Who Benefits from Surge Pricing?*

Juan Camilo Castillo†

October 26, 2020

Abstract

In the last decade, new technologies have led to a boom in real-time pricing. I analyze the most salient example, surge pricing in ride hailing. Using data from Uber, I develop an empirical model of spatial equilibrium to measure the welfare effects of surge pricing. The 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 1.59% of gross revenue. Welfare effects differ substantially across sides of the market: rider surplus increases by 5.25% of gross revenue, whereas driver surplus and platform profits decrease by 1.81% and 1.77% of gross revenue, respectively. Riders at all income levels benefit, while disparities in driver surplus are magnified.

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 Benkard for their invaluable advice and support. I am also grateful to Tim Bresnahan, Nick Buchholz, Emma Harrington, Caroline Hoxby, Brad Larsen, Jonathan Levin, Aviv Nevo, 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, Elizabeth Mishkin, Helin Zhu, and several other people at Uber whose support and feedback made this project possible. This research was supported by the Kapnick Foundation Fellowship through a grant to the Stanford Institute for Economic Policy Research.

†Economics Department, University of Pennsylvania. E-mail: jccast@upenn.edu

1 Introduction

Until about ten years ago, few companies adjusted prices in real time in response to supply and demand. Such cases were mostly limited to a couple of industries, such as airlines, hotels, and electricity. New technologies, however, have led to rapid changes. Companies can now use the internet and smartphones to communicate prices instantly, and they can use rich datasets to design better pricing algorithms. Consequently, more and more companies set prices in real time, especially in two-sided markets and e-commerce. This can be desirable from an efficiency point of view, since more flexible prices allow markets to clear. However, real-time pricing often hurts some market participants, raising concerns about redistribution.

Ride-hailing platforms like Uber and Lyft have become the most salient adopters of real-time pricing—or surge pricing, as Uber calls it. Despite potential efficiency gains,[1] surge pricing has received widespread criticism. Some people suggest that it can hurt riders, calling it a form of price discrimination, or even price gouging (Dholakia, 2015). Others suggest that it can hurt drivers, whose earnings might drop too low unless they carefully plan their actions around surge pricing (Goncharova, 2017). In response to these concerns, cities like Honolulu, New Delhi, and Singapore have banned or capped surge pricing (Puckett, 2018; Kazmin, 2016; Yee, 2018). Some ride-hailing companies have also voluntarily chosen to avoid surge pricing. DiDi—the largest platform in China—stopped using surge pricing, instead adopting potentially inefficient queuing mechanisms. 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 gains and who loses from surge pricing. The model allows me to measure the welfare effects—on riders, drivers, and the platform—of moving from uniform pricing, where prices only depend on trip distance and duration, to surge pricing. The model is composed of three main parts: demand, supply, and a matching technology. On the demand side, riders decide whether to open the app and whether to

[1] These potential gains have prompted cities like New York to consider surge pricing for taxis (Rosenthal, 2020).

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 who request trips are matched, and, thus, how long riders need to wait for pickup. The model accounts for high-resolution spatial and temporal heterogeneity as well as transient random shocks. I integrate all pieces into a model of spatial equilibrium that allows me to simulate market behavior under alternative pricing policies.

I estimate the model using Uber data from Houston in March-April 2017, a period during which Uber was the only ride-hailing platform. I observe highly detailed data on drivers and riders, which allows me to estimate supply and demand directly from the data.[2] I identify agents' short-run elasticities—how real-time price changes affect drivers' movements and riders' decisions 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 measure long-run elasticities—how riders' and drivers' decisions to log in to the app respond to changes in expected prices—from experiments run by Uber and by Angrist et al. (2020).

My estimates imply that riders are very inelastic in the short run. 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. High-income riders are even less price sensitive, and time is especially valuable to them. With regard to drivers, I find that they are more likely to move to areas with high surge pricing. 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 that match precisely the distribution of the surge multiplier, a scale factor for prices that determines surge pricing.

I find that surge pricing increases total welfare by 1.59% of gross revenue—or \$0.19 per trip—relative to uniform pricing. This reflects the fact that surge pricing brings efficiency gains to the market. However, the effects on different sides of the

[2]Previous works on taxi markets such as Buchholz (2018) and Frechette et al. (2019) do not observe riders, so they rely on structural assumptions to back out demand.

market are strikingly dissimilar. Whereas rider surplus increases by 5.25% of gross revenue, driver surplus and Uber's profits decrease by 1.81% and 1.77% of gross revenue, respectively.[3] The fact that Uber chooses to use surge pricing implies that it is willing to forgo short-run profits to increase rider surplus. That is plausible if Uber believes that shareholder value and long-run profits are more closely tied to short-run rider surplus—which drives customer retention—than to short-run profits and driver surplus. Consistent with this interpretation, I find that the overall price level in the data is well below the level that maximizes short-run profits, but it is exactly at the point that maximizes rider surplus.

The asymmetry in welfare effects across riders and drivers can be decomposed into three parts. First, surge pricing saves people's time as it mitigates imbalances between supply and demand: riders are picked up more quickly, and drivers wait less between trips. But these time savings are much more valuable for riders. The value of time to drivers is their average hourly earnings net of driving costs, which is slightly above minimum wages. Riders, on the other hand, request trips during time-sensitive moments, so their welfare gains from time savings are substantially larger than those of drivers.

Second, surge pricing allocates trips more efficiently, but this only benefits riders. At times of driver scarcity, uniform pricing allocates trips randomly: only riders who 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, increasing rider surplus. Drivers, on the other hand, see no benefit from a better allocation of trips. The value of getting a trip—relative to being idle—is fairly homogeneous since it is mainly driven by earnings from the trip. In my model, it only differs across drivers with the distance to the rider: drivers that are closer need less time to complete the trip. Surge pricing does not help reallocate trips towards drivers that are closer, so it does not improve the way trips are allocated to drivers.[4]

Third, surge pricing allows Uber to set lower average prices, decreasing driver surplus and platform profits. I assume that if Uber is not allowed to do surge

[3]These three numbers do not add up exactly to the 1.59% total welfare increase. The remaining 0.08% corresponds to a decrease in tax revenue from a 2% sales tax.

[4]Drivers 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.

pricing it chooses a uniform multiplier that maximizes rider surplus—consistent with the fact that, in the data, rider surplus is maximized at the overall price level. The optimal uniform multiplier is above the average multiplier in the data. 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 hurts all market participants when drivers are scarce (i.e., when prices are too low). With uniform pricing, the only way to avoid the matching failure is with a high multiplier at all times. With surge pricing, on the other hand, the platform can set a lower average price, and surge pricing automatically increases prices during high demand times, avoiding the problem.

I also analyze the distributional effects within riders and within drivers. Surge pricing does not hurt any riders, regardless of their income. In fact, low-income riders benefit the most. They gain from lower prices, shorter pickup times, and more reliable trips. High-income riders also benefit, but they would prefer higher prices, which would further shorten pickup times and make trips more reliable. On the drivers' side, I find that surge pricing makes 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.

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—longer pickup times and lower reliability without surge pricing—and because they are unaware that they would pay higher average prices without surge pricing. 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 welfare results are entirely driven by features of the matching frictions—lost time, an inefficient allocation of trips, and pricing to avoid a matching failure. This highlights that understanding two-sided markets in which search and matching are central—such as labor, home rental, and e-commerce platforms—requires modeling those frictions carefully. This stands in contrast with the theoretical lit-

erature on two-sided markets, which assumes simple reduced forms for matching, and concludes that most results are instead driven by elasticities and cross-market externalities (Rochet and Tirole, 2003; Armstrong, 2006; Weyl, 2010).

One limitation of this work is that it focuses on a market that has a single ride-hailing platform. This provides a clean environment in which to analyze surge pricing in the absence of competition, but I am not able to say how competition might affect the welfare effects of surge pricing. Answering that question would require either merging data from two platforms or assumptions about multi-homing behavior.

Related work A few related papers analyze the welfare effects of surge pricing. 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 times, nor do they model spatial heterogeneity, which limits the extent to which they can account for matching frictions. 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 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 design of the algorithm as a given and analyze its effect on 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. Ghili and Kumar (2020) and Shapiro (2018) focus on economies of density. 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 and whether they request trips; thus, I estimate demand directly from the data.

A number of papers estimate demand and supply in ride-hailing markets. 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. (2020) estimate the value of time for riders in Prague. Their estimates are lower than mine, but they are also above median wages and well above minimum wages. Papers that estimate supply elasticities include Angrist et al. (2020) and Chen et al. (2020), who use field experiments to estimate the degree to which drivers value flexibility, and Lu et al. (2018), who, on the basis of an outage in the Uber platform, estimate drivers' short-run response to surge pricing. 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. Most theoretical works assume simple reduced forms for matching frictions (Rochet and Tirole, 2003; Armstrong, 2006; Weyl, 2010). In contrast, my findings are driven by a form for matching frictions that I microfound and estimate from the data, as in empirical works like Cullen and Farronato (2018), Fradkin (2017), and Dinerstein et al. (2018).

Roadmap My analysis begins in section 2, where I introduce the Uber market and describe the data. I introduce the model in section 3. In section 4 I present descriptive evidence that shows the variation that identifies the 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.

2 Setting and data

I analyze the Uber market in Houston between March 16 and April 8, 2017.[5] I focus on trips that start in the area of central Houston shown in figure 1a. It covers 8.63% of the area of the city, but it accounts for 56.4% of trips and 78.4% of trips with surge pricing. 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 the period of analysis.[6] In total, the sample includes around half a million trips.

Lyft was not present in Houston between November 2014 and May 2017.[7] It was the one large American city where Uber was the only ride-hailing platform. Thus, Houston is a clean setting in which to analyze the welfare effects of surge pricing without having to consider competition between platforms. For that reason, my results speak to a market with only one ride-hailing platform.

Riders, drivers, and trips The raw data has a high temporal and spatial resolution: I observe riders and drivers—with anonymized identifiers—every few seconds whenever the app is open, and I observe their location up to the precision of the cell phone GPS. I aggregate the data at the level at which surge pricing varies: into two minute *periods*, which I index by $t \in T$, and into the hexagons in figure 1a, which I call *locations* and index by $l \in L$. There are 1681 such hexagons in the region of analysis; each one is roughly 400 meters across. This level of aggregation is much finer than in previous papers, such as Buchholz (2018) and Frechette et al. (2019).[8] A higher resolution is important to capture the short run, local imbalances surge pricing is meant to counteract.

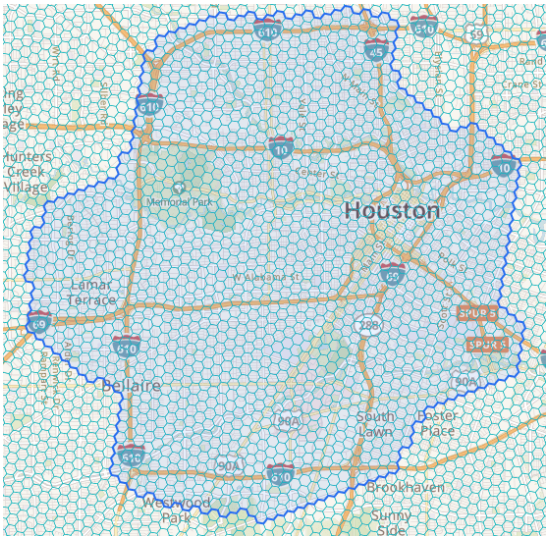
I observe riders whenever the app is open. I see the rider's location, the selected destination (if she has chosen one), the fare for a trip to her destination, and an

[5]Before this period, Uber did not record data on 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.

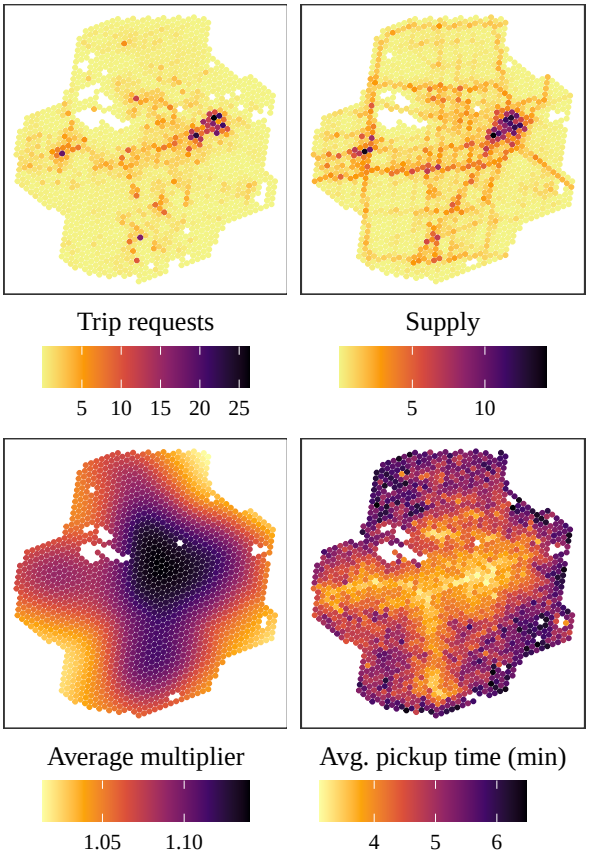
[6]Other Uber products, such as UberXL (larger cars) and UberBlack (luxury cars), accounted for less than 5% of trips.

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

[8]Frechette et al. (2019) split Manhattan into eight areas and use one-hour periods. Buchholz (2018) splits Manhattan into 48 areas and uses five-minute periods.



(a) Region of analysis



(b) Main spatial patterns in the data

Figure 1: Region of analysis and main spatial patterns

Note: Subfigure (a) shows a map of central Houston. The shaded area represents the region of analysis. Surge pricing varies spatially at the hexagon (or *location*) level. Subfigure (b) shows the main spatial patterns in the data. The upper panels show the number of trips that take place from every location and the average number of drivers waiting in every location (both are normalized to have mean 1). The bottom panels show the average surge multiplier and pickup time by location.

estimated time to pickup (or *pickup time*, for short). 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.[9] I take the multiplier, fare, and pickup time to be the last ones the rider observed before deciding to request or leave.

I do not observe rider demographics. However, I use the trip history to identify riders' home location, which I match to median income by census tract (see appendix D.1). I call this variable *income*, with the understanding that it is only a proxy. I identify with confidence the home of riders that account for 87.3% of trips.

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

These riders tend to take many trips and have a high conversion rate, so I call them *frequent riders*. The remaining trips were taken by *occasional riders*, who took too few trips to allow me to determine where they live.[10]

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 the 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.

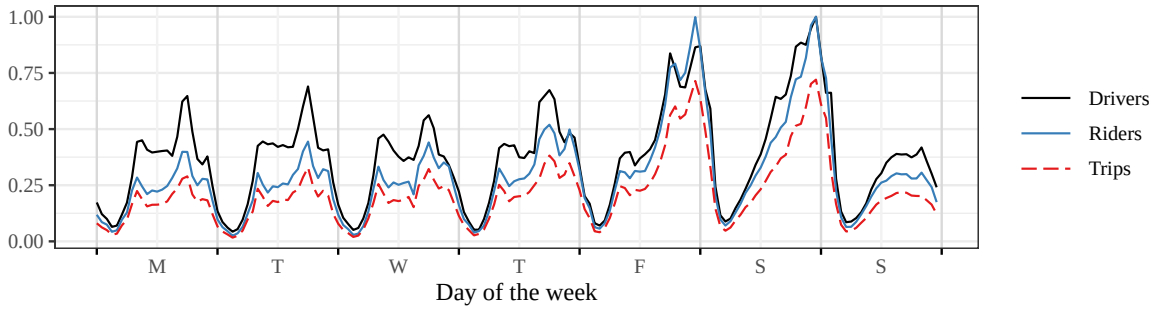


Figure 2: Weekly patterns in demand, supply, and trips

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 over the course of a week. Riders and drivers are normalized to have a maximum of one. Trips are on the same scale as riders.

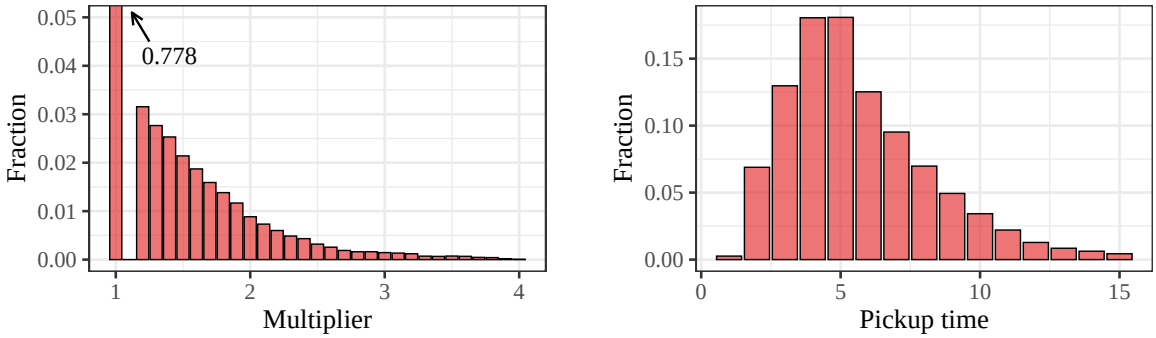
Figure 2 plots the average market behavior as the week goes by. There is low activity at night and high activity during the day, with spikes during rush hours. The least busy day is Sunday, and Friday and Saturday are the busiest days. All three variables behave similarly, with two noticeable differences: there is excess supply around noon during weekdays and excess demand during Friday and Saturday evenings.

The upper panels in subfigure 1b show spatial patterns for supply and demand. The area with most trip requests, towards the northeast, is Downtown. Two other high demand areas are The Galleria, a business area on the west, and an area surrounding two major sports complexes towards the south. Drivers tend to be in areas that have a large number of trip requests and along major highways.

[10]On average, frequent riders took 4.73 trips and their conversion rate is 81.2%. Occasional riders took 1.67 trips on average, and their conversion rate is 52.7%.

Trip request, matching, fares, and surge pricing The process before a trip takes place starts with the rider selecting pickup and destination points.[11,12] The app then displays a fare in dollars and a pickup time (see a screenshot in appendix H).[13] If the rider requests a trip, it is offered to the nearest available driver, who has a few seconds to accept. If the driver does not accept, the trip is offered to the second nearest available driver, and so on. The driver does not know the trip destination before the rider gets in the car. Once the trip is completed, the rider pays the fare by credit card. Uber takes a booking fee of $2.30 per trip. The rest of the fare is split between the driver and Uber, which takes a percent commission that varies between 24% and 28%, depending on the time the driver joined Uber.

The fare is the product of two components. The first is the *base* or *unsurged fare*, which is a linear function of the expected trip distance and duration (computed from 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) Distribution of multipliers

(b) Distribution of pickup times

Figure 3: Distribution of multipliers and pickup times

Note: These figures show histograms of the multipliers and pickup times observed by riders. Each observation represents a rider session.

Spatially, surge multipliers vary by the locations in figure 1a. They are updated every two minutes simultaneously across the whole city. Whenever a driver is

[11] It is possible to request a trip without a destination, but the interface makes it difficult.

[12] This is the process as it was in 2017. Some changes are that Uber now groups trips into batches that are matched every few seconds, surge pricing no longer uses rounding, and surge pricing is no longer multiplicative for drivers (they get an additive fare bonus, see Garg and Nazerzadeh, 2019).

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

available, he can see a map of all multipliers in the city (see a screenshot in appendix H). 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.

Figure 3 shows the distribution of surge multipliers and pickup times that passengers observe. 77.8% of passengers see a multiplier of 1 when they open the app. When the multiplier is greater than one, it is typically less than 2 and it is rarely above 3. The bottom two panels in subfigure 1b show the average multipliers and pickup times across space. Multipliers are highest around Downtown and towards the south. Pickup times 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 start by giving an overview of the model of a ride-hailing market. Subsections 3.1-3.4 explain each part of the model 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., open 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 pickup. 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 a pickup time, on the basis of 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.

In addition to agents' decisions, the platform takes some actions. It computes surge multipliers, fares, and pickup times, and it shows them to riders and drivers. It also assigns a nearby driver to riders that request a trip. Figure 4 is a timeline that

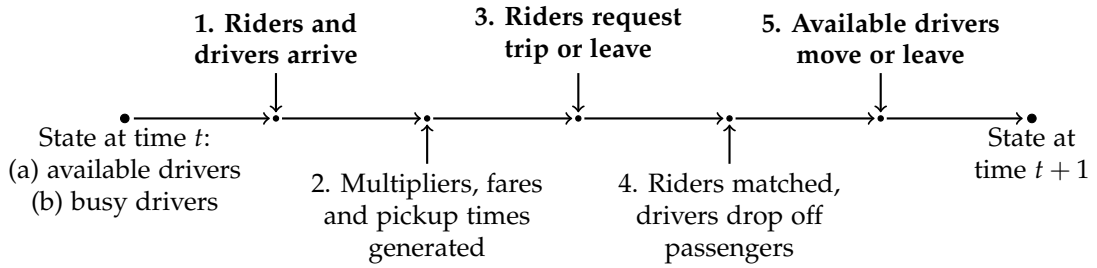


Figure 4: Model timeline

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.

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 decided by the platform.

3.1 Demand

Riders make two decisions. First, riders 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 pickup time that they observe in the app. I model riders as static decision makers. In reality, they can decide to wait and request a trip at a later time. However, waiting for a lower multiplier does not seem to be an important response to surge pricing in the data.[14]

I now explain how riders make those two choices in reverse order.

3.1.1 Trip request

Rider i, who opened the app, is in location l during hour of the week h and wants to go to a destination k. She is characterized by a vector of covariates x_i. The rider gets a "quote" that includes a price p_i and a pickup time before pickup w_i.

If she requests a trip, the rider gets utility

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

[14] As the multiplier goes from 1 to 1.5, the fraction of riders who request a trip or leave the platform in the first two minutes only decreases from 73% to 68%, and to 66% if the multiplier is above 1.5.

The first term captures patterns in the value of a trip by location, hour of the week, and covariates. The second term captures the disutility of paying, and the third term captures the disutility from waiting. Finally, ϵ_i is an error term. 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 pickup times. $\beta(x_i)$ measures the short run elasticity of demand, and $\gamma(x_i)$ measures the pickup time elasticity. I allow both to depend on covariates.[15] The ratio $\frac{\gamma(x_i)}{\beta(x_i)}$ measures the value of time.

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 pickup times. 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.

Let $U_{lhx_i} = \mathbb{E}\left[\frac{1}{\beta(x_i)} \max\{U_i, 0\} \,\middle|\, l, h, x_i\right]$ be rider i's ex-ante dollar value of opening the app. It is her expectation on the value of her best choice—either requesting a trip, with value $\frac{U_i}{\beta(x_i)}$, or the short-run outside option, with value 0—given what she knows *before* opening the app and observing the fare and the pickup time. This is an equilibrium quantity: it depends on the distribution of prices and pickup times that rider i faces in equilibrium. She opens the app if and only if $U_{lhx_i} > u_i$.

There is an arrival rate λ^0_{lhx} of riders with demographics x that could potentially open the app in location l during hour of the week h. Each potential entrant i chooses to open the app with the probability that $U_{lhx_i} > u_i$, so the actual rate at

[15] In principle, both coefficients could also depend on location and hour of the week, but I do not have enough power to estimate such fine heterogeneity.

which riders open the app is $\lambda_{lhx} = \lambda^0_{lhx} F^u(U_{lhx})$. I assume that $F^u(u) \propto u^\rho$.[16] Thus,

$$\lambda_{lhx} = A_{lhx} U^\rho_{lhx} \tag{2}$$

for some demand shifter A_{lhx}. The elasticity of the number of people who open the app with respect to U_{lhx} is constant and equal to ρ. Rider surplus relative to the long-run outside option is $\int_0^{U_{lhx}} (U_{lhx} - u) dF^u(u) = \frac{A_{lhx}}{1+\rho} U^{\rho+1}_{lhx}$.

The key parameter of this arrival model is ρ. It determines the long-run elasticity of demand—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 to start working. If they do, every time they are available they decide whether to keep on working or leave the platform. Finally, if they stay, they choose where to move before the beginning of the next period.

I assume that drivers' utility is equal to earnings minus the physical cost of driving (fuel, depreciation, and maintenance) minus the opportunity cost of working. Therefore, two drivers with an equal opportunity cost who get the same net earnings—net of physical costs—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.

I now describe drivers' decisions in detail, starting with movement choices.

3.2.1 Movement

Driver j is available in location l at 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 $\mathbf{s}_t = (l, h, \mathbf{m}_t)$.

[16] This CDF 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 pickup time.

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 future 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 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. The driver cannot move instantaneously to the other end of the city, for instance, and he is less likely to move to far-away locations during rush hour than he is at 4 am.

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

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

where 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.2). Thus, $v_k(\mathbf{s}_t)$ incorporates almost all 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, multipliers might go up or down, 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 $v(k, \mathbf{s}_t)$.

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

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

The first term inside the exponential is a fixed effect by origin, destination, and

hour of the week that 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 4 am, but not when there is rush hour traffic, then ω_{lkh} is higher at 4 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 surge multipliers directly, but they respond indirectly: higher multipliers lead to higher expected earnings, so 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.

Drivers are not fully rational—i.e., forward-looking and utility-maximizing—in this setup (as they are in Buchholz (2018) and Frechette et al. (2019), for instance). It is not feasible to model fully rational drivers in this setting because of the curse of dimensionality: the state involves all multipliers surrounding the driver. The movement rule from equation (4), however, is a tractable model of driver movement that allows me to capture the essence of 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.

This model departs from standard dynamic models of fully rational agents in three ways.[17] First, 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

[17] In standard dynamic discrete choice models, the assumption of full rationality is needed to infer utility from agents' choices. I am able to depart from that assumption because of an unusual feature: I observe an objective measure of utility, net earnings, which I use to estimate continuation values $v_k(\mathbf{s}_t)$ directly from the data.

the distribution specified by the movement rule. This third point stems from the assumption that drivers' utility is equal to net earnings minus 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.

To compute drivers' utility, I simulate drivers' behavior based on movement rule (4) to compute their net earnings. I then subtract opportunity costs, which I define in section 3.2.2.

3.2.2 Entry and session duration

Entry Suppose 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 option (i.e., the opportunity cost of working), which I denote by $\bar{W}_i$, is drawn from a distribution F^W. Before observing surge multipliers, the driver expects to get hourly earnings W_{lh} if he starts working. This is an equilibrium quantity: it depends on how the driver will behave if he starts working, and on how he expects all other agents to behave. The driver starts working if and only if $W_{lh} \geq \bar{W}_i$.

There is a rate μ^0_{lh} 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, so the actual rate at which drivers start working is $\mu^0_{lh} F^W(W_{lh})$. I make the functional form assumption $F^W(W_{lh}) \propto W^\sigma_{lh}$.[19] The entry rate is thus

$$\mu_{lh} = B_{lh} W^\sigma_{lh}, \tag{5}$$

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

The key parameter of this entry model is σ, which 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.

[18] Drivers 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.

[19] This unbounded distribution can be understood to be the left tail of a very large distribution.

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.2 that current multipliers can only predict around 15% of unexpected variation in future multipliers more than ten minutes into the future. Therefore, unexpected changes in multipliers convey little information about the total earnings drivers would get if they start working.[20]

Shift duration At the time that driver j arrives, which I denote by t_j^0, he draws an intended shift duration D_j from a distribution G_h, which varies by the hour of the week. The driver stops working the first time that he is available after $t_j^0 + D_j$.[21]

3.3 Uber's decisions

Matching Whenever a rider requests a trip, the matching process determines which driver picks her up as well as how long it takes the driver to pick her up. Thus, it implictly determines the magnitude of the matching inefficiencies. I now describe a model of matching that closely follows the procedure used by Uber.

In period t, let I_t^r be the set of riders that request a trip and let J_t^a be the set of drivers that are available to be matched—including drivers who will drop off a rider during the next four minutes. 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 (i.e., dropping off a rider), and the hour of the week h.

Riders in I_t^r are 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 he is not.[22] If he does not accept the trip, it is then offered to the next closest driver, who also accepts it with probabilities

[20] One exception are drivers who could start working immediately in high demand locations, but there are very few of them: 94% start working outside the area of analysis and drive in.

[21] Actual drivers could in principle respond to surge multipliers by working longer, but I do not see a signifiant response in the data (see appendix G.3). Thus, I abstract from this response margin.

[22] Drivers can accept or reject trips selectively, but Uber 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 inattention.

ϕ^b and ϕ^a, depending on whether he is busy. 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.

A distribution $G(\cdot|l_i, l_j, b_j, h)$ with higher realized pickup times means a larger matching inefficiency. Lower acceptance rates ϕ^b and ϕ^a result in matches with drivers who are farther away, and, therefore, a larger matching inefficiency.

Surge multipliers 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 feature is that it depends on the number of available drivers and on the number of riders who open 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.

3.4 Other parts of the model

In appendix A I present three additional parts of the model that are necessary to fully describe market behavior. First, I explain what determines the duration of trips and how Uber generates base fares and pickup times. These are mechanical steps in which no agent makes choices. Second, I present a model for the trip destination and trip duration. It is based on empirical distributions, and assumes that the trip destination 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 movements of drivers who are far from the region of analysis are 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 mean

future earnings. Those beliefs are determined by an equilibrium condition: they must be consistent with empirical averages.

Let $\mathbf{U}$ be the vector of riders' ex-ante utilities U_{lhx}, and let $\mathbf{W}$ be the vector of drivers' expected hourly earnings W_{lh}. Also 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 any triple of beliefs $\mathbf{x} = (\mathbf{U}, \mathbf{W}, \mathbf{v})$ belongs to $\mathcal{X}$.

Suppose the market behaves as described by the model under a certain pricing policy P. Let $f^P(\cdot) : \mathcal{X} \to \mathcal{X}$ 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' beliefs are $\mathbf{x}$. A market equilibrium for pricing policy P is characterized by a vector of beliefs $\mathbf{x}^* \in \mathcal{X}$ that is a fixed point of $f^P(\cdot)$:

$$\mathbf{x}^* = f^P(\mathbf{x}^*). \tag{6}$$

This means that beliefs are 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.

Market clearing This market clears through a hybrid mechanism that involves both waiting times and prices. If, for instance, there is excess demand in some location and period, then prices and pickup times go up, and drivers do not have to wait long between trips. That induces higher supply and lower demand.

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 the exact identification strategy I use and 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 $\beta(x_i)$ and $\gamma(x_i)$, which determine how prices and pickup times affect the probability

that riders request a trip. The main challenge is that prices and pickup times are endogenous.

Response to prices I estimate riders' response to 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* $\tilde{m}_{lt}$ for each location.
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_{lt} is the *surge multiplier*.

The final price shown to riders is $p_i = b + m_{lt}(\bar{p}_i - b)$, 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 price variation from which I identify the demand response to prices. Since the surge multiplier is a deterministic function of the vector of recommended multipliers $\tilde{\mathbf{m}}_t$, any correlation between demand shocks and prices must come through $\tilde{\mathbf{m}}_t$. Once I control for $\tilde{\mathbf{m}}_t$, 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 regression discontinuity design (RDD). I rely on stronger assumptions (which I state in section 5) that allow me to capture 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, since 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 2.2 or 2.3. My estimation procedure chains together all these small responses to find the demand response to large price changes without relying on extrapolation.

[23] Temporal smoothing takes place by not allowing multipliers to change too much from one period to the next. Spatial smoothing takes place by computing a weighted sum of nearby multipliers.

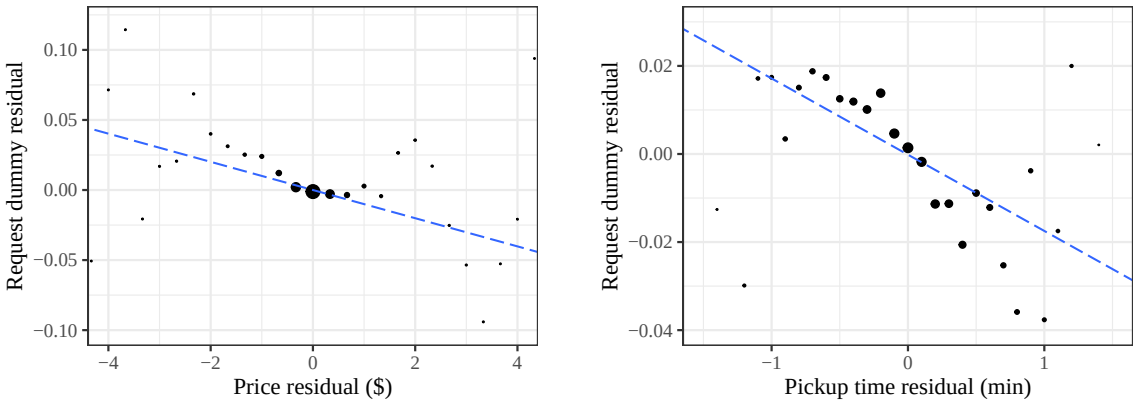
Response to pickup times The pickup time that passengers observe depends on the location of nearby drivers. The number of nearby drivers is correlated with demand shocks—if more people request trips, the number of available drivers goes down, increasing pickup times—and, thus, it is endogenous. But drivers' 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 a pickup time of three minutes while the second one sees a pickup time of one minute.

I use this variation to estimate the response of riders to pickup times. To do so, I include location by time-period fixed effects. 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, so fixed effects would clean them out.[24]

Residuals and value of time Figure 5 shows the raw variation in the data from which I identify riders' response to prices and pickup times, using the identification strategy I described. Vertical axes shows residuals of a dummy for whether riders request trips. In figure 5a, the horizontal axis represents the variation in multipliers that remains after controlling for the *unrounded multiplier*, an estimate of what the multiplier would have been if there was no rounding in the surge pricing algorithm. I explain how I compute it in appendix D.2. 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 a trip request. Figure 5b shows within-location-by-time-period variation in pickup times and trip request dummies. There is also a downward pattern, indicating that a higher pickup time decreases the probability of requesting a trip.

Comparing the slopes in subfigures 5a and 5b gives a sense of how riders trade off prices and pickup times. The pickup time response divided by the price re-

[24]One potential confounder is that riders that are close to major streets—and thus get low pickup times—might be systematically different from the rest. I run models that control for the accessibility of the rider's location and obtain almost identical coefficients.



(a) Price response

(b) Pickup time response

Figure 5: Residuals used to identify the response of riders to prices and pickup times

Note: Both figures are bin scatters in which the vertical axis is the residual of a dummy for trip request. The point size represents the number of observations. Subfigure (a) shows residuals from regressions on the unrounded multiplier. 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 pickup time deviations and show a small demand response.

sponse measures the value of time. Following this idea, figure 6 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 pickup times, where I control for the unrounded price and for the average pickup time by location by time period to get causal estimates. The average value of time for the whole sample is around $2 per minute. This value is high, but it is comparable with estimates in the literature.[25] There is significant variation across subsamples, all of which follows the patterns one would have expected: the value of time is higher for high-income riders and higher during the week than during weekends, and it is especially high for airport trips.

Several reasons might lead to the high implied value of time. First, riders tend to request trips during time sensitive moments. They might want to be in time for an appointment, or they 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. Second, some trips are requested by more than one rider, so the relevant value is the sum of

[25]The value of time is 4.6 times the mean wage in Houston of $26 per hour, and is similar to the values implied by Cohen et al. (2016) in other cities in the US. For Prague, Buchholz et al. (2020) find values that are on average 1.5 times the average wage of $9 per hour. In a field experiment, Kreindler (2018) estimates a value of time for commuters in India that is 5 times the average wage of $3 per hour.

the value of time for all riders. Third, business passengers who do not pay for the trip might have a limited response to prices.[26]

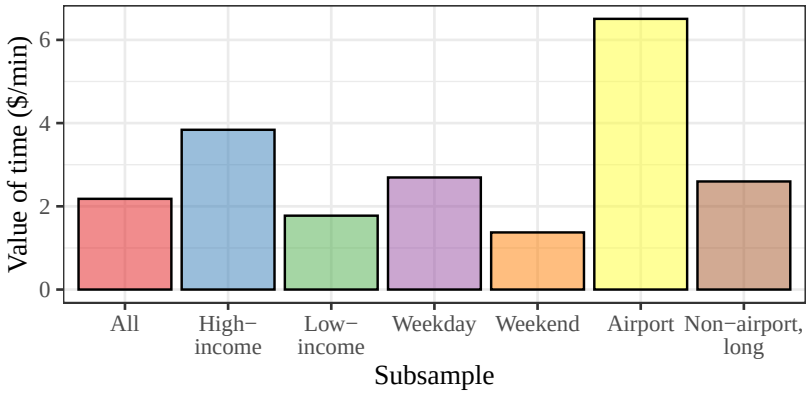


Figure 6: Average value of time for subsamples of the data

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 pickup time. I control for the unrounded price and for the average pickup time 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 pickup time coefficient divided by the price coefficient.

4.2 Short-run supply response

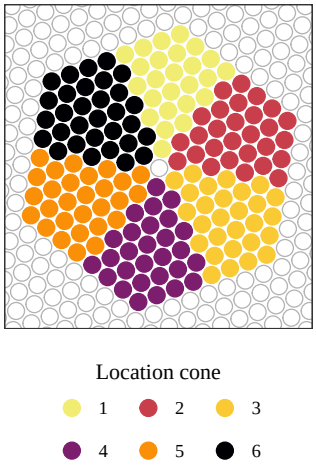
In the driver movement model from section 3.2.1, the main parameter I want to estimate is δ, the extent to which drivers move towards areas with high mean earnings. This is challenging because changes in mean earnings are mainly driven by surge multipliers, which are endogenous. If some shock induces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

[26]These values could also be driven by behavioral effects if riders overreact to pickup times. Behavioral effects as well as unresponsive business travelers complicate the interpretation of welfare estimates. Appendix F.2 shows that the main results hold if riders' true value of time is lower than measured.

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.

Table 1: Effect of multipliers on movement direction



	Dependent variable:					
	Dummy for moving to cone in parentheses					
	(1)	(2)	(3)	(4)	(5)	(6)
Avg. multiplier in cone 1	0.22***	−0.04	−0.16***	0.07	−0.04	−0.04
	(0.07)	(0.07)	(0.06)	(0.07)	(0.06)	(0.06)
Avg. multiplier in cone 2	−0.04	0.29***	0.08	−0.18**	−0.07	−0.07
	(0.07)	(0.08)	(0.06)	(0.07)	(0.07)	(0.07)
Avg. multiplier in cone 3	−0.05	−0.09	0.16**	−0.04	−0.01	0.04
	(0.07)	(0.08)	(0.07)	(0.08)	(0.07)	(0.07)
Avg. multiplier in cone 4	−0.09	−0.05	−0.05	0.37***	−0.06	−0.12*
	(0.06)	(0.07)	(0.07)	(0.09)	(0.08)	(0.07)
Avg. multiplier in cone 5	−0.08	0.04	−0.05	−0.15*	0.22***	0.01
	(0.06)	(0.07)	(0.06)	(0.08)	(0.08)	(0.07)
Avg. multiplier in cone 6	0.02	−0.13*	0.03	−0.07	−0.04	0.19***
	(0.06)	(0.07)	(0.05)	(0.07)	(0.07)	(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 shows how I aggregate the space surrounding a driver. In the table, each column represents a regression of a dummy for whether a driver moved to one cone on the average multiplier by cone. I control for the average recommended multiplier by cone, for the multiplier at the driver's location, and for location by hour of the week fixed effects. Standard errors are clustered by location and by 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 determine the log-run elasticities of demand and supply (σ and ρ). I determine their values using experimental data.

I estimate the long-run elasticity of demand using data from one-week experi-

ments that Uber ran in 2017 in five Latin American cities (Belo Horizonte, Guadalajara, Mexico City, Rio de Janeiro, and São Paulo). 177,349 riders were randomized into a control group and two treatment groups. Treated riders got 10% or 20% discounts for every trip they took during the experiment week. The experiment ran from Monday to Sunday, and treated riders were notified on Sunday before the experiment. I measure the demand elasticity by running a Poisson regression of the number of trips taken by each rider on the log price factor including city fixed effects.[27] I find a demand elasticity of -0.633, with robust standard error 0.059.[28]

I use the supply elasticity Angrist et al. (2020) estimate from an experiment they run in Boston. To a random sample of Uber drivers, they offer for one week the choice between a standard contract and a "taxi" contract, in which drivers pay a weekly lease but receive higher earnings per trip. Based on drivers' choices, they infer a supply elasticity of 1.2.

There are some potential problems using these experimental estimates for my model. First, they measure the response of people who had an Uber account, but not 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 buy or sell their car. Third, these elasticities were measured in cities that are not Houston—and, for demand, not even in the US. Many differences across cities can lead to different elasticities, such as the availability of public transportation or outside job opportunities. Because of these concerns, I run the main counterfactuals for a large range of values for σ and 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.2). This means that the main findings are not driven by long-run elasticities. Instead, they are mainly driven by short-run elasticities and by the matching technology.

5 Estimation and results

In this section I explain how I use the variation in the data and the identification strategies described in section 4 to estimate the parameters in the model.

[27] The log price factor is $\log(1)$ for the control group, and $\log(0.9)$ or $\log(0.8)$ for treatment groups.
[28] Appendix G.4 shows additional results with estimates by city and by treatment.

5.1 Demand

5.1.1 Trip request

I estimate the trip request model following the ideas from section 4.1: I identify the price response from rounding in surge multipliers and the pickup time response from within location by time period variation. I estimate the following equation:

$$U_i = \alpha(x_i, l, h) + \beta(x_i)p_i + \gamma(x_i)w_i + g(\tilde{\mathbf{m}}_t, x_i, w_{lt}^0; l, h) + \eta_i, \tag{7}$$

where $w_{lt}^0 = E[w_i|l, t]$ is the expected time to pickup by location and period. This is the original utility specification (1), except that I decompose the error as $\epsilon_i = g(\tilde{\mathbf{m}}_t, x_i, w_{lt}^0; l, h) + \eta_i$, where $g(\cdot)$ is a flexible function that controls for recommended multipliers, rider covariates, expected pickup time, location, and hour of the week.

I assume that the error η_i is orthogonal to $(p_i, w_i, \tilde{\mathbf{m}}_t, x_i, w_{lt}^0, l, h)$. That is true if $g(\cdot)$ captures all the correlation between ϵ_i and covariates, which is justified by the following intuition. First, 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_{lt}^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 $g(\cdot)$: it must be flexible enough to capture the true dependence. In appendix C.1, I show formally that if that is true, 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 pickup times within location by period—i.e., $(w_i - w_{lt}^0)$—is orthogonal to demand shocks, rider covariates, 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, pickup time variation is exogenous. The third assumption is necessary to identify price coefficients.[29] If it was not true, it would not be possible to tell apart discrete changes in recommended multipliers from rounding. It holds because the surge pricing algorithm computes recommended multipliers as smooth functions of market observables.

Functional form assumptions The main objects of interest are the coefficients $\beta(x_i)$ and $\gamma(x_i)$. The vector x_i includes a constant, log income, a dummy for occasional riders, the unsurged fare (a measure of traffic-adjusted trip distance), a dummy for airport trips, and a weekend dummy. I assume that the pickup time coefficient is linear, $\gamma(x_i) = \theta^w x_i$, and that the price coefficient takes the form $\beta(x_i) = s(\theta^p x_i)$, where $s(\cdot)$ is a function that behaves almost like the identity function for negative values but caps the coefficient at -0.005. Relative to a linear specification, $s(\cdot)$ only affects around 2% of observations, which correspond to very high income riders that are going far away (see appendix D.3). The affected observations would otherwise have a price coefficient that approaches zero or becomes positive, in which case the value of time diverges or becomes negative.

For the function $g(\cdot)$, I estimate a flexible model of recommended multipliers to predict the *unrounded price* $\hat{p}_i$, the fare the rider would have seen if there was no rounding (see appendix D.2). I then set $g(\cdot)$ as a sum of high-dimensional splines of $\hat{p}_i$, $\tilde{\mathbf{m}}_t$, and w^0_{lt} (see appendix D.3 for details).[30] Intuitively, since I control for $\hat{p}_i$ and w^0_{lt}, the price coefficient is identified from variation in $p_i - \hat{p}_i$, which arises from rounding, and the pickup time coefficient is identified from variation in $w_i - w^0_{lt}$. I assume the intercept term $\alpha(x_i; l, h)$ is additively separable into a linear function of x_i and a flexible function of (l, h) with 155 degrees of freedom (see appendix D.3 for details).

To estimate equation (14), I plug in $\bar{w}_{lt}$ for w^0_{lt}.[31] I assume that the error η_i is

[29] The third assumption plays the role of the main RDD assumption: that the conditional mean of each treatment group is continuous in the forcing variable. It is stronger in that (a) the functional form for conditional means of different treatment groups is the same and (b) $g(\cdot)$ must be correctly specified. Stronger assumptions let me use variation beyond a threshold δ around the discontinuity.

[30] I assume that the dependence on (l, h) is additively separable, and so it is captured by $\alpha(x_i; l, h)$.

[31] One potential concern is that there are few observations in each location by period, so $\bar{w}_{lt}$ does not converge asymptotically to w^0_{lt}. In appendix G.1, I present an alternative specification that avoids this problem, with very similar estimates.

distributed iid logistic. My model is thus a logit model with nonlinear coefficients that has a large number of covariates. I estimate it by maximum likelihood.

Table 2: Estimates of demand parameters

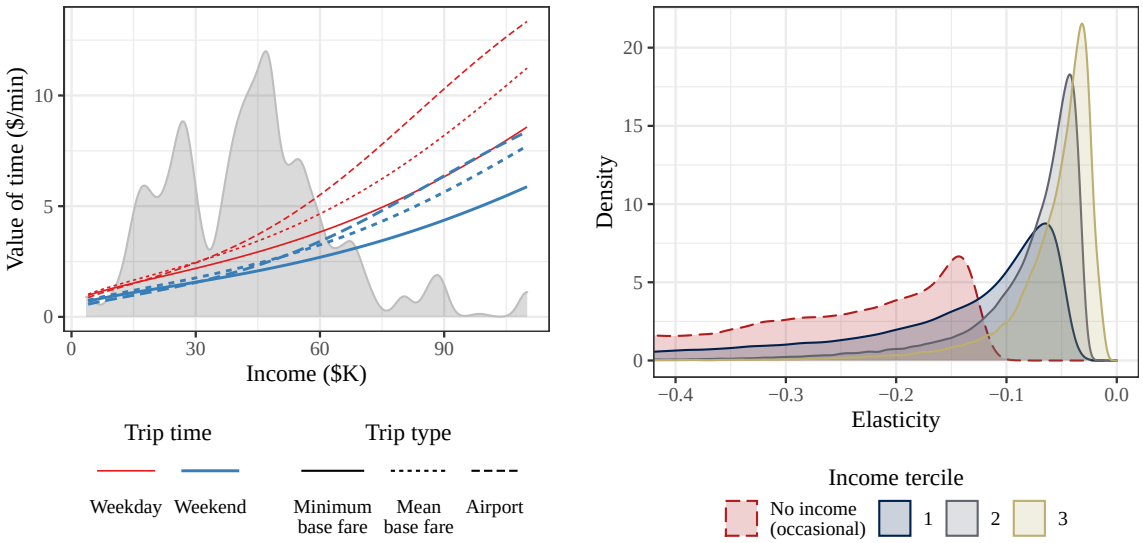
	Dependent variable: Trip requested					
	Coefficient dependence on:					
	Constant	Occasional	Log income	Base fare	Airport	Weekend
Price	-0.0476^{***}	-0.0090	0.0388	0.0016	0.0173	-0.0007
	(0.0183)	(0.0194)	(0.0240)	(0.0010)	(0.0249)	(0.0195)
ETA	-0.1164^{***}	0.0473^{**}	0.0003	-0.0002	0.0290	0.0342^{*}
	(0.0130)	(0.0186)	(0.0241)	(0.0013)	(0.0462)	(0.0187)
						Observations: 650,233

Note: Estimates of the main parameters of the trip request model (equation (7)), which includes the function $g(\cdot)$ that controls flexibly for recommended multipliers, rider covariates, and expected pickup time, and a term $\alpha(x_i, l, h)$ that controls for covariates, location, and hour of the week. The whole table presents estimates from one single model. The first row shows parameters for the price coefficient; the second row shows parameters for the pickup time coefficient. All covariates are demeaned. Standard errors are clustered by location and hour of the week.

Results Table 2 shows estimates of the main parameters. The first row shows the parameters of the price coefficient; the second row shows the parameters of the pickup time coefficient. All covariates are demeaned, so the first column measures the average price and pickup time coefficients. Both are negative and significant, as expected.

The heterogeneity parameters are somewhat imprecisely estimated, but they imply the patterns one would expect. Subfigure 7a shows that the value of time is higher for high-income riders, during weekdays, and when people are going far (i.e., the base fare is high), especially to the airport.[32] The average value of time across all sessions is $2.63 per minute, which is somewhat higher than in the analysis in section 4.1. Figure 7b shows the distribution of the price elasticity for occasional riders and for frequent riders by income tercile. Frequent riders, and especially high-income riders, have low elasticities. Occasional riders are more elastic, as one might expect given that they only use the Uber app sporadically. The average elasticity for the whole sample is 0.179. Although low, this value should not be surprising because it is a very short run elasticity. Furthermore, Houston has

[32]The value of time for occasional riders is between $0.48 and $1.65 per minute, resembling that of frequent riders with income around $15,000.



(a) Value of time

(b) Price elasticity

Figure 7: Value of time and price elasticity from the main demand model

Note: Subfigure (a) shows how the value of time varies with income, time of the week, and trip distance. The minimum base fare is $5.41, and the mean base fare is $10.65. The gray area in the background shows a kernel density plot of income. Subfigure (b) shows the density of the implied price elasticity for occasional riders as well as for frequent riders by income tercile.

few alternative transportation options: there is no competing ride-hailing app, and public transit is limited.[33]

5.1.2 Opening the app

The main parameter of the rider arrival model (section 3.1.2) is ρ, which measures the long run elasticity of demand. The model also has parameters A_{lhx}, which are demand shifters by location, hour of the week, and rider demographics. I set the values of A_{lhx} and ρ jointly so that (a) for every (l, h, x), the arrival rate is equal to the average arrival rate in the data given the average utilities U_{lhx} I compute from the data, and (b) the market-wide elasticity of demand is equal to the value of -0.633 from section 4.3.[34] This results in a value of $\rho = 1.80$.

A few trips are not in the demand dataset because they were requested from Google Maps or other external apps, so the rider did not interact with the Uber

[33]Cohen et al. (2016) find elasticities around 0.45, which are probably higher because they analyze cities like New York, San Francisco, and Chicago that have good transportation alternatives.

[34]To estimate U_{lhx} consistently, I aggregate combinations of (l, h, x) into larger groups because there are too many values of l and h, and because x is continuous (see appendix D.4).

app before requesting a trip. To account for this, I allow for a multiplicative scale factor ψ^d that shifts all of demand up or down. In 5.2.2 I describe a similar scale factor ψ^s for supply. I set both factors jointly such that simulations of the market equilibrium in the status quo result in the same number of trips and the average surge multiplier as in the data. This results in $\psi^d = 1.015$.

5.2 Supply

5.2.1 Movement

I now explain how I estimate the parameters from the driver movement model from section 3.2.1: road and traffic patterns $\omega(l, k, h)$ and riders' responsiveness to earnings δ. I identify δ from variation that arises from rounding in surge multipliers.

The model I estimate is

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

This is the movement model in equation (4), under the assumption that the unobserved term ζ_{kt} 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 from variation that arises from rounding. The intuition is similar to that 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 exogenous.

In appendix C.2 I justify this argument 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, and on the following two assumptions, which are entirely analogous to the assumptions in 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 that has discontinuities and, second, that those discontinuities induce discrete changes in mean future earnings. Part 2 is analogous to the assumption of smooth conditional means in RDDs, and it is a reasonable assumption because recommended multipliers depend smoothly on market observables.

Mean future earnings 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 net hourly earnings from time $t+1$ until $t+\bar{t}$ or the time he leaves, whichever comes earlier.[35] I set $v_k(\mathbf{s}_t)$ to be the prediction from the following model:[36]

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

where l' is the direction driver j moved to. The term $\alpha(l', h)$ is a flexible function that plays the role of location by hour-of-the-week fixed effects. The term $f(\mathbf{m}_{tl'})$ is a flexible function of $\mathbf{m}_{tl}$, the vector of multipliers surrounding location l. I use a relatively simple smooth function that is radially symmetric. Appendix D.5 explains the model in detail and shows that the fit follows the patterns one would expect it to follow. For instance, earnings are higher when multipliers are higher.

Functional form assumptions I set $g^M(\tilde{\mathbf{m}}_{tk}; l, h)$ to be a sum of high dimensional splines of the unrounded multiplier and the average recommended multiplier in the six nearest hexagon rings around the current location. The dependence on (l, h) is absorbed by $\omega(l, k, h)$, which is the sum of two terms. The first term is origin by destination fixed effects that capture the fact that drivers' movements are defined to a large extent by road patterns. The second is fixed effects for origin zone by hour of the week by movement trend (the general direction and distance of movement) that capture the fact that traffic patterns change over the week. I estimate the model by maximum likelihood. See further details in appendix D.5.

[35] I assume driving costs of \$0.26 per mile, the internal Uber estimate for the average UberX car in Houston. This estimate includes fuel, maintenance, repairs, and depreciation.

[36] With unlimited data, I would average Π_{jt} by the state and subsequent position to estimate $v_k(\mathbf{s}_t)$. However, there are far more states than observations because of the curse of dimensionality.

Results The estimate I obtain 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.[37]

5.2.2 Entry and shift length

Entry The main parameter of the driver arrival model described in section 3.2.2 is σ, which measures the long-run elasticity of supply. The model also has parameters B_{lh}, which are supply shifters by location and hour of the week. I set B_{lh} and σ jointly so that (a) for every (l, h), the arrival rate is equal to the average arrival rate in the data given the empirical W_{lh}, and (b) the market-wide elasticity of supply is equal to the value of 1.2 from Angrist et al. (2020). This results in $\sigma = 0.771$.[38]

I allow for a multiplicative scale factor ψ^s that shifts all of supply up or down—similar to the demand scale factor ψ^d. This corrects for some complications in the data that I do not account for in my model, such as the fact that some drivers work not only for UberX, but also for UberBlack, UberXL, or UberEats. As described in section 5.1.2, I set ψ^s and ψ^d jointly such that simulations of the market equilibrium in the status quo result in the same number of trips and the average surge multiplier in the data. This results in $\psi^s = 1.244$.

Shift duration I assume that the distribution G_h of the intended shift duration during week hour h is Gamma(α_h, β_h).[39] 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

[37] Compare these estimates with Lu et al. (2018), who estimate a multinomial logit model from a surge outage that affected drivers using iOS but not those using Android. Based on their estimates, the example above would shift the probability from 0.25 to 0.269, which is somewhat smaller.

[38] The number of combinations of (l, h) is too large to estimate W_{lh} consistently. Thus, I aggregate the data into larger groups (see appendix D.4).

[39] Appendix D.6 compares the fit with the data. The Gamma assumption seems well justified.

intended exit time between $\bar{D}_j$ and $\underline{D}_j$. Thus, the likelihood for driver j is given by

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

5.3 Matching

I need to estimate two elements in the matching model: the distribution of pickup times $G(\cdot|l_i, l_j, b_j, h)$ for rider-driver pairs, and the matching rates ϕ_a and ϕ_b.

For the distribution of pickup times $\hat{w}(X_i, X_j, h)$, I first fit a random forest of realized pickup times as a function of the rider's coordinates, the driver's coordinates when pickup started, the time of the day, and the hour of the week. Let $\hat{w}(X_i, X_j, h)$ be the prediction from 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 residual of this model by bins of the prediction $\hat{w}$. Let $\hat{sd}(\hat{w})$ 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. Second, I compute $\hat{w}(X_i, X_j, h)$ from the drawn coordinates and $\hat{sd}(\hat{w})$ from this prediction. Third, I draw the pickup time w_{ij}^P from a lognormal distribution with mean $\hat{w}$ and standard deviation $\hat{sd}(\hat{w})$.

I fit driver acceptance rates by the method of simulated moments. The two moments I match are the mean realized pickup time (3.52 minutes) and the fraction of trips that are assigned to busy drivers (13.7%). These two moments relate to the parameters in an intuitive way: 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 every period, after which I compute simulated 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 Uber uses internally. 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.

5.4 Equilibrium computation and model fit

With the parameter estimates I have described, I can simulate the behavior of the market given agents' beliefs $\mathbf{x}$. From those simulations I can compute market averages to obtain an unbiased estimate for $f^P(\mathbf{x}^*)$. Computing market equilibria—i.e., a fixed point of $f^P(\mathbf{x}^*)$—is challenging because naive iterative algorithms typically diverge. In appendix B I explain the algorithm I use and discuss why it converges.

In appendix E I compare the data with simulations of the status quo in equilibrium. The model captures well the high-resolution spatial patterns in the data as well as hourly patterns of supply and demand. The simulated distribution of surge multipliers is almost identical to the distribution in the data.

6 The effects of surge pricing

6.1 Measuring welfare

I compute rider surplus and driver surplus as sums of rider and driver surplus, as defined in section 3, over the whole market:[40]

$$RS = \sum_{lhx} \frac{A_{lhx}}{1+\rho} U_{lhx}^{\rho+1}, \qquad DS = \sum_{lh} \bar{D}_h \frac{B_{lh}}{1+\sigma} W_{lh}^{\sigma+1}. \tag{11}$$

I compute Uber's profit as

$$\Pi = \sum_{n} \Big((1-\tau-\nu) p_n - \pi_n - I_n \Big). \tag{12}$$

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 1% credit card transaction cost. As described in section 2, the driver receives the fare minus the booking fee and commission that Uber takes.[41] Uber also pays per-mile insurance whenever a driver is picking up or dropping off a rider. I use a cost of \$0.30 per mile, which is somewhat below

[40] The 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*.

[41] I use a fixed commission rate of 26.3%, which is the average in the data.

market rates for private customers in Houston.[42]

6.2 Welfare effect on riders, drivers, and Uber

In figure 8, I compare counterfactuals with 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 dotted lines cross at the status quo.

The solid line represents alternative surge pricing policies in which the multiplier is scaled up or down by a constant factor: If under certain market conditions the surge multiplier in the status quo would be m_{lt}, then the surge multiplier is am_{lt} under those same conditions if the scale factor is a. Increasing a thus entails an increase in average multipliers.[43] The 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 multipliers.

Surge pricing dominates uniform pricing in the sense that, for every level of the average multiplier, welfare is higher with surge pricing. This means that there are efficiency gains from surge pricing. The vertical distance is around 6.4% of gross revenue at the average multiplier in the status quo. Both for surge and uniform pricing, welfare has an inverted-U shape with a maximum at an average multiplier of 1.35-1.4. Uber is therefore pricing lower than is socially optimal.

The other three subfigures break down welfare into rider surplus, driver surplus, and profit.[44] Driver surplus and profit are increasing for the whole range of multipliers. This means that Uber prices well below short-run profit maximization. Instead, it prices exactly at the level that maximizes rider surplus. Lower prices hurt riders because of higher pickup times. More specifically, high income riders would prefer higher prices, whereas low-income and occasional riders would prefer lower

[42] The actual price, which they do not disclose, is the outcome of bargaining with insurers.

[43] I also counduct a similar exercise in which, instead of scaling the multiplier multiplicatively, I add or subtract some fixed quantity to the multiplier. This results in almost identical figures.

[44] I do not show tax revenue, which only accounts for a very small fraction of welfare.

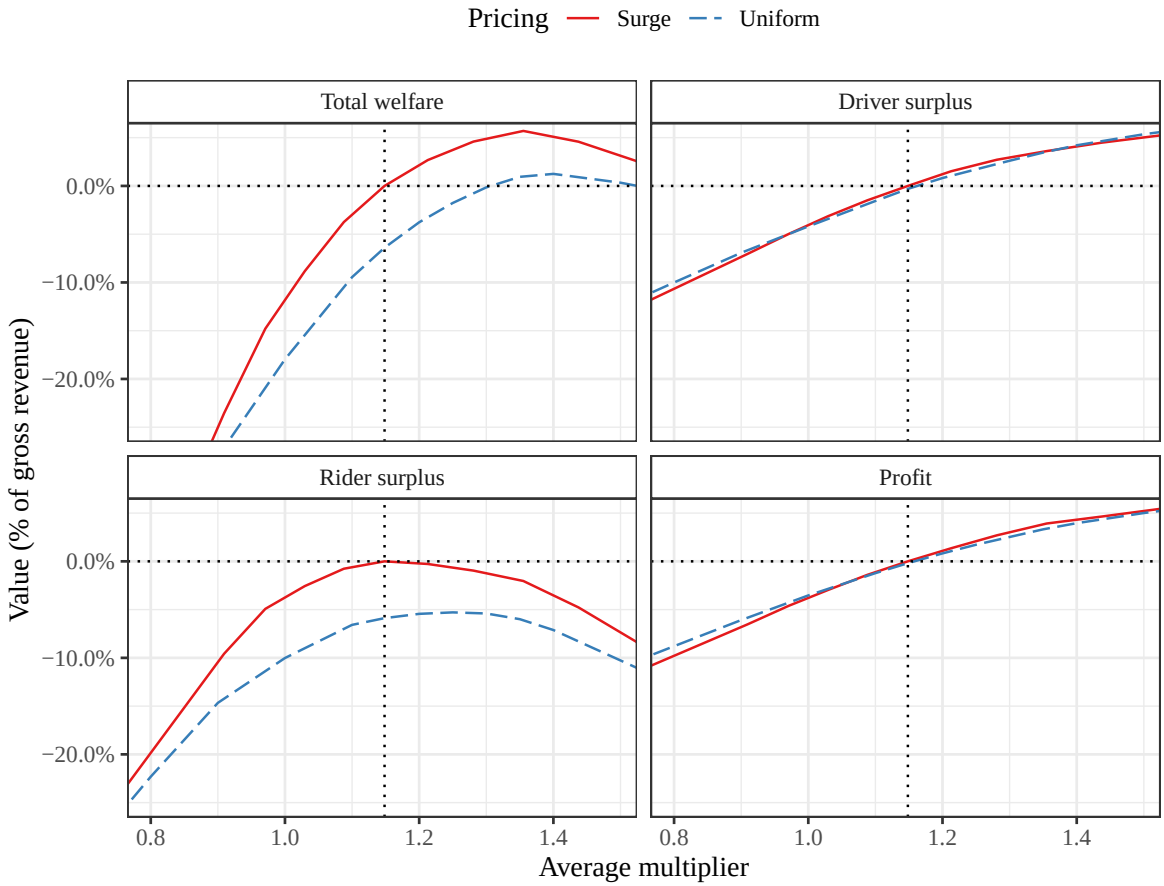


Figure 8: Welfare under different pricing policies

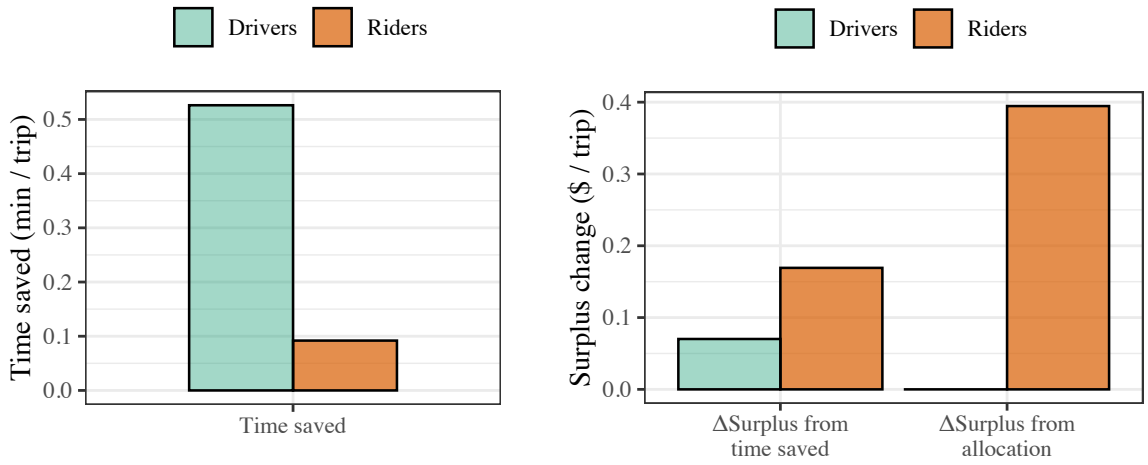
Note: These figures compare welfare and its components for different pricing policies. The horizontal axis represents the average surge multiplier. Dotted lines highlight the status quo. The vertical axis represents total welfare, rider surplus, driver surplus, or profits relative to the status quo.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

prices (see appendix F.3).

It might be surprising that Uber prices to maximize rider surplus instead of profit. However, this model computes profit on a time frame of a few weeks (the time frame of long-run elasticities). Uber's goal is not to maximize profits in the short run—in fact, it has lost money since it was founded. Instead, Uber aims to maximize shareholder value. Maximizing rider surplus is the right strategy if Uber believes the main determinant of long-run performance is consumer satisfaction.

This is a predominant belief among technology companies.[45,46]

Uniform and surge pricing with the same average multiplier One remarkable feature of figure 8 is that most of the welfare gap between surge and uniform pricing is accounted for by rider surplus. Driver surplus and profit are also higher for surge pricing, but only slightly. To understand why effects are so asymmetric, I compare surge and uniform pricing when the average multiplier is held constant at the status quo average. Figure 9 breaks down the welfare effect on riders and drivers of moving from uniform to surge pricing (the effect on profits is almost identical to the effect on driver surplus).



(a) Time saved

(b) Surplus change

Figure 9: Efficiency gains of surge pricing

Note: These figures break down the welfare effects of moving from uniform to surge pricing holding the average multiplier constant at the status quo level. Subfigure (a) shows how much time riders and drivers save. Subfigure (b) decomposes the change in rider and driver surplus into time savings (the decrease in pickup time times riders' value of time, and the decrease in time between trips times drivers' average hourly wage) and a better allocation (lower wasted surplus due to denied trips).

The welfare gains of riders are substantially larger than those of drivers for two reasons. First, surge pricing mitigates supply-demand imbalances, saving market participants' time. Riders wait less time to be picked up, and drivers spend less time waiting to be matched and picking up passengers. The bars in figure 9a measure

[45] As said by Jeff Bezos, founder and CEO of Amazon, "take a long-term view, and the interests of customers and shareholders align" (Bezos, 2013).

[46] Dinerstein et al. (2018) find that eBay's behavior is also consistent with maximization of consumer surplus instead of profit. Thus, they also assume its objective function is consumer surplus.

those time savings. Drivers save around five times as much time per trip as riders.[47] However, riders' valuation of one minute is much higher than drivers', and so in dollar terms, riders' time savings are more valuable. The bars at the left in figure 9b measure those welfare gains.

Second, 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 towards a price mechanism. While allocative efficiency benefits riders, it has no effect on drivers. The value of getting a trip (relative to remaining idle) only varies across available drivers with their distance to the rider: all drivers would get the same net earnings if they did the trip, but those who are closer would spend less time completing it. Surge pricing has no effect on the matching process, so it does not help reallocate trips towards drivers who are closer. Thus, there is no effect on driver surplus from trip allocation.[48] The bars at the right in figure 9b measure allocative welfare gains.

The value of completing a trip is homogeneous across drivers because of the simple form I assume for utility. I do not take into account the fact that actual drivers have different preferences for long or short trips, for trips to certain neighborhoods, or for trips that require driving in traffic. Surge pricing, however, might not help allocate trips more efficiently along these dimensions because drivers observe very little information before accepting a trip. They do not know the destination or the distance to the destination; all they know is the surge multiplier and how long it will take to pick up the rider.

Optimal uniform pricing If Uber can only do uniform pricing—because of regulation, for instance—it would reoptimize and set the multiplier at a different level from the status quo average multiplier. I assume that Uber would set it at the level that maximizes rider surplus, 1.253, consistent with the fact that the status quo price level maximizes rider surplus. In appendix F.1 I consider an alternative

[47] On average, surge pricing only saves a few seconds per trip for riders. Appendix F.4 shows that these savings are spread unevenly, and that the most clear effect is that surge pricing cuts down the upper tail of the pickup time distribution.

[48] Surge pricing lowers pickup times because it results in more available drivers at the times and places that riders request trips, but not because it reallocates trips among available drivers.

assumption—that Uber maximizes a weighted sum of rider surplus, driver surplus, and profits that rationalizes the status quo price level—with very similar results.

The average multiplier that maximizes rider surplus is higher with uniform pricing than with surge pricing. Castillo et al. (2018) explain why in detail. Driver scarcity is bad for riders because they are matched to far-away drivers. 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 that Castillo et al. call a *wild-goose chase*.

The implication for the market is that at any given time and place, it is much worse for the platform to set prices too low than too high. That is not a big issue with surge pricing: the algorithm takes care of avoiding prices that are too low. But with uniform pricing, it is optimal to set a high multiplier to avoid wild-goose chases and a drop in rider surplus. Appendix F.5 provides evidence that this phenomenon is why the optimal uniform multiplier is higher.

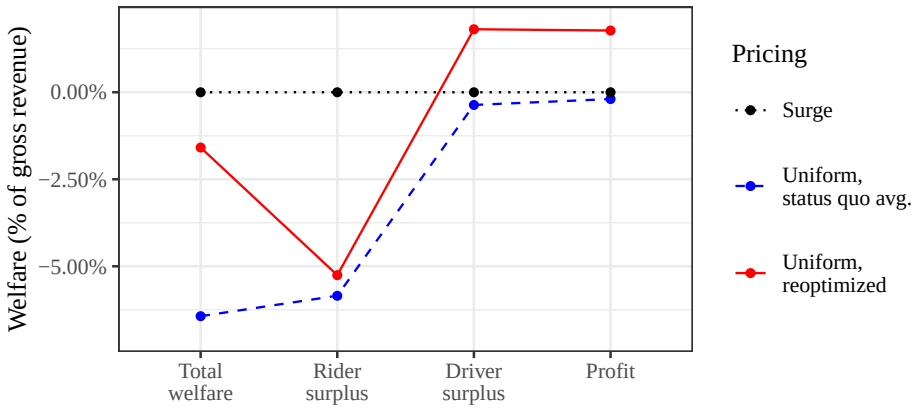


Figure 10: Welfare for different pricing policies

Note: This figure shows welfare for different pricing policies relative to surge pricing (the dotted, black line). The dashed, blue line represents a uniform multiplier at the average from the status quo. The solid, red line is for a uniform multiplier at the level that maximizes rider surplus.

Figure 10 shows the welfare effect of surge pricing on every side of the market. Moving from the dashed line to the dotted line measures effects if average multipliers are fixed—i.e., due to time savings and a better trip allocation. Moving from

the solid line to the dotted line additionally takes into account the fact that surge pricing results in lower average prices, reducing profits and driver surplus. Surge pricing increases rider surplus by 5.25% of gross revenue; a more efficient allocation and time savings substantially increase rider surplus, and lower prices lead to a tiny decrease. In contrast, there is a net decrease in driver surplus and profits of 1.81% and 1.77%, respectively, of gross revenue: the decrease from lower prices overtakes the small increase from time savings.

Appendix F.2 shows that these results are robust to higher long-run elasticities. Magnitudes differ, but the main qualitative results still hold.[49] I also show that the main findings also hold if the value of time for riders is lower.

6.3 Distributional effects

Riders One might be concerned that, despite the overall increase in rider surplus, low-income riders might be hurt by surge pricing. Figure 11a, which depicts rider surplus for different income levels, shows that that is not the case.[50]

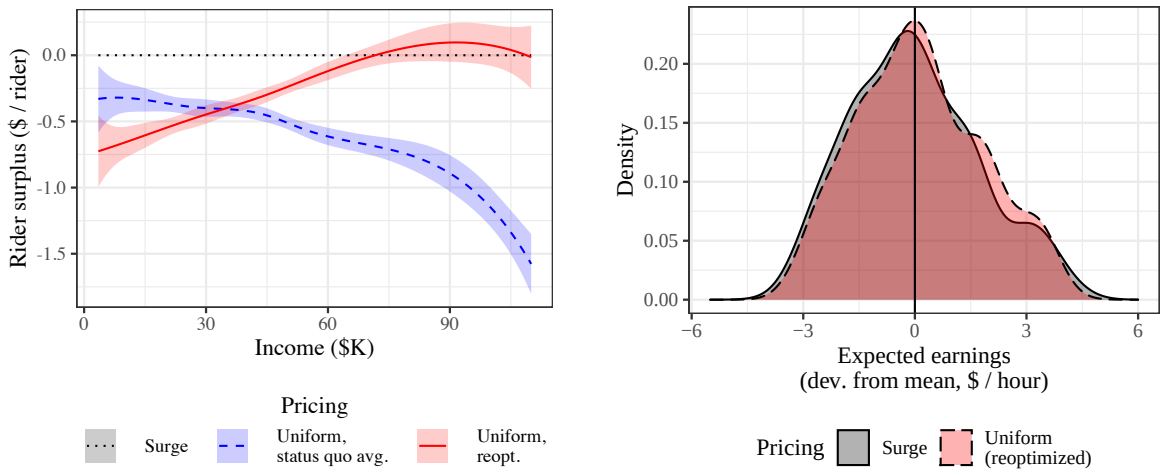
The effects of surge pricing due to a better allocation and time savings (moving from the dashed line to the dotted line) benefit riders across all income levels. High-income riders benefit most: they have a higher value of time and lose more when they are denied a trip. On the other hand, a lower average multiplier has different effects across income levels. Low-income riders are price sensitive, and thus prefer low prices. High-income riders prefer higher prices that skim other riders and bring in more drivers, resulting in more reliable trips with lower pickup times. Appendix F.3 shows further evidence on the effects by income.

The net effect of surge pricing—accounting for price variability and for lower average prices—can be seen by moving from the solid line to the dotted line. Low-income riders are better off. Benefits decline with income, and the highest income riders are roughly indifferent between uniform and surge pricing.[51] The decrease

[49] When demand elasticity is high, driver surplus and profits are somewhat higher with surge pricing—the effect from lower prices is smaller than the effect from better matching. As in the main results, the net effect is smaller in magnitude than the increase in rider surplus.

[50] The effects on occasional riders resemble those on low-income riders. Relative to surge pricing, their surplus with uniform pricing is lower by $0.24 per rider with the average multiplier from the status quo, and by $0.47 per rider with the multiplier that maximizes rider surplus.

[51] In fact, the net effect of surge pricing is higher average pickup times and a worse trip allocation:



(a) Rider surplus by income level

(b) Drivers' expected earnings

Figure 11: Heterogeneity in welfare effects within riders and drivers

Note: Subfigure (a) shows rider surplus by income level, relative to the status quo. For each pricing policy, I fit a cubic spline with six degrees of freedom based on simulated data for 45 weeks. The lines are differences between these fits. Shaded areas represent 95% confidence intervals. Subfigure (b) shows kernel density plots of drivers' expected hourly earnings. Each observation is the average earnings of drivers who started working in one entry location during one hour of the week.

in average prices is strong enough that the main beneficiaries of surge pricing are low-income riders. This evidence suggests that redistribution within riders may not be a first-order concern.

Drivers Figure 11b shows the distribution of drivers' expected hourly earnings, both in the status quo and with uniform pricing at the level that maximizes rider surplus. Surge pricing not only reduces average hourly earnings, it also increases the dispersion in earnings. With surge pricing, multipliers go down at times of low demand, precisely when hourly earnings are lowest. 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.

price effects wash out all the gain. Rider surplus is higher only because riders pay less.

7 Conclusion

In the debate about the desirability of surge pricing, standard arguments about efficiency gains are challenged by concerns that individual market participants might be hurt. I 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 across all income levels. 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. Given that earnings are only slightly above minimum wages, even small effects on drivers might be a concern.

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 assumptions about agents' multi-homing behavior might shed light on these issues.

References

Angrist, Joshua D., Sydnee Caldwell, and Jonathan Hall, "Uber vs. Taxi: A Driver's Eye View," *Forthcoming, AEJ: Applied Economics*, 2020.

Armstrong, Mark, "Competition in Two-Sided Markets," *The RAND Journal of Economics*, 2006, *37* (3), 668–691.

Besbes, Omar, Francisco Castro, and Ilan Lobel, "Surge Pricing and its Spatial Supply Response," *Working paper*, 2019.

Bezos, Jeffrey P., 2012 Letter to Shareholders, April 2013.

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: Evidence from Auctioned Cab Rides," *Working paper*, 2020.

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, Kuan-Ming, Ning Ding, John A List, and Magne Mogstad, "Reservation Wages and Workers' Valuation of Job Flexibility: Evidence from a Natural Field Experiment," *Working paper*, 2020.

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.

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. Click here to open URL.

Dinerstein, Michael, Liran Einav, Jonathan Levin, and Neel Sundaresan, "Consumer Price Search and Platform Design in Internet Commerce," *American Economic Review*, July 2018, *108* (7), 1820–59.

Fradkin, Andrey, "Search, matching, and the role of digital marketplace design in enabling trade: Evidence from Airbnb," *Working paper*, 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.

Ghili, Soheil and Vineet Kumar, "Spatial Distribution of Supply and the Role of

Market Thickness: Theory and Evidence from Ride Sharing," *Working paper*, 2020.

Goncharova, Masha, "Ride-Hailing Drivers Are Slaves to the Surge," *The New York Times*, January 2017. Click here to open URL.

Kazmin, Amy, "New Delhi bans Uber 'surge pricing'," *Financial Times*, April 2016. Click here to open URL.

Korolko, Nikita, Dawn Woodard, Chiwei Yan, and Helin Zhu, *Dynamic Pricing and Matching in Ride-Hailing Platforms*, Working Paper, 2018.

Kreindler, Gabriel E, "The welfare effect of road congestion pricing: Experimental evidence and equilibrium implications," *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.

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.

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.

Rosenthal, Brian M., "New York Is Urged to Consider Surge Pricing for Taxis," *The New York Times*, January 2020. Click here to open URL.

Shapiro, Matthew H, "Density of Demand and the Benefit of Uber," *Working paper*, 2018.

Weyl, E. Glen, "A Price Theory of Multi-Sided Platforms," *American Economic Review*, 2010, *100* (4), 1642–1672.

Yee, Jovic, "Commuters protest Grab's high fare; TNC firm denies surge pricing," *Inquirer*, April 2018. Click here to open URL.

Online Appendix

Appendix A Remaining details about the model

A.1 Pickup duration and distance

Rider i's pickup time is a function $w(\mathbf{a}_t, l, h)$ of the rider's location and the hour of the week, and of $\mathbf{a}_t$, the number of available drivers in every nearby location.

If the rider requests a trip and is matched to driver j, the pickup duration that enters utility is the one that was generated in the matching process (section 3.3). The pickup distance, which is relevant to compute driver costs, is 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 $w(\mathbf{a}_t, l, h)$ is the prediction from a random forest of pickup times 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 location. G^{pickup} is a shifted Gamma distribution with minimum one. I set the distribution parameters so that the average and variance match their empirical counterparts.

A.2 Rider destination, trip distance and duration, and base fare

Rider i opened the app at time t during hour of the week h in 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.

The trip distance is equal to the straight-line distance between the origin and destination times a factor that is drawn from a distribution $G^{dist}(\cdot|l, k, h)$. The trip duration is equal to the trip distance times a factor drawn from a distribution $G^{duration}(\cdot|l, k, h)$. The base fare is a function of the trip distance and duration, using the fare structure used by Uber at the time: 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.

$G^{dest}(\cdot|l,h,\tilde{r})$ 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 $\frac{\nu_k \mu_{o_l d_k} \lambda_{h d_k}}{\sum_{k' \in K_{l\tilde{r}}} \nu_{k'} \mu_{o_l d_{k'}} \lambda_{h d_{k'}}}$. $\mu_{o_l d_k}$ is the fraction of trips originating in o_l that go to d_k. $\lambda_{h d_k}$ is 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 estimate it as the empirical probability.

$G^{dist}(\cdot|l,k,h)$ is a lognormal distribution with parameters $(\mu^{dist}_{lkh}, \sigma^{dist}_{lkh})$. I estimate μ_{lkh} as a linear model of the log ratio between trip distance and straight-line distance on origin group by destination group and hour fixed effects. I estimate σ^{dist}_{lkh} with a linear model of residual standard deviation by bins of μ^{dist}_{lkh}. $G^{duration}(\cdot|l,k,h)$ is a lognormal distribution with parameters $(\mu^{duration}_{lkh}, \sigma^{duration}_{lkh})$, which I estimate just as with $G^{dist}(\cdot|l,k,h)$, from a model of the log ratio between trip duration and distance.

A.3 Outside behavior

I split Houston into three regions. The first one is the main region of analysis, where my full model applies. The second one is a buffer zone surrounding the main region, which has roughly the same area, in which I model drivers' movements and match riders and drivers just as in the inside region, but I take demand to be exogenous. Finally, the third area includes all locations outside of the buffer area. I do not model drivers' movements in this area. Instead, I model a pool of outside drivers that are be matched to exogenous demand.

Let l be a location outside the central region (the whole outermost region is 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 $\nu_{lh\tilde{r}}$. 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 $G^{pickup,buffer}$.

The movement model applies to the buffer area. Moving outside is one more option in the choice set. Drivers that are outside move to location l in the buffer area with probability $p^{movein}(l;h)$ and stay outside otherwise. Drivers that drop off passengers after trips that end in the outside area join the pool of outside drivers.

Estimation $G^{dest,out}(\cdot|l,h,\tilde{r})$, $G^{dist,out}(\cdot|l,k,h)$, $G^{duration,out}(\cdot|l,k,h)$, $G^{pickup,out}(\cdot|h)$, 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 take $p^{movein}(l;h)$ to be empirical frequencies.

Appendix B Details about equilibrium

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

Proof. $f^P(\cdot)$ is continuous: all functions that are involved are continuous. $\mathcal{X}$ is bounded below by the greatest possible loss for a driver (moving back and forth between the two farthest locations) and by zero utility, and above by the earnings a driver gets if he always gets the most profitable trip possible immediately and by riders' utility if prices and pickup times are zero. $\mathcal{X}$ is thus a convex, compact space, so by Brouwer's fixed point theorem $f^P(\cdot)$ has a fixed point. □

The following assumption is necessary to prove uniqueness:

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

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

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

Proof. Consider $g_\gamma : \mathcal{X} \to \mathcal{X}$ defined by $g_\gamma(\mathbf{x}) = (1-\gamma)\mathbf{x} + \gamma f^P(\mathbf{x})$, where $0 < \gamma < 1$. The set of fixed points of f^P 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 $\frac{||f^P(\mathbf{x}) - f^P(\mathbf{x}')||}{||\mathbf{x}-\mathbf{x}'||} < \beta$. For all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ we have that $||g_\gamma(\mathbf{x}) - g_\gamma(\mathbf{x}')||^2 = (1-\gamma)^2||\mathbf{x}-\mathbf{x}'||^2 + 2\gamma(1-\gamma)\langle f^P(\mathbf{x}) - f^P(\mathbf{x}'), \mathbf{x}-\mathbf{x}'\rangle + \gamma^2||f^P(\mathbf{x}) - f^P(\mathbf{x}')||^2 < \left[(1-\gamma)^2 + 2\gamma(1-\gamma)\delta + \gamma^2\beta^2\right]||\mathbf{x}-\mathbf{x}'||^2$. A Taylor expansion about $\gamma = 0$ of the term in brackets is $(1-\gamma)^2 + 2\gamma(1-\gamma)\delta + \gamma^2\beta^2 = 1 + 2(\delta-1)\gamma + O(\gamma^2)$. This 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 $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$. □

Stability Suppose that only a fraction γ of agents update their beliefs in every iteration, or that in every iteration riders only give weight γ to new observations. The mapping $g_\gamma(\cdot)$ represents this belief update process. Since $g_\gamma(\cdot)$ is a contraction mapping, the market is stable according to this process.

Computation I can only compute by simulation an estimator $\hat{f}^P(\mathbf{x})$ such that $\mathrm{E}[\hat{f}^P(\mathbf{x})] = f^P(\mathbf{x})$, so it is hard to assess convergence because of simulation randomness. Additionaly, a naive iterative algorithm, where a sequence of beliefs $(\mathbf{x}_n)_{n=1}^\infty$ is generated according to $\mathbf{x}_{n+1} = \hat{f}^P(\mathbf{x}_n)$, often diverges.

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

$$\mathbf{x}_{n+1} = (1-\gamma_n)\mathbf{x}_n + \gamma_n \hat{f}^P(\mathbf{x}_n), \qquad \gamma_n \propto n^{-b}, \tag{13}$$

where $b \in (0,1)$. New beliefs are a convex combination of old beliefs and the new empirical average. I need the following assumption to guarantee convergence:

Assumption 2. *Draws from simulation averages take the form $\hat{f}^P(\mathbf{x}) = f^P(\mathbf{x}) + \varepsilon$, where ε is a vector of mean-zero independent random variables with bounded variance.*

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

Proposition 3 (Convergence). *Let $(\mathbf{x}_n)_{n=0}^{\infty}$ be a sequence over $\mathcal{X}$ defined iteratively by equation* (13)*, where $\mathbf{x}_0 \in \mathcal{X}$. Under assumptions 1 and 2, $\mathbf{x}_n \xrightarrow{p} \mathbf{x}^*$.*

Proof. Note that $\mathrm{E}\left[||\mathbf{x}_n - \mathbf{x}^*||^2\right] = \mathrm{E}\left[||(1-\gamma_n)\mathbf{x}_{n-1} + \gamma_n f_n^P(\mathbf{x}_{n-1}) - \mathbf{x}^*||^2\right] + \gamma_n^2 \mathrm{Var}\left[\epsilon\right]$ (by the independence of ε). The algebra in the proof of proposition 2 implies that $||\mathbf{x}_n - \mathbf{x}^*||^2 < \left[(1-\gamma_n)^2 + 2\gamma_n(1-\gamma_n)\delta + \gamma_n^2\beta^2\right]||\mathbf{x}_{n-1} - \mathbf{x}^*||^2$ pointwise, so $V_n < \left[(1-\gamma_n)^2 + 2\gamma_n(1-\gamma_n)\delta + \gamma_n^2\beta^2\right]V_{n-1} + \gamma_n^2\Gamma$, where $V_n = \mathrm{E}\left[||\mathbf{x}_n - \mathbf{x}^*||^2\right]$ and $\Gamma < \infty$ is such that $\Gamma > \mathrm{Var}\left[\epsilon\right]$. This recursion over V_n converges to zero (see Broadie et al., 2011, online appendix), so $\mathbf{x}_n$ converges in probability to $\mathbf{x}^*$. □

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 $b = 0.65$, which typically leads to stable beliefs after around 10 iterations.

Appendix C Formal results about the identification strategy

C.1 Demand (trip request model)

My identification strategy relies on the following assumption:

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

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 $(\epsilon_i, x_i, \tilde{\mathbf{m}}_t)$.*

Under assumption 3, the causal effect of prices and pickup times can be isolated by controlling for a 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*

$$y_i = \alpha(x_i, l, h) + \beta(x_i)p_i + \gamma(r_i)w_i + g(\tilde{\mathbf{m}}_t, x_i, w_{lt}^0; l, h) + \eta_i, \tag{14}$$

where $g(\tilde{\mathbf{m}}_t, x_i, w_{lt}^0, l, h) = E[\epsilon_i|\tilde{\mathbf{m}}_t, x_i, w_{lt}^0; l, h]$ and $E[\eta_i|p_i, w_i, \tilde{\mathbf{m}}_t, x_i, w_{lt}^0, l, h] = 0$.

Proof. We can write $\epsilon_{it} = g(\tilde{\mathbf{m}}_t, x_i, w_{lt}^0; l, h) + \eta_i$, where $E[\eta_i|\tilde{\mathbf{m}}_t, x_i, w_{lt}^0, l, h] = 0$ by the definition of conditional mean. This yields equation (14). Since p_i is a deterministic function of $(\tilde{\mathbf{m}}_t, \bar{p}_i, l)$, $\bar{p}_i$ is part of x_i, and $w_i = w_{lt}^0 + \xi_i$, then $E[\eta_i|p_i, w_i, \tilde{\mathbf{m}}_t, x_i, w_{lt}^0, l, h] =$

$E[\eta_i|w_i, \tilde{\mathbf{m}}_t, x_i, l, h, \xi_i] = E[\epsilon_i|w_i, \tilde{\mathbf{m}}_t, x_i, l, h, \xi_i] - E[E[\epsilon_i|\tilde{\mathbf{m}}_t, x_i, w^0_{lt}; l, h]|w_i, \tilde{\mathbf{m}}_t, x_i, l, h, \xi_i]$. By assumption, ξ_i is orthogonal to $(\epsilon_i, x_i, \tilde{\mathbf{m}}_t)$, and by its definition, it is orthogonal to (w^0_{lt}, l, t). The dependence on ξ_i can thus be dropped out of both terms in the last expression, which is then zero. □

I estimate equation (14). I can estimate it by standard regression techniques by proposition 4: there are no remaining endogeneity issues after including $g(\cdot)$.

The following assumption is necessary to ensure $\beta(x_i)p_i$ is 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. *$E[\epsilon_i|\tilde{\mathbf{m}}_t, x_i, \bar{w}_{lt}, l, h]$ is continuously differentiable in $\tilde{\mathbf{m}}_t$.*

The first part 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 $X(\tilde{\mathbf{m}}_t, x_i, w^0_{lt}; l, h)$ that is continuously differentiable in $\tilde{\mathbf{m}}_t$ and such that $E[\epsilon_i|\tilde{\mathbf{m}}_t, x_i, \bar{w}_{lt}, l, h] \in \text{Span}(X(\tilde{\mathbf{m}}_t, x_i, w^0_{lt}; l, h))$. For any such vector, and for any nonzero function $f(x_i)$, $f(x_i)p(\tilde{\mathbf{m}}_t; \bar{p}_i, l) \notin \text{Span}(X(\tilde{\mathbf{m}}_t, x_i, w^0_{lt}; l, h))$.*

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

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

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 $l_{j,t+1} = \text{argmax}_k\, y^k_{j,t+1}$, where

$$y^k_{j,t+1} = \omega(l,k,h) + \delta v_k + \zeta^k_{j,t+1} \tag{15}$$

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 controlling for a function that depends on $\tilde{\mathbf{m}}_t$:

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

$$y^k_{j,t+1} = \omega(l,k,h) + \delta v_k + g^M(\tilde{\mathbf{m}}_{tk}; l, h) + \psi^k_{j,t+1}, \tag{16}$$

where $g^M(\tilde{\mathbf{m}}_{tk}; l, h) = E[\zeta^k_{j,t+1} | \tilde{\mathbf{m}}_{tk}; l, h]$ and $E[\psi^k_{j,t+1} | v_k, \tilde{\mathbf{m}}_{tk}, l, h] = 0$.

Proof. We can write $\zeta^k_{j,t+1} = g^M(\tilde{\mathbf{m}}_{tk}; l, h) + \psi^k_{j,t+1}$, where $E[\psi^k_{j,t+1} | \tilde{\mathbf{m}}_{tk}; l, h] = 0$ by the definition of conditional mean. This yields equation (16). Since v_k is a deterministic function of $\tilde{\mathbf{m}}_{tk,}$, $E[\psi^k_{j,t+1} | v_k, \tilde{\mathbf{m}}_{tk}, l, h] = E[\psi^k_{j,t+1} | \tilde{\mathbf{m}}_{tk}, l, h] = 0$. □

I estimate equation (16). I can estimate it by standard regression techniques by proposition 6: there are no remaining endogeneity issues after including $g^M(\cdot)$. If $\psi^k_{j,t+1}$ are iid EV type I random variables, then this model reduces to equation (8).

The following assumption is necessary to ensure δv_k is identified:

Assumption 6. *The latent variable model for drivers' movement, equation* (15)*, 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. *$E[\zeta^k_{j,t+1} | \tilde{\mathbf{m}}_{tk}; l, h]$ is continuously differentiable in $\tilde{\mathbf{m}}_t$.*

The first part is satisfied 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 reasonable 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 $E[\epsilon_i | \tilde{\mathbf{m}}_{tk}, l, h] \in \text{Span}(Z(\tilde{\mathbf{m}}_{tk}; l, h))$. For any such vector, $v_k \notin \text{Span}(Z(\tilde{\mathbf{m}}_{tk}; l, h))$.*

Proof. $Z(\tilde{\mathbf{m}}_{tk}; l, h) = E[\zeta^k_{j,t+1} | \tilde{\mathbf{m}}_{tk}, l, h]$ is continuously differentiable in $\tilde{\mathbf{m}}_t$ and $E[\zeta^k_{j,t+1} | \tilde{\mathbf{m}}_{tk}, l, h]$ 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 $v_k \notin \text{Span}(Z(\tilde{\mathbf{m}}_{tk}; l, h))$. □

Just as with demand, a practical concern is finding a vector $Z(\tilde{\mathbf{m}}_{tk}; l, h)$ whose span contains $E[\zeta^k_{j,t+1} | \tilde{\mathbf{m}}_{tk}, l, h]$. I use a combination of high order splines.

Appendix D Data and estimation

D.1 Identifying riders' home location

I start with all trips between February 18 and May 31, 2017. I aggregate each rider's origins and destinations into clusters of points that are within 200 meters of each other.[52] I give each cluster a score for how likely it is to be the home location.[53] For each rider, I pick the cluster with the largest score among those clusters that have at least two points that contribute to the score and call it the home address. I then assign to each rider the median income for the home address census tract.[54]

D.2 Unrounded multipliers and prices

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. This model has the same form as the surge pricing algorithm (section 4.1), except that I omit rounding steps.

Let $\bar{m}_{lt}$ be the *bounded multiplier*, which is the recommended multiplier subject to upper and lower bounds that the surge pricing algorithm sets on multipliers to

[52] I run an agglomerative hierarchical clustering algorithm with complete linkage.

[53] The score is the sum of (a) how many trips started 4-10 am during the week, (b) 0.5 times how many trips ended 3-6 pm during the week, (c) 0.3 times how many trips ended 7 pm-midnight during weekdays (except for Friday), and (d) 0.2 times how many trips started 6 am-noon during weekends or ended between 7 pm-midnight Friday through Sunday.

[54] I create alternative income variables based on scores that use different weights, that focus only on morning weekday trips, and that require a larger number of origins or destinations to call it a home address. All income variables have over 97% correlation within the sample of riders they assign a home to, but they differ in the fraction of riders that get assigned an income variable.

avoid abrupt multiplier changes. I estimate a model of the following form:

$$m_{lt} = \alpha_{0,l} h_0(\bar{m}_{lt}) + \sum_{r=1}^{r^{max}} \alpha_{r,l} \sum_{k \in R_{r,l}} h_r(\bar{m}_{kt}) + \phi_{lt}. \tag{17}$$

$R_{r,l}$ are the locations at a distance r from l. The α coefficients are weights given to multipliers at a given distance from l. $h_r(\cdot)$ represent a flexible functional form for the dependence on bounded multipliers at a distance r.

This model takes into account the fact that the surge pricing algorithm gives the same weight to all nearby locations that are at the same distance. α coefficients vary by location: more dense areas give weights to a smaller number of nearby locations.[55] The function for each distance is a spline with 10 degrees of freedom that is the same for all locations at the same distance.[56]

Figure 12 shows the residual variation ϕ_{lt} that identifies riders' and drivers' response to prices. There is substantial variation, especially around the gap created by the fact that the multiplier cannot be 1.1.

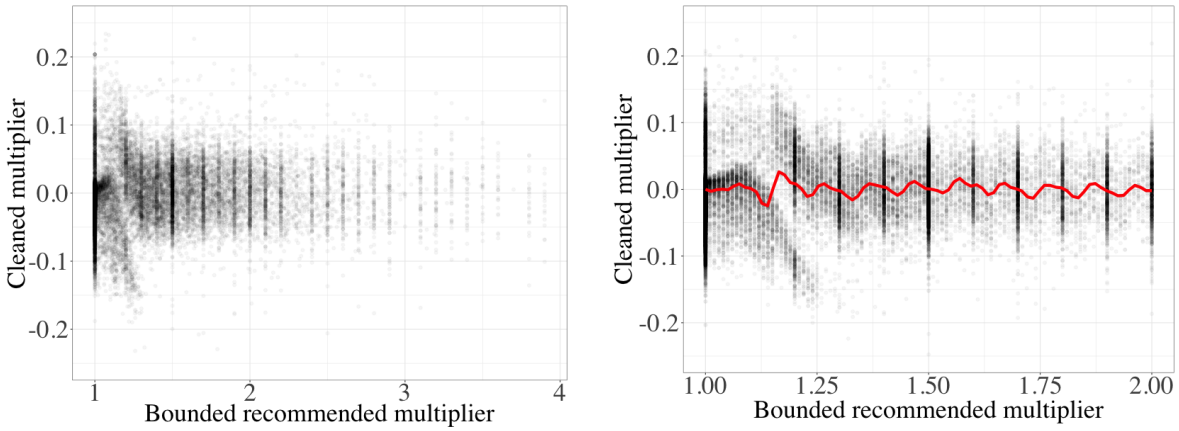


Figure 12: Residual variation in multipliers

Note: Residual of regression (17) 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 nonparametric fit.

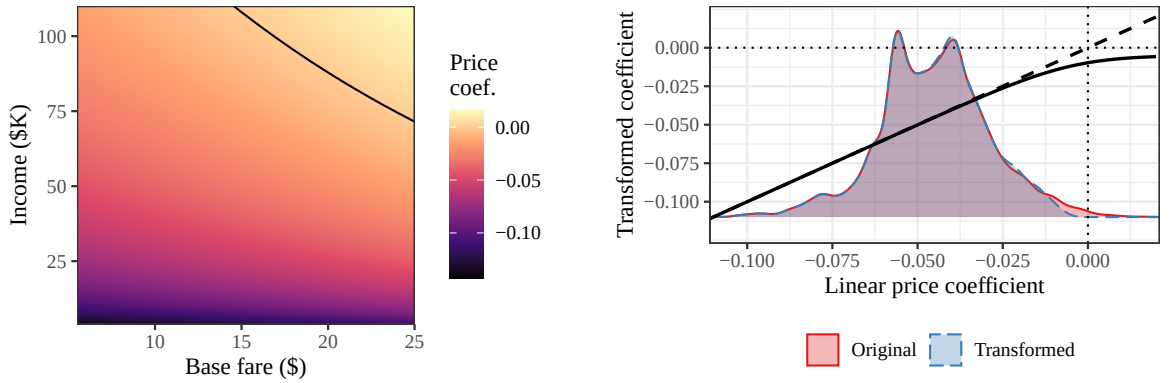
After estimating the unrounded multiplier, I also compute the *unrounded price* $\hat{p}_i = b_i + \hat{m}_{lt}(\bar{p}_i - b_i)$, which is the fare the rider would have seen without rounding.

[55] Locations also have different number of nearby locations, so weights vary with that number.
[56] Increasing the degrees of freedom has essentially no effect.

D.3 Demand estimation—functional form assumptions

Price coefficient functional form Figure 13a shows the price coefficient for weekday, non-airport trips that is estimated with a linear specification, $\beta(x_i) = \theta^p x_i$. The coefficient is negative in the region above the black line, which has only a few observations: only 0.8% of the sample has a negative price coefficient.[57]

I adjust the coefficient by setting $\beta(x_i) = s(\theta^p x_i)$, where $s(x) = -\frac{1}{a}\log(e^{-ab} + e^{-ax})$. I set $a = 100$ and $b = -0.005$ (varying a and b only has minor effects on my estimation). The black, solid line in figure 13b shows $s(\cdot)$, with the identity function as a dashed line for reference. This is essentially equivalent to a linear specification as long as $\theta^p x_i$ is below -0.025. From that point on, $s(\cdot)$ adjusts the coefficient so it asymptotes to b. The solid-line histogram shows the distribution of the original coefficient $\theta^p x_i$ before applying $s(\cdot)$. The dashed-line histogram shows the distribution after the adjustment. The only noticeable difference is that the right tail, which crosses zero, is cut down to ensure the coefficient is always negative.



(a) Price coefficient with a linear specification

(b) Nonlinear specification for price coefficient

Figure 13: Price coefficient and adjustments

Note: Subfigure (a) shows the price coefficient for weekday, non-airport trips based on a model with linear coefficients. It is negative above the black line. Subfigure (b) shows how I adjust the coefficient to ensure it is never positive. The black, solid line is the transformation I apply. The histograms show the distribution of the linear coefficient and the adjusted coefficient.

Control function I set $g(\tilde{\mathbf{m}}_t, x_i, w^0_{lt}; l, h) = \tilde{\alpha}(x_i; l, h) + g^1(\hat{p}_i) + g^2(\tilde{\mathbf{m}}_t) + g^3(w^0_{lt})$, where all individual functions are cubic splines with knots placed evenly at quan-

[57] They are negative because of extrapolation, which could be solved with a more flexible specification, but I do not have enough power to identify a more flexible model.

tiles 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. Neither omitting $g^2(\cdot)$ nor increasing the order of the splines has any noticeable impact on my results. I also include $g^2(\tilde{\mathbf{m}}_t)$ to control for any variation that may not have been captured by $\hat{p}_i$. The term $\tilde{\alpha}(x_i, l, h)$ is absorbed by the original intercept term $\alpha(x_i; l, h)$.

Fixed effects I assume $\alpha(x_i; l, h)$ is additively separable into a linear function of x_i and a function of (l, h). The latter is flexible to capture broad demand patterns. Fixed effects by (l, h) would result in too many parameters (1681×168). Instead, I 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 total, this specification has 155 degrees of freedom.[58]

D.4 Aggregating locations and hours of the week

Rider arrival To compute U_{lhx}, I aggregate incomes into *income groups*—quantiles plus a separate group for occasional riders. I also aggregate distances into *distance groups*—quantiles plus a separate group for airport trips. Let $\tilde{y}$ and $\tilde{r}$ denote the income and distance groups x belongs to. I aggregate locations into 16 zones a with similar number of arrivals, and I aggregate hours of the week into 14 groups g.[59] I set $U_{lhx} = \tilde{U}(a(l), g(h), \tilde{y}, \tilde{r})$, where $a(l)$ is the zone that contains l, $g(h)$ is the hour group that contains h, and $\tilde{U}(a, g, \tilde{y}, \tilde{r})$ denotes average utility within $(a, g, \tilde{y}, \tilde{r})$.

I set $A_{hl\tilde{y}\tilde{r}} = \psi^d \tilde{A}_{a(l)g(h)\tilde{y}\tilde{r}} \chi_h^{g(h)} \chi_l^{a(l)}$. The term ψ^d is a uniform scale factor for all demand. $\tilde{A}_{a(l)g(h)\tilde{y}\tilde{r}}$ is a demand shifter, and it is the term I vary when I set $A_{lh\tilde{y}\tilde{r}}$ and ρ jointly. $\chi_h^{g(h)}$ 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.

[58] Linear models using this methodology and fixed effects both lead to very similar estimates.
[59] The groups are the same as in footnote 60.

Driver entry 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 60. I then set $W_{lh} = \tilde{W}_{s(l)g(h)}$, the average hourly earnings for all drivers who start working in the zone l belongs to $s(l)$ during the hour group h belongs to $g(h)$.

I set $B_{lh} = \psi^s \tilde{B}_{s(l)g(h)} \varphi_h^{g(h)} \varphi_l^{s(l)}$. The first term, ψ^s, is a uniform scale factor for supply. $\tilde{B}_{s(l)g(h)}$ is a supply shifter at the zone by hour group level, and it is the term I vary when I set B_{lh} and σ jointly. $\varphi_h^{g(h)}$ 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.

D.5 Movement model estimation

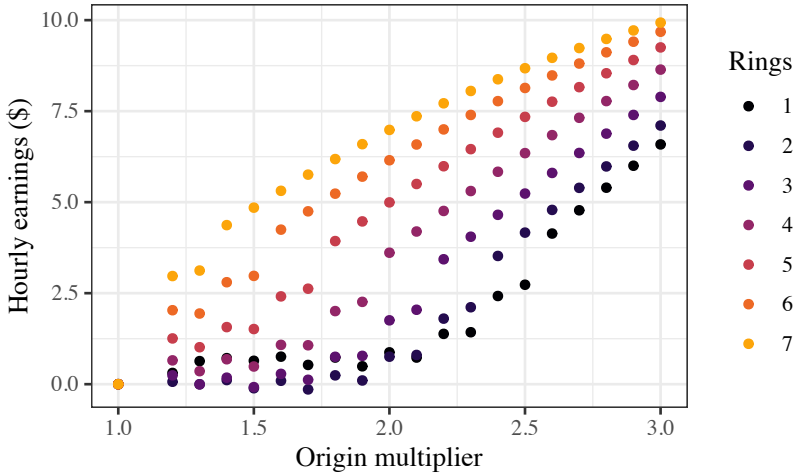


Figure 14: Expected hourly earnings as a function of multipliers

Note: Expected hourly earnings for the next 90 minutes as a function of surrounding surge multipliers, relative to all multipliers equal to one. Each series increases the multiplier at the origin. "Rings" refers to how far from the origin the multiplier starts to change. For rings=4, e.g., all multipliers up to a distance 4 are equal to the one at the origin. Multipliers further than 4 decay smoothly to one.

Mean future earnings fit $\alpha(l', h)$ plays the role of location by hour of the week fixed effects. I use the specification from the trip request model (appendix D.3).

The goal of $f(\mathbf{m}_{tk})$ is not to predict earnings as accurately as possible; instead, I want to capture an intuitive functional form that reflects drivers' expectations. 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. I constrain the coefficient for each term to be linear in the distance to the current location. Figure 14 shows how the estimated $f(\mathbf{m}_{tk})$ varies as multipliers change. Earnings increase as multipliers increase, and as more locations have higher surge levels. A surge multiplier of 1.5 in all surrounding hexes, for instance, results in expected earnings that are \$4.85 per hour higher than if all multipliers are 1.

Functional form assumptions I set $g^M(\tilde{\mathbf{m}}_{tk}; l, h) = \tilde{\omega}(l, h) + \tilde{g}^M(\tilde{\mathbf{m}}_{tk})$. $\tilde{g}^M(\tilde{\mathbf{m}}_{tk})$ 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.

$\tilde{\omega}(l, h)$ is absorbed by $\omega(l, k, h)$. I set $\omega(l, k, h) = \zeta_l^k + \chi_{z_l, g_h}^{n_k}$. ζ_l^k are origin by destination fixed effects, which capture the fact that drivers' movements are defined to a large extent by road patterns. $\chi_{z_l, g_h}^{n_k}$ models the fact that traffic patterns change over the week. It has 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 capture the general direction and distance of moving from l to k.[60,61]

I limit possible destinations k to the set of the most frequent destinations from each origin l. I include as many destinations as it is necessary to account for over 98% of movements from l. This results, on average, in 25 possible destinations.

Movement model estimation algorithm I estimate the movement model by maximum likelihood, which is computationally challenging because of the large number of fixed effects. I split the optimization problem into an outer loop that maximizes over δ and $g^M(\tilde{\mathbf{m}}_{tk})$ and an inner loop that maximizes over fixed effects.

[60] The 9 zones are the interaction of three quantiles for latitude and three quantiles for longitude. The 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. 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 distance group includes drivers who stay in the same location.

[61] The large number of fixed effects could lead to an incidental parameters problem, but Monte Carlo simulations show that the bias in δ is only around 2% of the parameter value.

The likelihood of one individual observation is $L_{jt} = \frac{\lambda^c_{jlt}}{\Lambda_{jlt}}$, where $\lambda^k_{jlt} = \exp(\alpha^k_l + \gamma^{m_k}_{z_l h_t} + \beta x^k_{jlt})$, and $\Lambda_{jlt} = \sum_{k \in K_l} \lambda^k_{jlt}$. The term βx^k_{jlt} represents $\delta v_k(\mathbf{s}_t) + g^M(\tilde{\mathbf{m}}_{tk}; l, h)$, 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. The likelihood maximization problem I wish to solve is $(\hat{\alpha}, \hat{\gamma}, \hat{\beta}) = \text{argmax}_{(\alpha,\gamma,\beta)} \sum_{jt} \log L_{jt}(\alpha, \gamma, \beta)$. The vector $(\hat{\alpha}, \hat{\gamma}, \hat{\beta})$ is high dimensional, so a standard nonlinear optimization algorithm takes too many iterations to converge. I follow an algorithm that finds quickly the optimal value of (α, γ) given β. Formally, I solve $\max_\beta \max_{(\alpha,\gamma)} \sum_{jt} \log L_{jt}(\alpha, \gamma, \beta)$. For the outer problem, I follow a Quasi-Newton algorithm. I now describe how I solve the inner problem.

Let f^k_l be the fraction of drivers that move to location k after starting in location l, and let p^k_{jlt} 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^k_l takes the form $f^k_l = \frac{1}{N_l} \sum_{jt} p^k_{jlt}$, where the sum is over all observations where the driver starts at location l. This means that sample fractions are equal to predicted probabilities. Similarly, if f^m_{zh} 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^m_{jlt} is the predicted probability of driver j moving to a location corresponding to m, the first order condition for γ^m_{zh} is $f^m_{zh} = \frac{1}{N_{zh}} \sum_{jlt} p^m_{jlt}$, where the sum is over all observations staring in z during h.

I compute fixed effects, conditional on the main model parameters, using both conditions. The probability p^k_{jlt} is $\frac{\exp(\alpha^k_l)\exp(\theta X^o_{jlt})}{\sum_o \exp(\alpha^o_l)\exp(\theta X^o_{jlt})}$, where θX_{jlt} corresponds to all variables that are not related to origin-destination fixed effects—but including movement trend fixed effects. Thus, $f^k_l = \frac{1}{N_l} \sum_{jt} p^k_{jlt}$ determines all fixed effects implicitly, and they can be computed by iterating on s for $\exp(\alpha^{k,s+1}_l) = \frac{f^k_l}{\frac{1}{N_l} \sum_{jt} \frac{\exp(\theta X^k_{jlt})}{\sum_o \exp(\alpha^{o,s}_l)\exp(\theta X^o_{jlt})}}$.

I follow an analogous process to compute γ^m_{zh}. The estimates of $\alpha^{k,s+1}_l$ and γ^m_{zh} depend on one another, so I iterate back and forth between them until convergence.

D.6 Shift length fit

Figure 15 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 the

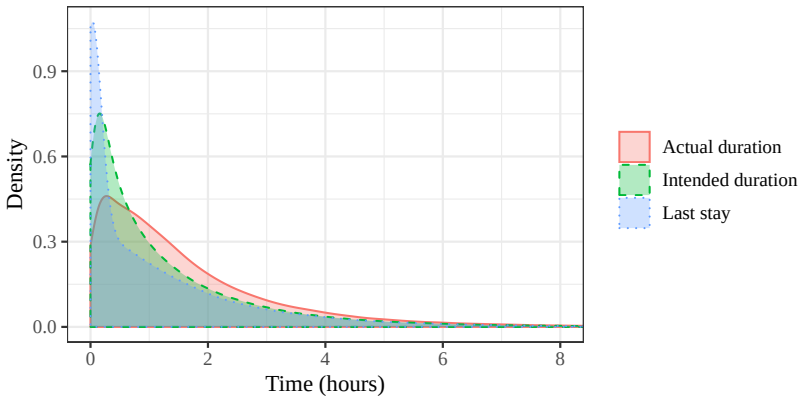


Figure 15: Empirical and estimated distribution of shift length

Note: The actual duration represents how long the driver worked. 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 other two variables.

other two. It shows that the assumption of a gamma distribution seems reasonable.

Appendix E Model fit

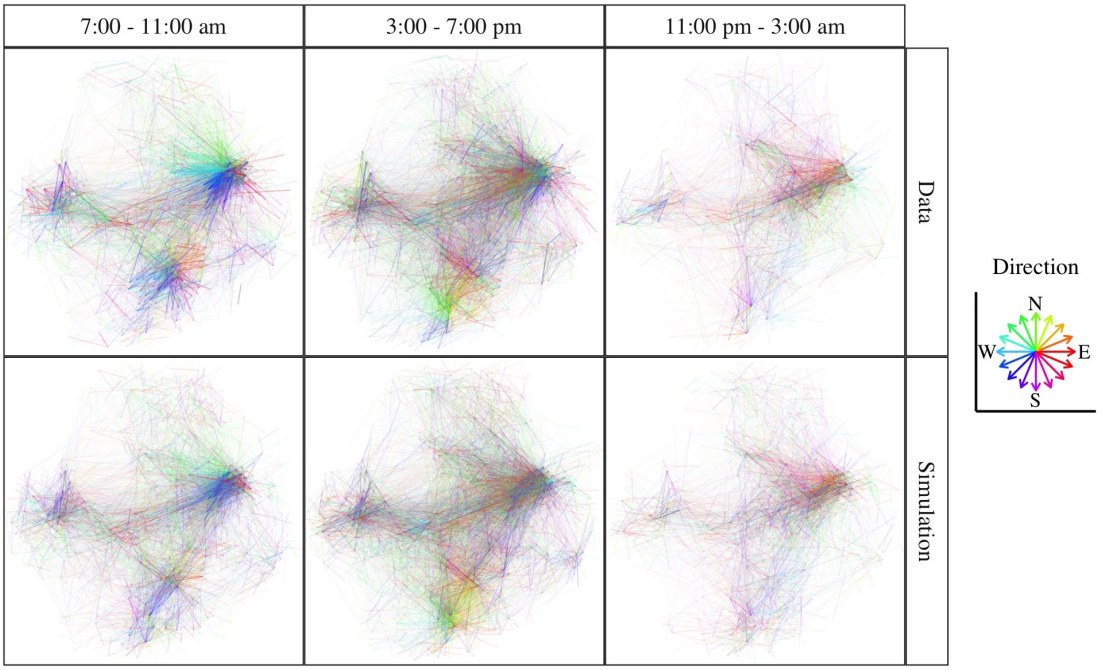
Figure 16 shows that the spatial patterns from simulations fit the data well. I focus on specific times of the week that have salient patterns, but other times also show a good fit. Figure 17 shows temporal patterns of supply, demand, and trips. Demand and the number of trips fit the data very well. The number of drivers is somewhat less precise, but it still follows the broad patterns in the data. Figure 18 shows that the simulated distribution of surge multipliers is almost identical to the data.

Appendix F Additional counterfactual results

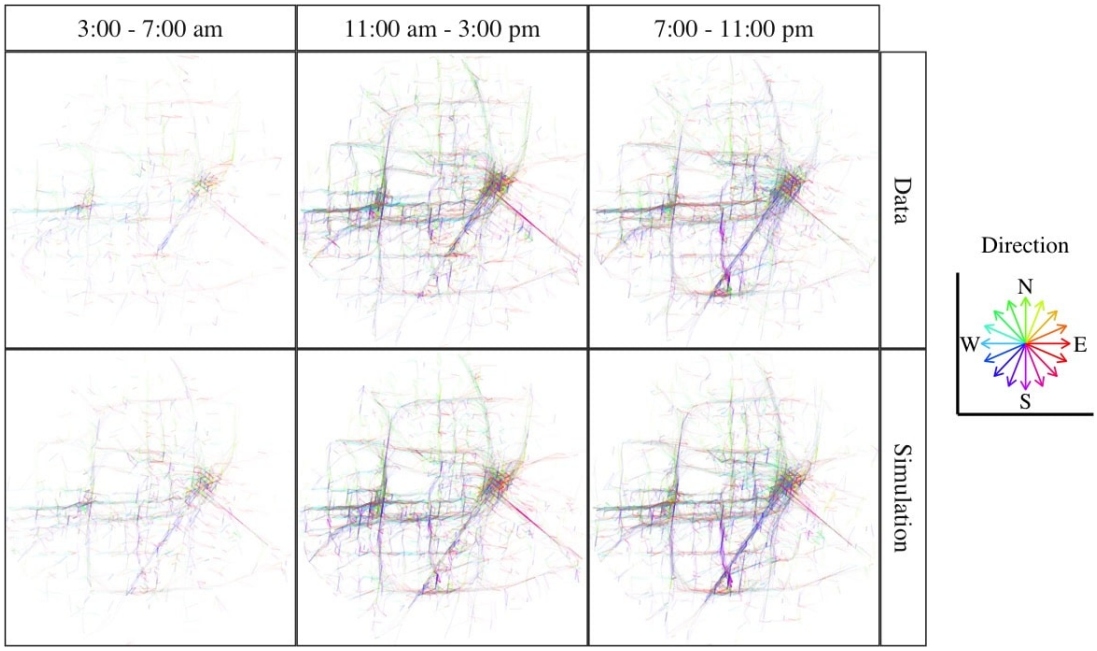
F.1 Alternative objective functions for Uber

In this section, I assume that to maximize long-run profits Uber maximizes a weighted sum of short-run profits, rider surplus, and driver surplus. It gives different weights to occasional riders, as well as to low, middle, and high income riders.

Let $\mathcal{P}$ be the set of policies the platform chooses from. Let $\Pi(P)$ and $DS(P)$ denote short-run profits and driver surplus, respectively, and $RS^G(P)$ denote rider surplus for group $G \in \{O, L, M, H\}$ of riders (occasional, low income, middle in-



(a) Trips



(b) Driver movements

Figure 16: Trips and driver movements in simulations and in the data

Note: Subfigure (a) shows a 20% random sample of trips at some 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.

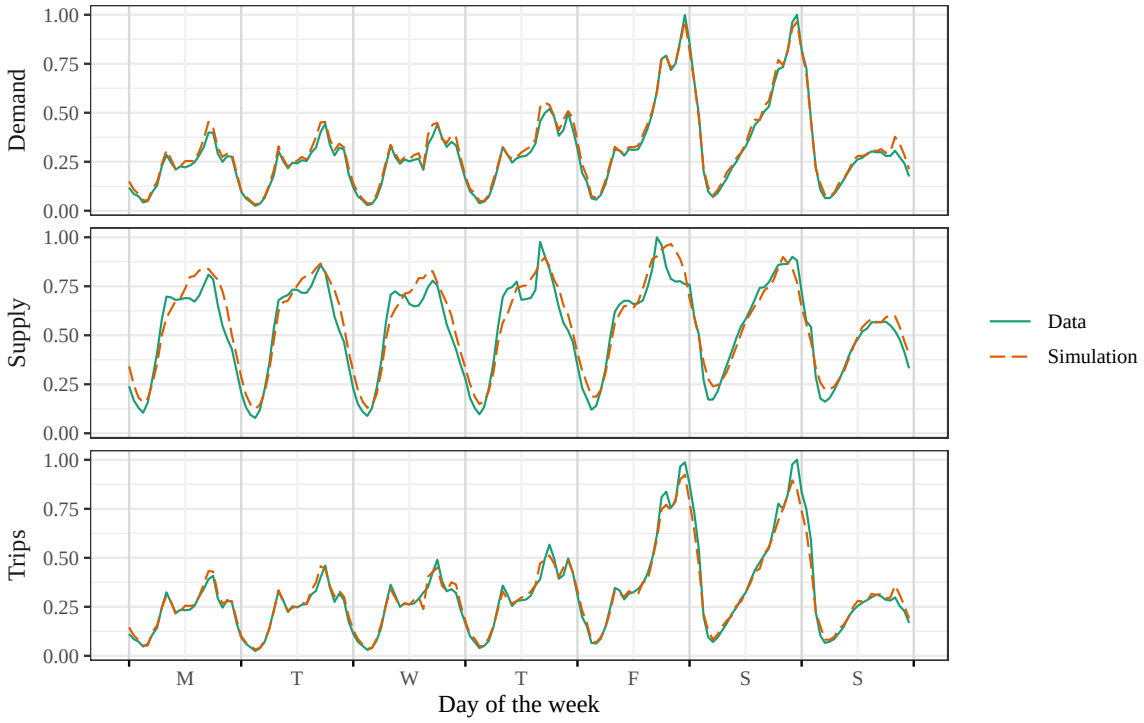


Figure 17: Temporal patterns in simulations and in the data

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.

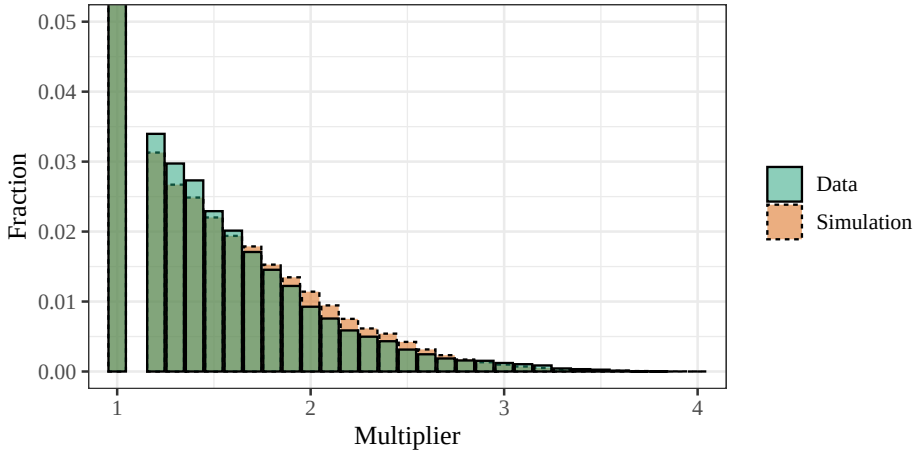


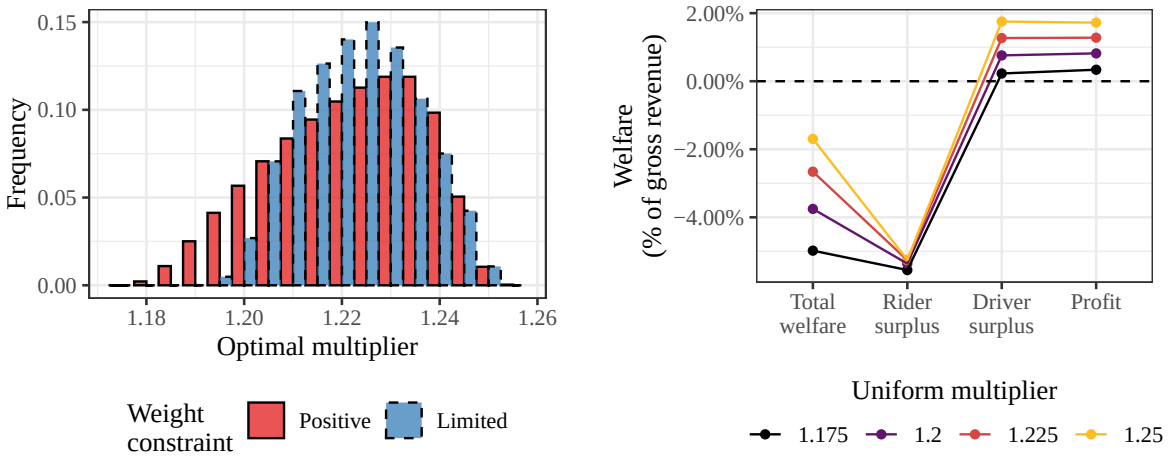
Figure 18: Distribution of surge multipliers in simulations and in the data

Note: Histograms of surge multipliers in simulations and in the data. The fraction of observations with multiplier 1 is 77.8% in the data and 79.2% in simulations.

come, and high income) with pricing policy $P \in \mathcal{P}$. The platform's problem is

$$\max_{P \in \mathcal{P}} \alpha^{\Pi} \Pi(P) + \alpha^{R,O} RS^{O}(P) + \alpha^{R,L} RS^{L}(P) + \alpha^{R,M} RS^{M}(P) + \alpha^{R,H} RS^{H}(P) + \alpha^{D} DS(P). \tag{18}$$

I consider all combinations of weights α such that (a) the sum is one, (b) no weight is negative, and (c) the objective function is maximized at the status quo price level. I sometimes consider a fourth condition, that (d) none of the rider surplus weights is greater than twice the sum of the other three rider surplus weights.



(a) Optimal uniform multipliers

(b) Welfare with different uniform multipliers

Figure 19: Uniform pricing with alternative objective functions

Note: Subfigure (a) shows the distribution of all the optimal uniform multipliers that result from different combinations of welfare weights that rationalize pricing in the status quo. A "positive" weight constraint refers to no weights being negative. A "limited" constraint also adds the restriction that no group of riders can have more weight than twice the weight for all other rider groups. Subfigure (b) shows welfare with different uniform multipliers, measured relative to the status quo.

For every combination of weights that satisfy the constraints, I compute the uniform multiplier that maximizes the objective function. Figure 19a shows that all weight combinations result in an optimal uniform multiplier between 1.177 and 1.255. When I also impose constraint (d), all optimal uniform multipliers are between 1.191 and 1.253. Figure 19b shows welfare with uniform multipliers in these ranges. In every case, the main qualitative findings from figure 10 hold.

F.2 Robustness of welfare effects

To explore how sensitive my main results are to changes in some model parameters, I set some of the model parameters to alternative values. I then rerun all the counterfactuals and recompute the effects of surge pricing. I compute welfare effects assuming Uber sets the uniform multiplier at the median of the distribution of all

possible values using the procedure in appendix F.1 with the constraint that none of the rider surplus weights accounts for more than 2/3 of rider surplus weight.

Table 3: Welfare effects 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)
	Panel A: Baseline							
Baseline	-0.63	1.2	1	1	-2.74%	-5.27%	1.23%	1.24%
	Panel B: Higher long-run elasticities							
Alternative supply elasticity	–	0.6	–	–	-4.92%	-7.84%	1.70%	1.18%
	–	1.8	–	–	-5.54%	-6.44%	0.29%	0.59%
Higher demand elasticity	-1	–	–	–	-8.34%	-6.84%	-0.95%	-0.48%
	-1.5	–	–	–	-8.95%	-7.01%	-1.11%	-0.75%
Alternative supply and higher demand elast.	-1	0.6	–	–	-7.21%	-7.09%	-0.48%	0.41%
	-1	1.8	–	–	-10.33%	-7.68%	-1.28%	-1.28%
	Panel C: Lower value of time for riders							
Lower ETA coefficient	–	–	–	0.75	-2.14%	-5.44%	1.58%	1.65%
	–	–	–	0.5	-4.70%	-6.00%	0.52%	0.77%
Higher fare coefficient	–	–	1.5	–	-3.61%	-6.49%	1.49%	1.34%
	–	–	2	–	-5.22%	-7.78%	1.27%	1.25%
Higher fare coef. and lower ETA coef.	–	–	1.5	0.75	-4.15%	-8.16%	1.99%	1.94%
	–	–	2	0.5	-5.53%	-7.74%	1.21%	0.98%

Note: This table shows welfare with the reoptimized uniform multiplier relative to surge pricing (as in the red, solid line in figure 10). Every row represents different market parameters. Panel A presents the baseline market. Panel B shows alternative markets with different long-run elasticities. Panel C shows alternative markets with a lower value of time for riders. For every alternative market, I rerun all the counterfactuals that are necessary to measure welfare effects. Columns (2)-(5) describe how parameters are modified relative to the baseline. A dash means that parameters are the same as in baseline market (elasticities are unchanged in panel A, and coefficient factors are one in panel B). Columns (6)-(9) measure welfare effects.

Long-run elasticities I explore robustness to alternative long-run supply and demand elasticities (table 3). Uber, for instance, uses a long-run demand elasticity of -1.2 in its internal models. Panel A shows numbers for the baseline market.[62] Every row in panel B represents one version of the market with alternative elasticities. 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,

[62]The numbers are not identical to the main results in figure 10 because, for consistency across rows, I assume that Uber maximizes a weighted sum of welfare components instead of rider surplus.

but by small amounts, since there are two opposing effects: an increase because of matching that reduces drivers' idle time, and a decrease because of lower prices.

Value of time I also welfare effects with different parameters of the request model so that the value of time is lower. I scale up the price coefficient $\beta(x_i)$ by a factor of 1.5 or 2, or I scale down the pickup time coefficient $\gamma(x_i)$ by a factor of 0.75 or 0.5. Panel C in table 3 shows these results. The main qualitative takeaways still hold.

F.3 Rider surplus by income

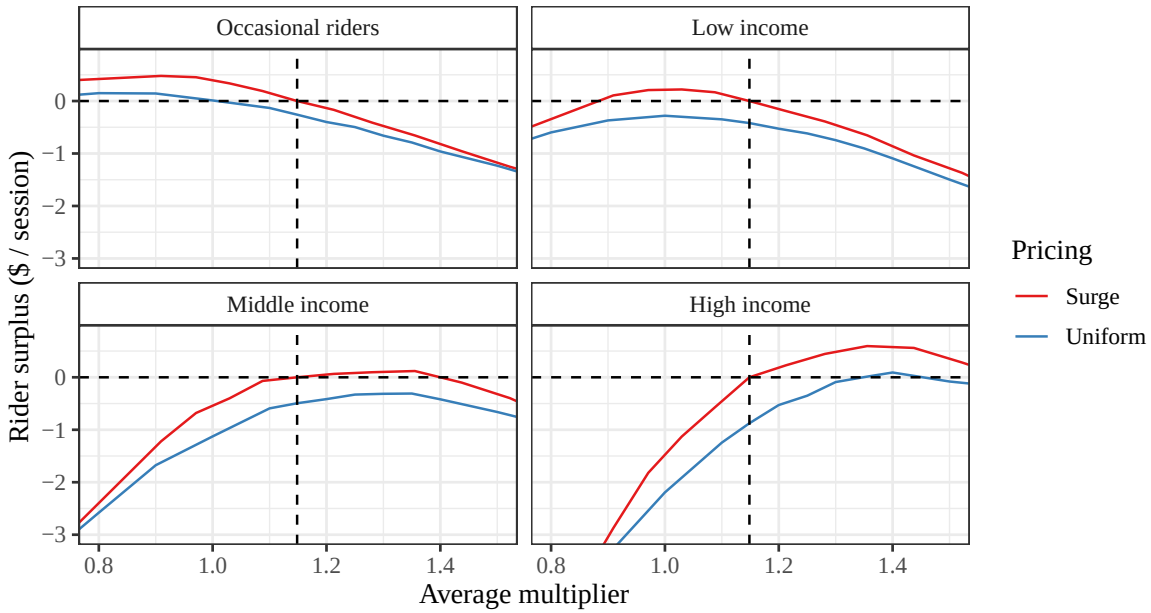


Figure 20: Rider surplus under different pricing policies

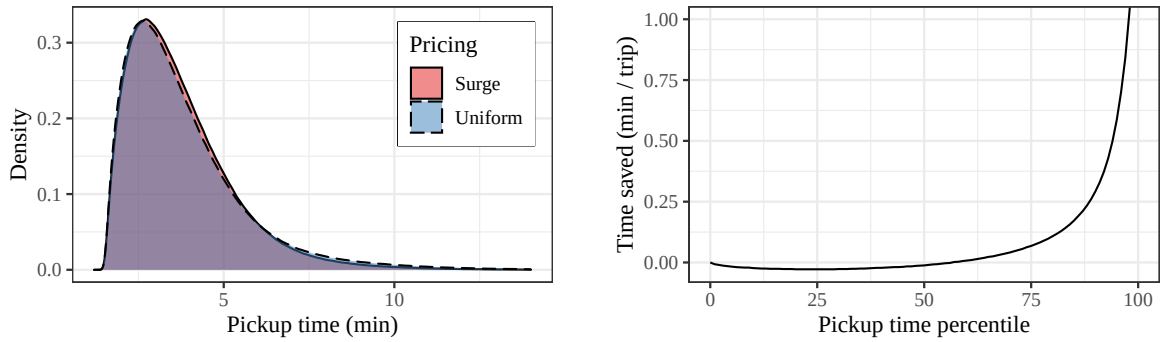
Note: These figures compare rider surplus by income groups for different pricing policies. The horizontal axis represents the average surge multiplier. The vertical axis represents rider surplus per session relative to the status quo. Curves for surge pricing represent policies in which multipliers are computed as in the status quo, but are scaled up or down by a factor that is constant across the whole market. Curves for uniform pricing represent a unique multiplier at different levels.

Figure 20 breaks down the lower left panel of figure 8 (rider surplus) by income groups. For a fixed average multiplier, all income groups prefer surge pricing. The benefits are larger for higher income riders—they put the largest value on reliable trips and low pickup times. Occasional and low-income riders prefer lower prices than in the status quo, whereas middle- and high-income riders prefer higher prices: the first two groups are price sensitive, whereas the second two groups value

low pickup times and reliable trips, and so prefer high prices that skim other riders and bring more drivers to the market.

F.4 Effect of surge pricing on the distribution of pickup times

Figure 21 compares the distribution of pickup times with surge pricing and with uniform pricing with the same average multiplier. Subfigure 21a shows that surge pricing reduces the variance of the distribution, with a small effect on the mean. Subfigure 21b highlights that surge pricing cuts the upper tail of the distribution.



(a) Distribution of pickup time

(b) Effect across the distribution

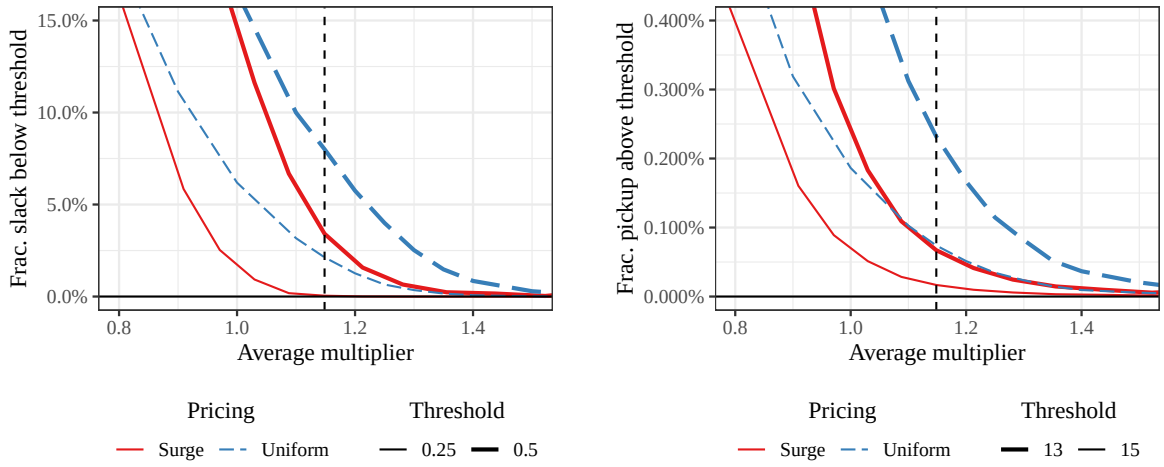
Figure 21: Effect of surge pricing on pickup time

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.

F.5 Evidence of wild-goose chases

I show evidence that wild-goose chases (WGCs) (Castillo et al., 2018) are the reason why rider surplus is maximized at a higher average multiplier with a uniform price than with surge 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 threshold that is between 0.25 and 0.5. Figure 22a shows the fraction of time that slack is below 0.25 and 0.5. WGCs—times with low slack—start taking place at higher prices with uniform pricing than with surge pricing.

Another tell-tale sign of WGCs is that some riders get extremely high pickup times. Figure 22b shows the fraction of pickups that are above 13 and 15 minutes,



(a) Low slack

(b) High pickup times

Figure 22: 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.

with a similar pattern to figure 22a. As Castillo et al. emphasize, WGCs result in a steep decline in rider surplus. Right when figure 22 suggests WGCs start taking place, the gap between surge and uniform pricing starts to widen (figure 8). Thus, rider surplus is maximized at a higher average multiplier with uniform pricing.

Appendix G Additional empirical evidence

G.1 Hierarchical demand model

Suppose that riders' utility follows equation 7, where $w_{lt}^0 \sim N(\mu_{kh}, \nu)$ and $w_i \sim N(w_{lt}^0, \sigma)$. The parameter μ_{kh} is a mean by location group and hour of the week, and w_{lt}^0 is a random effect. This model avoids using $\bar{w}_{lt}$ as an estimator for w_{lt}^0 as in my main model, which might create some bias because of the combination of two elements: (a) the model is nonlinear, and (b), the number of observations by group lt is small so I cannot rely on consistency of $\bar{w}_{lt}$. This new model, however, has the drawback that it assumes w_{lt}^0 is independent conditional on μ_{kh}

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. In the second step, I

maximize the conditional likelihood

$$L(\theta, \hat{\sigma}, \hat{v}, \hat{\mu}_{kh}) = \prod_{lt} \int \prod_{i \in lt} \Lambda(u_i(w_{lt}^0))^{y_i} (1 - \Lambda(u_i(w_{lt}^0)))^{1-y_i} \, dF(w_{lt}^0 | \mathbf{w}, \hat{\sigma}, \hat{v}, \hat{\mu}_{kh}), \quad (19)$$

where y_i is an indicator variable for whether the rider requested a trip, and Λ is the standard logistic function. I compute the integral by Gaussian quadrature.

Table 4: Estimates of the parameters of the hierarchical demand model

	Dependent variable: Trip requested					
	Coefficient dependence on:					
	Constant	Occasional	Log income	Base fare	Airport	Weekend
	(1)	(2)	(3)	(4)	(5)	(6)
Price	-0.0438^{**}	-0.0072	0.0371	0.0012	0.0124	-0.0013
	(0.0204)	(0.0224)	(0.0291)	(0.0013)	(0.0274)	(0.0214)
ETA	-0.1230^{***}	0.0869^{***}	-0.0135	-0.0022^{**}	-0.1389^{***}	-0.0463^{***}
	(0.0101)	(0.0144)	(0.0173)	(0.0010)	(0.0330)	(0.0150)
Observations: 650,233						

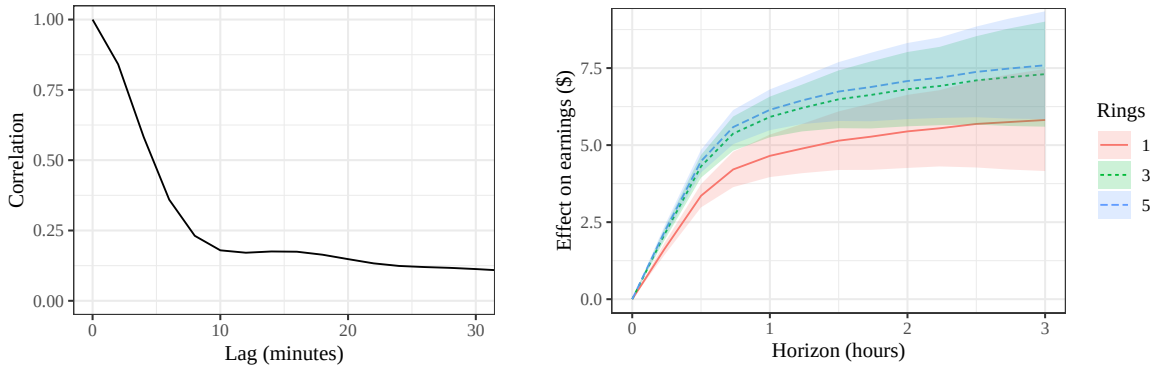
Note: Estimates of the main parameters of the hierarchical demand model. The price and pickup time coefficients are evaluated at the mean of the base fare.

Table 4 reports the estimates from this model. Most parameters are very close to the ones for the main model in table 2. The only noticeable difference is that the airport parameter in the pickup time coefficient becomes negative and significant.

G.2 Multipliers: correlation and impact on expected earnings

Figure 23a shows the persistence of unexpected variation 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. There is significant correlation for the first ten minutes. After that, it settles down at around 0.15.

I also 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 by hour of the week fixed effects. Figure 23b plots the main coefficient. The effect is larger for the three closest rings than only for the local multiplier, and only a little bit larger for five 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 starts to level off after one hour.



(a) Temporal correlation of multipliers

(b) Effect of multipliers on expected earnings

Figure 23: Patterns in multipliers and earnings

Note: Subfigure (a) shows an autocorrelation plot of the residuals of surge multipliers after controlling for location by hour of the week fixed effects. Confidence intervals are narrower than the line in the figure. Subfigure (b) shows the main coefficient for regressions of hourly earnings for the next h hours on current multipliers. 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. Standard errors are computed with two-way clustering by location and hour of the week.

G.3 Impact of multipliers on driver exit

In this section I show that there is no empirical evidence that drivers respond to unexpected changes in multipliers by leaving the market. I run a regression of whether open drivers decide to leave as a function of the multiplier in his location. I also run similar regressions where the main covariate is the average multiplier in all locations within three or five hexagons (with 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.

The estimates I obtain for the effect of the multiplier on leaving are 0.018 (s.e.=0.028) for the local multiplier, -0.010 (s.e.=0.036) for the three nearest rings, and 0.003 (s.e.=0.036) for the five nearest rings. None of the estimates is significant. The number of observations is 1,218,286 for all three regressions.

G.4 Additional results from demand elasticity experiment

Table 5 reports the result of regressions of trips per rider on treatment dummies. There is heterogeneity across cities. Mexico city shows an unexpected positive elasticity for the 10% treatment group. I thus exclude it from the main regression used

Table 5: Average treatment effects in demand experiment

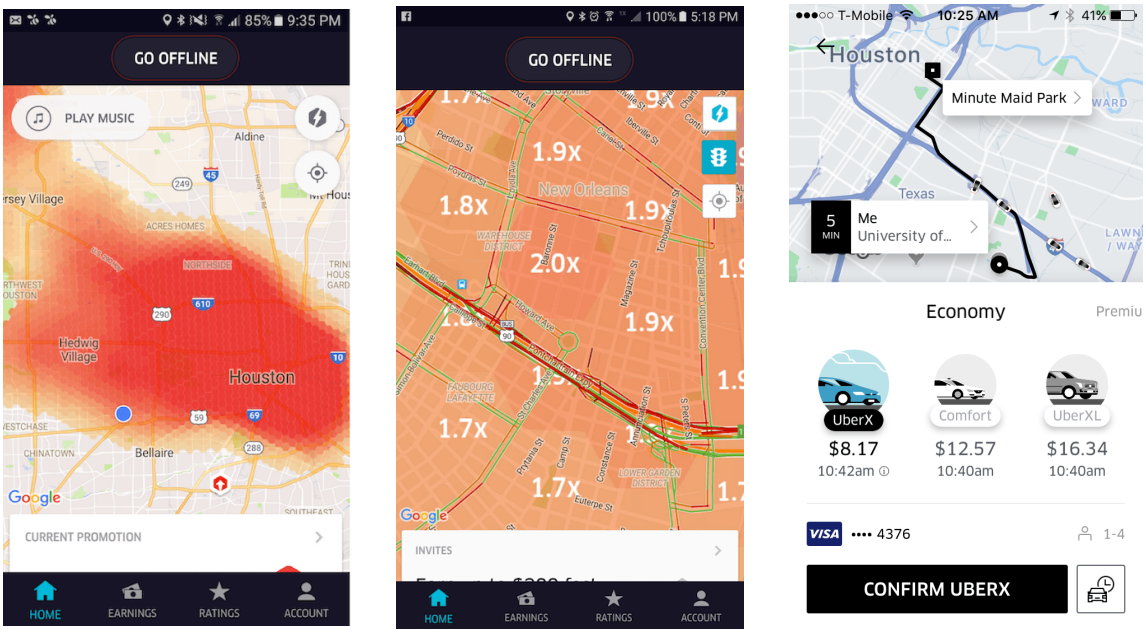
	Dependent variable: Trips per rider				
	Mexico City	Guadalajara	Rio	Sao Paulo	Belo Horizonte
	(1)	(2)	(3)	(4)	(5)
Constant	1.0470***	1.2558***	0.9200***	0.7142***	0.8179***
	(0.0125)	(0.0138)	(0.0110)	(0.0098)	(0.0104)
10% discount	−0.0666**	0.0795**	0.0624**	0.0247	0.0534**
	(0.0264)	(0.0331)	(0.0255)	(0.0228)	(0.0244)
20% discount	0.0459	0.1616***	0.1883***	0.0422*	0.1800***
	(0.0282)	(0.0341)	(0.0280)	(0.0226)	(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.

to calibrate the long run demand parameter.

Appendix H App interface

Figure 24 shows screenshots of the rider and driver app.



(a) Driver interface (b) Driver interface - zoom in (c) Rider interface

Figure 24: 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.