivers, whose earnings might be too low unless they carefully plan their actions around surge pricing (Goncharova, 2017).

Because of these concerns, cities like Honolulu, Manila, New Delhi, and Singapore have banned or capped surge pricing (Puckett, 2018; Kazmin, 2016; Yee, 2018; Yusof, 2018). In the US, ongoing litigation might result in a ruling that surge pricing is a form of price fixing (Katz, 2016). Some ride-hailing companies have also voluntarily chosen to avoid surge pricing. DiDi—the largest platform in China—stopped using dynamic pricing, instead adopting potentially inefficient queuing mechanisms (Xinyu, 2017). Determining whether moving away from surge pricing actually benefits riders and drivers requires a firm understanding of the welfare effects of surge pricing. So far the evidence has been limited.

In this paper I develop an empirical model of ride hailing to determine who are the winners and losers from surge pricing. My model allows me to measure the welfare effects—on riders, drivers, and the platform—if the market moves from uniform pricing to surge pricing.¹ The model is composed of three main parts: demand, supply, and a matching technology. On the demand side, riders decide

¹By uniform pricing I refer to prices that are a function of distance and duration only, as with standard taxis, but not of market conditions.

whether to open the app and whether to request a trip. On the supply side, drivers decide when to start and stop working and where to move when they are available. The matching technology determines the drivers to whom riders are matched when riders request trips, and, thus, how long riders need to wait for pickup. I then integrate all three parts in a model of spatial equilibrium to simulate market behavior under alternative pricing policies.

I estimate my model using Uber data from Houston in March-April 2017. Uber was the only ride-hailing platform in Houston at the time; thus, my results speak to surge pricing in a market that has only one platform. To match supply and demand patterns realistically, the model accounts for high-resolution spatial and temporal heterogeneity as well as randomness. I identify agents' short-run elasticities—how real-time price changes affect drivers' movements and riders' decision to request a trip—by exploiting rounding in the surge pricing algorithm, as in Cohen et al. (2016). I identify riders' response to pickup times, which I use to back out the value of time, from variation that arises from drivers' exact position relative to that of riders. Finally, I use data from Uber-run experiments to estimate long-run elasticities—how riders' and drivers' decisions to log in to the app respond to changes in expected prices.

My estimates imply that riders are very inelastic in the short run, in terms of both prices and pickup times. They are more responsive to prices in the long run, but elasticities are also below one. Riders highly value their time. This is consistent with trips taking place during time sensitive moments: riders need to be in time for an appointment, or they need to get to the airport in time for a flight. With regard to drivers, I find evidence that they are more likely to move to areas with high surge multipliers. When I put together all these estimates in an equilibrium model, I obtain simulations that fit spatial and temporal patterns of market behavior well and match the distribution of surge pricing precisely.

I find that surge pricing increases total welfare by 3.53% of gross revenue—or \$0.41 per trip—relative to uniform pricing. This reflects the fact that surge pricing brings efficiency gains to the market. However, surge pricing has strikingly dissimilar effects on different sides of the market. Whereas rider surplus increases by 6.98% of gross revenue, driver surplus and short-run profits decrease by 1.97% and 1.42% of gross revenue, respectively.²

The cause of the asymmetry in welfare effects can be decomposed into three

²These three numbers do not add up to the 3.53% total welfare increase. The remaining 0.06% of gross revenue corresponds to a decrease in revenue from a 2% sales tax.

parts. First, surge pricing brings allocative efficiencies, which only benefit riders. At times of scarcity, uniform pricing allocates trips randomly: only riders that are lucky to be near a driver get a trip. With surge pricing, trips are allocated to riders who have a high willingness to pay, which increases rider surplus. Drivers, on the other hand, see no benefit from a better allocation of trips. Their value for a trip is fairly homogeneous; in my model, it is simply equal to earnings from the trip minus the physical cost of completing it, which is the same for every driver.³ Thus, driver surplus cannot increase from a better allocation.

Second, riders have a much higher value of time than drivers. Besides allocating trips better, surge pricing further increases welfare because of time-saving matching efficiencies: riders are picked up more quickly, and drivers wait less between trips. I find that drivers save more time than riders, but they value these savings less. Whereas the value of time to drivers is their average hourly earnings net of driving costs, which is slightly above minimum wages, riders who request trips are very time sensitive. The welfare gain from time savings is thus substantially higher for riders than for drivers.

Third, surge pricing allows the platform to set lower prices on average, transferring welfare from drivers and the platform towards riders. I assume that to maximize long-run profits, Uber maximizes a weighted sum of its short-run profits, rider surplus, and driver surplus. I back out the weights as the values that rationalize the commission rate and the average price in the data. I find that when constrained to a uniform multiplier, the platform sets it above the average multiplier with surge pricing. This is the case because, for a given time and place, it is worse to err by setting prices too low than too high, in part because of a matching failure—which Castillo et al. (2018) analyze theoretically—that takes place when drivers are scarce. With uniform pricing, the only way to avoid prices that are too low is with a high price. With surge pricing, on the other hand, the platform can set a low average price, and surge pricing automatically avoids the problem.

I also analyze the distributional effects within riders and within drivers. I find that surge pricing makes driver earnings more unequal. Drivers who work during busy times—and thus have high earnings—are even better off with surge pricing because of higher prices. During off-peak hours, in contrast, prices and earnings are lower. On the riders' side, I find that the vast majority of riders benefit, including

³Drivers might have idiosyncratic preferences for particular types of trips. However, Uber as a policy reveals very little information about trips to drivers—nothing beyond the rider's name and location; thus, there is limited scope for surge pricing to improve the allocation to drivers.

those who own expensive phones, those who request trips from low income areas, and those that have a low willingness to pay.⁴ The only riders who are hurt are those who want to request a trip during a couple of hours on Friday afternoon and Saturday midday, when prices are highest.

The public debate about the desirability of surge pricing has emphasized its negative effects on riders and drivers. My results suggest that riders' complaints are not well-founded. Their confusion might arise because they do not account for equilibrium effects—higher pickup times and lower reliability without surge pricing—and because they are unaware that without surge pricing, they would pay higher average prices. On the other hand, my findings suggest that drivers might have good reason to complain. Given that their hourly earnings are not much higher than the minimum wage, even the small effects I find might be a concern.

My analysis focuses on a market that has a single ride-hailing platform, which provides a clean environment in which to analyze surge pricing in the absence of competition. Thus, I am not able to say how competition might affect the welfare effects of surge pricing. An answer to that question would require either data from two platforms or strong assumptions about multi-homing behavior. Another limitation is that I do not allow heterogeneity in the response of agents to prices. If I observed income data, for instance, I could allow riders' response to depend on their income. My findings about overall welfare on riders, drivers, and Uber would be unlikely to change, but I could measure the distributional effects within riders more accurately.

My analysis begins in section 2, where I introduce the Uber market and describe the data I use. I introduce my model in section 3. In section 4 I present descriptive evidence that shows the variation in the data that drives my main model parameters. In section 5 I explain my identification strategy and show parameter estimates. I analyze the welfare effects of surge pricing in section 6, and I conclude in section 7.

Related work

A few related papers analyze the welfare effects of surge pricing. For example, Cachon et al. (2017) propose a theoretical model without matching frictions. Ming et al. (2019) build an empirical model using DiDi data; they do not observe waiting

⁴I do not observe data on individual riders' income, so I use phone price and income by trip request location as proxies for income.

times, nor do they model spatial heterogeneity, which limits the extent to which they can account for matching frictions.⁵ Both papers find that riders, drivers, and platforms benefit from surge pricing, although riders might be hurt at times. In a theoretical analysis, Castillo et al. (2018) point out important matching inefficiencies ("wild-goose chases") that arise with excess demand. Those inefficiencies can be avoided with a high uniform price, or with surge pricing and lower average prices. Lower prices mean that surge pricing potentially hurts *drivers*. I confirm empirically that surge pricing hurts drivers, which underscores the importance of matching frictions.

Many computer science and operations research papers have analyzed surge pricing (e.g., Bimpikis et al., 2019; Besbes et al., 2019; Ma et al., 2018; Garg and Nazerzadeh, 2019). A survey by Korolko et al. (2018) gives a detailed overview. Their main goal is to improve the design of surge pricing algorithms. In contrast, I take the current design of the algorithm as a given and analyze its effect on different market participants.

Methodologically, this work relates to empirical papers on matching and spatial equilibrium in transportation. The first few contributions analyze taxi markets—and thus they have little to say about dynamic pricing, which is not used by taxis. Lagos (2003) analyzes entry restrictions and fares; Frechette et al. (2019) also analyze entry restrictions, as well as the adoption of Uber-like matching. My notion of equilibrium is similar to the one used by Buchholz (2018), who finds that the structure of taxi fares can be modified to decrease search inefficiencies. Bian (2018) and Shapiro (2018) consider the interaction between taxis and ride hailing. In one counterfactual, Bian shuts down surge pricing, which results in less efficient matching. None of these papers has data on riders—only on drivers and trips—and so they rely on structural assumptions to back out demand. In contrast, I observe riders, whether they request trips; thus, I estimate demand directly from the data.

A number of papers estimate demand and supply in ride-hailing markets.⁶ Cohen et al. (2016) and Lam and Liu (2017) estimate rider surplus. My identification strategy is closely related to Cohen et al.'s: we both exploit rounding in the surge algorithm to identify agents' response to prices. Buchholz et al. (2019) estimate the

⁵To account for matching frictions, they assume that the utility of riders depends directly on the number of available drivers and that the utility of drivers depends directly on the number of riders.

⁶A fast-growing literature analyzes other aspects of ride hailing. Some studies focus on labor supply (Hall et al., 2019; Chen et al., 2017; Cook et al., 2018), efficiency vs. taxi markets (Cramer and Krueger, 2016), the matching algorithm (Afeche et al., 2017), market thickness and competition (Nikzad, 2017), and the customer experience vs. taxis (Athey et al., 2018; Liu et al., 2018).

value of time for riders in Prague. Their findings are lower than the ones I estimate, but they are also above median wages and well above minimum wages.⁷ Papers that estimate supply elasticities include Angrist et al. (2017), who design an experiment to estimate drivers' value for flexibility, and Lu et al. (2018), who, on the basis of an outage in the Uber platform, estimate drivers' short-run response to surge multipliers. Relative to these works, I estimate both sides of the market and put them together in an equilibrium model to analyze welfare effects.

My paper also adds to the literature on two-sided platforms. The seminal papers in this literature find that welfare effects are mainly determined by elasticities and cross-market externalities (Rochet and Tirole, 2003; Armstrong, 2006; Weyl, 2010). They do not consider matching frictions, which are the main drivers of my results. Some empirical papers analyze markets in which matching plays a major role, such as Airbnb (Fradkin, 2017) and online labor platforms (Arnosti et al., 2018; Cullen and Farronato, 2018). Finally, my paper relates to papers that analyze dynamic pricing in various settings. In airlines, some consumers benefit and some are hurt, with ambiguous net effects (Lazarev, 2013; Williams, 2017). Hendel and Nevo (2013) find that temporary sales in retail increase overall welfare.

2 The Uber market

I analyze the Uber market in Houston between March 16 and April 8, 2017. Before this period, Uber did not record data about riders who did not request a trip. After this period, Uber stopped using rounding in its surge pricing algorithm, which I rely on for my identification strategy. I focus on trips starting in the area of central Houston shown in figure 1. It covers 8.63% of the area within the city limits of Houston, but it accounts for 56.4% of all trips and 78.4% of trips with surge pricing. In total, my sample includes around half a million trips.

I focus on UberX, Uber's main product, which matches passengers to independent drivers. UberPool, which also matches passengers going in similar directions, was not available in Houston during my period of analysis. Other Uber products, such as UberXL (larger cars) and UberBlack (luxury cars), accounted for less than 5% of Uber trips.⁸

⁷The value of time ranges between \$0.40/min during peak weekday hours and \$0.05/min during off-peak weekend hours. The minimum wage was \$2.79/h and the mean wage was \$10.30/h.

⁸UberXL and UberBlack drivers can also be matched to people who request UberX or UberPool trips, in which case they get paid the fare for the product requested by the driver.

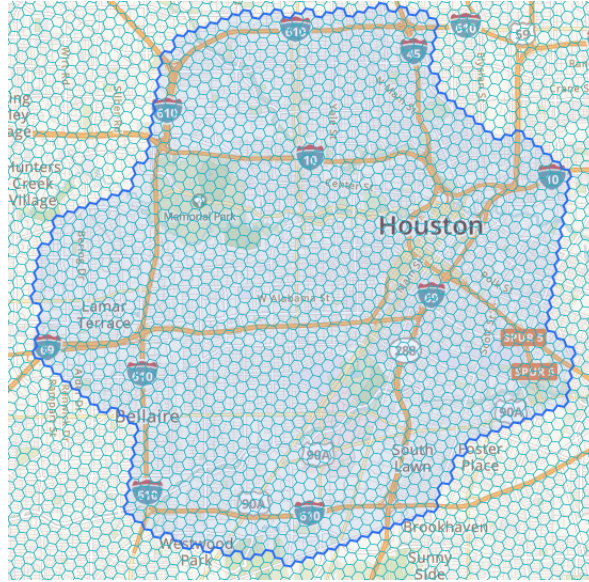


Figure 1: Region of analysis

Note: Central Houston. The shaded area represents the region of analysis. Surge pricing varies spatially at the hexagon (or *location*) level.

Lyft was not present in Houston between November 2014 and May 2017.⁹ Houston was the one large American city where Uber was the only ride-hailing platform. Thus, Houston presents a clean setting in which to analyze the welfare effects of surge pricing without having to consider the effects of competition between platforms. For that reason, my main results speak to a market with only one ride-hailing platform.

Riders, drivers, and trips The raw data I observe has a high temporal and spatial resolution: I observe market participants every few seconds whenever the app is open, and I observe their location up to the precision of their cell phone GPS. I aggregate the data at the level at which surge pricing varies: into two minute periods,¹⁰ which I index by $t \in T$, and into the hexagons in figure 1, which I call *locations* and index by $l \in L$. Each hexagon is roughly 400 meters (one quarter mile) across. This level of aggregation is much finer than in previous papers, such as Buchholz (2018) and Frechette et al. (2019).¹¹ A higher resolution is important to

⁹Lyft decided to exit Houston after the City Council passed an ordinance requiring extensive background checks for drivers, including fingerprinting and a physical exam.

¹⁰Surge multipliers are not updated exactly every two minutes, but over 99% of the time they are updated after between 100 and 140 seconds. I define periods to start and end whenever multipliers are updated.

¹¹Frechette et al. (2019) split Manhattan into eight areas. Buchholz (2018) splits Manhattan into 48 areas. Both papers use one-hour periods.

capture the short run, local imbalances surge pricing is meant to counteract.

I observe riders whenever the app is open. I can see the rider’s location, the selected destination (if she has chosen one), the fare for a trip to her destination, and an estimated time of arrival (ETA) before pickup. I aggregate rider data by *session*, defined as a period of activity with no gaps of half an hour or longer. Sessions can end in two ways: the rider can request a trip, or she can be inactive for half an hour, after which I say the rider *leaves* the app.¹² A new session begins if the rider opens the app and chooses a destination after the end of a session. I take the multiplier, fare, and ETA to be the last ones that were observed by the rider before deciding to request or leave.

I also observe detailed data on drivers. I can see their location whenever they are logged in, as well as whether they are available to be matched (33.5% of the time), on their way to pick up a passenger (20.6% of the time), or taking a passenger to her destination (45.9% of the time). Finally, I observe trip statistics, including the time and location during request, pickup, and drop-off, as well as the trip fare and how it was split between Uber and the driver.

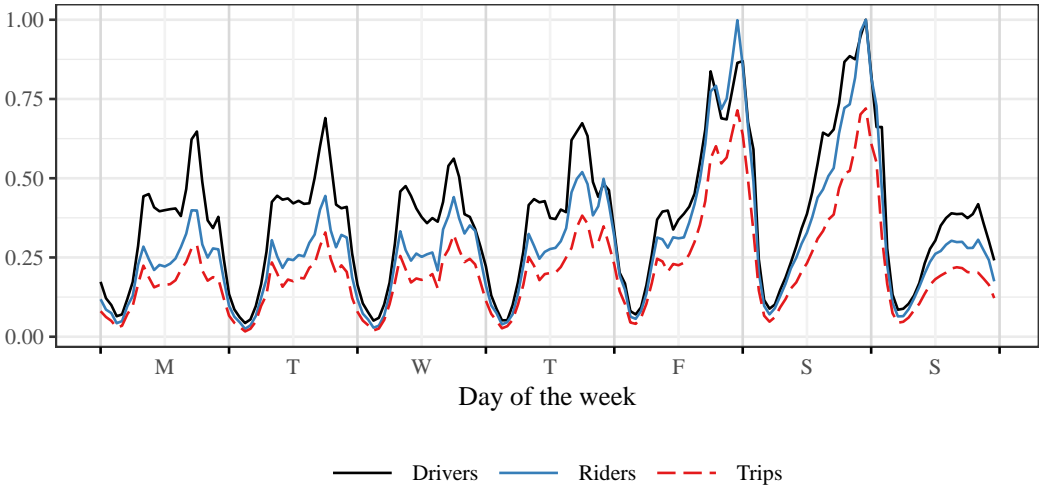


Figure 2: Weekly patterns in demand, supply, and trips

Note: The three time series represent the average number of drivers working, the average number of riders that open the app, and the average number of trips that take place. Riders and drivers are normalized to have a maximum of one. Trips are on the same scale as riders.

Figure 2 plots the average market behavior as the week goes by. The main daily patterns are low activity at night and high activity during the day. During weekdays, a dip takes place around noon, and a big spike occurs around the afternoon

¹²If the rider requests a trip but cancels it before pickup, I assume the rider never requested it.

rush hour and the evening. The least busy day is Sunday, and Friday and Saturday are the most busy days. All three variables behave very similarly. The most noticeable difference is some excess supply around noon during weekdays, as well as higher demand relative to supply during Friday and Saturday evenings.

Figure 3 shows spatial patterns. The area with most trip requests, towards the northeast, is Downtown. Another high demand area is The Galleria, a business area on the west. There is a third area with high demand towards the south, around NRG Stadium and the Astrodome, two major sports complexes. Drivers tend to be in areas that have a large number of trip requests and along major highways.

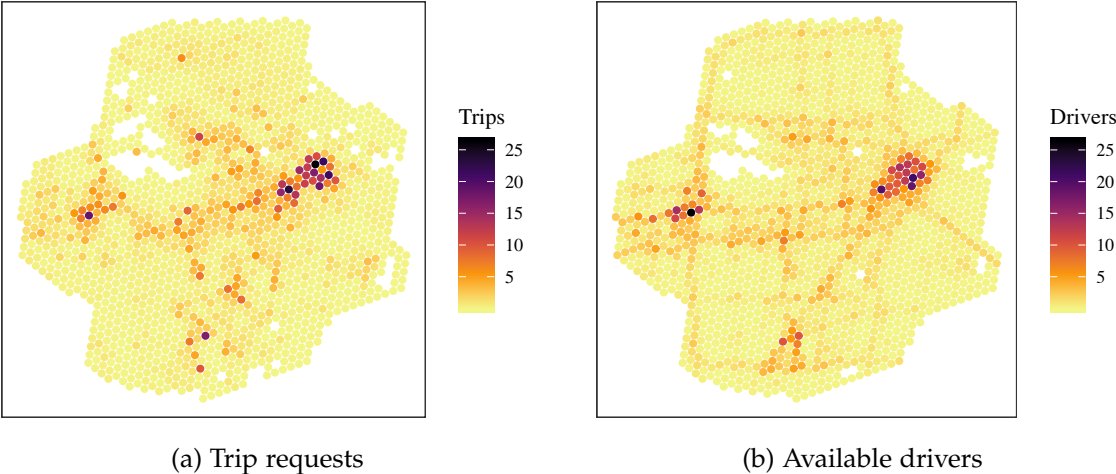


Figure 3: Spatial patterns in demand and supply

Note: Number of trip requests and available drivers by location for the whole sample. Color scales are normalized to have mean one.

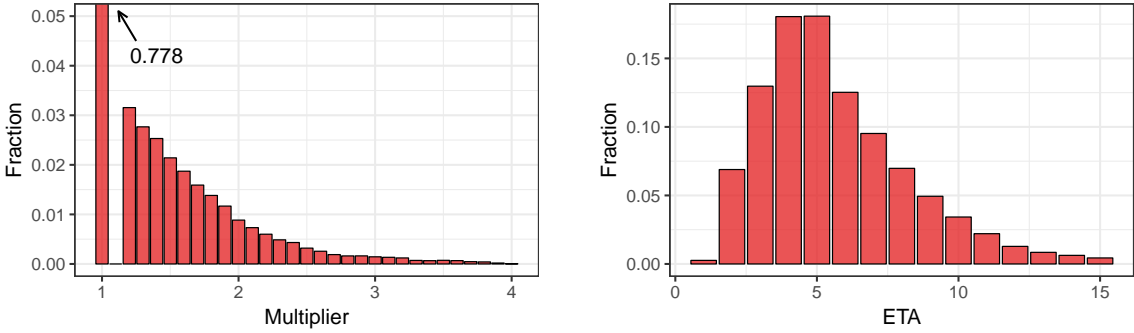
Matching and surge pricing I explain now how matching and surge pricing work. This description is specific to my period of analysis; some of these features have changed since the period of study.¹³ The matching process starts with the rider opening the app and selecting pickup and destination points.¹⁴ The app then displays a fare in dollars and an ETA before pickup (see a screenshot in appendix H.2).¹⁵ If the rider decides to request a trip, she is matched to the nearest available driver, who then has a few seconds to accept the trip. If the driver does not accept the trip, the rider is then matched to the second nearest available driver, and so on.

¹³Some changes are: Uber now groups trips into batches that are matched every few seconds. Surge pricing no longer uses rounding. On drivers' side, surge pricing is no longer multiplicative; instead, they get a bonus per trip, no matter how long (Garg and Nazerzadeh, 2019).

¹⁴It is possible to request a trip without a destination, but the interface makes it very difficult.

¹⁵Until mid 2016 the rider was shown a surge multiplier instead of a fare to the destination.

The fare shown to the customer is the product of two components. The first is an underlying fare—the *base* or *unsurged fare*—which is a linear function of the expected trip distance and duration given the pickup and dropoff coordinates and the hour of the week. It does not change in real time. The second is a *surge multiplier* that responds in real time to supply and demand.

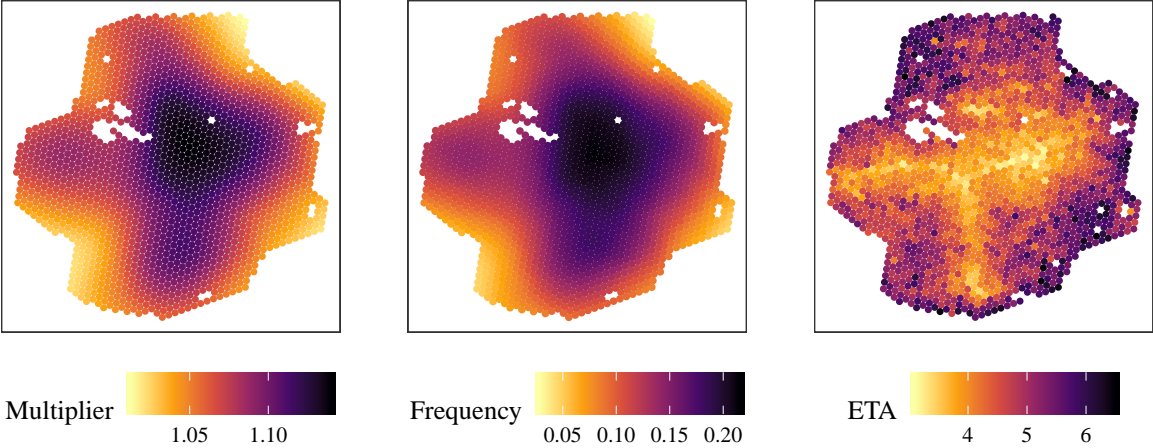


(a) Observed multiplier distribution

(b) ETA distribution

Figure 4: Distribution of multipliers and ETAs observed by riders

Spatially, multipliers vary by location (see figure 1). They are updated every two minutes simultaneously across the whole city. Whenever a driver is available, he can observe a map showing all multipliers in the city (see appendix H.2). Appendix H.1 shows how multipliers behaved during one typical Tuesday afternoon. Less than half of the variation in prices is predictable: a regression of the surge multiplier on half hour of the week by location fixed effects has an R^2 of 0.288.



(a) Average multiplier

(b) Multiplier > 1

(c) Average ETA

Figure 5: Spatial patterns in surge multipliers and ETAs

Figure 4 shows the distribution of surge multipliers and ETAs that passengers

observe. 77.8% of passengers see a multiplier of 1 when they open the app.¹⁶ When the multiplier is greater than one, it is typically less than 2 and it is rarely above 3. Figure 5 shows the average behavior of multipliers and ETAs across space. Multipliers are highest around Downtown and towards the south. ETAs tend to be lowest in areas with most trip requests, where most available drivers are located, and they tend to be highest in peripheral areas.

3 Model

I now present a model of a ride-hailing market. I start by giving a brief overview of the whole model. Subsections 3.1-3.4 explain each part in detail.

Agents make two types of decisions: long-run and short-run decisions. In the long run, they decide whether to enter the market—i.e., log in to the app—based on expectations. Riders choose whether to open the app given what they expect prices to be and how long they expect to wait before being picked up. Drivers decide if they want to start working depending on how much they expect to earn should they decide to work. These decisions are based on expectations because agents must make adjustments ahead of time. A driver might have to arrange for someone to take care of his children, for instance, and a rider might only open the app at the end of the workday if she decided not to drive to work in the morning.

In the short run, agents who are already in the market make choices using the information they observe in the app. Riders observe a fare and an ETA, based on which they decide whether they want to request a trip. Drivers that are available observe a map with all surge multipliers in the city. Using the information from that map, they decide where they want to move.

Besides agents' decisions, the platform also takes some actions. It computes surge multipliers, fares, and ETAs, and it shows them to riders and drivers. It also assigns a nearby driver to riders that request a trip. Figure 6 is a timeline that clarifies the mechanics of the interaction between riders, drivers, and the platform. It shows everything that happens during one two-minute period. Steps during which riders and drivers make decisions are highlighted in bold above the timeline. Steps below the timeline are mechanical actions taken by the platform.

¹⁶I weight multipliers by the number of passengers who observe them. Without weighting, 85.7% of periods have a surge multiplier of 1.

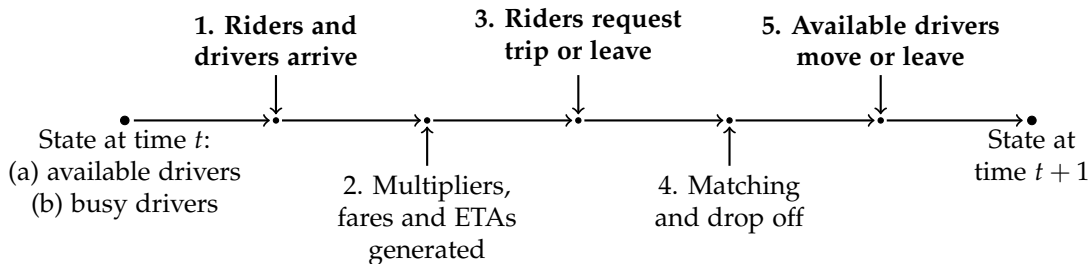


Figure 6: Model timeline

Note: Timeline of events that take place during every period. The initial state is the set of available and busy drivers. Each available driver is in a certain location, and every busy driver will drop off a passenger and become available during some future period in a certain location. Riders’ and drivers’ decisions are highlighted in bold.

3.1 Demand

Riders make two decisions. First, a number of riders decide to open the app and choose a destination. Second, after having opened the app, riders decide whether to request a trip or not, based on the fare and ETA that they observe in the app.

3.1.1 Trip requests

Rider i , who already decided to open the app, is in location l at time t (during hour of the week h) and wants to go to a destination at a distance r_i . She gets a “quote” that includes a price p_i and an ETA before pickup w_i .

If she requests a trip, the rider gets utility

$$U_i = \alpha(r_i, l, h) + \beta(r_i)p_i + \gamma(r_i)w_i + \epsilon_i. \quad (1)$$

The first term captures patterns in the intrinsic value of a trip by how far the rider is going, by location, and by hour of the week. The second term captures the disutility of paying, and the third term captures the disutility from waiting. Finally, ϵ_i is an error that captures all remaining heterogeneity. Utility is measured relative to a short-run outside option the rider chooses when she does not request a trip. It may be an alternative form of transportation—e.g., driving herself, biking, or taking a bus—or simply not going to her destination. The rider requests a trip if $U_i > 0$.

The key parameters in this model are the coefficients on prices and ETAs. $\beta(r_i)$ measures the short run elasticity of demand, and $\gamma(r_i)$ measures the ETA elasticity. I allow both coefficients to change with trip distance: one would expect different elasticities for someone taking a short, \$5 trip, and for someone taking a \$30 trip to

the other side of the city. The ratio $\frac{\gamma(r_i)}{\beta(r_i)}$ measures the value of time for rider i .

In this model, riders are static decision makers. In reality, riders can decide to wait and request a trip at a later time. However, dynamic decisions do not seem to play a major role in the data. Only 21% of trip requests take place more than 2 minutes after choosing a destination. Thus, I do not model explicitly the decision to wait. I only take into account the final decision the rider made, whether to request a trip or leave the app. Appendix G.2 shows that previous prices and ETAs observed by riders have no effect on their decisions, suggesting that modeling riders' behavior as a static decision does not bias estimates of their elasticities.

3.1.2 Opening the app

Besides their short-run outside option, riders have a long-run outside option that requires some planning. Thus, they can only choose it before they observe prices and ETAs. For instance, riders might buy a car or coordinate to carpool with coworkers when they expect high prices. Let u_i be the value of this outside option for rider i , relative to her short-run outside option. It is drawn from a distribution F^u .

Suppose rider i would want to go from l to a destination at a distance r_i during time of the week h . I aggregate distances into quantiles, which I call *distance groups*. Let \tilde{r} be the distance group r_i belongs to. Then $U(l, h, \tilde{r}) = \mathbb{E} \left[\frac{1}{\beta(r_i)} \max \{U_i, 0\} \mid l, h, \tilde{r} \right]$ is the rider's ex-ante dollar value of opening the app. It represents her expectation on the value of her best choice—either requesting a trip, with value $\frac{U_i}{\beta(r_i)}$, or the short-run outside option, with value 0—given what she knows *before* opening the app and observing the fare and the ETA. This is an equilibrium quantity, as it depends on the distribution of prices and ETAs that the rider faces in equilibrium. Rider i opens the app if and only if $U(l, h, \tilde{r}) > u_i$.

There is an arrival rate $\lambda_{lh\tilde{r}}^0$ of riders that could potentially open the app to go to a destination in distance group \tilde{r} in location l during hour of the week h . Each one of them chooses to open the app with probability $\Pr(U(l, h, \tilde{r}) > u_i) = F^u(U(l, h, \tilde{r}))$. Thus, the actual rate at which riders open the app is $\lambda_{lh\tilde{r}} = \lambda_{lh\tilde{r}}^0 F^u(U(l, h, \tilde{r}))$. I assume that $F^u(x) \propto x^\rho$.¹⁷ With this functional form, riders open the app at a rate

$$\lambda_{lh\tilde{r}} = A_{lh\tilde{r}} U(l, h, \tilde{r})^\rho \quad (2)$$

for some demand shifter $A_{lh\tilde{r}}$. Thus, the elasticity of the number of people who

¹⁷ x^ρ is unbounded, but it can be interpreted as the left tail of the actual distribution: at any particular time, the vast majority of people would not request a trip under any price and ETA.

open the app with respect to $U(l, h, \tilde{r})$ is constant and equal to ρ . Rider surplus relative to the long-run outside option is $\int_0^{U(l, h, \tilde{r})} (U(l, h, \tilde{r}) - u) dF^u(u) = \frac{A l h \tilde{r}}{1 + \rho} U(l, h, \tilde{r})^{\rho + 1}$.

The key parameter of this arrival model is ρ . It determines the long run elasticity of demand—i.e., how the number of requests changes if there is a change in prices that riders know of in advance.

3.2 Supply

Drivers make three decisions. At the beginning of a shift, they decide whether they want to start working. If they do, during every period they are available—i.e., waiting to be matched—they decide whether they want to keep on working or leave the platform before the end of the current period. Finally, if they decide to stay, they must also choose where to move before the beginning of the next period.

A standard approach is to model drivers as fully rational utility maximizers, as in Buchholz (2018) and Frechette et al. (2019). This approach, however, is substantially more complicated in this setting given my goal of measuring the effects of surge pricing. First, I want to capture the short-run, local imbalances that surge pricing aims to correct. Thus, my model has a much higher temporal and spatial resolution than previous papers. Second, the state space must incorporate all multipliers that surround the driver, which means it is high dimensional. Computing the value function of drivers is not feasible due to the curse of dimensionality.

I, thus, need to make some simplifying assumptions that result in a feasible problem. The first assumption is that utility is equal to earnings minus the physical cost of driving (fuel, depreciation, and maintenance) minus the opportunity cost of working. Thus, two drivers with an equal opportunity cost who get the same net earnings are equally well off, regardless of how they get those earnings. In particular, it does not matter where in the city they drove, whether they had to drive in traffic, or whether they were busy or idle most of the time.

3.2.1 Movement

Driver j is available in location l during some time t during hour of the week h , and he observes surge multipliers \mathbf{m}_t . The state—the information observed by the driver that influences his behavior—is $\mathbf{s}_t = (l, h, \mathbf{m}_t)$.

I assume that the driver’s location in period $t + 1$ follows a stochastic movement rule that depends on two things. First, it depends on mean earnings given the state: the driver is more likely to move to locations where, on average, earnings

are higher. Mean earnings do not depend on any private information the driver has about what he might do next. Instead, they are simply an empirical average of the earnings drivers get for the next \bar{t} periods if they move to some location k after being in state \mathbf{s}_t . Second, the movement rule follows road and traffic patterns. For instance, the driver cannot move instantaneously to the other end of the city, and he is less likely to move to far-away locations during rush hour than he is at 3 am in the morning.

Let $v_k(\mathbf{s}_t)$ be *mean future earnings*, the mean of the sum of the net earnings drivers get during period $t + 1$ until $t + \bar{t}$ if they move to location k after being in state \mathbf{s}_t :

$$v_k(\mathbf{s}_t) = \mathbb{E} \left[\sum_{s=t+1}^{t+\bar{t}} \pi_s \mid \mathbf{s}_t, l_{t+1} = k \right] - c_l^k \quad (3)$$

In this equation π_t denotes net earnings during period t —i.e., earnings from trips minus physical driving costs. I denote the location at time t by l_t , and c_l^k represents the physical cost of moving from location l to k . I set $\bar{t} = 45$ because, in the data, most of the effect of surge multipliers on earnings takes place during the first 90 minutes (see appendix G.4). Thus, $v_k(\mathbf{s}_t)$ incorporates almost all the information about future earnings drivers can infer from multipliers.

Mean future earnings $v_k(\mathbf{s}_t)$ are an empirical average of market behavior in equilibrium. There are many possible future outcomes for drivers who move to k after being in state \mathbf{s}_t . They might subsequently move north or south; they might be matched immediately, or they might have to wait to be matched. The probability of each one of these outcomes depends on equilibrium behavior—how drivers behave after moving to k , how every other driver in the market behaves, and how riders behave. The mean of net earnings over all these possibilities is $v(k, \mathbf{s}_t)$. Although this quantity only depends directly on the driver’s movement during the current period, it implicitly accounts for future movements.

Let $l_{j,t+1}$ be the location to which driver j moves. The movement rule is

$$\Pr(l_{j,t+1} = k \mid \mathbf{s}_t) = \frac{\exp(\omega_{lkh} + \delta v_k(\mathbf{s}_t) + \zeta_{kt})}{\sum_{k'} \exp(\omega_{lk'h} + \delta v_{k'}(\mathbf{s}_t) + \zeta_{k't})}. \quad (4)$$

The first term inside the exponential is a fixed effect by origin, destination, and hour of the week. It captures road and traffic patterns. If locations l and k are very far from each other or have no roads connecting them (e.g., a river separates them), then ω_{lkh} is very low, and so the probability of moving to k in one period is negligible. If it is easy to move from l to k in one period at 3 am, but not when

there is rush hour traffic, then ω_{lkh} is higher at 3 am than during rush hour.

The middle term captures the fact that drivers are more likely to go to locations that have higher mean future earnings. The key parameter in this model is δ , which measures the extent to which drivers are more likely to move towards areas with high earnings. Drivers do not respond to changes in surge multipliers directly; however, higher multipliers lead to higher expected earnings, so they respond to multipliers indirectly. For that reason, δ also measures whether drivers are more likely to move to areas with higher surge multipliers.

Finally, ζ_{kt} is an unobserved term that captures systematic shocks that cause drivers to flock towards specific locations. For instance, if drivers know that an event will end soon, they might move systematically towards the event location.

Movement rule (4) departs from a standard model of a fully rational—i.e., forward-looking and utility-maximizing—driver in three ways. First, $v_k(\mathbf{s}_t)$ only accounts for earnings during the next \bar{t} periods, so drivers do not respond to any earnings they might get more than \bar{t} periods into the future. Second, all drivers respond to earnings until period $t + \bar{t}$, even if they plan to stop working before then. Third, $v_k(\mathbf{s}_t)$ does not account for future fixed effects ω_{lkh} , unobserved terms ζ_{tk} , or random draws from the distribution specified by the movement rule. This third point stems from the assumption that drivers' utility only depends on net earnings and opportunity costs. My model, thus, views the fixed effects and error terms as capturing constraints imposed by roads or traffic. They might also capture mistakes from inattention or limited knowledge.

In this setting, it is not feasible to compute value functions for fully rational drivers due to the curse of dimensionality. The simpler model I propose allows me to model drivers in a tractable way, while still capturing the essence of fully rational behavior. Concretely, drivers are more likely to move towards high-earnings areas in a forward-looking manner: when deciding where to move based on mean future earnings, drivers implicitly consider where they will move in subsequent periods, whether surge multipliers might go up, and whether they will get a trip quickly.

3.2.2 Entry

Driver j is considering whether to start working in location l at time t .¹⁸ He has an outside option that represents, for instance, leisure, or working at a different job. The hourly value of this outside options, \bar{W}_i , is drawn from a distribution F^W .

¹⁸Drivers have no choice over l and t . They simply log in whenever they finish doing whatever they were doing before in the location they were.

If he starts working, the driver expects hourly earnings W_{lh} *before observing surge multipliers*. This is an equilibrium quantity: it depends on how the driver expects all other agents to behave. In equilibrium, it must be consistent with empirical averages. The driver starts working if and only if $W_{lh} \geq \bar{W}_i$.

There is a rate μ_{lh}^0 of potential entrants to the market in location l during hour of the week h . A fraction $\Pr(W_{lh} \geq \bar{W}_i) = F^W(W_{lh})$ of them start working. Thus, the actual rate at which drivers start working is $\mu_{lh}^0 F^W(W_{lh})$. I make the functional form assumption $F^W(W_{lh}) \propto W_{lh}^\sigma$.¹⁹ The entry rate is thus

$$\mu_{lh} = B_{lh} W_{lh}^\sigma, \quad (5)$$

where B_{lh} is a horizontal demand shifter, and driver surplus for (l, h) is $B_{lh} \int_0^{W_{lh}} (W_{lh} - W') dF^W(W') = \frac{B_{lh}}{1+\sigma} W_{lh}^{\sigma+1}$ dollars per hour.

The parameter σ represents the elasticity of entry to hourly earnings. It is a measure of the long run elasticity of supply: it determines the extent to which the number of drivers who start working responds to expected changes in earnings.

This model does not allow drivers to start working in response to unexpectedly high multipliers. Although that might happen to some extent in real life, it is unlikely to have a large effect. I show in appendix G.3 that multipliers can only predict around 15% of unexpected variation in multipliers—i.e., residuals from location by hour of the week fixed effects—more than ten minutes into the future. Therefore, unexpected changes in multipliers convey little information about the total earnings drivers would get if they decide to start working.²⁰

3.2.3 Exit

At the time that driver j arrives, which I denote by t_j^0 , he draws an intended shift duration D_j from distribution G_h , which varies by the hour of the week. The driver stops working the first time that he is available after $t_j^0 + D_j$. Modeling the exit decision accurately is important to get a good approximation of the number of drivers who are working at any given time.

In reality, drivers might to some extent respond to surge multipliers by working a while longer. If I run regressions of a dummy for leaving on nearby multipliers, I

¹⁹As with riders' opportunity cost, this distribution is unbounded, but it can be understood to be the left tail of a very large distribution.

²⁰One exception are drivers who could start working immediately in high demand locations. However, there are very few of them: only 6% of drivers start working in the main area of analysis. The remaining 94% start working outside and drive in.

do not see a significant effect (see appendix G.5), so I abstract from this margin of response.

3.3 Matching technology

In period t , the set of riders that request a trip is I_t^r and the set of drivers that are available is J_t^a . Uber sometimes matches drivers who are completing a trip; thus, J_t^a includes riders who will drop off a rider during the next two periods. The matching technology determines which drivers in J_t^a are matched to each rider in I_t^r .

Matches take place as follows. First, the platform computes a pickup time w_{ij}^p for every pair of a rider $i \in I_t^r$ and a driver $j \in J_t^a$. The pickup time is drawn from a distribution $G(\cdot | l_i, l_j, b_j, h)$ that depends on the rider's location l_i , the driver's location l_j , whether the driver is busy b_j , and the hour of the week h .

Riders in I_t^r are then matched sequentially in a random order. For every rider, her trip is offered first to the driver with the lowest pickup time, who accepts it with probability ϕ^b if he is busy and probability ϕ^a if not.²¹ If he does not accept the trip, it is then offered to the next closest driver, who also accepts it with probabilities ϕ^b and ϕ^a , depending on whether he is available. The process goes on until the rider is eventually matched or until no available drivers remain within 10 km, in which case the rider does not get a trip and is forced to take her outside option.

The matching technology depends on the distribution $G(\cdot | l_i, l_j, b_j, h)$, as well as on parameters ϕ^b and ϕ^a . Together, they all determine the distribution of realized pickup times—in other words, the size of the matching inefficiency. A distribution $G(\cdot | l_i, l_j, b_j, h)$ with higher values means a larger matching inefficiency. Lower acceptance rates result in matches with drivers who are farther away, and, therefore, a larger matching inefficiency.

3.4 Surge multipliers and other parts of the model

One essential part of the model is the algorithm Uber uses to generate surge multipliers. In simulations with surge pricing, I generate multipliers using the exact same algorithm Uber used in Houston during the period of analysis. I cannot disclose the algorithm because it is proprietary. Its most important features are that it depends on the number of available drivers and on the number of riders who open

²¹In principle, drivers could accept or reject trips selectively. Uber, however, punishes drivers with low acceptance rates. I assume acceptance is an exogenous process. This is the case if the main reason drivers fail to accept trips is because they were not paying attention.

the app. Both quantities are aggregated over nearby locations and over the last few minutes. I specify a few more details necessary for my identification strategy in section 4.1. In other counterfactuals, Uber simply sets a uniform multiplier.

In appendix A I present three additional parts of my model that are necessary to fully describe market behavior. First, I explain what determines the duration of trips and how Uber generates fares and ETAs. These are simply mechanical steps where no agent makes choices. Second, I present a model for the exact trip destination, which I assume is exogenous. Third, I need to account for the behavior of drivers who are outside the region of analysis (since there is one unique pool of drivers in all of Houston). I assume that the behavior of drivers who are far from the region of analysis is unchanged in counterfactuals.

3.5 Equilibrium

Agents' choices depend on their beliefs. Rider arrival depends on expected utilities, driver entry depends on expected earnings, and driver movements depend on how they believe surge multipliers map into expected earnings. To fully describe market behavior, I need to specify how those beliefs are determined. I do so based on an equilibrium condition: beliefs must be consistent with empirical averages.

Let \mathbf{U} be the vector of riders' ex-ante utilities $U(l, h, \tilde{r})$. Let \mathbf{W} be the vector of drivers' expected hourly earnings $W(l, h)$ before observing surge multipliers. Finally, let \mathbf{v} be the vector of drivers' mean future earnings $v_k(\mathbf{s}_t)$. I define \mathcal{X} as the set of all possible beliefs, so that some triple of beliefs $\mathbf{x} = (\mathbf{U}, \mathbf{W}, \mathbf{v})$ belongs to \mathcal{X} .

Suppose the market behaves as described by my model, with a set of parameters θ (which includes all the parameters from demand, supply, and matching) and under a certain pricing policy P . Let $f^P(\cdot, \theta) : \mathcal{X} \rightarrow \mathcal{X}$ be the function that maps a vector of beliefs \mathbf{x} into the vector of beliefs that is equal to market averages, given that agents behave according to \mathbf{x} . A market equilibrium for parameters θ and pricing policy P is characterized by a vector of beliefs $\mathbf{x}^* \in \mathcal{X}$ that is a fixed point of $f^P(\cdot, \theta)$. In other words, it satisfies

$$\mathbf{x}^* = f^P(\mathbf{x}^*, \theta). \quad (6)$$

This means that beliefs must be consistent with market averages. Appendix B proves that an equilibrium exists and that under an additional assumption (which is always satisfied in my simulations), the equilibrium is unique and stable.²²

²²Informally, the assumption is that, as beliefs move from \mathbf{x} to \mathbf{x}' , market averages do not move

4 Descriptive evidence and identification

In this section I give an informal explanation of my empirical strategy and I show the variation in the data from which I identify the main model parameters. In section 5 I explain my exact identification strategy and I justify it formally.

4.1 Short-run demand response

In the trip request model in section 3.1.1, the main parameters I want to identify are $\beta(r_i)$ and $\gamma(r_i)$, which determine how prices and ETAs affect the probability that riders request a trip. The main challenge is that prices and ETAs are endogenous.

Response to prices I estimate riders' response to short-run price changes by exploiting a feature of the surge pricing algorithm, whose main steps are:

1. After analyzing supply and demand in the whole city, the algorithm computes a continuous *recommended multiplier* \tilde{m}_{it} for each location for the next period.
2. Recommended multipliers are rounded to the nearest tenth (or to 1 if <1.15).
3. Rounded multipliers are smoothed out in space and time.²³
4. Smoothed multipliers are rounded to the nearest tenth (or to 1 if <1.15). The outcome m_{it} is the *surge multiplier*.

The final price shown to riders is $p_i = b + m_{it}(\bar{p}_i - b)$, where \bar{p}_i is the *unsurged fare*, the price for the trip if the multiplier was one, and b is a \$2.30 *booking fee*.

Steps 2 and 4 provide a source of exogenous variation from which I identify the effect of prices on demand. Since the final surge multiplier is a deterministic function of the vector of recommended multipliers $\tilde{\mathbf{m}}_t$, any correlation between demand shocks and prices comes through $\tilde{\mathbf{m}}_t$. Once I control for recommended multipliers, the residual variation, which arises solely from rounding, is uncorrelated with demand shocks.

Cohen et al. (2016) also exploit rounding to estimate Uber demand, using a RDD.²⁴ I rely on stronger assumptions (which I state in section 5) that me to capture

too far in the direction $\mathbf{x}' - \mathbf{x}$. One would expect averages to move in the *opposite* direction: if beliefs on earnings increase, for instance, the number of drivers increases, decreasing average earnings.

²³Temporal smoothing takes place by not allowing multipliers to change too abruptly. Spatial smoothing takes place by computing a weighted sum of nearby multipliers.

²⁴There is no spatial smoothing in their data, allowing them to set up a clean RDD. A big limitation of their data is that Uber did not record the destination chosen by riders who did not request a trip.

variation beyond the immediate neighborhood of the discontinuities; otherwise I would not have enough power to estimate the main demand model.

Rounding generates small variation in prices. This might seem problematic, given that my goal is to estimate agents' response to larger changes in prices that are generated by changes in pricing policies. However, I observe variation at different price levels: sometimes the multiplier is rounded to 1 or 1.2, but it is also often rounded to 2 or 2.3 (see in figure 4). My estimation procedure in practice chains together all these small responses to find the demand response to large price changes without relying on extrapolation.²⁵ In section 5.1.1 I show that this procedure is valid if (a) the demand response is continuous and (b) the expectation of demand shocks conditional on recommended multipliers is a continuous function.²⁶

Response to pickup times The ETA that passengers observe depends on which drivers are nearby. The number of nearby drivers is correlated with demand shocks—if more people request trips, the number of available drivers goes down, increasing ETAs—and, thus, it is endogenous. But drivers' exact location relative to riders is unlikely to be correlated with any demand shocks. For instance, two riders may be half a block apart. A driver can pick up one of them right away but has to go around the block to pick up the second one. The first rider could see an ETA of three minutes while the second one sees an ETA of one minute.

I use this variation in ETAs to estimate the response of riders to pickup times. I include fixed effects so that I only use variation within location by time period. In essence, I compare pairs of riders who are in the same location and in the exact same time period. Any systematic demand shocks that might cause endogeneity would affect both riders equally, but fixed effects would clean them out.

Residuals and value of time Figure 7 shows the main variation in the data from which I identify riders' response to prices and ETAs, using the identification strategy I described. I also show linear fits that emphasize the main trends. These are not the actual residuals from my model. Instead, they are simply meant to give a sense of the patterns in the data that give rise to my demand coefficients.

In figure 7a, the horizontal axis represents the variation in multipliers that remains after controlling for the *unrounded multiplier*, an estimate of what the multi-

²⁵Cohen et al. (2016) use this idea to build a demand curve and compute rider surplus.

²⁶To be precise, section 5.1.1 shows that this procedure is valid for one particular continuous demand response, but the same argument holds for an arbitrary continuous demand response.

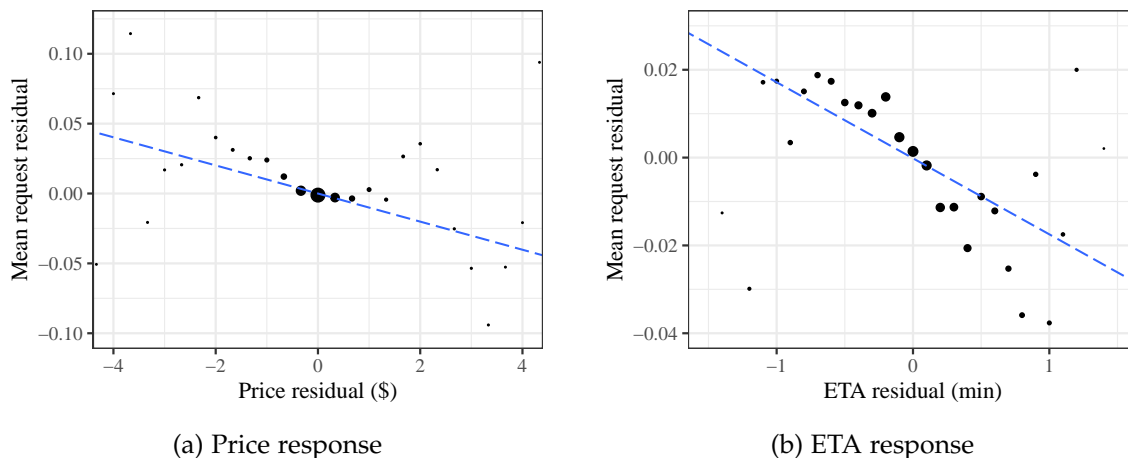


Figure 7: Residuals used to identify the response of riders to prices and ETAs

Note: In both figures, the variable in the vertical axis is the residual of a dummy for trip request. Points represent the mean of this residual by bin of the horizontal axis variable, while point size represents the number of observations. Subfigure (a) shows residuals from regressions on the unrounded multiplier—i.e, what the multiplier would be if there was no rounding in the surge pricing algorithm. Subfigure (b) presents residuals from regressions on location by period fixed effects; I omit observations between 11 pm and 7 am, which are unusually likely to have large ETA deviations and show a small demand response.

plier would have been if there was no rounding in the surge pricing algorithm. I explain how I compute it in appendix D.1. Observations to the right thus represent times when the multiplier was rounded up, whereas observations to the left take place when the multiplier was rounded down. The downward pattern means that an increase in surge multipliers decreases the probability of trip request.

Figure 7b shows within location by time period variation in ETAs and trip request dummies. There is also a downward pattern, indicating that a higher ETA decreases the probability of requesting a trip.

Comparing the slopes in subfigures 7a and 7b gives a sense of how riders trade off higher multipliers and pickup times. The ETA coefficient divided by the price coefficient is a rough measure of the value of time for riders. Following this idea, figure 8 shows the average value of time for different subsamples of the data. I compute it based on linear regressions of trip request on prices and ETAs, where I control for the unrounded price and average ETA by location by time period to get causal estimates. I include location by hour of the week fixed effects, and I allow heterogeneous coefficients by trip distance.

The value of time for the whole dataset is around \$2 per minute. One possible explanation for this high value is that many riders request trips during time sensitive moments. They might want to be in time for an appointment, or they simply

might not want someone they are meeting to wait. In the extreme case, they do not want to miss a flight when they go to the airport. The figure shows significant variation in the value of time for different subsamples: it is higher during the week than during weekends, and it is especially high for airport trips.

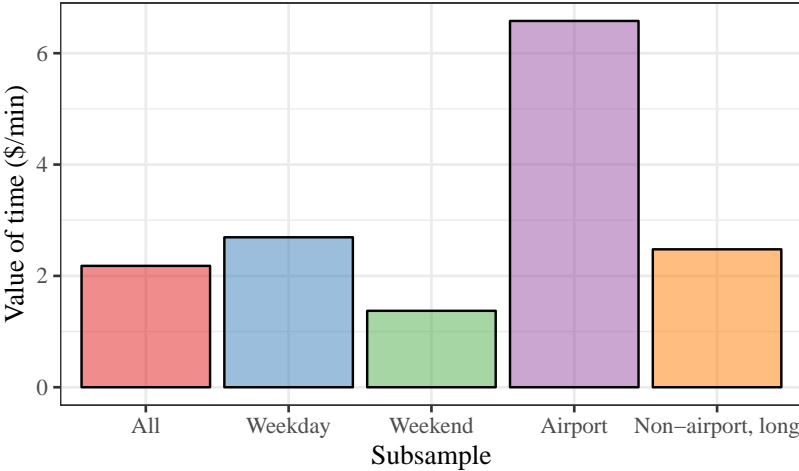


Figure 8: Value of time for subsamples of the data

Note: Each bar represents the average value of time for a subsample of the data. For each subsample, I estimate a linear regression of trip request on price and ETA. I control for the unrounded price and for the average ETA by location by time period. I allow all coefficients to vary linearly in the base fare, which is a proxy for trip distance. I also include location by hour of the week fixed effects. For each observation, I compute the value of time as the ETA coefficient divided by the price coefficient.

These values are similar to those implied by the elasticities that Cohen et al. (2016) estimate for Uber riders using discontinuities in prices and ETAs. Buchholz et al. (2019) estimate values for riders in Prague that, as a multiple of average wages, are roughly one third of the values I estimate.²⁷

My estimates could also be driven by behavioral effects: riders might overreact to ETAs beyond their true value of time. I explain in my welfare analysis (section 6) that my main qualitative findings still hold if riders’ true value of time is lower.

4.2 Short-run supply response

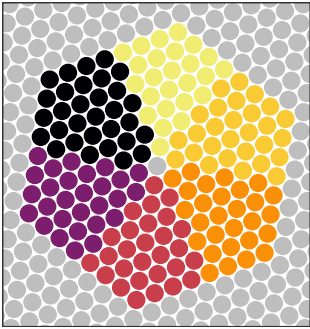
In the model of drivers’ movements from section 3.2.1, the main parameter I want to estimate is δ : the extent to which drivers move towards areas with high mean earnings. This is challenging because changes in mean earnings are mainly driven by surge multipliers, which are endogenous. If some shock induces drivers to move

²⁷They find average values between \$0.20 and \$0.30 per minute. The average wage in Prague was Kc 41,851 per month (around \$10 per hour). The average wage in Houston was around \$26 per hour.

towards a certain area—for instance, if they expect an event to end—higher supply induces lower multipliers.

Just as when estimating demand, I exploit the exogenous variation that arises from rounding in the surge pricing algorithm to estimate drivers’ response. In order to measure broad patterns in the data, I aggregate the space surrounding every available driver into six direction cones, as in the figure to the left of table 1. I then run six regressions. In each one of them, the outcome variable is a dummy for whether the driver moved to one of these cones. I regress the dummy on the average multiplier in every one of the six cones. To obtain a causal estimate, I control for the average recommended multiplier in each cone; thus, my estimates are identified from variation that arises from rounding. I also control for the multiplier in the driver’s location as well as for location by hour of the week fixed effects.

Table 1: Effect of multipliers on movement direction



	<i>Dependent variable:</i>					
	Dummy for moving to cone in parentheses					
	(1)	(2)	(3)	(4)	(5)	(6)
Avg. multiplier in cone 1	0.22*** (0.07)	-0.04 (0.07)	-0.16*** (0.06)	0.07 (0.07)	-0.04 (0.06)	-0.04 (0.06)
Avg. multiplier in cone 2	-0.04 (0.07)	0.29*** (0.08)	0.08 (0.06)	-0.18** (0.07)	-0.07 (0.07)	-0.07 (0.07)
Avg. multiplier in cone 3	-0.05 (0.07)	-0.09 (0.08)	0.16** (0.07)	-0.04 (0.08)	-0.01 (0.07)	0.04 (0.07)
Avg. multiplier in cone 4	-0.09 (0.06)	-0.05 (0.07)	-0.05 (0.07)	0.37*** (0.09)	-0.06 (0.08)	-0.12* (0.07)
Avg. multiplier in cone 5	-0.08 (0.06)	0.04 (0.07)	-0.05 (0.06)	-0.15* (0.08)	0.22*** (0.08)	0.01 (0.07)
Avg. multiplier in cone 6	0.02 (0.06)	-0.13* (0.07)	0.03 (0.05)	-0.07 (0.07)	-0.04 (0.07)	0.19*** (0.07)
Observations	645,133	645,133	645,133	645,133	645,133	645,133

Note: *p<0.1; **p<0.05; ***p<0.01

Note: The figure on the left shows how I aggregate the space surrounding a driver into six location cones. In the table on the right, each column reports estimates from a regression of a dummy for whether a driver moved to one of the cones on the average multiplier in every cone. I control for the average recommended multiplier by cone, for the multiplier in the driver’s location, and for location by hour of the week fixed effects. Standard errors are clustered by location and hour of the week.

Table 1 shows the estimates from these regressions. The estimates on the diagonal are all positive and significant. Off-diagonal terms are noisier, but they tend to be negative. Thus, as multipliers increase in one cone, drivers are more likely to move towards that cone and less likely to move to the other five cones. This is

evidence that drivers tend to move towards areas that have high multipliers.

4.3 Long-run response

Two parameters in my model determine the long run elasticities of demand and supply (σ and ρ). My data from Houston was all generated under the same surge pricing policy, so I do not observe any variation I can use to identify agents' long-run response. Instead, I use data from one-week experiments that Uber ran in 2017 in six Latin American cities (Belo Horizonte, Guadalajara, Mexico City, Rio de Janeiro, and São Paulo) with the purpose of estimating long-run elasticities.

On the demand side, riders were split into a control group and two treatment groups. Treated riders got 10% or 20% discounts for every trip they took during the experiment week. On the supply side, riders were split into a control group and a treatment group that got 10% higher earnings for every trip they made during the week. The experiment ran from Monday to Sunday, and treated agents were notified on Sunday before the experiment started.

I measure the elasticity of demand with a Poisson regression of the number of trips taken by each rider on the log price factor ($\log(1)$ for the control group, $\log(0.9)$ and $\log(0.8)$ for treatment groups). To measure the elasticity of supply, I run a Poisson regression of a dummy for working on a given day on the log earnings factor ($\log(1)$ for the control group, $\log(1.1)$ for the treatment group).²⁸ Table 2 shows the estimates from these regressions.²⁹ Appendix G.6 shows some additional results, including average treatment effects by city.

There are some potential problems with these experiments. First, they only measure the response of people who had an Uber account; they do not measure the effect of new riders and drivers. Second, they only measure the response during one week; thus, they do not measure the response, for instance, of people who decide to sell their car. This could be tackled with longer run experiments, but that would be prohibitively costly.

Third, these elasticities were not measured in Houston. Uber had at least 80% of the market share in all these cities at the time of the experiment, alleviating the concern that cities with alternative ride-hailing apps should have higher elasticities. There remain, however, many other ways in which cities differ. They are located in

²⁸This regression measures whether a driver works in a given day, but not whether he works more hours. Appendix G.6 shows that a regression of hours worked yields almost identical estimates.

²⁹I exclude Mexico City from the demand sample because of an unexpected pattern: the 10% discount group suggests a positive demand slope. See the details in appendix G.6.

Table 2: Long-run elasticities

(a) Demand		(b) Supply	
<i>Dependent variable:</i>		<i>Dependent variable:</i>	
Trips per rider		Worked dummy	
Log price factor	−0.633*** (0.059)	Log earnings factor	0.383*** (0.099)
City FE	✓	City × week day FE	✓
Observations	177,349	Observations	266,000
<i>Note:</i>	*p<0.1; **p<0.05; ***p<0.01 Robust s.e.	<i>Note:</i>	*p<0.1; **p<0.05; ***p<0.01 Robust s.e. clustered by driver

Note: Estimates of the long-run elasticities of demand and supply based on experimental data. For demand, I estimate a Poisson regression for the number of trips requested by driver on the log of the price factor—e.g., $\log(0.9)$ if the rider was in the treatment group with a 10% discount. For supply, I estimate a Poisson regression for whether the driver worked during every day on the log of the earnings factor—e.g., $\log(1.1)$ if the driver was in the treatment group with 10% higher earnings.

different countries, for instance, and the availability of public transportation differs. Because of these concerns, I run my counterfactuals using a large range of values for ρ to check for robustness. I find that the magnitude of the welfare effects change, but the main qualitative results remain the same (appendix F.1).

5 Estimation and results

5.1 Demand

5.1.1 Trip request

My goal is to estimate equation (1). My identification strategy follows the ideas from section 4.1: I identify the price response from rounding in surge multipliers and the ETA response from within location by time period variation.

I estimate the following equation:

$$U_i = \alpha(r_i, l, h) + \beta(r_i)p_i + \gamma(r_i)w_i + g(\tilde{\mathbf{m}}_t, \bar{p}_i, w_{it}^0; l, h) + \eta_i \quad (7)$$

where $w_{it}^0 = E[w_i|l, t]$ is the expected ETA at the location by period level. Relative to the original utility specification (1), this equation decomposes the error as $\epsilon_i = g(\tilde{\mathbf{m}}_t, \bar{p}_i, w_{it}^0; l, h) + \eta_i$, where $g(\tilde{\mathbf{m}}_t, \bar{p}_i, w_{it}^0; l, h)$ is a flexible control function that depends on the vector of all recommended multipliers, the base fare, the expected ETA, the location, and the hour of the week.

I assume that the error η_i is orthogonal. That is true as long as the control func-

tion captures all the correlation between ϵ_i and covariates, which can be justified intuitively. First, the fare p_i is fully determined by (i) recommended multipliers, (ii) the base fare, and (iii) rounding in the surge algorithm. Thus, if one controls flexibly for recommended multipliers and the base fare, the only variation in prices that remains comes from exogenous rounding. Second, if one controls for w_{it}^0 , only the variation in w_i within location by period remains, which, I have argued, is exogenous.

This argument relies on a correct specification of the control function. In other words, it must be flexible enough to capture the true dependence. In appendix C, I show formally that if the control function is correctly specified, and if the following assumptions (which I state formally in the appendix) are satisfied, then the error η_i is indeed orthogonal and the price coefficient is identified:

1. The fare p_i is a deterministic function of recommended multipliers and the base fare that has at least one discontinuity in $\tilde{\mathbf{m}}_t$.
2. Variation in ETAs within location by period—i.e., $(w_i - w_{it}^0)$ —is orthogonal to demand shocks, base fares, and recommended multipliers.
3. Demand shocks and recommended multipliers are related smoothly.

The first assumption is a property of the surge pricing algorithm. The discontinuities provide the variation I use to identify the price response. The second assumption states that within location by period, variation is exogenous. The third assumption is necessary to identify price coefficients. If it was not true, it would not be possible to tell apart discrete changes in recommended multipliers and rounding in the surge multiplier.³⁰ It is true because the surge pricing algorithm computes recommended multipliers as smooth functions of market observables.

I now describe the functional form I specify for $g(\cdot)$. First, I estimate a model based on high dimensional splines of recommended multipliers to predict the *unrounded multiplier* \hat{m}_{it} , which is what the surge multiplier would have been if there was no rounding. My fit depends on all multipliers within the closest seven rings of hexagons that surround location i . I allow the weights given to locations in every one of the seven rings to vary by location. The full details of my specification are presented in appendix D.1. Then I compute the *unrounded price*

³⁰Part 3 plays the role of the main RDD assumption, which is that the conditional mean of each treatment group must be continuous in the forcing variable. It is stronger in that (a) the functional form for conditional means of different treatment groups must be the same and (b) $g(\cdot)$ must be correctly specified. Stronger assumptions allow me to use variation beyond a small threshold around the discontinuity.

$\hat{p}_i = b_i + \hat{m}_{it}(\bar{p}_i - b_i)$, which is the fare the rider would have seen without rounding.

I specify $g(\tilde{\mathbf{m}}_t, \bar{p}_i, w_{it}^0; l, h) = \tilde{\alpha}(\bar{p}_i; l, h) + g^1(\hat{p}_{it}) + g^2(\tilde{\mathbf{m}}_t) + g^3(w_{it}^0)$, where the terms $g^i(\cdot)$ are high dimensional splines.^{31,32} With this functional form, the price coefficient is identified from variation in $p_i - \hat{p}_i$, which arises from rounding—as long as my model for \hat{p}_i is correct—and the ETA coefficient is identified from variation in $w_i - w_{it}^0$. I also include $g^2(\tilde{\mathbf{m}}_t)$ to control for any variation that was not captured by my model for \hat{p}_i .

I assume that $\beta(r_i)$ and $\gamma(r_i)$ depend on r_i linearly. I set $r_i = \bar{p}_i$: I use the unsurged fare as a measure of traffic-adjusted trip distance. The term $\tilde{\alpha}(\bar{p}_i, l, h)$ is then absorbed by the original intercept term $\alpha(r_i; l, h)$, which I assume is additively separable into a cubic function of distance and a function of (l, h) . The latter should be as flexible as possible to capture broad demand patterns. I cannot use fixed effects at the location by hour of the week level: it would result in 1025×168 parameters, which is of the order of magnitude of the number of observations. Instead, I include two flexible, high order splines that account for variation somewhat more smoothly, with 155 degrees of freedom.³³

To estimate equation (15), I plug in \bar{w}_{it} for w_{it}^0 .³⁴ I assume that the error η_i is distributed iid logistic. My model is thus simply a logit model with a large number of covariates, which I estimate by maximum likelihood.

Results Table 3 shows estimates for the main demand parameters. The price and ETA coefficients are negative, as expected. The average over all drivers of the value of time—the ETA coefficient divided by the price coefficient—is \$1.84 per minute, which is consistent with the analysis in section 4.1.

Figure 9a shows how the price coefficient, the ETA coefficient, and the value of

³¹All functions are cubic splines with knots placed evenly at quantiles of the distribution of the variable the spline depends on. g^1 is of order 8, g^2 is of order 5, and g^3 is of order 5.

³²In practice, I find that neither omitting $g^2(\cdot)$ nor increasing the order of the splines has any noticeable impact on my results.

³³I include a tensor product spline of latitude and longitude with five degrees of freedom on each coordinate, interacted with a quadratic function of how busy the hour of the week is. I also interact a sixth order spline of the hour of the day with a function that behaves linearly from Monday to Friday, as well as dummies for Saturday and Sunday, all of which I interact with a quadratic function of how busy the location is. In appendix G.7, I compare linear models using this methodology and fixed effects to show that they both lead to very similar estimates.

³⁴This could be problematic. The number of observations in each location by time period group is small, so I cannot rely on asymptotics to state that \bar{w}_{it} converges to w_{it}^0 . In appendix G.1, I present an alternative specification that avoids this problem, but assumes that w_{it}^0 are uncorrelated with covariates. I find very similar parameter estimates with both models.

time vary with the base fare. As the base fare increases, riders are less responsive to both prices and ETAs: people who want to go far away do not care as much about one dollar or one minute. The price coefficient approaches zero faster than the ETA coefficient, so the value of time is higher for riders who want to go far away.

Table 3: Estimates of demand parameters

	<i>Dependent variable:</i>
	Request
Base fare	-0.0253*** (0.0027)
Fare	-0.0471*** (0.0141)
Fare \times base fare	0.0010 (0.0007)
ETA	-0.0847*** (0.0135)
ETA \times base fare	0.0004 (0.0011)
Observations	650,233
<i>Note:</i>	* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

Note: Estimates of the parameters of the main demand model (equation (7)), which measures how the probability that a rider requests a trip responds to changes in prices and ETAs. The price and ETA coefficients are evaluated at the mean of the base fare.

Figure 9b shows how the price elasticity varies with the base fare. It is very low for short trips, which account for the majority of the observations. It increases until an elasticity is reached of around 0.35 for trips around \$30, which is the typical fare for trips from the city center to airports. Thus, the price coefficient decreases in absolute value as the base fare increases, but not so fast that demand becomes less elastic. This has an intuitive interpretation: people respond more to a 1% price change in a \$30 trip than in a \$5 trip, but they respond more to a \$1 change in a \$5 trip than in a \$30 trip.

The average price elasticity is 0.17. This value is low, but it should not be surprising because it is a very short run elasticity. Furthermore, Houston has few alternative transportation options: there is no competing ride-hailing app, and public transit is limited. One useful benchmark is the estimates from Cohen et al. (2016). They find price elasticities around 0.45. Their estimates are probably higher because they analyze cities like New York, San Francisco, and Chicago, all of which have good transportation alternatives.

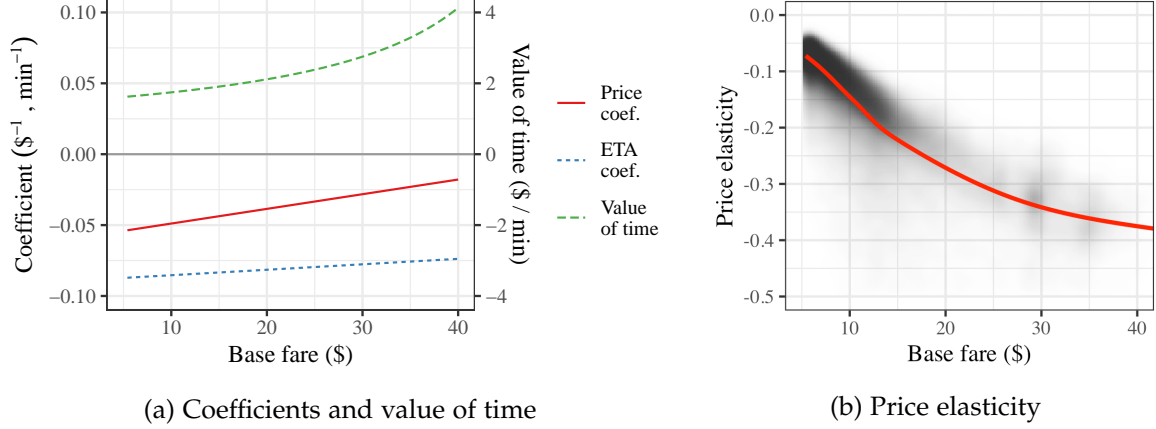


Figure 9: Coefficients, value of time, and price elasticity from the main demand model

Note: Subfigure (a) plots estimates for the price coefficient, the ETA coefficient, and the value of time as a function of the base fare. Subfigure (b) shows implied price elasticities as a function of the base fare. Every observation represents one rider. Tones of grey represent a two-dimensional kernel density. The red line is a nonparametric fit.

5.1.2 Opening the app

The main parameter of the rider arrival model in section 3.1.2 is ρ , which measures the long run elasticity of demand. I now explain how it set its value.

In practice, the number of combinations of distance group, location, and hour of the week (\tilde{r}, l, h) is too large to estimate $U(\tilde{r}, l, h)$ consistently. Thus, I aggregate the data into larger groups. I define \tilde{r} as five distance quintiles, I aggregate locations into 32 zones a with similar number of arrivals, and I aggregate hours of the week into 14 groups g .³⁵ I set $U(\tilde{r}, l, h) = \tilde{U}(\tilde{r}, a(l), g(h))$, where $a(l)$ is the zone that contains l , $g(h)$ is the hour group that contains h , and $\tilde{U}(\tilde{r}, a, g)$ denotes average utility by distance group, zone, and hour group.

I model $A_{\tilde{r}hl}$ based on the decomposition $A_{\tilde{r}hl} = \psi^d \tilde{A}_{\tilde{r}a(l)g(h)} \chi_h^{g(h)} \chi_l^{a(l)}$. The term ψ^d is a uniform scale factor for all demand. $\tilde{A}_{\tilde{r}a(l)g(h)}$ is a demand shifter at the distance group by zone by hour group level. $\chi_h^{g(h)}$ is a factor that captures hourly patterns; I set it to be equal to the fraction of arrivals during hour group g that take place during hour h . Finally, $\chi_l^{a(l)}$ allows me to model spatial patterns precisely. It is equal to the fraction of arrivals to zone z that take place in location l .

³⁵The groups are: early morning weekday (7-9 am), late morning weekday (9-11 am), midday weekday (11 am-1 pm), early afternoon weekday (1-4 pm), mid afternoon Mo-Thu (4-6 pm), late afternoon Mo-Thu (6-8 pm), early evening Mo-Thu (8-10 pm), late evening Mo-Thu (10 pm-1 am), Friday afternoon (4-8 pm), evening Fridays and Saturdays (8 pm-12 am), bar hours Fridays and Saturdays (12 am-3 am), Saturday and Sunday morning (9 am-2 pm), and Saturday and Sunday afternoon (2-8 pm). All remaining hours are off-peak hours.

I set the values of \tilde{A}_{rag} and ρ jointly so that (a) for every (\tilde{r}, a, g) , the total arrival rate is equal to the average arrival rate in the data given the $\tilde{U}(\tilde{r}, a, g)$ I compute from the data and (b) the market-wide elasticity of demand is equal to the value of -0.633 from table 2. This system is exactly identified. It results in a value of $\rho = 1.72$.

I set ψ^d and a similar uniform scale factor ψ^s for supply at the values such that simulations of the market equilibrium in the status quo result jointly in the same number of trips and the average surge multiplier that I observe in the data. This results in a factor $\psi^d = 1.034$. This scale factor is different from one because of some complications in the data that I do not account for in my model. For instance, a small percentage of trips are requested from Google Maps or other external apps, in which case the rider is not in my demand dataset.

5.2 Supply

5.2.1 Movement

I now explain how I estimate the parameters from the movement model from section 3.2.1: $\omega(l, k, h)$ and δ in equation (4). As with my trip request estimation, I identify δ only from variation that arises from rounding in surge multipliers.

The model I estimate is

$$\Pr(l_{j,t+1} = k | \mathbf{s}_t) = \frac{\exp(\omega(l, k, h) + \delta v_k(\mathbf{s}_t) + g^M(\tilde{\mathbf{m}}_{tk}; l, h))}{\sum_{k'} \exp(\omega(l, k', h) + \delta v_{k'}(\mathbf{s}_t) + g^M(\tilde{\mathbf{m}}_{tk}; l, h))}. \quad (8)$$

This amounts to assuming that the unobserved term ζ_{kt} from equation (4) is equal to $g^M(\tilde{\mathbf{m}}_{tk}; l, h)$, which is a function of the recommended multipliers surrounding location k .

This allows me to estimate δ based on variation that arises from rounding. The intuition is the same as in the trip request model. Mean future earnings $v_k(\mathbf{s}_t)$ are a function of multipliers, which, in turn, are a function of recommended multipliers. If g^M is flexible enough, it captures all of the variation in $v_k(\mathbf{s}_t)$ that arises from recommended multipliers. Thus, the residual variation that identifies δ comes solely from rounding and is arguably exogenous.

In appendix C I justify this procedure formally under the assumption that the movement rule arises from a latent variable model. My argument relies on g^M being specified correctly, on the errors in the latent variables having an extreme value type I distribution, as well as on the following two assumptions, which are entirely analogous to the assumptions used for the trip request model:

1. Mean future earnings are a function of recommended multipliers that have at least one discontinuity.
2. Supply shocks and recommended multipliers are related smoothly.

Part 1 is necessary because it provides the variation I need to identify δ . It can be justified by noting, first, that surge multipliers are a function of recommended multipliers with discontinuities, and, second, that those discontinuities should induce discrete changes in mean future earnings. Part 2 is analogous to the assumption of smooth conditional means in RDDs. This is a reasonable assumption because recommended multipliers depend smoothly on market observables.

The functional form that I use for the control function is $g^M(\tilde{\mathbf{m}}_{tk}; l, h) = \tilde{\omega}(l, h) + \tilde{g}^M(\tilde{\mathbf{m}}_{tk})$. The term $\tilde{\omega}(l, h)$ is then absorbed by $\omega(l, k, h)$. $\tilde{g}^M(\tilde{\mathbf{m}}_{tk})$ is the sum of twelve splines, one for the average unrounded multiplier in each one of the six nearest hexagon rings around the current location, and one for the average recommended multiplier in each one of the six nearest hexagon rings. I thus assume that there is no interaction between terms.

The value of $v_k(\mathbf{s}_t)$ that I use for estimation arises from data averages. Let Π_{jt} be driver j 's realized hourly earnings from time $t + 1$ until $t + \bar{t}$ or the time he leaves, whichever comes earlier.³⁶ I would want to average Π_{jt} by the state and subsequent position to obtain an estimate of $v_k(\mathbf{s}_t)$. However, the state space is so big that there are far more states than observations. Instead, I fit the following model:

$$\Pi_{jt} = \alpha(l', h) + f(\mathbf{m}_{tl'}) + \chi_{jt}, \quad (9)$$

where l' is the direction driver j moved to. The term $\alpha(l', h)$ plays the role of location by hour of the week fixed effects. I use the same specification from the trip request model, which has 155 degrees of freedom.

The term $f(\mathbf{m}_{tl'})$ is a flexible function of $\mathbf{m}_{tl'}$, the vector of multipliers surrounding location l . My goal is not to predict earnings as accurately as possible; to do that I would use a machine learning methodology. Instead, I want to capture an intuitive functional form that models drivers' expectations well.

I use a relatively simple smooth function that is radially symmetric. It includes all multipliers and their squares up to a distance 7 from the current location, a dummy for whether each multiplier is greater than one, and the maximum multiplier at each distance. To simplify the model, I constrain the coefficient for each

³⁶A naive choice would be to discard all drivers that leave before $t + \bar{t}$. This, however, would bias my estimates of the choice-specific value: drivers that are matched quickly are more likely to stay.

one of these terms to be linear in the distance to the current location. I show in appendix D.2 that my fit follows the two main patterns one would expect it to follow: expected earnings increase with surge multipliers, and they increase as more neighboring locations have surge pricing. A surge multiplier of 1.5 in all surrounding hexes, for instance, results in expected earnings that are \$4.85 higher.

I assume driving costs of \$0.26 per mile. This is the internal Uber estimate for the average UberX car in Houston, including fuel, maintenance, repairs, and depreciation.³⁷ For available drivers, I assume they drive 1.6 times the straight-line distance between their origin and destination locations. This approximates the average ratio of driven to straight-line distance when picking up.

I set $\omega(l, k, h) = \zeta_l^k + \chi_{z_l, g_h}^{n_k}$. The term ζ_l^k represents origin by destination fixed effects, and it captures the fact that drivers' movements are defined to a large extent by road patterns. The term $\chi_{z_l, g_h}^{n_k}$ represents fixed effects that model the fact that traffic patterns change over the week. To capture this, $\chi_{z_l, g_h}^{n_k}$ are fixed effects for the cartesian product of 9 zones z_l for the origin location, 15 hour of the week groups g_h , and 19 movement trends n_{lk} that take into account the general direction and distance of moving from l to k .³⁸

If driver j is in location l , I limit the driver's possible destinations to the set K_l of the n_l most frequent movement destinations from l . It includes as many destinations as it is necessary to account for over 98% of the movements that started in l . This results, on average, in 25 possible movement choices for each origin location l .

I estimate my model by maximum likelihood. The optimization is complicated by the large number of fixed effects. Some fixed effects only affect a small number of observations, which might lead to an incidental parameters problem. Some Monte Carlo simulations, however, show that the bias from this problem is negligible.³⁹ The large number of fixed effects also means that maximizing the likelihood function with a standard nonlinear optimization algorithm would need too many iterations iterations because of the large parameter space. To reduce the number of iterations, I divide the problem into an outer loop that maximizes over δ and

³⁷Drivers are covered by insurance paid by Uber while they are working

³⁸The 9 zones are the interaction of three quantiles for latitude and three quantiles for longitude. The hours of the week are the same as in footnote 35. Movement trends are the product of 6 directions according to the hexagonal lattice and three distance groups. The middle group consists of movements between $\frac{5}{6}$ and $\frac{6}{5}$ of the average movement distance by direction and location. An additional group includes drivers who stay in the same location.

³⁹I simulate data assuming my maximum likelihood estimator ($\hat{\delta} = 0.0878$) is correct, and maximize the likelihood based on the simulated dataset. The mean estimate for δ over twenty simulations is 0.0897 with standard deviation 0.006.

$g^M(\tilde{\mathbf{m}}_{tk})$ and an inner loop that maximizes over the fixed effects ζ_l^k and $\lambda_{z_l, g_h}^{n_k}$ (see appendix D.3).

Results The estimate I obtain for the main parameter is $\hat{\delta} = 0.0878$ (s.e. = 0.016, $N = 1,094,729$). To interpret this quantity, consider a driver who is equally likely to move to one of four destinations surrounding him in the next period. If earnings in one of the four location increase by \$3, which roughly corresponds to an increase in surge multipliers from 1 to 1.5 in all surrounding hexes, the probability that the driver goes in that direction increases from 0.25 to 0.302.

A reference point to which one can compare this estimate is the results provided by Lu et al. (2018), who estimate a similar multinomial logit model based on a surge pricing outage that affected Uber drivers using iOS but not those using Android. They find that, on average, the example above would cause an increase in the probability of going towards the location with higher earnings from 0.25 to 0.269. The effect they measure is smaller, but it is of the same order of magnitude as the one I measure.⁴⁰

5.2.2 Entry

The main parameter of the driver arrival model in section 3.2.2 is σ , which measures the long-run elasticity of supply. I now explain how it set its value.

Just as with the arrival of riders, the number of groups (l, h) is too large to allow me to estimate W_{lh} consistently for each group. I aggregate locations into 32 zones s with a similar number of driver entries, and I aggregate hours of the week into the same groups defined in footnote 35. I then set $W_{lh} = \tilde{W}_{s(l)g(h)}$, which is the average hourly earnings for all drivers who start working in the zone to which l belongs $s(l)$ during the hour group to which h belongs $g(h)$.

I model B_{lh} based on the decomposition $B_{lh} = \psi^s \tilde{B}_{s(l)g(h)} \varphi_h^{g(h)} \varphi_l^{s(l)}$. The first term, ψ^s , is a uniform scale factor for all of supply. $\tilde{B}_{s(l)g(h)}$ is a supply shifter at the zone by hour group level. $\varphi_h^{g(h)}$ captures hourly patterns. I set it to be equal to the fraction of drivers who start working during hour group $g(h)$ that do so during hour h . Finally, $\varphi_l^{s(l)}$ captures fine spatial patterns. It is equal to the fraction of drivers who enter to zone z that do so in location l .

⁴⁰Two reasons might explain the differences. First, their model has no fixed effects, so the error term has higher variance than it would if they had included fixed effects. Second, earnings are more correlated than multipliers, so earnings differences across locations are relatively small.

I set σ and all \tilde{B}_{sg} jointly so that: (a) for every (s, g) , the total arrival rate is equal to the average arrival rate in the data given the observed \tilde{W}_{sg} ; and (b) the market-wide elasticity of supply is equal to the value of 0.383 from table 2. This system is exactly identified. It results in $\sigma = 0.246$.

As I explain in section 5.1.2, I set ψ^s and ψ^d at the values such that simulations of the market in the status quo result jointly in the same number of trips and the average surge multiplier that I observe in the data. This results in $\psi^s = 1.274$. This scale factor is different from one because of some complications in the data that I do not account for in my model, such as the fact that some drivers decide to work both for UberX and UberBLACK, UberXL, or UberEATS (a food delivery service).

5.2.3 Exit

I assume that the distribution G_h of the intended shift duration during week hour h is $\text{Gamma}(\alpha_h, \beta_h)$. I estimate the parameters by maximum likelihood. Let \bar{D}_j be the actual shift duration for driver j , and let $t_j^0 + \underline{D}_j$ be the last time that the driver was available but did not leave. The driver must have had an intended exit time between \bar{D}_j and \underline{D}_j . Thus, the likelihood for driver j is given by

$$\mathcal{L}_j(\alpha_h, \beta_h, \bar{D}_j, \underline{D}_j) = F(\bar{D}_j; \alpha_h, \beta_h) - F(\underline{D}_j; \alpha_h, \beta_h). \quad (10)$$

Appendix D.4 compares the overall distribution I fit for D_j with the distributions of \bar{D}_j and \underline{D}_j . The assumption of a Gamma distribution seems well justified.

5.3 Matching technology

I need to estimate two elements in the matching model from section 3.3: (i) the distribution of ETAs $G(\cdot | l_i, l_j, b_j, h)$ for individual rider-driver pairs; and (ii) the matching rates ϕ_a and ϕ_b .

For the distribution of ETAs, I first fit a random forest of all realized pickup times as a function of pickup coordinates, the driver's coordinates at the time pickup started, the time of the day, and the hour of the week.⁴¹ Let $\hat{w}(x_i, x_j, h)$ be the predicted pickup time of this model, where x_i and x_j represent the rider and driver coordinates, respectively. I also fit a linear model of the standard deviation of the

⁴¹The random forest includes transformations of the variables to improve out-of-sample MSE: 45° rotations of coordinates, latitude and longitude differences and 45° rotations, straight-line distance, the ratio between coordinate differences and straight-line distances, and sines and cosines of the hour of the day and day of the week.

residual of this model by bins of the prediction \hat{w} . Let $s\hat{d}(\hat{w})$ be the fit from this model.

Based on these two elements, I follow a three-step process to generate draws from $G(\cdot|l_i, l_j, b_j, h)$. First, I draw x_i from the empirical distribution of all pickup coordinates in location l_i , and I draw x_j from the empirical distribution of all coordinates of available drivers in l_j .⁴² Second, I compute $\hat{w}(x_i, x_j, h)$ from these coordinates and $s\hat{d}(\hat{w})$ from this prediction. Third, I draw the pickup time from a lognormal distribution with mean \hat{w} and standard deviation $s\hat{d}$.

I fit driver acceptance rates by the simulated method of moments. The two moments I match are the average pickup time and the fraction of trips that are assigned to busy drivers. How the moments relate to the parameters is intuitive: higher acceptance rates result in matches with lower pickup times, and higher ϕ_b relative to ϕ_a leads to a higher fraction of trips assigned to busy drivers. I simulate moments from all of the trip requests and available drivers in the data. For parameters (ϕ_b, ϕ_a) , I run my matching model for every period, after which I compute moments.

Results The estimates I obtain are $\hat{\phi}_a = 0.816$ (s.e.=0.037) and $\hat{\phi}_b = 0.104$ (s.e.=0.009). The value for $\hat{\phi}_a$ is close to 0.8, which is the value in some models used internally by Uber. The value of $\hat{\phi}_b$ is low because only a small fraction of trips are assigned to drivers who are dropping off a passenger, despite the fact that the number of drivers that will drop off a passenger is about the same as the number of available drivers.

5.4 Model fit

With the parameter estimates I have described, I can simulate the behavior of the market given agents' beliefs. Using simulations to compute market equilibria is challenging. In appendix B.4 I explain the algorithm I use to compute equilibria and discuss how it solves those challenges.

To show that my model results in a good fit, I now compare the data with results from simulations of the status quo in equilibrium. My model involves spatial heterogeneity at a high resolution: I model trip origin and destination, as well as driver movements, at the hexagon level. Figure 10 shows that these patterns in

⁴²Fitting times directly on the coordinates of the midpoints of locations l_i and l_j creates a selection issue: the riders in l_i and drivers l_j that are matched tend to be those who are close to each other.

the model fit the data well. I focus on specific times of the week that have salient patterns, but I show in appendix E that other times also have a good fit. Appendix E also shows that the simulations fit high-resolution temporal patterns in the data.

Figure 11 shows that the simulated distribution of surge multipliers is close to the one in the data. My model thus captures heterogeneity in a way that results in the right amount of variation in supply-demand imbalances.⁴³

6 Welfare effects

To measure the welfare effects of surge pricing I compare market simulations under different pricing policies. I start by defining the components of welfare.

I compute rider surplus and driver surplus as sums of rider and driver surplus, as defined in section 3, over the whole market:⁴⁴

$$RS = \sum_{lh\tilde{r}} \frac{A_{lh\tilde{r}}}{1 + \rho} U(l, h, \tilde{r})^{\rho+1}, \quad DS = \sum_{lh} \bar{D}_h \frac{B_{lh}}{1 + \sigma} W_{lh}^{\sigma+1}. \quad (11)$$

I compute Uber's short-run profit as

$$\Pi = \sum_n \left((1 - \tau - \nu) p_n - \pi_n - I_n \right). \quad (12)$$

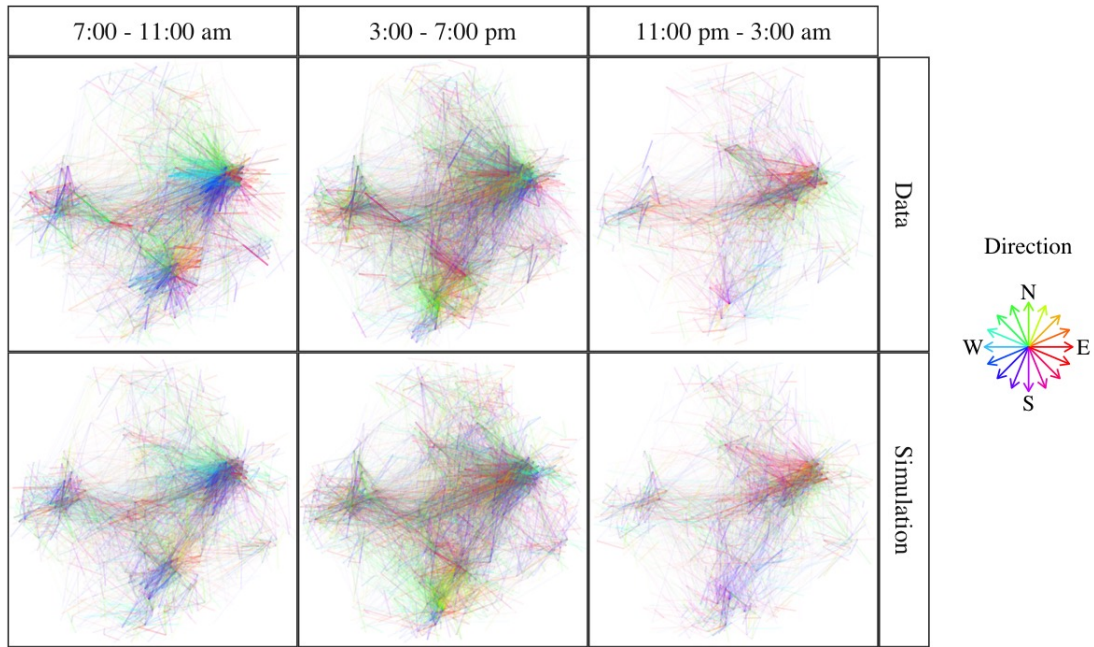
In this sum n indexes requested trips; p_n is the trip fare, π_n is the payment to the driver, and I_n is insurance costs. Uber only gets a fraction $(1 - \tau - \nu)$ of the fare, where τ is a 2% sales tax and ν is a credit card transaction cost that I set at 1%. Uber pays $\pi_n = (1 - \kappa) [(1 - \tau) p_n - b]$ to the driver, where b is a \$2.30 booking fee and κ is the commission rate. In the data, κ varies between 24% and 28% among drivers, depending on how long ago they became Uber drivers. I set it at its average, 26.3%.

Uber pays per-mile insurance whenever a driver is picking up or dropping off a rider. The price they pay is the outcome of bargaining with insurers, and it is a number they are not willing to disclose. In my model, I use a value of \$0.30 per mile, which is somewhat below market rates for private customers in Houston.⁴⁵

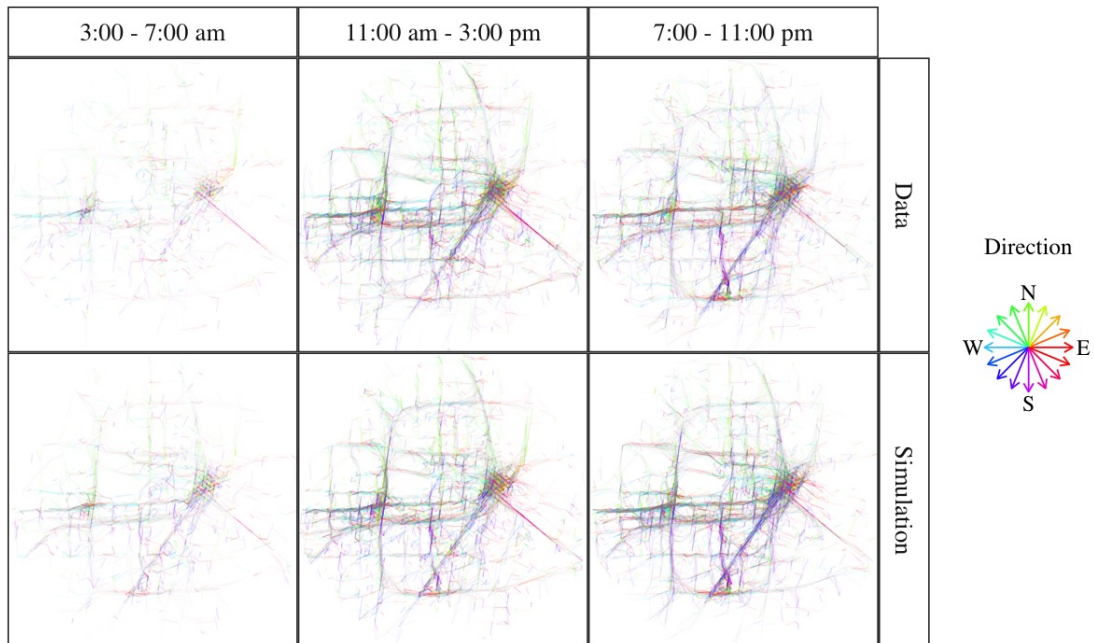
⁴³The mean of both distributions must be the same because I fit model parameters to that moment.

⁴⁴The term inside the sum for driver surplus has a factor of \bar{D}_h , which is the average shift length for drivers who start working during hour of the week h , since $\frac{B_{lh}}{1 + \sigma} W_{lh}^{\sigma+1}$ is surplus per hour.

⁴⁵At this rate, insurance costs are 18% of gross revenue. Uber's financial statements imply a value of only 10%, but that is because insurance costs in the US are much higher than in other markets.



(a) Trips



(b) Driver movements

Figure 10: Trips and driver movements in simulations and in the data

Note: Subfigure (a) shows a 20% random sample of trips at certain hours from Monday to Thursday for one week. Each line connects the origin and destination of a trip. Similarly, subfigure (b) shows a random sample of 5% of the movements of available drivers. Each line connects the initial and final location of an available driver during one period. Colors represent the direction of movement.

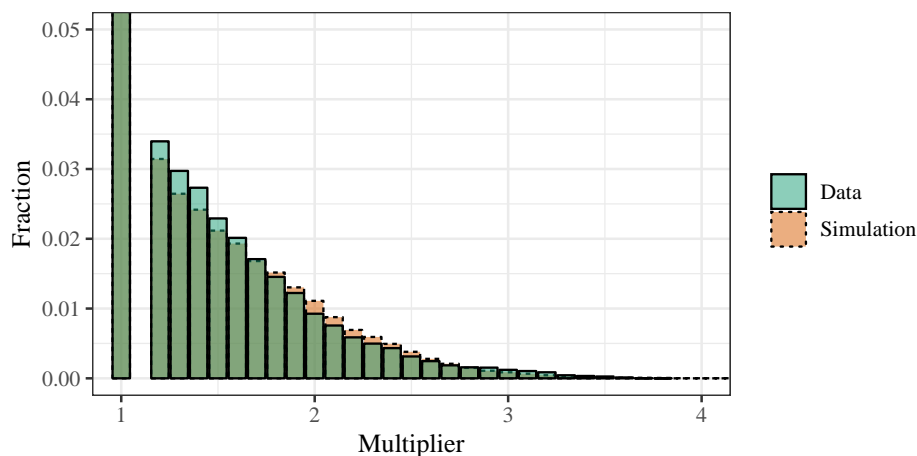


Figure 11: Distribution of surge multipliers in simulations and in the data

Note: Histograms of surge multipliers in simulations and in the data. The fraction of observations with multiplier 1 is 77.5% in the data and 79.2% in simulations.

6.1 Net welfare effect on riders, drivers, and Uber

In figure 12, I compare different surge pricing and uniform pricing policies. The horizontal axis represents the average surge multiplier for the whole market. In the upper left subfigure, the vertical axis represents total welfare—the sum of rider surplus, driver surplus, profit, and tax revenue—relative to the status quo. The point at which both dotted lines cross represents the status quo. The red, solid line represents alternative surge pricing policies in which the multiplier is scaled up or down by a constant scale factor: If under certain market conditions the surge multiplier with the status quo is m_{lt} , when the scale factor is f , the surge multiplier is $f m_{lt}$. Increasing f thus entails an increase in average multipliers.⁴⁶⁴⁷

The blue, dashed line represents policies in which there is a uniform surge multiplier that applies to all times of the week and all locations. The horizontal axis simply represents the level at which the uniform multiplier is set. Thus, moving vertically from a uniform pricing policy towards the surge pricing policy right above it represents a mean preserving spread of the surge multiplier.

Welfare has an inverted-U shape: it is low both with high and low prices. In the case of surge pricing, it is maximized at an average multiplier that is slightly above one. This is not mechanical: Uber in fact prices at a level that is close to welfare maximizing. For every level of the average multiplier, surge pricing results

⁴⁶The average multiplier is not proportional to f because of equilibrium effects.

⁴⁷I could conduct a similar exercise in which, instead of scaling the multiplier up or down, I add or subtract some fixed quantity to the multiplier. This results in almost identical figures.

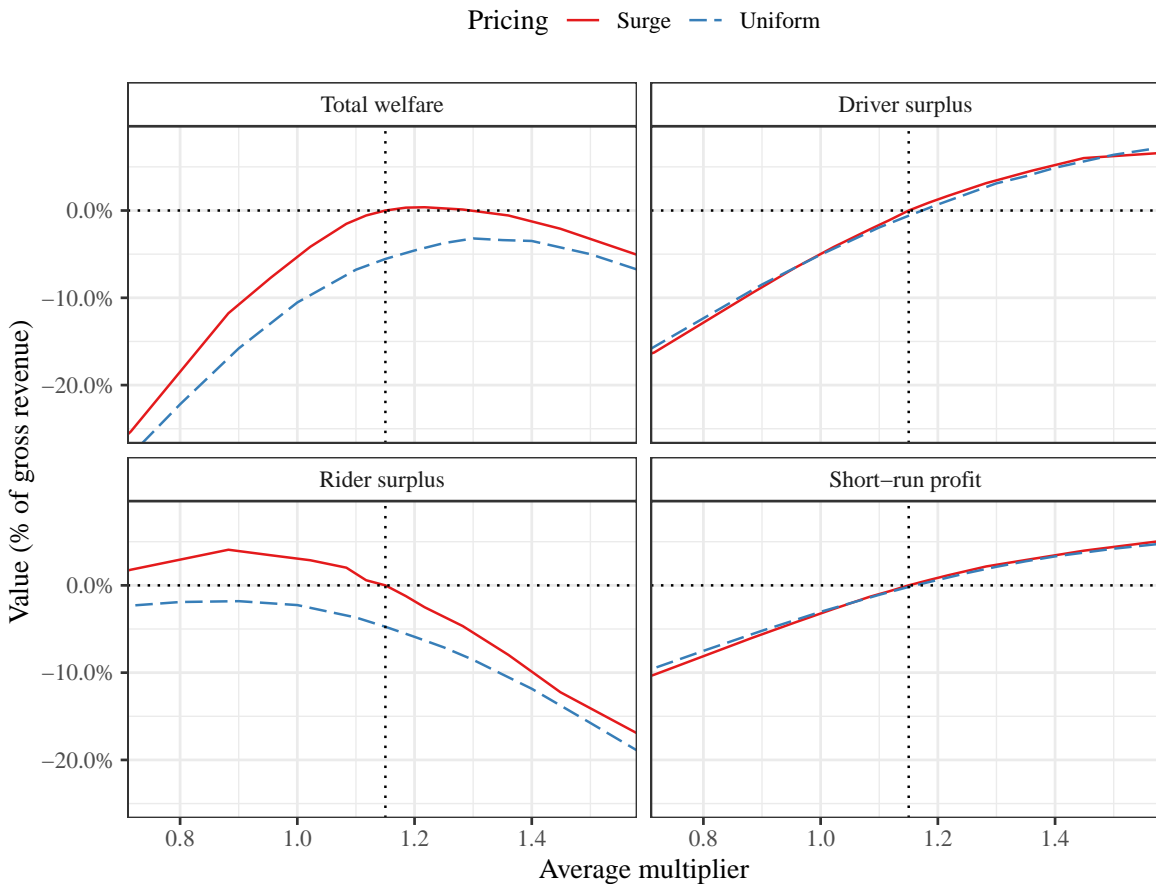


Figure 12: Welfare under different pricing policies

Note: These figures compare welfare and its components (rider and driver surplus, short-run profits) for different pricing policies. The horizontal axis represents the average surge multiplier for the whole market. The vertical dotted line is the average surge multiplier in the status quo. The vertical axis represents total welfare, rider surplus, driver surplus, or short-run profits relative to the status quo. Curves for surge pricing represent different policies in which multipliers are computed just as in the status quo, but then they are scaled up or down by a factor that is constant across the whole market, which leads to different levels of average multipliers. Curves for uniform pricing represents policies that have a unique multiplier for the whole market that is set at different levels.

in higher welfare. This means that there are efficiency gains from surge pricing. The vertical distance is around 5.5% of gross revenue at the average multiplier in the status quo.

The other three subfigures break down welfare into rider surplus, driver surplus, and profit. Remarkably, most of the welfare gap between surge pricing and uniform pricing is accounted for by rider surplus. Driver surplus and profit are also higher for surge pricing, but only slightly. As expected, rider surplus is decreasing, except for low prices, and driver surplus and profit are increasing. Uber prices well below

short-run profit maximization.

There are two reasons for the large difference in the welfare gains of riders and drivers. First, allocative efficiencies only benefit riders. When there are few available drivers, trips are allocated randomly to those riders lucky to be close to a driver, while unlucky riders do not get a trip. Surge pricing largely avoids this by increasing prices, moving from a random allocation mechanism to a price mechanism. While allocative efficiency benefits riders, it has no effect on drivers: every driver’s valuation for a trip is the same—earnings from the trip minus the driving cost. The bars at the left in figure 13a show how large those welfare gains are when the status quo is compared to a uniform multiplier with the same average.

Second, surge pricing also improves welfare through time savings. Riders wait less time to be picked up, and drivers spend less time picking up passengers and waiting to be matched. The bars in figure 13b show how big those time savings are. Drivers save almost four times as much time as riders. However, riders’ valuation of one minute is much higher than drivers’, and so in dollar terms, riders’ time savings are significantly more valuable. The bars at the right in figure 13a show the size of those welfare gains.

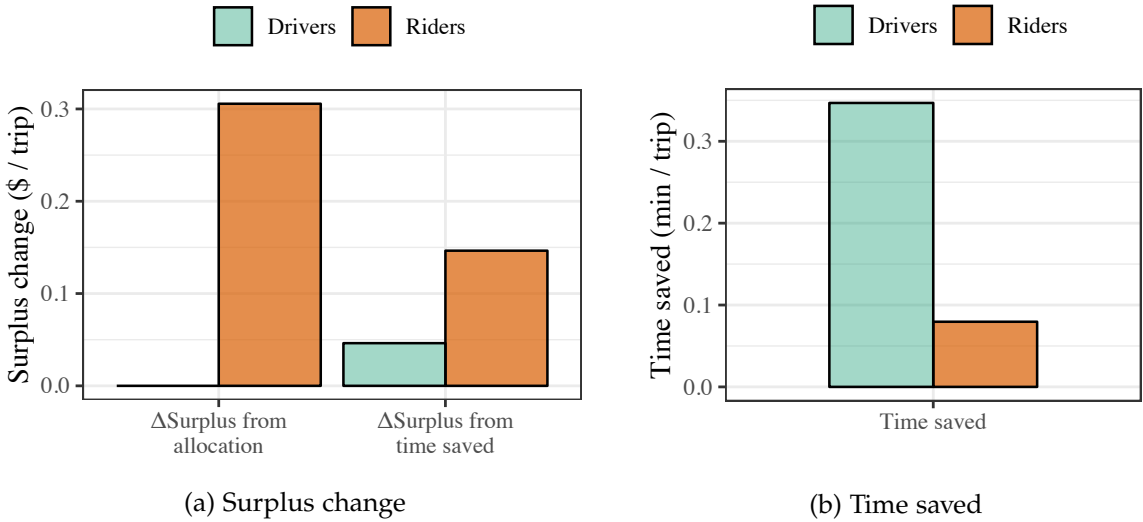


Figure 13: Efficiency gains of surge pricing

Note: These figures break down the efficiency gains of surge pricing, relative to a uniform multiplier at the average level for the status quo. Subfigure (a) decomposes the change in rider and driver surplus into better allocation (lower wasted surplus due to denied trips) and time saved (the decrease in pickup time times riders’ value of time, and the decrease in time between trips times drivers’ average value of time). Subfigure (b) shows how much time riders and drivers save.

On average, surge pricing only saves a few seconds per trip for riders, but the

savings are spread unevenly. The variance of pickup times goes down (subfigure 14a): surge pricing balances the market, making very low and very high pickup times less likely. The most clear effect is that the upper tail of the distribution is cut down (subfigure 14b). At the 95th percentile, pickup times go down by over half a minute, and they go down by over one minute at the 99th percentile.

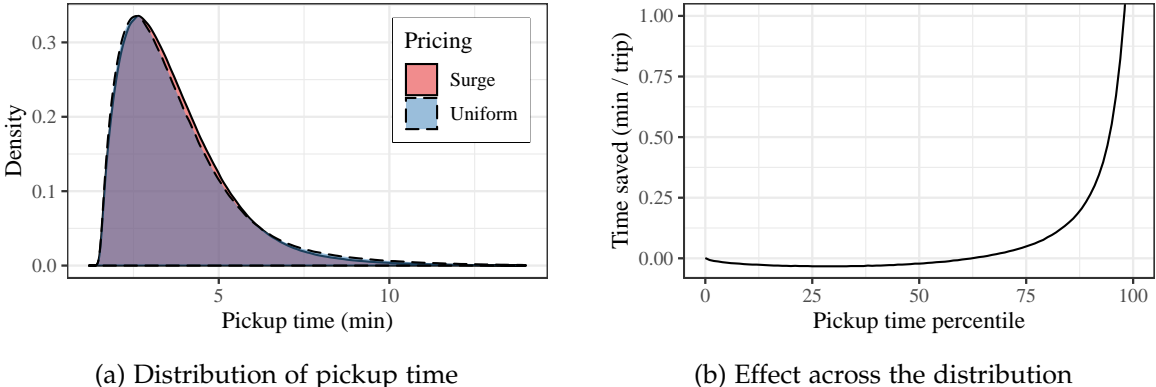


Figure 14: Effect of surge pricing on pickup time

Note: Subfigure (a) shows the distribution of pickup times, both for surge pricing and for a uniform multiplier at the average level for the status quo. Subfigure (b) shows the reduction in pickup times for each percentile of the distribution as the market moves from uniform to surge pricing.

Optimal uniform pricing So far I have analyzed the effect of surge pricing relative to a uniform multiplier at the status quo average. But if, for instance, a regulator constrained Uber to set a uniform multiplier, Uber would reoptimize and set the multiplier at a different level. I assume that to maximize long-run profits Uber maximizes a weighted sum of its short-run profits, rider surplus, and driver surplus. Uber cares about rider and driver surplus because its goal is to maximize investors' value in the long run. Higher rider and driver surplus means satisfied consumers who are likely to return to the platform.

Let \mathcal{P} be the set of policies the platform chooses from. Let $\Pi(P)$, $RS(P)$, and $DS(P)$ be short-run profits, rider surplus, and driver surplus, respectively, with pricing policy $P \in \mathcal{P}$. Then the platform's problem is

$$\max_{P \in \mathcal{P}} \Pi(P) + \alpha^R RS(P) + \alpha^D DS(P). \tag{13}$$

To find weights α^R and α^D , I assume that in the status quo Uber selects a pricing policy by choosing two parameters: the percent commission they take from every trip, and a scale factor f for the surge multiplier. The weights are such that Uber's

first-order conditions for both parameters are equal to zero.

I find welfare weights $\alpha^R = 0.804$ and $\alpha^D = 0.262$: Uber values rider and driver surplus, but not as much as it values its short-run profit. It also cares more about rider surplus than driver surplus. Figure 15 plots Uber’s objective function for the policies analyzed in figure 12. With surge pricing, this function is maximized at the status quo. This should not be surprising: I chose welfare weights based on the first order condition for this function. Uniform pricing leads to a lower value for all average multipliers.

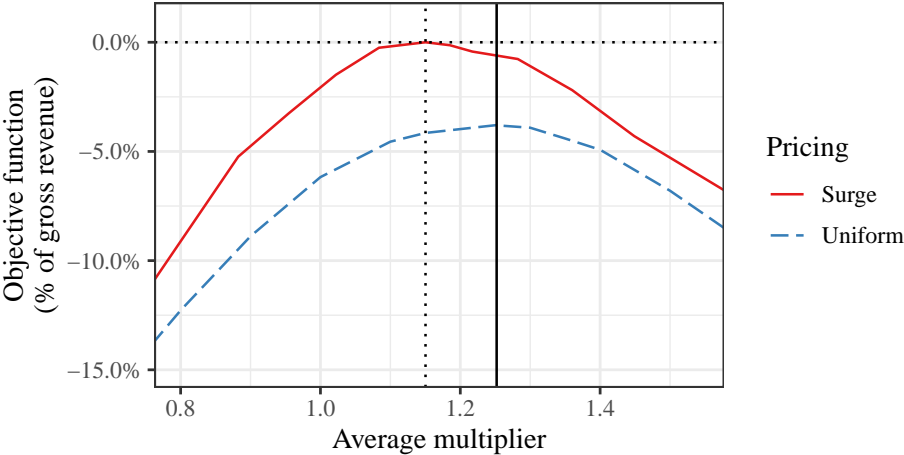


Figure 15: Uber’s objective function under different pricing policies

Note: These figures compare the platform’s objective function and the average multiplier for different pricing policies. The solid vertical line indicates the optimal uniform multiplier. The set of policies I consider are the same as those examined in figure 12.

The most important feature of figure 12 is that the optimum average multiplier is higher for uniform pricing than for surge pricing. A detailed explanation is provided by Castillo et al. (2018). Driver scarcity is bad for riders because they are matched to drivers who are far away. But it is also bad for drivers, who must spend a long time picking up riders. Drivers end up inefficiently using their time picking up riders far away, right when their time is most needed. A negative feedback loop starts, wherein driver scarcity leads to inefficient driver time use, further fueling driver scarcity. Rider surplus, driver surplus, and profit all decrease in a situation Castillo et al. call a *wild-goose chase*.

The implication for the market is that at any given time and place, it is much more painful for the platform to set prices too low than too high. That is less of an issue with surge pricing: the algorithm takes care of avoiding prices that are too low. But with a uniform multiplier, it is optimal to set a high multiplier to avoid wild-

goose chases and a drop in the objective function. Appendix F.3 shows evidence that this phenomenon is the reason why uniform pricing results in a higher optimal multiplier.

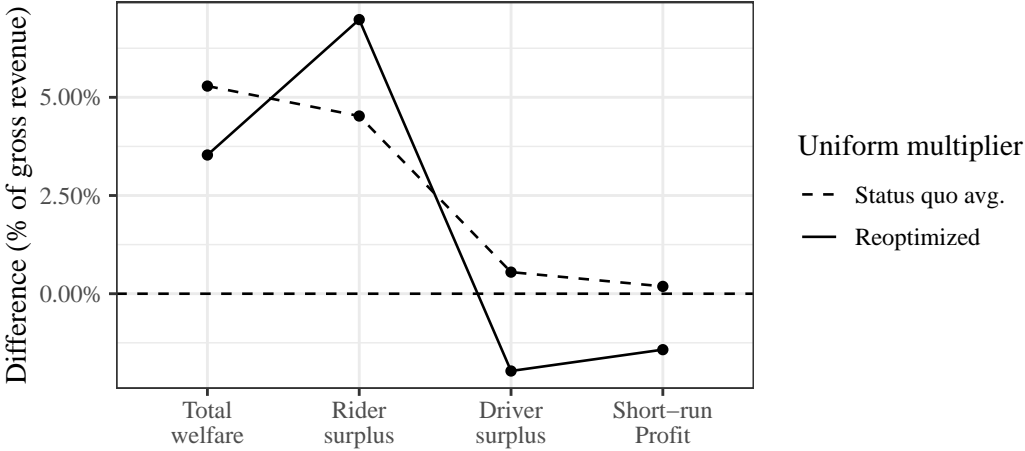


Figure 16: Welfare effect of surge pricing

Note: The dashed line measures the welfare effect from surge pricing relative to a uniform multiplier at the average from the status quo. The solid line measures the welfare effect from surge pricing relative to a uniform multiplier at the level that maximizes Uber’s objective function.

Figure 16 shows the welfare effect of surge pricing on every side of the market. The dashed line compares the status quo to a uniform multiplier at the average of the status quo; thus, it only measures welfare changes due to efficiency gains. The solid line compares the status quo to uniform pricing at the optimum from figure 15. It also takes into account the fact that surge pricing results in lower average prices, and, thus, it transfers welfare from drivers and Uber towards riders. Rider surplus increases with surge pricing, both because of efficiency gains and lower prices. In contrast, there is a net decrease in driver surplus and short-run profits: the decrease from lower prices overtakes the small increase from efficiency gains.

Appendix F.1 shows that these results are robust to higher long-run elasticities. Magnitudes differ, but the main qualitative results still hold. The only change is that, when demand elasticity is high, driver surplus and short-run profits are somewhat higher with surge pricing because the effect from lower prices is not enough to overcome the effect from better matching. I also show that the main qualitative results also hold with versions of the trip request model in which the value of time is lower.

6.2 Welfare effects within drivers and riders

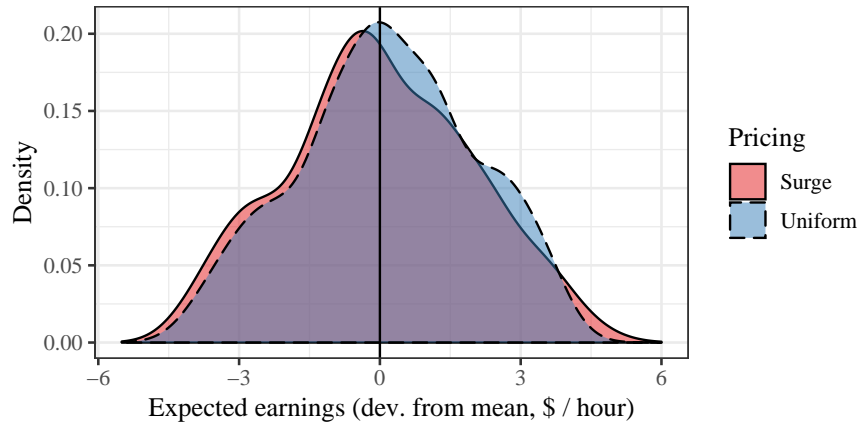


Figure 17: Distribution of drivers' expected earnings

Note: Kernel density plots of drivers' expected hourly earnings by entry location and hour of the week. Each observation represents the average earnings of drivers who started working in that location and time period, regardless of how long they kept on working.

Drivers Figure 17 shows the distribution of drivers' ex-ante hourly earnings, both in the status quo and with uniform pricing at the level that maximizes Uber's objective function. Surge pricing not only reduces average hourly earnings; it also increases the dispersion in earnings, and, as a consequence, the density of drivers that get low earnings is higher. This is perhaps not surprising. With surge pricing, multipliers go down at times of low demand, precisely when hourly earnings are lowest (see appendix F.2). The opposite happens during high-demand times.

Some critics (e.g., Goncharova, 2017) argue that surge pricing is undesirable because it forces drivers to plan their actions carefully around surge pricing. If drivers do not, they get earnings that are too low to cover their costs. My findings suggest these critics might be correct. My results also suggest an explanation for the structure of surge pricing, which has a low multiplier most of the time and higher levels a small fraction of the time. If Uber lowered prices even further at times of low demand, the negative effect on low-earning drivers would be magnified.

Riders Surge pricing might have undesirable distributional effects within riders. Rich, high willingness to pay riders might be better off because they reliably get a trip with a low pickup time—for which they sometimes have to pay a high price—

while poor, low willingness to pay riders are priced out of the market when multipliers are high.

In appendix F.2 I analyze how the effect of surge pricing varies by the price of the rider's cell phone and by the median income at the census tract where the trip originated. I do not find any evident pattern, and, in particular, I do not find that welfare gains are higher for riders with a more expensive cell phone or in high-income areas. I also analyze how welfare effects vary by how much riders are willing to pay. Riders who are willing to pay the most get the largest benefit from surge pricing, but low willingness to pay riders are also better off. Sometimes they are priced out, but they benefit when prices are low. Furthermore, they get lower ETAs and their trip is less likely to be denied. The net effect is that they are ex-ante better off.

I find heterogeneity in the welfare effects of surge pricing along other dimensions, such as time of the week and trip distance (also see appendix F.2). In all cases I find that all riders are better off, with one exception: riders who want to request a trip during a couple of hours on Friday evening and Saturday midday each week, when surge multipliers are highest. All this evidence suggests that concerns about redistribution within riders are not well justified.

One limitation of my model is that it does not capture heterogeneity in riders' behavior by income. Presumably, higher income riders are less elastic and have a higher value of time; surge pricing could therefore affect riders differently depending on their income level. I do not have access to data on riders' income, so I cannot measure heterogeneity by income level directly. One possible extension of my model, however, would be to allow for heterogeneity by the price of riders' cell phone, which can be thought of as a proxy for riders' income.

7 Conclusion

In the debate about the desirability of surge pricing, economists' standard arguments about efficiency gains are challenged by concerns that individual market participants might be hurt. My results provide evidence that supports both sides of the debate. I find efficiency gains that lead to higher welfare. I also find that riders benefit substantially from surge pricing, and I do not find evidence that any riders are worse off. Riders' frequent complaints might arise because they are not aware that, without surge pricing, they would have to wait longer for less reliable trips. On the other hand, my findings about drivers' earnings—a small overall decrease

and an increase in variance—suggest that drivers might be right to complain about surge pricing. This is especially true given their low average earnings, which are only slightly above minimum wages.

A question left unanswered by my paper is what the effects of surge pricing would be if there was competition between platforms. This is a complicated issue given that platforms compete for both riders and drivers, some of whom might multi-home. It could be, for instance, that at times of scarcity higher multipliers would induce riders to switch to a competing platform that would then deplete a common pool of multi-homing drivers. Platforms would then be more reluctant to increase prices, even if it would improve the efficiency of the market. The main challenge that prevents me from tackling these issues is data availability, but certain assumptions about agents' multi-homing behavior might shed light on these questions.

References

- Afeche, Philipp, Zhe Liu, and Costis Maglaras**, "Ride-Hailing Networks with Strategic Drivers: The Impact of Platform Control Capabilities on Performance," *Working paper*, 2017.
- Angrist, Joshua D., Sydnee Caldwell, and Jonathan Hall**, "Uber vs. Taxi: A Driver's Eye View," *Working Paper*, 2017.
- Armstrong, Mark**, "Competition in Two-Sided Markets," *The RAND Journal of Economics*, 2006, 37 (3), 668–691.
- Arnosti, Nick, Ramesh Johari, and Yash Kanoria**, "Managing congestion in matching markets," *Available at SSRN 2427960*, 2018.
- Athey, Susan, Juan Camilo Castillo, and Dan Knoepfle**, "Service Quality in the Gig Economy: Empirical Evidence about Driver Safety at Uber," *Working paper*, 2018.
- Berry, Steven, James Levinsohn, and Ariel Pakes**, "Automobile Prices in Market Equilibrium," *Econometrica*, 1995, 63 (4), 841–890.
- Besbes, Omar, Francisco Castro, and Ilan Lobel**, "Surge Pricing and its Spatial Supply Response," *Working paper*, 2019.

- Bian, Bo**, “Search Frictions, Network Effects and Spatial Competition: Taxis versus Uber,” *Working paper*, 2018.
- Bimpikis, Kostas, Ozan Candogan, and Saban Daniela**, “Spatial pricing in ride-sharing networks,” *Operations Research*, 2019, 67 (3), 744–769.
- Broadie, Mark, Deniz Cicek, and Assaf Zeevi**, “General Bounds and Finite-Time Improvement for the Kiefer-Wolfowitz Stochastic Approximation Algorithm,” *Operations Research*, 2011, 59 (5), 1211–1224.
- Buchholz, Nicholas**, “Spatial Equilibrium, Search Frictions and Efficient Regulation in the Taxi Industry,” *Working paper*, 2018.
- , **Laura Doval, Jakub Kastl, Filip Matjka, and Tobias Salz**, “The Value of Time,” *Working paper*, 2019.
- Cachon, Gérard P., Kaitlin M. Daniels, and Ruben Lobel**, “The Role of Surge Pricing on a Service Platform with Self-Scheduling Capacity,” *Manufacturing & Service Operations Management*, 2017, 19 (3), 368–384.
- Castillo, Juan Camilo, Dan Knoepfle, and Glen Weyl**, “Surge Pricing Solves the Wild Goose Chase,” *Working Paper*, 2018.
- Chen, M Keith, Judith A Chevalier, Peter E Rossi, and Emily Oehlsen**, “The value of flexible work: Evidence from Uber drivers,” Technical Report, National Bureau of Economic Research 2017.
- Cohen, Peter, Robert Hahn, Jonathan Hall, Steven Levitt, and Robert Metcalfe**, “Using Big Data to Estimate Consumer Surplus: The Case of Uber,” Technical Report, National Bureau of Economic Research 2016.
- Cook, Cody, Rebecca Diamond, Jonathan Hall, John A. List, and Paul Oyer**, “The Gender Earnings Gap in the Gig Economy: Evidence from over a Million Rideshare Drivers,” *Working paper*, 2018.
- Cramer, Judd and Alan B. Krueger**, “Disruptive Change in the Taxi Business: The Case of Uber,” *American Economic Review*, May 2016, 106 (5), 177–82.
- Crilly, Rob**, “Customers complain about Uber’s surge pricing on New Year’s Eve,” *The Telegraph*, January 2016. URL: <https://www.telegraph.co.uk/news/worldnews/northamerica/usa/12078264/Customers-complain-about-Ubers-surge-pricing-on-New-Years-Eve.html>.

- Cullen, Zoë and Chiara Farronato**, “Outsourcing tasks online: Matching supply and demand on peer-to-peer internet platforms,” *Working Paper*, 2018.
- Dholakia, Utpal M.**, “Everyone Hates Uber’s Surge Pricing—Here’s How to Fix It,” *Harvard Business Review*, December 2015. URL: <https://hbr.org/2015/12/everyone-hates-ubers-surge-pricing-heres-how-to-fix-it>.
- Fradkin, Andrey**, “Search, matching, and the role of digital marketplace design in enabling trade: Evidence from airbnb,” *Matching, and the Role of Digital Marketplace Design in Enabling Trade: Evidence from Airbnb (March 21, 2017)*, 2017.
- Frechette, Guillaume, Alessandro Lizzeri, and Tobias Salz**, “Frictions in a Competitive, Regulated Market: Evidence from Taxis,” *Forthcoming, American Economic Review*, 2019.
- Garg, Nikhil and Hamid Nazerzadeh**, “Driver Surge Pricing,” *Working Paper*, 2019.
- Goncharova, Masha**, “Ride-Hailing Drivers Are Slaves to the Surge,” *The New York Times*, January 2017. URL: <https://www.nytimes.com/2017/01/12/nyregion/uber-lyft-juno-ride-hailing.html>.
- Hall, Jonathan, John Horton, and Dan Knoepfle**, “Pricing Efficiently in Designed Markets: The Case of Ride-Sharing,” *Working paper*, 2019.
- Hendel, Igal and Aviv Nevo**, “Intertemporal Price Discrimination in Storable Goods Markets,” *American Economic Review*, December 2013, 103 (7), 2722–51.
- Katz, Elai**, “Uber Algorithm Alleged To Constitute Price-Fixing,” *New York Law Journal*, 2016, 225 (124).
- Kazmin, Amy**, “New Delhi bans Uber ‘surge pricing’,” *Financial Times*, April 2016. URL: <https://www.ft.com/content/742d189a-0785-11e6-96e5-f85cb08b0730>.
- Korolko, Nikita, Dawn Woodard, Chiwei Yan, and Helin Zhu**, *Dynamic Pricing and Matching in Ride-Hailing Platforms*, Working Paper, 2018.
- Lagos, Ricardo**, “An Analysis of the Market for Taxicab Rides in New York City,” *International Economic Review*, 2003, 44 (2), 423–434.
- Lam, Chungsang Tom and Meng Liu**, “Demand and Consumer Surplus in the On-Demand Economy: The Case of Ride Sharing,” *Working paper*, 2017.

- Lazarev, John**, “The Welfare Effects of Intertemporal Price Discrimination: An Empirical Analysis of Airline Pricing in U.S. Monopoly Markets,” *Working paper*, 2013.
- Liu, Meng, Erik Brynjolfsson, and Jason Dowlatabadi**, “Do digital platforms reduce moral hazard? The case of Uber and taxis,” *Working Paper*, 2018.
- Lu, Alice, Peter Frazier, and Oren Kislev**, “Surge Pricing Moves Uber’s Driver Partners,” *Working paper*, 2018.
- Ma, Hongyao, Fei Fang, and David C Parkes**, “Spatio-Temporal Pricing for Ridesharing Platforms,” *arXiv preprint arXiv:1801.04015*, 2018.
- Ming, Liu, Tunay I Tunca, Yi Xu, and Weiming Zhu**, “An Empirical Analysis of Market Formation, Pricing, and Revenue Sharing in Ride-Hailing Services,” *Working paper*, 2019.
- Nikzad, Afshin**, “Thickness and Competition in Ride-sharing Markets,” *Working paper*, 2017.
- Puckett, Jessica**, “Honolulu Limits Surge Pricing for Uber and Lyft,” *The Points Guy*, June 2018.
- Rochet, Jean-Charles and Jean Tirole**, “Platform Competition in Two-sided Markets,” *Journal of the European Economic Association*, 2003, 1 (4), 990–1029.
- Shapiro, Matthew H**, “Density of Demand and the Benefit of Uber,” 2018.
- Weyl, E. Glen**, “A Price Theory of Multi-Sided Platforms,” *American Economic Review*, 2010, 100 (4), 1642–1672.
- Williams, Kevin R**, “Dynamic airline pricing and seat availability,” *Working Paper*, 2017.
- Xinyu, Zhang**, “How Didi Works,” *Pingwest*, August 2017. URL: <https://en.pingwest.com/a/112>.
- Yee, Jovic**, “Commuters protest Grab’s high fare; TNC firm denies surge pricing,” *Inquirer*, April 2018. URL: <https://newsinfo.inquirer.net/982744/commuters-protest-grabs-high-fare-tnc-firm-denies-surge-pricing>.

Yusof, Amir, “New measures protect Grab users against ‘excessive’ surge prices, but no restoring of discounts expected,” *Channel News Asia*, September 2018. URL: <https://newsinfo.inquirer.net/982744/commuters-protest-grabs-high-fare-tnc-firm-denies-surge-pricing>.

Appendix A Remaining details about the model

In this section I describe the parts of the model I did not explain in the main text. I first explain how I model them, and then how I fit them to the data.

A.1 Rider destination, trip distance and duration, and base fare

Consider rider i , who opened the app at time t , during hour of the week h , at location l , and wants to go to a destination in distance group \tilde{r} . His destination k is drawn from a distribution $G^{dest}(\cdot|l, h, \tilde{r})$ over locations in distance group \tilde{r} from l , which varies by hour of the week.

The dropoff distance (i.e., how many miles of driving are needed to get from the origin to the destination) is equal to the straight-line distance between the origin and destination times a factor that is drawn from a distribution $G^{dist}(\cdot|l, k, h)$ that varies by origin, destination, and hour of the week. The dropoff duration (i.e., how much driving time it takes to get from the origin to the destination) is equal to the dropoff distance times a factor drawn from a distribution $G^{duration}(\cdot|l, k, h)$ that varies by origin, destination, and hour of the week. The dropoff distance and duration are then used to compute the base fare using the fare structure used by Uber at the time: the sum of a \$2.30 commission plus a \$1.00 fixed rate plus \$0.87 per mile and \$0.11 per minute.

Estimation

I split rider locations into 128 similarly sized origin groups o and into 128 similarly sized destination groups d . For each pair, I compute which distance group \tilde{r} the distance between the midpoints lies in.

I assume that $G^{dest}(\cdot|l, h, \tilde{r})$ is generated as follows. Let $K_{l\tilde{r}}$ be the set of all locations in destination groups d that are at a distance \tilde{r} from o_l , the origin group where l is. Location k is drawn with probability $\frac{v_k \mu_{o_l d_k} \lambda_{hd_k}}{\sum_{k' \in K_{l\tilde{r}}} v_{k'} \mu_{o_l d_{k'}} \lambda_{hd_{k'}}$. $\mu_{o_l d_k}$ is a factor that measures how frequent is it to go from o_l to d_k . I estimate it as the fraction of

trips from o_l that go to d_k . λ_{hd_k} measures how likely are people to go to locations in d_k during h , relative to other times of the week. I estimate it as the ratio between the fraction of trips going to d_k during h and the fraction of trips going to d_k at all times of the week. ν_k is a measure of how likely trips going to d_k go to k . I simply estimate it as the empirical probability.

I assume that $G^{dist}(\cdot|l, k, h)$ is a lognormal distribution with parameters $(\mu_{lkh}^{dist}, \sigma_{lkh}^{dist})$. I estimate μ_{lkh} from a model of the log ratio between dropoff distance and straight-line distance on origin group by destination group and hour fixed effects. I estimate σ_{lkh}^{dist} with a linear model of residual standard deviation by bins of μ_{lkh}^{dist} .

I assume that $G^{duration}(\cdot|l, k, h)$ is a lognormal distribution with parameters $(\mu_{lkh}^{duration}, \sigma_{lkh}^{duration})$. I estimate these parameters as above, starting from a model of the log ratio between trip duration and dropoff distance.

A.2 ETAs and actual pickup distance and duration

The ETA shown to rider i is a function $w(\mathbf{a}_t, l, h)$ of the rider's location and the hour of the week, as well as on the number of available drivers in every nearby location, denoted by \mathbf{a}_t .

If the rider requests a trip and is matched to driver j , the actual pickup duration is the one that was generated in the matching process (section 3.3). The pickup distance, which is relevant to compute driver costs, is equal to the straight-line distance between the midpoints of the request location the location where the pickup starts (either the driver's location, or, if he is busy, the dropoff location) times a factor drawn from a distribution G^{pickup} that has support $[1, \infty)$.

Estimation

I generate the ETAs shown to riders from a random forest of ETAs on the coordinates of the midpoint of the rider's location, the hour of the week, and the number of available drivers in every location relative to the rider's location. As in my model to predict ETAs for rider-driver pairs, I include transformations of the variables to improve the model fit (see footnote 41).

In order to use the functional form I fit in my counterfactuals, this relation must be causal. The exogeneity assumption is justified if I assume that, after controlling for the number of available drivers in every surrounding location, variation in waiting times arise only because of the idiosyncratic location of drivers around a rider,

which is uncorrelated with any market characteristics.

I take G^{pickup} to be a shifted Gamma distribution (so that the minimum value is one). I set the rate and scale parameters so that the average and variance of this distribution match the empirical moments of the distribution of the ratio between driving and straight line distances for pickups.

A.3 Outside behavior

I focus on an area of central Houston. However, I cannot simply ignore the behavior in surrounding areas. The main problem is that few drivers work a whole shift without ever exiting this central area.

In order to solve this issue, I split all of Houston into three regions. The first one is my main region of analysis, where my full model applies. The second one is a buffer zone surrounding the central area, which has roughly the same area as the central region, in which I model drivers' movements and match riders and drivers just as in the inside region, but I do not model the response of demand to prices and ETAs. Finally, the third area includes all locations outside of the buffer area. I do not model drivers' movements in this area. I simply model a single pool of outside drivers that might be matched to riders who mechanically request trips in this outside area.

Let l be a location outside the area of analysis. Regarding trip requests, I consider the outside area to be one such location. During hour of the week h , riders request trips from location l to a destination in distance group \tilde{r} at a rate $v_{lh\tilde{r}}$. For every such request, the destination is chosen from a distribution $G^{dest,out}(\cdot|l, h, \tilde{r})$, and the distance and duration are drawn as with inside trips, using distributions $G^{dist,out}(\cdot|l, k, h)$ and $G^{duration,out}(\cdot|l, k, h)$. For trips in the buffer area, pickup times are the ones generated in the matching process. For those in the outside area, pickup times are drawn from a distribution $G^{pickup,out}(\cdot|h)$ that varies by hour of the week. For those in the buffer area, it is drawn as for inside trips, with a distribution $G^{pickup,buffer}$.

Drivers' movement model applies to the buffer area. Moving outside is just one more option in the choice set they can take. Drivers that are outside move to location l in the buffer area with probabilities $p^{movein}(l; h)$ that depend on the hour of the week, and stay outside otherwise. Drivers that drop off passengers after trips that end in the outside area join the pool of outside drivers.

Estimation

I assume that $G^{dest,out}(\cdot|l,h,\tilde{r})$, $G^{dist,out}(\cdot|l,k,h)$, $G^{duration,out}(\cdot|l,k,h)$, $G^{pickup,out}(\cdot|h)$, and $G^{pickup,buffer}$ are generated by the same process as their equivalents for inside trips. I fit their parameters the same way, using the sample of trips that take place outside. I estimate $p^{movein}(l;h)$ as empirical frequencies.

Appendix B Details about equilibrium

B.1 Existence

Proposition 1. $f^P(\cdot, \theta)$ has at least one fixed point $\mathbf{x}^* \in \mathcal{X}$.

Proof. $f^P(\cdot, \theta)$ is a continuous function, since all functions that are involved are continuous. \mathcal{X} can be bounded below by the greatest loss a driver can get (moving back and forth between the two locations that are furthest away) and by zero utility, and above by the earnings a driver would get if he was matched immediately in all subsequent periods to the most profitable trip possible and by drivers' utility if prices and ETAs were both always zero. \mathcal{X} is thus a convex, compact space, and by Brouwer's fixed point theorem Π has a fixed point. \square

B.2 Uniqueness

Assumption 1. $f^P(\cdot, \theta)$ is uniformly continuous, and there exists some $\delta < 1$ such that $(f^P(\mathbf{x}, \theta) - f^P(\mathbf{x}', \theta)) \cdot (\mathbf{x} - \mathbf{x}') < \delta \|\mathbf{x} - \mathbf{x}'\|^2$ for every $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$.

I cannot prove that assumption 1 is true, but it holds for every single pair $(\mathbf{x}, \mathbf{x}')$ I have tried. A simple intuition justifies the second part. As beliefs change from \mathbf{x} to \mathbf{x}' , drivers would tend to move towards locations and times with higher earnings, and more riders would request trips at times and locations with higher utilities. Those crowded locations and times should then get lower earnings/utilities, suggesting that the new vector \mathbf{x} is negatively correlated with the old vector, i.e., $(f^P(\mathbf{x}, \theta) - f^P(\mathbf{x}', \theta)) \cdot (\mathbf{x} - \mathbf{x}') < 0$. This condition is stronger than assumption 1. It would only be violated if there are strong complementarities between riders and drivers.

Proposition 2. Under assumption 1, $f^P(\cdot, \theta)$ has a unique fixed point.

Proof. Consider the mapping $g_\gamma : \mathcal{X} \rightarrow \mathcal{X}$ defined by $g_\gamma(\mathbf{x}) = (1 - \gamma)\mathbf{x} + \gamma f^P(\mathbf{x}, \theta)$, where $0 < \gamma < 1$. It is straightforward to see that the set of fixed points of $f^P(\cdot, \theta)$

and g_γ is the same. I will show that there exists some γ such that g_γ is a contraction mapping, which implies, by the contraction mapping theorem, that g_γ has a unique fixed point.

By uniform continuity, there exists some $\beta < \infty$ such that $\frac{\|f^P(\mathbf{x}, \theta) - f^P(\mathbf{x}', \theta)\|}{\|\mathbf{x} - \mathbf{x}'\|} < \beta$. For all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ we have that $\|g_\gamma(\mathbf{x}) - g_\gamma(\mathbf{x}')\|^2 = (1 - \gamma)^2 \|\mathbf{x} - \mathbf{x}'\|^2 + 2\gamma(1 - \gamma) \langle f^P(\mathbf{x}, \theta) - f^P(\mathbf{x}', \theta), \mathbf{x} - \mathbf{x}' \rangle + \gamma^2 \|f^P(\mathbf{x}, \theta) - f^P(\mathbf{x}', \theta)\|^2 < [(1 - \gamma)^2 + 2\gamma(1 - \gamma)\delta + \gamma^2\beta^2] \|\mathbf{x} - \mathbf{x}'\|^2$. A Taylor expansion about $\gamma = 0$ of the term in brackets yields $(1 - \gamma)^2 + 2\gamma(1 - \gamma)\delta + \gamma^2\beta^2 = 1 + 2(\delta - 1)\gamma + O(\gamma^2)$. This quantity is less than one for small enough $\gamma > 0$, and since β is bounded, there exists some $\gamma > 0$ and some $\delta \in (0, 1)$ such that $\|g_\gamma(\mathbf{x}) - g_\gamma(\mathbf{x}')\| \leq \delta \|\mathbf{x} - \mathbf{x}'\|$ for all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$. This means that g_γ is a contraction mapping. \square

B.3 Stability

The most common notion of stability requires $f^P(\cdot, \theta)$ to be a contraction mapping (at least in some subset of \mathcal{X} , the basin of attraction of the equilibrium). The idea is that, after some perturbation, the market is brought back to equilibrium by repeated belief updating according to $f^P(\cdot, \theta)$. This is not necessarily true in my model. In an extreme case, if drivers are too sensitive to earnings, they will all herd to the most profitable locations, which will then get very small earnings according to $f^P(\cdot, \theta)$. The next iteration would result in the opposite: drivers would herd *away* from those regions.

The mapping g_γ I used in the proof above corresponds to stability in terms of a different belief update process. Suppose that only a fraction γ of agents update their beliefs in every iteration. Alternatively, suppose that agents' beliefs have some inertia so that in every iteration they only give weight γ to new observations. The mapping $g_\gamma(\cdot)$ represents this kind of belief update process. Since $g_\gamma(\cdot)$ is a contraction mapping, the market is stable according to this belief updating process.

B.4 Computation

Computing market equilibria has some complications. First, I cannot compute $f^P(\mathbf{x}, \theta)$ directly. Instead, I can only compute an estimator $\hat{f}^P(\mathbf{x}, \theta)$ such that $E[\hat{f}^P(\mathbf{x}, \theta)] = f^P(\mathbf{x}, \theta)$ by simulation. Second, a naive iterative algorithm, where a sequence of beliefs $(\mathbf{x}_n)_{n=1}^\infty$ is generated according to $\mathbf{x}_{n+1} = \hat{f}^P(\mathbf{x}_n, \theta)$, often diverges. Third, even when this algorithm converges, it is hard to assess convergence because of the randomness in the simulation.

In order to solve these issues, I compute market equilibria by iterating on

$$\mathbf{x}_{n+1} = (1 - \gamma_n)\mathbf{x}_n + \gamma_n \hat{f}^P(\mathbf{x}_n, \theta), \quad \gamma_n \propto n^{-b}, \quad (14)$$

where $b \in (0, 1)$. New beliefs are not empirical averages under old beliefs, but a convex combination between old beliefs and the empirical average.

In essence, this algorithm consists of iteratively applying the contraction mapping g_γ I defined to prove proposition 2, with a decaying value of γ . Beyond assumption 1, in order to prove convergence I need the following assumption:

Assumption 2. *Draws from simulation averages can be written as $\hat{f}^P(\mathbf{x}, \theta) = f^P(\mathbf{x}, \theta) + \varepsilon$, where ε are mean-zero independent random variables with bounded variance.*

Note that this specification for $\hat{f}^P(\mathbf{x}, \theta)$ is very flexible: I allow the distribution of ε to depend on (\mathbf{x}, θ) arbitrarily, as long as it has mean zero and the variance is bounded.

Proposition 3. *Let $(\mathbf{x}_n)_{n=0}^\infty$ be a sequence over \mathcal{X} defined iteratively by equation (14), where $\mathbf{x}_0 \in \mathcal{X}$. Under assumptions 1 and 2, $\mathbf{x}_n \xrightarrow{P} \mathbf{x}^*$.*

Proof. First, note that $E[||\mathbf{x}_n - \mathbf{x}^*||^2] = E[||(1 - \gamma_n)\mathbf{x}_{n-1} + \gamma_n f_n^P(\mathbf{x}_{n-1}, \theta) - \mathbf{x}^*||^2] + \gamma_n^2 \text{Var}[\varepsilon]$, where both terms can be separated by the independence of the error terms ε . The algebra in the proof of proposition 2 implies that $||\mathbf{x}_n - \mathbf{x}^*||^2 < [(1 - \gamma_n)^2 + 2\gamma_n(1 - \gamma_n)\delta + \gamma_n^2\beta^2] ||\mathbf{x}_{n-1} - \mathbf{x}^*||^2$ pointwise, so $V_n < [(1 - \gamma_n)^2 + 2\gamma_n(1 - \gamma_n)\delta + \gamma_n^2\beta^2] V_{n-1} + \gamma_n^2\Gamma$, where $V_n = E[||\mathbf{x}_n - \mathbf{x}^*||^2]$ and $\Gamma < \infty$ is such that $\Gamma > \text{Var}[\varepsilon]$. This recursion over V_n converges to zero (see Broadie et al., 2011, online appendix), which implies that $E[||\mathbf{x}_n - \mathbf{x}^*||^2]$ converges to zero and \mathbf{x}_n converges in probability to \mathbf{x}^* . \square

This process resembles maximization of an objective function by stochastic gradient descent, with a decaying step γ_n . The fact that it decays ensures that the sequence converges, since noise variance decreases with the weight. It converges to the fixed point since $\sum_{n=1}^\infty \gamma_n = \infty$, which means that the starting point eventually becomes irrelevant.

There is a tradeoff when setting b . If it is too low, noise variance decays too slowly. If it is too high, it takes a long time to incorporate information from new runs. I set $b = 0.8$, which typically leads to stable beliefs after around 10 iterations.

Appendix C Formal results about the identification strategy

C.1 Demand (trip request)

My identification strategy relies on the following assumption:

Assumption 3. *Rider i 's utility is given by equation (1), which satisfies the following properties:*

1. $p_i = p(\tilde{\mathbf{m}}_t; \bar{p}_i, l)$, where $p(\cdot)$ is a deterministic function.
2. $w_i = w_{lt}^0 + \xi_i$, where $w_{lt}^0 = E[w_i|l, t]$, and ξ_i is orthogonal to $(\epsilon_i, \bar{p}_i, \tilde{\mathbf{m}}_t)$.

The rider requests a trip if $U_i > 0$.

Under assumption 3, the causal effect of prices and ETAs can be isolated by adding a control function that depends on $(\tilde{\mathbf{m}}_t, w_{lt}^0)$:

Proposition 4. *Under assumption 3, the rider's utility (1) can be rewritten as*

$$y_i = \alpha(r_i, l, h) + \beta(r_i)p_i + \gamma(r_i)w_i + g(\tilde{\mathbf{m}}_t, \bar{p}_i, w_{lt}^0; l, h) + \eta_i, \quad (15)$$

where $g(\tilde{\mathbf{m}}_t, \bar{p}_i, w_{lt}^0; l, h) = E[\epsilon_i | \tilde{\mathbf{m}}_t, \bar{p}_i, w_{lt}^0; l, h]$ and the error η_i has zero conditional mean: $E[\eta_i | p_i, w_i, \tilde{\mathbf{m}}_t, \bar{p}_i, w_{lt}^0, l, h] = 0$.

Proof. We can write $\epsilon_{it} = g(\tilde{\mathbf{m}}_t, \bar{p}_i, w_{lt}^0; l, h) + \eta_i$, where $E[\eta_i | \tilde{\mathbf{m}}_t, \bar{p}_i, w_{lt}^0, l, h] = 0$ by the definition of conditional mean. This yields equation (15). Since p_i is a deterministic function of $(\tilde{\mathbf{m}}_t, \bar{p}_i, l)$ and $w_i = w_{lt}^0 + \xi_i$, $E[\eta_i | p_i, w_i, \tilde{\mathbf{m}}_t, \bar{p}_i, w_{lt}^0, l, h] = E[\eta_i | w_i, \tilde{\mathbf{m}}_t, \bar{p}_i, l, h, \xi_i] = E[\epsilon_i | w_i, \tilde{\mathbf{m}}_t, \bar{p}_i, l, h, \xi_i] - E[E[\epsilon_i | \tilde{\mathbf{m}}_t, \bar{p}_i, w_{lt}^0; l, h] | w_i, \tilde{\mathbf{m}}_t, \bar{p}_i, l, h, \xi_i]$. By assumption, ξ_i is orthogonal to $(\epsilon_i, \bar{p}_i, \tilde{\mathbf{m}}_t)$, and by its definition, it is orthogonal to (w_{lt}^0, l, t) . ξ_i can thus be dropped out of both terms in the last expression, which is then equal to zero. \square

My identification strategy relies on estimating equation (15). This equation can be estimated by standard regression techniques by proposition 4, which states there are no remaining endogeneity issues after including the control function $g(\cdot)$.

The following assumption is necessary to ensure $\beta(r_i)p_i$ and $g(\tilde{\mathbf{m}}_t, \bar{p}_i, w_{lt}^0; l, h)$ are identified:

Assumption 4. *Agents' utility (1) satisfies the following properties:*

1. $p(\tilde{\mathbf{m}}_t; \bar{p}_i, l)$ has at least one discontinuity in $\tilde{\mathbf{m}}_t$.

2. $E[\epsilon_i | \tilde{\mathbf{m}}_t, \bar{p}_i, \bar{w}_{lt}, l, h]$ is continuously differentiable in $\tilde{\mathbf{m}}_t$.

The first part of this assumption is a property of the surge pricing algorithm. The second part states that recommended multipliers, unsurged fares, and the number of available drivers are related smoothly with demand shocks. This is a reasonable assumption since all the variables that are used as inputs to the surge pricing algorithm and to define unsurged fares enter through smooth functional forms.

Proposition 5. *Under assumption 4, there exists a finite vector $X(\tilde{\mathbf{m}}_t, \bar{p}_i, w_{lt}^0; l, h)$ that is continuously differentiable in $\tilde{\mathbf{m}}_t$ and such that $E[\epsilon_i | \tilde{\mathbf{m}}_t, \bar{p}_i, \bar{w}_{lt}, l, h] \in \text{Span}(X(\tilde{\mathbf{m}}_t, \bar{p}_i, w_{lt}^0; l, h))$. For any such vector, and for any nonzero function $f(r_i)$, $f(r_i)p(\tilde{\mathbf{m}}_t; \bar{p}_i, l) \notin \text{Span}(X(\tilde{\mathbf{m}}_t, \bar{p}_i, w_{lt}^0; l, h))$.*

Proof. $X(\tilde{\mathbf{m}}_t, \bar{p}_i, w_{lt}^0; l, h) = E[\epsilon_i | \tilde{\mathbf{m}}_t, \bar{p}_i, \bar{w}_{lt}, l, h]$ is continuously differentiable in $\tilde{\mathbf{m}}_t$ and $E[\epsilon_i | \tilde{\mathbf{m}}_t, \bar{p}_i, \bar{w}_{lt}, l, h]$ is in its span, which proves existence. Every linear combination of $X(\tilde{\mathbf{m}}_t, \bar{p}_i, w_{lt}^0; l, h)$ is continuously differentiable in $\tilde{\mathbf{m}}_t$, whereas $f(r_i)p(\tilde{\mathbf{m}}_t; \bar{p}_i, l)$ has at least one discontinuity in $\tilde{\mathbf{m}}_t$, so $f(r_i)p(\tilde{\mathbf{m}}_t; \bar{p}_i, l) \notin \text{Span}(X(\tilde{\mathbf{m}}_t, \bar{p}_i, w_{lt}^0; l, h))$. \square

A practical concern is finding a vector $X(\tilde{\mathbf{m}}_t, \bar{p}_i, w_{lt}^0; l, h)$ whose span contains $E[\epsilon_i | \tilde{\mathbf{m}}_t, \bar{p}_i, \bar{w}_{lt}, l, h]$. A combination of high order splines should approximate it well.

C.2 Supply (movement model)

Consider driver j in location l at time t during hour of the week h . Suppose that the driver movement rule is given by $l_{j,t+1} = \text{argmax}_k y_{j,t+1}^k$, where

$$y_{j,t+1}^k = \omega(l, k, h) + \delta v_k + \zeta_{j,t+1}^k \quad (16)$$

My identification strategy relies on the following assumption:

Assumption 5. $v_k = v_k(\mathbf{m}_t(\tilde{\mathbf{m}}_t); l, h)$ is a deterministic function of $(\tilde{\mathbf{m}}_t; l, h)$.

Under this assumption, the causal effect of v_k can be isolated by adding a control function that depends on $\tilde{\mathbf{m}}_t$:

Proposition 6. *Under assumption 5, $y_{j,t+1}^k$ can be rewritten as*

$$y_{j,t+1}^k = \omega(l, k, h) + \delta v_k + g^M(\tilde{\mathbf{m}}_{tk}; l, h) + \psi_{j,t+1}^k, \quad (17)$$

where $g^M(\tilde{\mathbf{m}}_{tk}; l, h) = E[\zeta_{j,t+1}^k | \tilde{\mathbf{m}}_{tk}; l, h]$ and the error $\psi_{j,t+1}^k$ has zero conditional mean: $E[\psi_{j,t+1}^k | v_k, \tilde{\mathbf{m}}_{tk}, l, h] = 0$.

Proof. We can write $\zeta_{j,t+1}^k = g^M(\tilde{\mathbf{m}}_{tk}; l, h) + \psi_{j,t+1}^k$, where $E[\psi_{j,t+1}^k | \tilde{\mathbf{m}}_{tk}; l, h] = 0$ by the definition of conditional mean. This yields equation (17). Since v_k is a deterministic function of $\tilde{\mathbf{m}}_{tk}$, $E[\psi_{j,t+1}^k | v_k, \tilde{\mathbf{m}}_{tk}, l, h] = E[\psi_{j,t+1}^k | \tilde{\mathbf{m}}_{tk}, l, h] = 0$. \square

My identification strategy relies on estimating equation (17). It can be estimated by standard regression techniques by proposition 6, which states there are no remaining endogeneity issues after including the control function $g^M(\cdot)$. Note that, if $\psi_{j,t+1}^k$ are iid random variables with an extreme value type I distribution, then this model reduces to movement rule (8).

The following assumption is necessary to ensure δv_k and $g^M(\tilde{\mathbf{m}}_{tk}; l, h)$ are identified:

Assumption 6. *The latent variable for drivers' movement (16) satisfies the following properties:*

1. $v_k = v_k(\mathbf{m}_t(\tilde{\mathbf{m}}_t); l, h)$ has at least one discontinuity in $\tilde{\mathbf{m}}_t$.
2. $E[\zeta_{j,t+1}^k | \tilde{\mathbf{m}}_{tk}; l, h]$ is continuously differentiable in $\tilde{\mathbf{m}}_t$.

The first part of this assumption is justified if (a) $\mathbf{m}_t(\tilde{\mathbf{m}}_t)$ has discontinuities, and (b) those discontinuities translate into discrete jumps in v_k . The second part states that recommended multipliers are related smoothly with supply shocks. This is a reasonable assumption since all the variables that are used as inputs to the surge pricing algorithm and to define unsurged fares enter through smooth functional forms.

Proposition 7. *Under assumption 6, there exists a finite vector $Z(\tilde{\mathbf{m}}_{tk}; l, h)$ that is continuously differentiable in $\tilde{\mathbf{m}}_t$ and such that $E[\epsilon_i | \tilde{\mathbf{m}}_{tk}, l, h] \in \text{Span}(X(\tilde{\mathbf{m}}_{tk}; l, h))$. For any such vector, $v_k \notin \text{Span}(X(\tilde{\mathbf{m}}_{tk}; l, h))$.*

Proof. $Z(\tilde{\mathbf{m}}_{tk}; l, h) = E[\zeta_{j,t+1}^k | \tilde{\mathbf{m}}_{tk}, l, h]$ is continuously differentiable in $\tilde{\mathbf{m}}_t$ and $E[\zeta_{j,t+1}^k | \tilde{\mathbf{m}}_{tk}, l, h]$ is in its span, which proves existence. Every linear combination of $Z(\tilde{\mathbf{m}}_{tk}; l, h)$ is continuously differentiable in $\tilde{\mathbf{m}}_t$, whereas v_k has at least one discontinuity in $\tilde{\mathbf{m}}_t$, so $v_k \notin \text{Span}(Z(\tilde{\mathbf{m}}_{tk}; l, h))$. \square

A practical concern is finding a vector $Z(\tilde{\mathbf{m}}_{tk}; l, h)$ whose span contains $E[\zeta_{j,t+1}^k | \tilde{\mathbf{m}}_{tk}, l, h]$. A combination of high order splines should approximate it well.

Appendix D Estimation

D.1 Unrounded multipliers

I estimate a model of the form $m_{lt} = h(\tilde{\mathbf{m}}_t) + \phi_{lt}$ to compute $\hat{m}_{lt} = h(\tilde{\mathbf{m}}_t)$, the *unrounded multiplier*, the multiplier that would have been set if there was no rounding in the surge pricing algorithm. The basic idea of this model is that it has the same form as the surge pricing algorithm (described in section 4.1), except that I omit the steps in which there is rounding. The recommended multipliers are simply smoothed out in space and time to obtain unrounded multipliers. I allow the dependence on recommended multipliers to vary very flexibly by location, giving me confidence that I am really accounting for everything except for the effect of rounding.

First, let \bar{m}_{lt} be the *bounded multiplier*, which is the recommended multiplier subject to the upper and lower bounds that the surge pricing algorithm sets on multipliers, depending on the previous surge multiplier, to avoid abrupt multiplier changes. These bounded multipliers are smoothed out spatially using the following function:

$$m_{lt} = \alpha_{0,l}h_0(\bar{m}_{lt}) + \sum_{r=1}^{r^{max}} \alpha_{r,l} \sum_{k \in R_{r,l}} h_r(\bar{m}_{kt}) + \phi_{lt} \quad (18)$$

$R_{r,l}$ is the set of locations at a distance r from location l . The α coefficients represent the weight given to multipliers at a given distance from the current location. Functions $h_r(\cdot)$ represent a flexible functional form that gives the dependence of multipliers on bounded multipliers at a distance r .

The functional form for takes into account the fact that the surge pricing algorithm gives the same weight to all nearby locations that are at the same distance. It is very important to allow α coefficients to vary by location because different locations have different number of nearby locations. For instance, the northernmost locations have no neighbors to the north, so they have to give higher weights to locations that are south of them. Furthermore, more dense areas give weights to a smaller number of nearby locations.

I restrict the function for each distance to be the same for all locations. I use a spline basis with 10 degrees of freedom to generate these functions.⁴⁸ This results in a nonlinear model which I estimate by iterating back and forth between two

⁴⁸I select knots that are closer to each other for lower values of the bounded multiplier, where most of the data lies. I also saw that further increasing the degrees of freedom has essentially no effect.

linear estimations. First, I estimate the weights given by each location to nearby rings given some value of the splines. Then I estimate the splines given some weights, and I use these splines to compute new weights. In other words, I follow a coordinate descent algorithm until convergence.

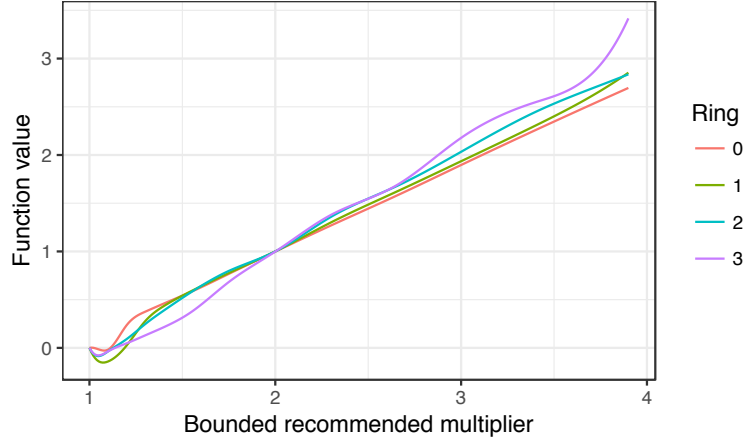


Figure 18: Functional form of $h(\cdot)$ functions.

Results Figure 18 shows the functional form of the estimated $h(\cdot)$ for the first three rings. They are all very close to the identity function, except for an important nonlinearity close to 1. The reason for this is the gap due to the fact that the multiplier is never 1.1. Figure 19 shows the residual variation ϕ_{lt} that I use to identify the demand response to prices as a function of the bounded multiplier at the observed location. There is substantial variation, especially around the gap created by the fact that the multiplier cannot be 1.1.

D.2 Mean earnings fit

Figure 20 shows how $f(\mathbf{m}_{tk})$ varies as multipliers change. Earnings increase as multipliers increase, and as more locations have higher surge levels. I plug in $v_k(\mathbf{s}_t) = \alpha(k, h) + f(\mathbf{m}_{tk}) - c_l^k$ to estimate the supply model.

D.3 Algorithm for estimation of the movement model

The likelihood of one individual observation is

$$L_{jt} = \frac{\lambda_{jlt}^c}{\Lambda_{jlt}} \quad (19)$$

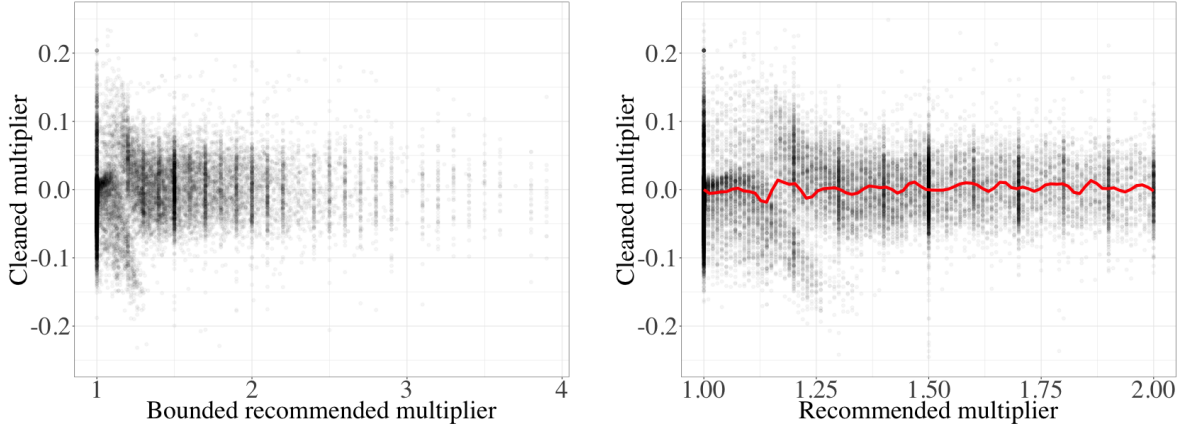


Figure 19: Residual variation in multipliers

Note: Residual of regression (18) as a function of the bounded multiplier for a random subsample of the data. The red line in the subfigure on the right represents a local fit.

where $\lambda_{jlt}^k = \exp(\alpha_l^k + \gamma_{z_l h_t}^{m_k} + \beta x_{jlt}^k)$, and $\Lambda_{jlt} = \sum_{k \in K_l} \lambda_{jlt}^k$. The term βx_{jlt}^k is equal to $\delta v_k(\mathbf{s}_t) + g^M(\tilde{\mathbf{m}}_{tk}; l, h)$, written out as a linear combination of variables. In all these expressions, $k = c$ represents the action chosen by the driver in the current observation.

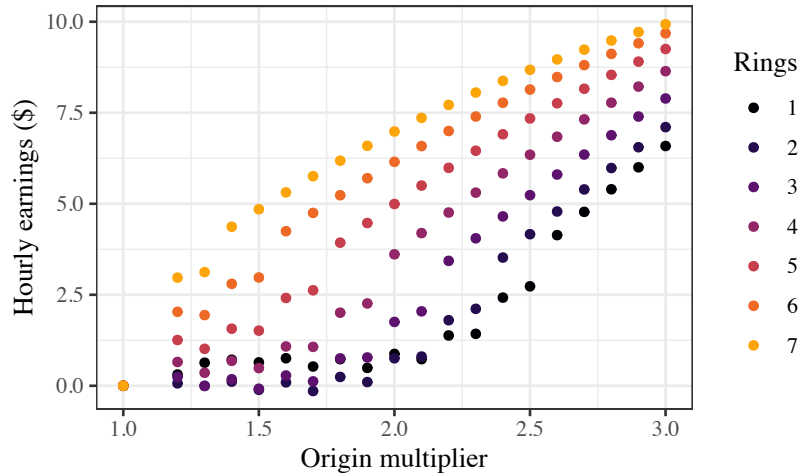


Figure 20: Expected hourly earnings as a function of multipliers

Note: Expected hourly earnings for the next 90 minutes as a function of surrounding surge multipliers, relative to earnings when all multipliers are one. Each series increases the multiplier at the origin. Rings refers to how many locations around the origin have surge pricing. For the series with rings=4, for instance, all multipliers up to a distance 4 are equal to the multiplier at the origin. Multipliers further than 4 decay smoothly all the way down to one.

The likelihood maximization problem I wish to solve is

$$(\hat{\alpha}, \hat{\gamma}, \hat{\beta}) = \underset{(\alpha, \gamma, \beta)}{\operatorname{argmax}} \sum_{jt} \log L_{jt}(\alpha, \gamma, \beta). \quad (20)$$

The vector $(\hat{\alpha}, \hat{\gamma}, \hat{\beta})$ is very high dimensional, so a standard nonlinear optimization algorithm takes too many iterations to converge. Instead, I follow an algorithm that finds quickly the optimal value of (α, γ) given some value of β (I “concentrate” the likelihood function). In other words, I solve

$$\max_{\beta} \max_{(\alpha, \gamma)} \sum_{jt} \log L_{jt}(\alpha, \gamma, \beta), \quad (21)$$

where the outer problem is a standard quasi-Newton algorithm for nonlinear optimization. I now describe how I solve the inner problem.

Let f_l^k be the fraction of drivers that move to location k after starting at location l , and let p_{jlt}^k be the probability that driver j in location l at time t moves to destination k . The first order condition for likelihood maximization with respect to α_l^k takes the form $f_l^k = \frac{1}{N_l} \sum_{jt} p_{jlt}^k$, where the sum is over all observations where the driver starts at location l . In other words, the value for fixed effects is such that the predicted fraction of movements from l to k is equal to the sample fraction. This intuitive condition arises from the assumption of extreme value type 1 errors.

Similarly, if f_{zh}^m is the fraction of drivers that start in a location in z and at a time in h that follow a movement trend in m and p_{jlt}^m is the predicted probability of driver j moving to a location corresponding to m , the first order condition for γ_{zh}^m is that $f_{zh}^m = \frac{1}{N_{zh}} \sum_{jlt} p_{jlt}^m$, where the sum is over all observations starting in a location in z and at a time in h .

These two conditions can be used to compute the values of the fixed effects that maximize the likelihood, conditional on values of the main model parameters. The probability p_{jlt}^k can be written as $\frac{\exp(\alpha_l^k) \exp(\theta X_{jlt}^o)}{\sum_o \exp(\alpha_l^o) \exp(\theta X_{jlt}^o)}$, where θX_{jlt} is the term corresponding to all variables that are not related to origin-destination fixed effects (and it includes movement trend fixed effects). Thus, $f_l^k = \frac{1}{N_l} \sum_{jt} p_{jlt}^k$ determines all fixed effects implicitly, and they can be computed by iterating on s for the following equation:

$$\exp(\alpha_l^{k,s+1}) = \frac{f_l^k}{\frac{1}{N_l} \sum_{jt} \frac{\exp(\theta X_{jlt}^k)}{\sum_o \exp(\alpha_l^{o,s}) \exp(\theta X_{jlt}^o)}} \quad (22)$$

This process is a sample analogue to the iterative process in Berry et al. (1995) to

compute effects that are consistent with observed market shares, and converges for the same reason: it is a contraction mapping. An analogous process allows me to compute the set of fixed effects γ_{zh}^m .

My full algorithm thus looks as follows:

```

Set initial value of main model parameters and initial value for fixed effects;
while magnitude of gradient is greater than tolerance do
  while maximum overall share gap is greater than tolerance do
    while maximum gap between predicted and actual shares for
      origin-destination is greater than tolerance do
      | Compute origin-destination fixed effects;
    end
    while maximum gap between predicted and actual shares for movement
      trends is greater than tolerance do
      | Compute movement trends fixed effects;
    end
    Compute overall gap as the maximum of origin-destination and
      movement trend gaps;
  end
  Compute gradient;
  Compute new parameters based on gradient and inverse Hessian;
end

```

D.4 Shift length fit

Figure 21 compares the empirical distribution of \bar{D}_j and \underline{D}_j with the distribution of generated shift lengths. The estimated distribution is intermediate between both distributions. Most importantly, it shows that the assumption of a gamma distribution seems reasonable.

Appendix E Model fit

In the next few figures I compare the output of my simulations and the data. Figure 22 shows temporal patterns of supply, demand, and the number of trips. Demand and the number of trips fit the data very well. The number of drivers is somewhat

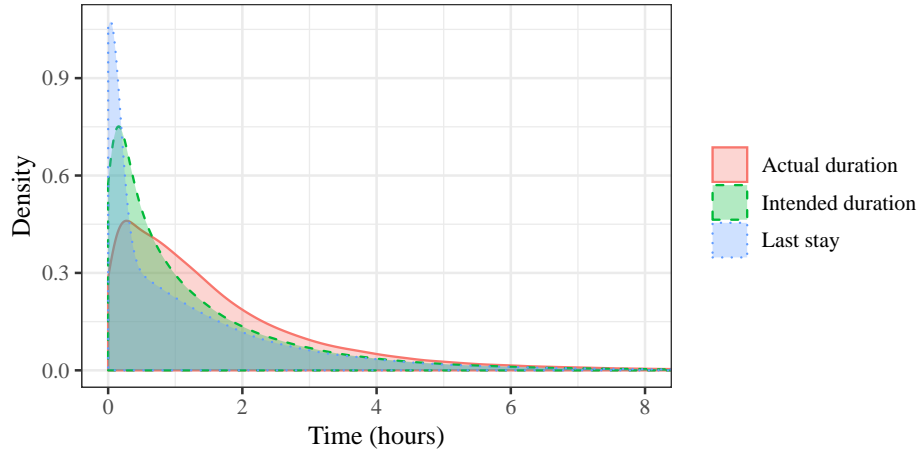


Figure 21: Empirical and estimated distribution of shift length

Note: Distributions related to shift length. The actual duration represents how long the driver worked before logging out. The last stay represents the last time the driver was available and chose not to stop working. The intended duration represents the estimated distribution for the shift length, which must be between the last stay and the actual duration.

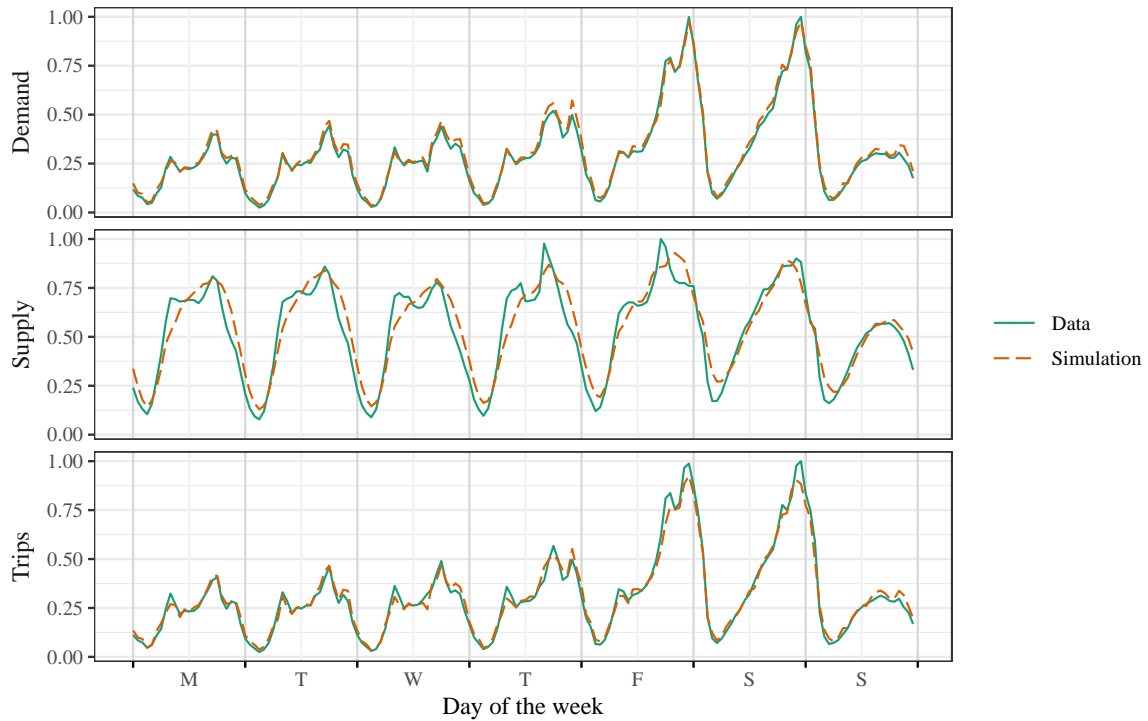
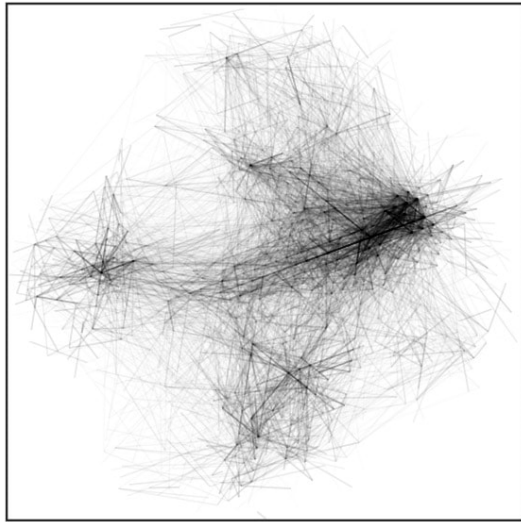


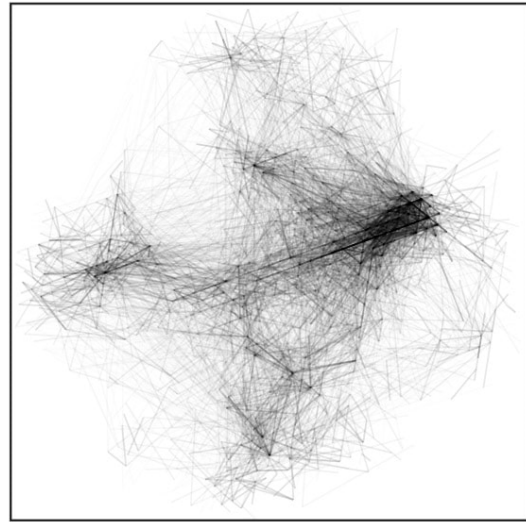
Figure 22: Temporal patterns in simulations and in the data

Note: Temporal patterns for supply, demand, and number of trips in simulations and in the data. Demand is the number of sessions. Supply is the number of drivers working. Trips is the number of trips that take place. All three figures are normalized so that the maximum in the data is one.

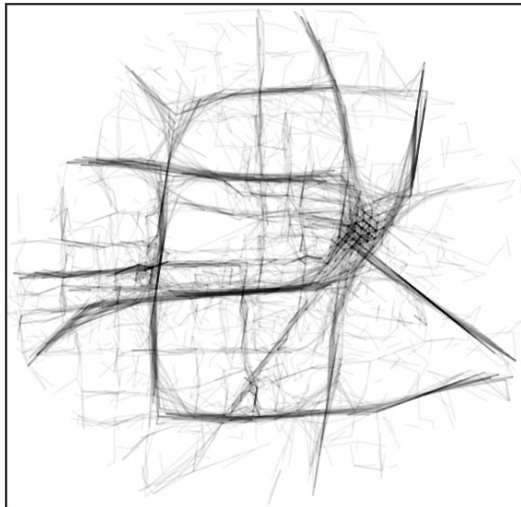
less precise, but it still follows the broad patterns in the data.⁴⁹



(a) Trips, data



(b) Trips, simulation



(c) Driver movements, data



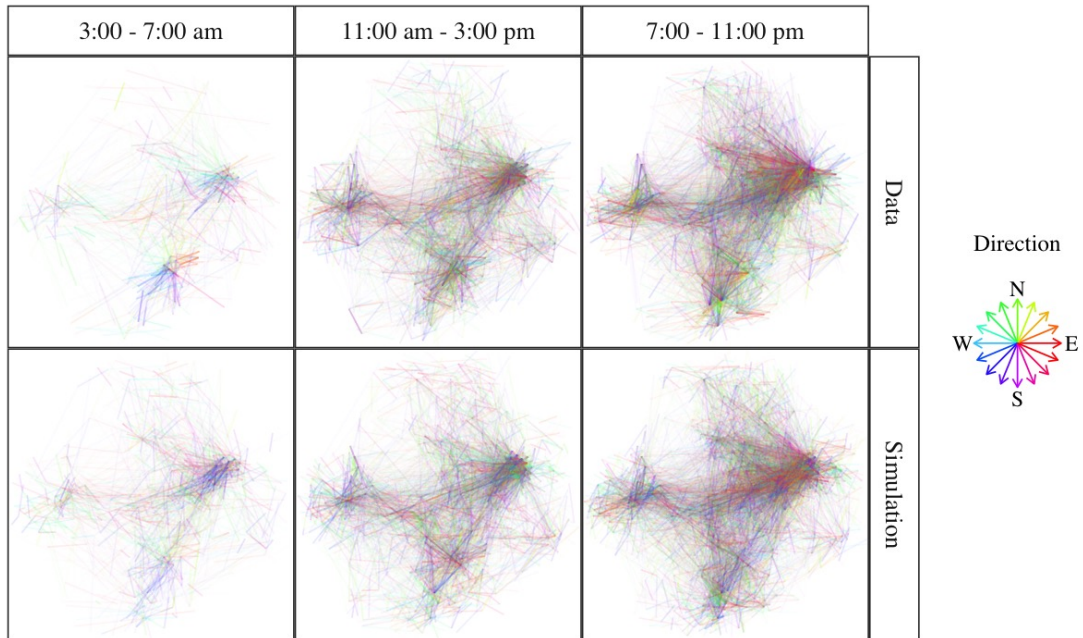
(d) Driver movements, simulation

Figure 23: Spatial patterns in the data and simulations

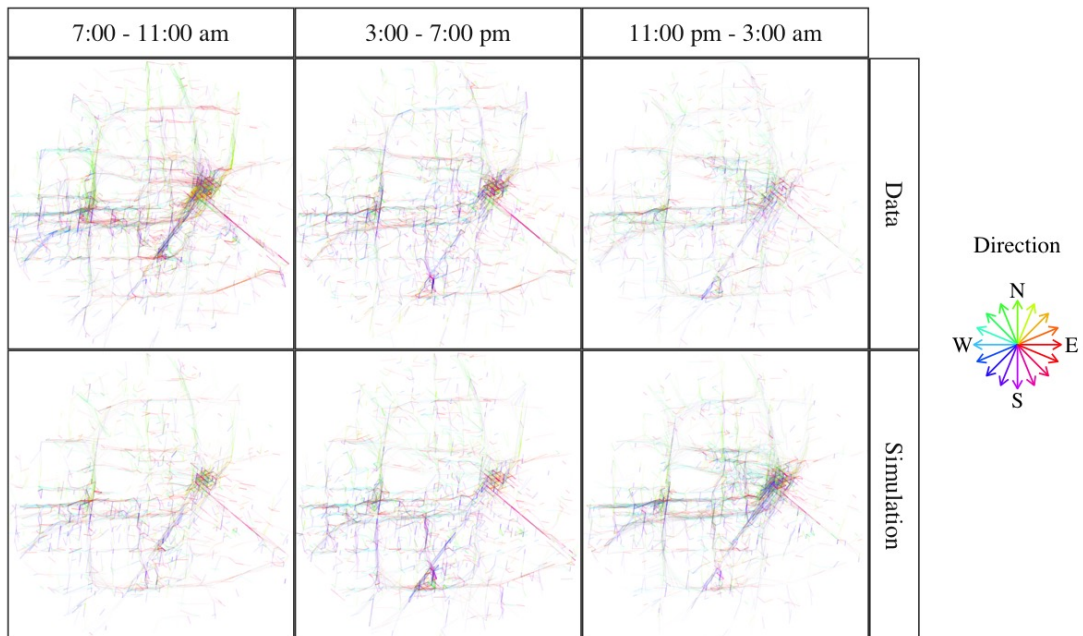
Note: Each figure shows a random subsample of 10,000 observations. Figures showing the simulation are for the status quo. For the first two figures, each line connects the origin and destination of a trip. For the first two figures, each line connects the locations in which an available driver was at the beginning and at the end of a two-minute period.

Figure 23 compares the main spatial patterns from my model, aggregated over time. The first two subfigures show the main pattern of trips. The last two sub-

⁴⁹It is much harder to match the supply in the data because it is a stock variable. It is not enough to include fixed effects in the entry process, as with demand, since the exit process depends endogenously on market behavior in equilibrium.



(a) Trips



(b) Driver movements

Figure 24: Trips and driver movements in simulations and in the data

Note: Subfigure (a) shows a 20% random sample of trips at certain hours from Monday to Thursday for one week. Each line connects the origin and destination of a trip. Similarly, subfigure (b) shows a random sample of 5% of the movements of available drivers. Each line connects the initial and final location of an available driver during one period. Colors represent the direction of movement.

figures show the main movement patterns followed by drivers. In both cases, the simulation follows the data closely.

Figure 24 shows the spatial patterns during the week at times that were not shown in figure 10. Similar patterns can be seen for weekend trips. The number of trips for the periods 7-11 am and 11 pm-3 am has a certain degree of mismatch. The reason is that it depends on the number of *available* drivers, which are only a small fraction of the total drivers. Any small mismatch in the number of drivers is amplified when measuring the number of available drivers.

Appendix F Additional counterfactual results

F.1 Robustness of welfare effects

Long-run elasticities In order to explore how sensitive my main results are to the values of the long-run elasticities, I fit the parameters of my model to higher elasticities. Based on the new fits, I rerun all the counterfactuals in order to obtain new estimates of the welfare effects of surge pricing.

Panel B in table 4 shows the result from that process. Every row represents one version of the market that has different values of the elasticities. These results can be compared to welfare results from the baseline market (panel A). The numbers vary across rows, but the main qualitative takeaways do not change. Surge pricing increases total welfare, and the main beneficiaries are riders. Driver surplus and Uber's short-run profits sometimes increase and sometimes decrease, but by small amounts. Its sign varies because of two opposing effects: an increase because of matching that reduces drivers' idle time, and a decrease because of lower prices. In most cases the price effect is more important, but with high demand elasticity the idle-time effect takes over.

Value of time I also measure welfare effects when modifying the parameters from the request model so that the value of time is lower. In some counterfactuals I scale up the price coefficient $\beta(r_i)$ by a factor of 1.5 or 2, and in some counterfactuals I scale down the ETA coefficient $\gamma(r_i)$ by a factor of 0.75 or 0.5.

Panel C in table 4 shows these results. Every row represents one version of the market that has different values of the trip request coefficients. The main qualitative takeaways all remain the same. Total welfare increases, and the main beneficiaries are riders. Driver surplus and Uber's short-run profits decrease by a small amount.

Table 4: Welfare effects of surge pricing with different model parameters

Market (1)	Demand elasticity (2)	Supply elasticity (3)	Fare coef. factor (4)	ETA coef. factor (5)	Total welfare (6)	Rider surplus (7)	Driver surplus (8)	Short-run profits (9)
<i>Panel A: Baseline</i>								
Baseline	-	-	-	-	3.53%	6.98%	-1.97%	-1.42%
<i>Panel B: Higher long-run elasticities</i>								
Higher supply elast.	-	1	-	-	4.37%	7.16%	-1.43%	-1.31%
	-	1.5	-	-	9.05%	10.20%	-0.34%	-0.81%
Higher demand elast.	-1	-	-	-	4.20%	4.87%	-0.25%	-0.44%
	-1.5	-	-	-	6.62%	6.26%	1.00%	-0.69%
Higher supply and demand elasticities	-1	1	-	-	6.27%	7.62%	-0.54%	-0.83%
	-1.5	1.5	-	-	12.87%	8.35%	2.41%	1.86%
<i>Panel C: Lower value of time for riders</i>								
Higher fare coef.	-	-	1.5	-	5.73%	8.69%	-1.77%	-1.16%
	-	-	2	-	6.14%	9.03%	-1.75%	-1.11%
Lower ETA coef.	-	-	-	0.75	4.88%	6.55%	-0.89%	-0.76%
	-	-	-	0.5	5.53%	7.45%	-1.05%	-0.85%
Higher fare coef. and Lower ETA coef.	-	-	1.5	0.75	5.46%	7.46%	-1.11%	-0.87%
	-	-	2	0.5	6.97%	8.88%	-1.07%	-0.83%

Note: This table shows the welfare effects of surge pricing when setting different values for some model parameters. Every row represents one version of the market with different parameters. Panel A represents the baseline market. Rows in panel B represent alternative markets with higher long-run elasticities. Rows in panel C represent alternative markets with a lower value of time for riders, in which the price coefficient is scaled up by some factor or the ETA coefficient is scaled down by some factor. For every alternative market I rerun all the counterfactuals that are necessary to measure the welfare effects shown in figure 16. Column (1) describes how the model differs from the baseline market. Columns (2)-(5) describe in detail how parameters are modified relative to the baseline market. A dash means that parameters are unchanged relative to the baseline market—i.e., elasticities are those from table 2, and coefficient factors are one. Columns (6)-(9) show effects on total welfare, rider and driver surplus, and the platform’s short-run profits as the market moves from an optimal uniform multiplier to the status quo.

F.2 Redistribution within riders and drivers

Drivers Figure 25 measures changes in drivers’ hourly earnings at different times of the week. Drivers who work during busy times that have higher prices, especially during Friday and Saturday, benefit from surge pricing. Drivers who work during less busy times with low prices, in particular Monday-Wednesday, are hurt by surge pricing.

Riders I run nonparametric fits of riders’ realized utility ($\max\{U_i, 0\}$) as a function of several variables for different counterfactuals. I then compare the model I fit for different counterfactuals to measure how the welfare effects of surge pricing (figure 26) vary along certain dimensions. I start by running models to measure

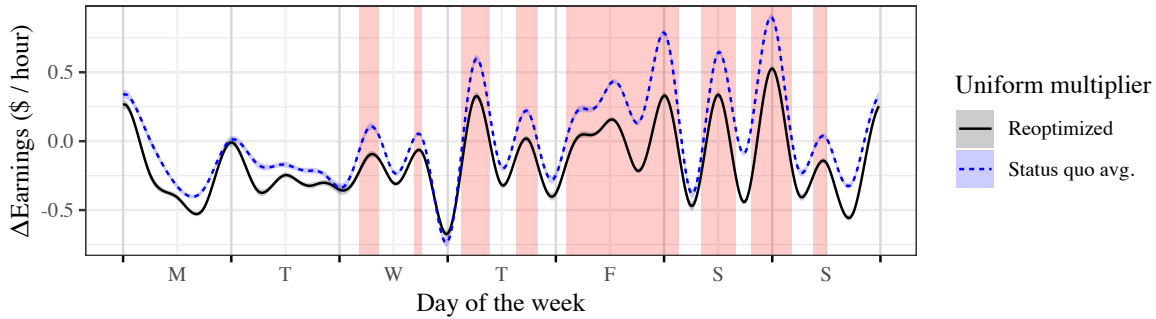


Figure 25: Heterogeneity in the welfare effect of surge pricing on drivers

Note: The black, solid line represents changes in earnings as pricing moves to the status quo from a uniform multiplier at Uber’s preferred level. The blue, dashed line represents changes in earnings as pricing moves to the status quo from uniform pricing at the average multiplier in the data. The lines are the differences between nonparametric fits. I run models of hourly earnings on the time of the week (the tensor product of cyclic cubic regression splines for time of the week and time of the day, using a cross-validated penalty). A shaded background indicates times when the average multiplier with surge pricing is above the mean of the full sample. For each counterfactual, I run a model based on simulated data for 15 weeks.

heterogeneity along proxies for drivers’ income. The first proxy is the price of the rider’s smartphone in November 2019 (see upper left panel). The second proxy is the median income in the census tract where the trip was requested (see upper right panel). This is a good measure of rider income if rich people tend to request trips from higher income neighborhoods. I find that although there is some slight heterogeneity along both variables, there is no increasing trend, as would be the case if surge pricing especially benefitted rich riders.

I run similar models to capture heterogeneity along riders’ willingness to pay for a trip with ETA zero— $\alpha(r_i, l, h) + \epsilon_i$ in equation (1). That allows me to tell which riders are willing to pay the most *within a given location and time period*. The middle left panel shows that riders with a high willingness to pay benefit most from surge pricing, but there is no level of willingness to pay at which riders are worse off. The effect at a very low willingness to pay is zero—these riders would not request a trip anyway. As willingness to pay approaches actual trip prices and people start requesting trips, the benefits from surge pricing start to increase. Some of these riders are priced out when multipliers are high, but they get a higher utility when prices are low. Furthermore, they get lower ETAs and they are less likely to be denied a trip. The net effect is that they are all ex-ante better off. Benefits then increase along the whole distribution.

The middle right panel of figure 26 shows how the effect of surge pricing on riders’ realized utility varies by base fare. Utility is measured as a percent of the

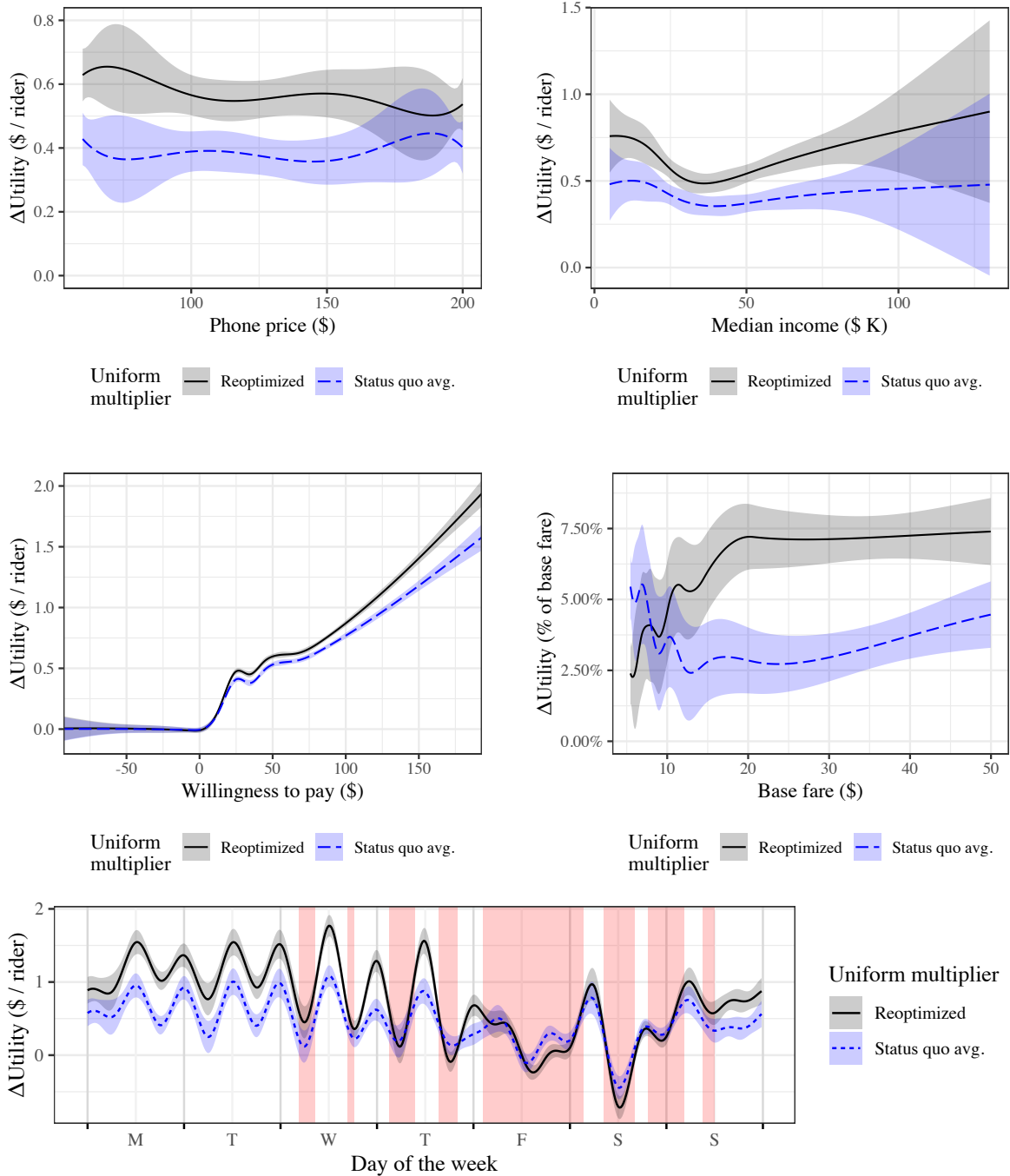


Figure 26: Heterogeneity in the welfare effect of surge pricing on riders

Note: These figures show changes in riders' realized utility as a function of several different variables. Black, solid lines represent changes as pricing moves to the status quo from a uniform multiplier at Uber's preferred level. Blue, dashed lines represent changes as pricing moves to the status quo from uniform pricing at the average multiplier in the data. The lines are the differences between nonparametric fits (see details in the main text). For each counterfactual, I fit a model based on simulated data for 15 weeks.

trip price. Before taking into account average price changes, riders with short trips benefit most: lower ETAs (which do not depend on the base fare) have a bigger impact on them as a percent of the base fare. After considering average price changes, riders who are going far are better off, because price effects take over.

To find heterogeneity by time of the week, I run models of realized utility on a tensor product of two cyclic cubic regression splines, one for the time of the week and one for the hour of the week. The subfigure on the bottom of figure 26 shows how the effect of surge pricing on riders' realized utility varies by the hour of the week. The vast majority of riders are better off. Only a few riders who want to request trips during a couple of hours when prices are highest, on Friday afternoon and Saturday midday, are hurt by surge pricing.

E.3 Evidence of wild-goose chases

In this section I show some evidence that the wild-goose chases (WGCs) (Castillo et al., 2018) are the reason why welfare and Uber's objective function are maximized at a higher average multiplier with a uniform price than with uniform pricing. Castillo et al. show that WGCs can be diagnosed using a simple descriptive statistic: *slack*, the ratio of available drivers to drivers that are picking up riders. It is a measure of the availability of drivers. WGCs take place when slack goes below a certain threshold that depends on the matching technology, but which is between 0.25 and 0.5.

Figure 27a shows the fraction of time that slack is below some threshold for different pricing policies. In order to compute these fractions, I aggregate the outcome of my simulation into half-hour periods. I only show results for thresholds of 0.25 and 0.5 for clarity, but similar, intermediate results can be seen for thresholds between these two values.

Regardless of the threshold, it is evident that WGCs (i.e., times with low slack) start taking place at higher prices with uniform pricing than with surge pricing. As Castillo et al. emphasize, WGCs result in a decline in welfare. Thus, my findings are consistent with welfare and Uber's objective function starting to decrease quickly after decreasing multipliers beyond a certain level that is higher for uniform pricing than for surge pricing. Furthermore, the gap between surge pricing and uniform pricing starts to widen precisely at the price level at which welfare and Uber's objective function start to drop down quickly relative to surge pricing (figures 12 and 15). This supports my claim that Uber sets higher prices with uniform pricing

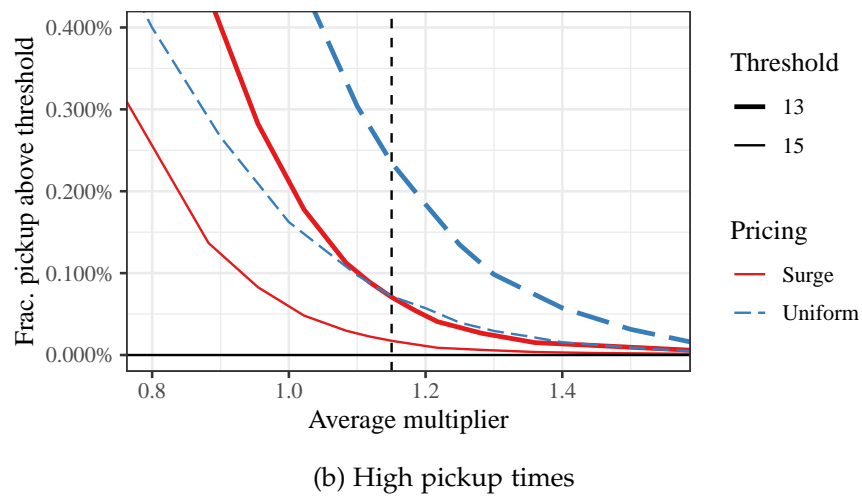
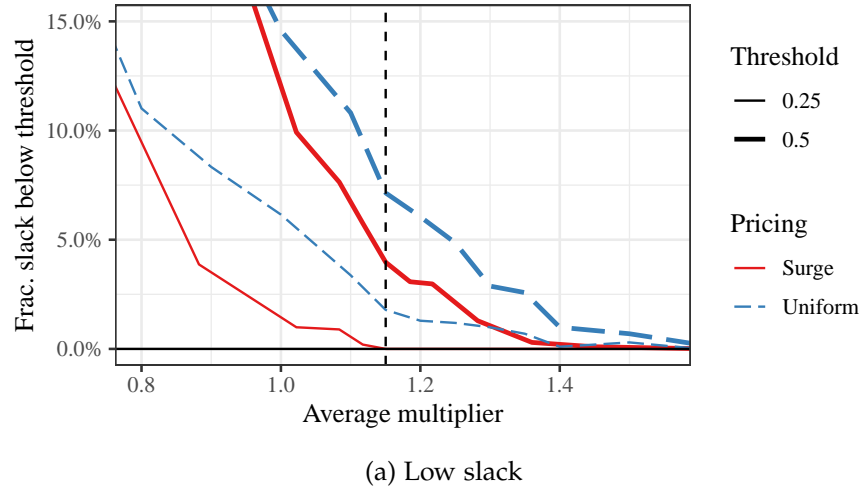


Figure 27: Evidence of wild-goose chases

Note: These figures show, for different pricing policies, how often slack—the ratio of available drivers to drivers that are picking up riders—is below some threshold and how often pickup times are above some threshold. Both types of event are telltale signs of wild-goose chases.

than with surge pricing because it wants to avoid entering WGCs.

I also show further evidence based on the number of extremely high pickup times, which quickly start becoming more frequent as the market enters WGCs. Figure 27b shows a similar pattern to figure 27a: high pickup times are extremely infrequent with high multipliers. But as prices go down they suddenly become frequent. The point at which they start taking place is higher with surge pricing than with uniform pricing.

Appendix G Additional empirical evidence

G.1 Hierarchical demand model

Suppose that riders' utility follows equation 7, where

$$w_{lt}^0 \sim N(\mu_{kh}, \nu), \quad w_i \sim N(w_{lt}^0, \sigma). \quad (23)$$

The mean μ_{kh} represents some mean by location group k and hour of the week h , and w_{lt}^0 represents a location by time period random effect.

This model avoids plugging in \bar{w}_{lt} as an estimator for w_{lt}^0 as in my main model, which creates some bias because of the combination of two elements: (a) the main model is nonlinear, and (b), the number of observations within every group lt is small so I cannot rely on \bar{w}_{lt} being a consistent estimator for w_{lt}^0 . This new model, however, has the drawback that it assumes random effects. In other words, it assumes no correlation between w_{lt}^0 and my main model variables.

I estimate this new model by two-step maximum likelihood. In the first step, I estimate σ^2 , ν^2 , and μ_{kh} based on the observed values of w_i , i.e.,

$$\hat{\mu}_{kh} = \sum_{i \in kh} w_i, \quad \hat{\sigma}^2 = \frac{\sum_{lt} (n_{lt} - 1) s_{lt}^2}{\sum_{lt} (n_{lt} - 1)}, \quad \hat{\nu}^2 = s^2 - \hat{\sigma}^2 - \frac{1}{n} \sum_i (\bar{w}_{kh} - \bar{w})^2 \quad (24)$$

where $\bar{w} = \frac{1}{n} \sum_i w_i$, $\bar{w}_{lt} = \frac{1}{n_{lt}} \sum_{i \in lt} w_i$, $s^2 = \frac{1}{n} \sum_i (w_i - \bar{w})^2$, $s_{lt}^2 = \frac{1}{n_{lt} - 1} \sum_{i \in lt} (w_i - \bar{w}_{lt})^2$, n is the total number of observations, and n_{lt} is the number of observations within group lt .

In the second step, I maximize the conditional likelihood

$$L(\theta, \hat{\sigma}, \hat{\nu}, \hat{\mu}_{kh}) = \sum_{lt} \int \sum_{i \in lt} \Lambda(u_i(w_{lt}^0))^{y_i} (1 - \Lambda(u_i(w_{lt}^0)))^{1-y_i} dF(w_{lt}^0 | \mathbf{w}, \hat{\sigma}, \hat{\nu}, \hat{\mu}_{kh}), \quad (25)$$

where y_i is an indicator variable for whether the rider requested a trip, and Λ is the logistic function. The conditional distribution of w_{lt}^0 is normal with mean $\frac{\hat{\mu}_{kh} + \frac{n_{lt} \bar{w}_{lt}}{\sigma^2}}{\frac{1}{\nu^2} + \frac{n_{lt}}{\sigma^2}}$ and variance $\frac{1}{\frac{1}{\nu^2} + \frac{n_{lt}}{\sigma^2}}$. I compute the integral by quadrature. In order to ensure the optimization converges quickly—which is challenging given the high dimensional parameter space—I use an iteratively reweighted least squares algorithm.

Table 5 reports the estimates from this model. Most of the parameters are very close to the ones for the main model in table 3. The only noticeable difference is that the coefficient for the interaction of ETA and the base fare changes sign. This

Table 5: Estimates of the parameters of the hierarchical demand model

<i>Dependent variable:</i>	
Request	
Base fare	-0.0288*** (0.0027)
Fare	-0.0466*** (0.0122)
Fare × base fare	0.0011 (0.0008)
ETA	-0.0947*** (0.0127)
ETA × base fare	-0.0035 (0.0021)
Observations	650,233
<i>Note:</i> *p<0.1; **p<0.05; ***p<0.01	

Note: Estimates of the main parameters of the hierarchical demand model. The price and ETA coefficients are evaluated at the mean of the base fare.

means that riders' value of time increases more quickly as the trip length increases.

G.2 Effect of past prices and ETAs on requests

I model rider's behavior as a static decision. One important question is to what extent this simplifying assumption biases my estimates for their response to price and ETA changes. In order to measure this, I build a dataset with one observation for every two minute period during which a rider interacted with the app. I focus on the 79% of sessions in which the rider interacted at least twice with the app.

My goal is to determine what affects whether riders decide to request a trip, leave the app, or wait until a later period (i.e., interact again with the app in the following half hour). I run regressions of dummies for each one of these decisions on my main demand model variables, as well as their lags, i.e., the last value that the rider observed. In order to make sure I am measuring a causal relation, I include a flexible function of the recommended multipliers surrounding the rider, the average ETA at the location by time period level, as well as lags of both of these variables.

Table 6 shows the result of this exercise. Columns (1), (3), and (5) show regressions that do not include lags. Higher prices and ETAs both lead to fewer requests and a higher probability of staying. In other words, it seems that at least some riders do choose to wait until later times. The coefficients on leaving are not significant. Columns (2), (4), and (6) also include the lags of the main covariates. The

coefficients are almost identical to those without lagged variables, which suggests that not modeling the full dynamic process does not have a major effect on the estimates of my model.

Table 6: Effect of previous prices and ETAs on probability of request

	<i>Dependent variable:</i>					
	Request		Leave		Stay	
	(1)	(2)	(3)	(4)	(5)	(6)
Base fare	0.0057*** (0.0006)	0.0060*** (0.0006)	-0.0071*** (0.0006)	-0.0073*** (0.0006)	0.0014** (0.0007)	0.0013* (0.0007)
Fare	-0.0082*** (0.0030)	-0.0079*** (0.0029)	-0.0045 (0.0034)	-0.0045 (0.0034)	0.0127*** (0.0038)	0.0124*** (0.0038)
Fare lag		0.0007 (0.0034)		-0.0023 (0.0034)		0.0016 (0.0034)
Fare × base fare	0.0004 (0.0003)	0.0004 (0.0003)	-0.0001 (0.0002)	-0.0001 (0.0002)	-0.0004 (0.0003)	-0.0004 (0.0003)
Fare lag × base fare		-0.00003 (0.0002)		-0.00002 (0.0003)		0.0001 (0.0003)
ETA	-0.0210*** (0.0031)	-0.0205*** (0.0030)	0.0039 (0.0033)	0.0039 (0.0033)	0.0171*** (0.0031)	0.0166*** (0.0031)
ETA lag		-0.0032 (0.0025)		0.0025 (0.0023)		0.0007 (0.0032)
ETA × base fare	-0.00003 (0.0002)	-0.00004 (0.0002)	-0.0001 (0.0002)	-0.0001 (0.0002)	0.0001 (0.0002)	0.0001 (0.0002)
ETA lag × base fare		-0.00001 (0.0003)		0.0001 (0.0002)		-0.00005 (0.0002)
Control function	✓	✓	✓	✓	✓	✓
Control fn. of lags		✓		✓		✓
Loc. × week hour FE	✓	✓	✓	✓	✓	✓
Observations	291,504	291,504	291,504	291,504	291,504	291,504

Note:

*p<0.1; **p<0.05; ***p<0.01

Note: Linear regressions of indicators for whether a rider requests a trip, leaves the app, or waits to make a decision later. I only include the subsample of observations in which drivers had already interacted with the app at least once during the previous half hour. Columns (1), (3), and (5) include the main covariates in my demand model. Columns (2), (4), and (6) also include the lags of the main covariates (prices and ETAs), i.e., the last value the rider observed. To measure causal effects, all regressions include a flexible function of nearby recommended multipliers and average ETAs by location and period, and regressions with lags include a flexible function of the lags of these variables. Standard errors are computed by clustering at the location by hour of the week level.

G.3 Temporal correlation of multipliers

Figure 28 shows the temporal correlation of multipliers. Each point is the coefficient of a regression of the surge multiplier on a lag of itself and location by hour of the

week fixed effects. Thus, it measures the persistence of unexpected variation in multipliers. There is significant correlation for the first ten minutes. After that, the correlation settles down at around 0.15.

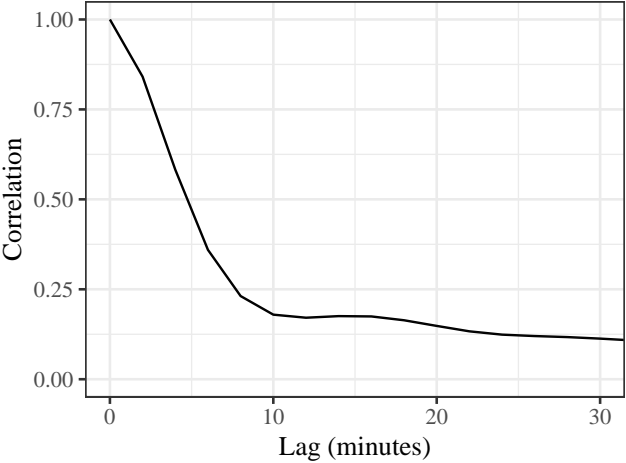


Figure 28: Temporal correlation of multipliers

Note: This figure shows an autocorrelation plot of the residuals of surge multipliers after controlling for location by hour of the week fixed effects. All correlations are so precisely estimated that confidence intervals are within the line for the point estimate.

G.4 Effect of multipliers on expected earnings

In this section I present evidence of the effect of surge multipliers on expected earnings. In order to do so, I estimate regressions of drivers’ net earnings for the next h hours as a function of the multiplier in the driver’s location and nearby locations. I include location and hour of the week fixed effects in every regression.

Figure 29 plots the main coefficient from these regressions. The effect is larger for the three closest hexagon rings than only for the local multiplier, suggesting that the local multiplier does not capture all information. But the difference is very small if one also includes two additional rings, suggesting that the three closest rings capture most of the information.

For all three series, the effect increases quickly as the time horizon increases, but it starts to level off after one hour.

G.5 Short-run entry and exit

In this section I show that there is no empirical evidence that drivers respond to unexpected changes in multipliers by entering or leaving the market. I first run

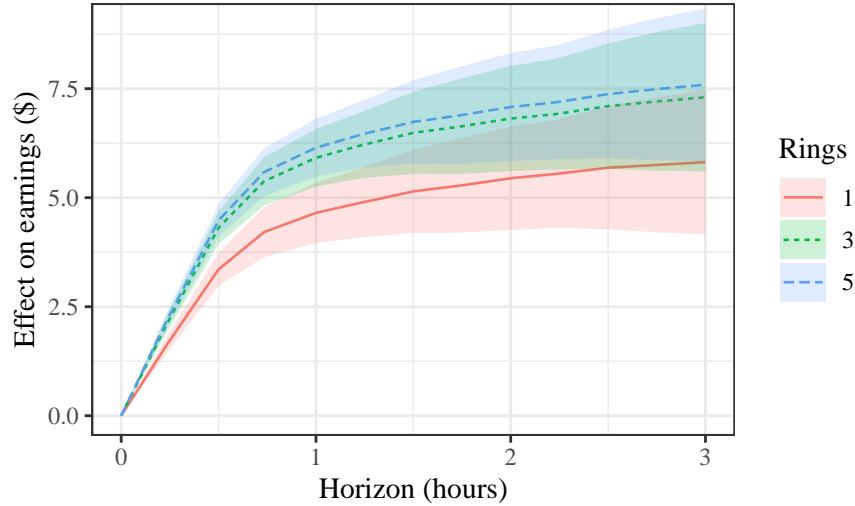


Figure 29: Effect of multipliers on expected earnings

Note: This figure shows the main coefficient for regressions of hourly earnings for the next h hours (the *horizon*) on current multipliers. When rings=1, the covariate is the local multiplier. For other values of rings, the covariate is the average multiplier among locations within a certain number of hexagon rings. All regressions include location by hour of the week fixed effects, and standard errors are computed with two-way clustering by location and hour of the week.

a regression for whether open drivers decide to leave or keep on working as a function of the multiplier in his location. I also run similar regressions, where instead of the multiplier in the driver’s location, the main covariate is the average multiplier in all locations within three or five hexagons. The averages give higher weights to the nearest locations. In order to measure a causal effect, I control for the recommended multipliers and for the unrounded multipliers in nearby locations. Table 7 shows that there is no evidence of a causal effect. For reference, the fraction of available drivers who leave in the whole data is 10.6%.

G.6 Additional results from elasticity experiments

Table 8 reports the result of regressions of trips per rider on treatment dummies. The estimates are consistent for the average treatment effect given that drivers were assigned randomly to their groups. As we can see, there is some heterogeneity across cities. We can also see that Mexico city shows an unexpected positive elasticity for the 10% treatment group. I thus exclude it from the main Poisson regression used to calibrate the long run demand elasticity parameter ρ .

Table 9 reports the result of regressions of a dummy for working and hours worked on treatment dummies. Columns (1)-(5) report results for each city. The

Table 7: Effect of multipliers on probability of driver exit

	Dependent variable: Left		
	(1)	(2)	(3)
Multiplier	0.018 (0.028)		
3-ring avg. mult.		-0.010 (0.036)	
5-ring avg. mult.			0.003 (0.036)
Loc. × week hour FE	✓	✓	✓
Rec. mult. controls	✓	✓	✓
Unrounded mult. controls	✓	✓	✓
Observations	1,218,286	1,218,286	1,218,286
Note:	*p<0.1; **p<0.05; ***p<0.01 Standard errors clustered by location and hour of the week		

Note: Linear regressions of an indicator for whether available drivers decide to leave as a function of multipliers. I control for recommended and unrounded multipliers, so that the main coefficient is identified from variation due to rounding in the surge pricing algorithm. The main covariate in column (1) is the multiplier at the driver’s location. In columns (2) and (3), the main covariate is an average of the multipliers within 3 and 5 hexagon rings surrounding the driver’s location. Standard errors are clustered by location and hour of the week.

Table 8: Average treatment effects in demand experiment

	Dependent variable: Trips per rider				
	Mexico City (1)	Guadalajara (2)	Rio (3)	Sao Paulo (4)	Belo Horizonte (5)
Constant	1.0470*** (0.0125)	1.2558*** (0.0138)	0.9200*** (0.0110)	0.7142*** (0.0098)	0.8179*** (0.0104)
10% discount	-0.0666** (0.0264)	0.0795** (0.0331)	0.0624** (0.0255)	0.0247 (0.0228)	0.0534** (0.0244)
20% discount	0.0459 (0.0282)	0.1616*** (0.0341)	0.1883*** (0.0280)	0.0422* (0.0226)	0.1800*** (0.0265)
Observations	44,070	44,228	44,359	44,382	44,380
Note:	*p<0.1; **p<0.05; ***p<0.01				

Note: Regressions of the number of trips taken by each rider on dummies for treatment group.

results are only significant for Rio de Janeiro. However, they are significant once I pool all cities.

Table 10 shows the main supply elasticity estimate, as well as a Poisson regression of the number of hours worked on the log earnings factor. As we can see, both coefficients are almost identical, suggesting that drivers do not respond by working

Table 9: Average treatment effects in supply experiment

	<i>City:</i>						
	Mexico City (1)	Guadalajara (2)	Rio (3)	Sao Paulo (4)	Belo Horizonte (5)	All cities (6)	All cities (7)
<i>Panel A. Dependent variable: Worked dummy</i>							
Constant	0.4823*** (0.0041)	0.4686*** (0.0042)	0.4079*** (0.0046)	0.3708*** (0.0045)	0.4065*** (0.0047)		
Treated	0.0160* (0.0084)	0.0136 (0.0085)	0.0173** (0.0086)	-0.0051 (0.0079)	0.0191** (0.0086)	0.0125*** (0.0038)	0.0125*** (0.0038)
City FEs						✓	
City × day of week FEs							✓
Observations	73,500	73,500	59,500	52,500	59,500	318,500	318,500
<i>Panel B. Dependent variable: Hours worked</i>							
Constant	3.2920*** (0.0343)	2.6268*** (0.0290)	2.2418*** (0.0312)	2.0771*** (0.0322)	2.4149*** (0.0365)		
Treated	0.0807 (0.0703)	0.0835 (0.0596)	0.1508** (0.0594)	-0.0712 (0.0548)	0.1030 (0.0657)	0.0713** (0.0280)	0.0713** (0.0280)
City FEs						✓	
City × day of week FEs							✓
Observations	73,500	73,500	59,500	52,500	59,500	318,500	318,500
<i>Note:</i>	*p<0.1; **p<0.05; ***p<0.01 Robust standard errors clustered by driver						

Note: Panel A shows estimates of regressions of a dummy for whether the driver worked during a given day on indicators for the treatment group. Panel A shows estimates of regressions of the number of hours worked during a given day on indicators for the treatment group.

more hours per day.

Table 10: Long run demand elasticities

	<i>Dependent variable:</i>	
	Worked dummy (1)	Hours worked (2)
Log of earnings factor	0.383*** (0.099)	0.401*** (0.123)
City × day of week FE	✓	✓
Observations	266,000	266,000
<i>Note:</i>	*p<0.1; **p<0.05; ***p<0.01 Robust s.e. clustered by driver	

Note: Poisson regressions of a dummy for whether the driver worked during a given day and of the number of hours worked during a given day. The dependent variable is the log of the earnings factor, i.e., $\log(0.9)$ for treated drivers and $\log(1)$ for control drivers.

G.7 Flexible controls

Table 11 shows linear regressions for a dummy of requesting a trip as a function of the surge multiplier and the ETA. I estimate equation (7), with the only difference that the left hand side variable is a dummy for requesting a trip. It is clear that the high dimensional spline in my main demand estimation and location by hour of the week fixed effects result in very similar coefficients.

Table 11: Linear regressions for demand

	<i>Dependent variable:</i>		
		Request	
	(1)	(2)	(3)
Base fare	-0.0026*** (0.0009)	-0.0046*** (0.0010)	-0.0039*** (0.0006)
Fare	-0.0048*** (0.0006)	-0.0098** (0.0050)	-0.0076** (0.0033)
Fare × base fare	-0.00001 (0.00002)	0.0002 (0.0003)	0.0001 (0.0002)
ETA	0.0029** (0.0014)	-0.0171*** (0.0047)	-0.0172*** (0.0032)
ETA × base fare	-0.00004 (0.0005)	-0.0001 (0.0004)	-0.0002 (0.0003)
Control function		✓	✓
Coord. and week hour spline	✓	✓	
Loc. × week hour FE			✓
Observations	650,233	650,233	650,233

Note: *p<0.1; **p<0.05; ***p<0.01

Note: Estimates of linear regressions of the form of 7. Column (1) omits the control function. Column (2) uses the same form as in main demand estimation. Column (3) uses location by hour of the week fixed effects instead of splines. Standard errors are computed by two-way clustering by location and hour of the week.

Appendix H Surge pricing and app interface

H.1 Typical multiplier pattern

Figure 30 shows the progression of multipliers during one particular Tuesday afternoon.

H.2 App interface

Figure 31 shows what riders and drivers observe in the version of the app that was active in the period of analysis.

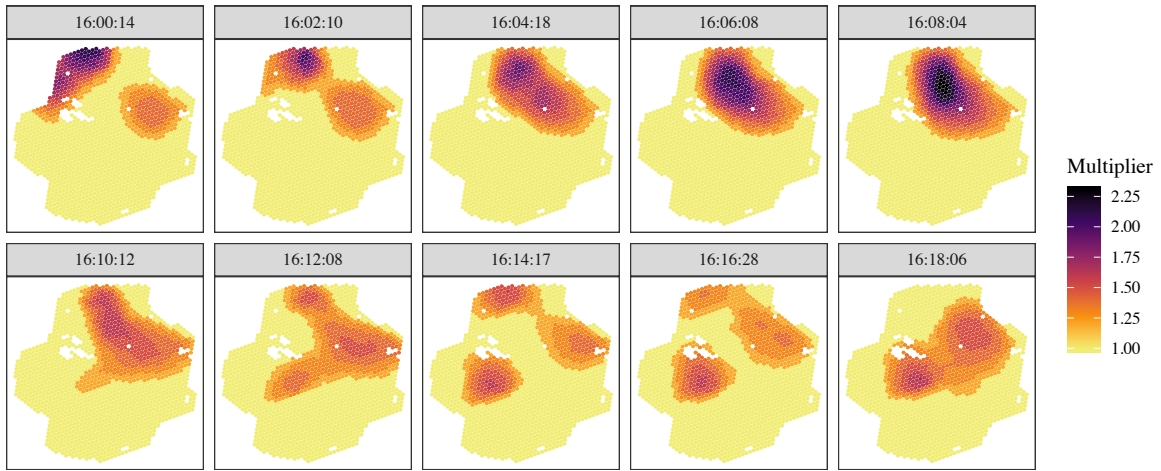
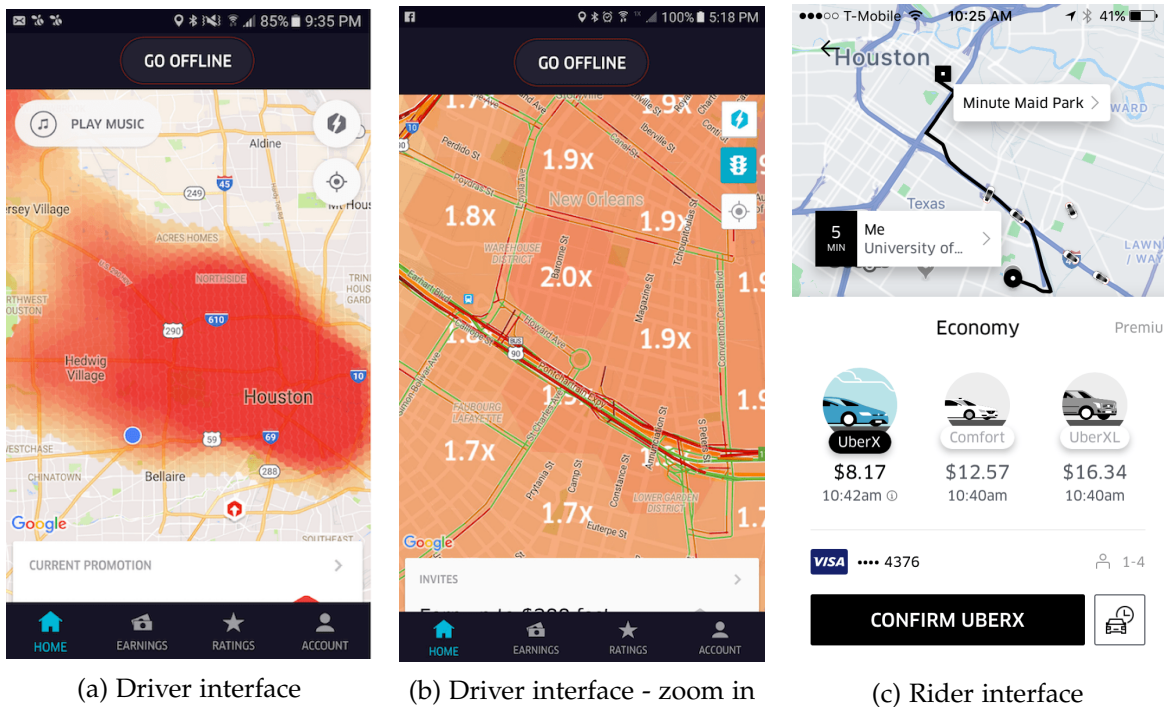


Figure 30: Surge multipliers during the afternoon of March 21, 2017



(a) Driver interface

(b) Driver interface - zoom in

(c) Rider interface

Figure 31: Screenshots of the app interface

Note: Subfigure (a) shows what drivers observe when they are available. Subfigure (b) shows how it looks when they zoom in. Subfigure (c) shows what riders see when they choose a destination.