Reliability and Pricing in Cloud Computing*

James Brand[1] Juan Camilo Castillo[2] Chinmay Lohani[2]

Leon Musolff[2]

November 2025

Abstract

To match volatile demand with fixed capacity, cloud computing platforms employ tiered reliability—offering discounted spot compute services from which users can be "evicted" (i.e., interrupted) with little warning when capacity tightens. We study this market design using proprietary data from a major cloud platform, exploiting a price experiment and the quasi-random nature of evictions to estimate a structural model. The price elasticity of demand is -0.5, and evictions persistently reduce usage by 40%, indicating a strong revealed preference for reliability. More usage increases eviction rates, consistent with congestion. We interpret these facts through a model where heterogeneous users choose the compute reliability for each workload, while learning about eviction risk through experience. On the supply side, evictions arise endogenously given fixed capacity. Preliminary counterfactual results suggest that tiered reliability provides Pareto gains relative to simply allowing the market to clear through congestion.

*James Brand: Microsoft. Juan Camilo Castillo: University of Pennsylvania. Chinmay Lohani: University of Pennsylvania. Leon Musolff: University of Pennsylvania. We gratefully acknowledge support from *Analytics at Wharton* (AAW) and *AI and Analytics for Business* (AIAB). We thank Will Wang and Chris Conlon for many valuable comments. We are grateful to Wanxi Zhou and Huaihan Shan for excellent research support.

1 Introduction

Cloud computing has transformed how companies satisfy their computing needs. Instead of investing in their own computing infrastructure—requiring substantial capital expenditures and lengthy lead times—companies can now rent computing power and data storage. This gives them a significant degree of flexibility, enabling them to scale up quickly and experiment with new technologies. The importance of cloud computing has only grown with the recent rise of artificial intelligence, which relies heavily on costly, specialized hardware, such as graphical processing units (GPUs), that few companies are willing or able to purchase outright.

The most popular computing products offered by major cloud providers like Amazon Web Services, Microsoft Azure, and Google Cloud are virtual machines (VMs), systems that emulate a stand-alone computer. Providers have a fixed stock of servers that they rent out in the form of VMs to a large customer base whose demand for VMs fluctuates over time. To serve this fluctuating demand with a fixed capacity, platforms thus must implement mechanisms to handle short-run mismatches between capacity and demand. This same challenge is present in several other industries such as electricity, airlines, broadband internet, restaurants, and urban transportation. Companies in those sectors employ strategies like peak-load pricing, rationing during peak demand, or simply allowing congestion to clear the market.

Over time, cloud providers have converged on a different strategy to address the mismatch between supply and demand, which we refer to as *tiered reliability*. Cloud providers engage in a form of second-degree price discrimination by offering two quality-differentiated products. *On-demand* VMs guarantee continued availability once deployed and command a high hourly price. To ensure reliability, providers must install enough capacity to satisfy peak demand for these high-reliability VMs. They sell the remaining capacity as *spot VMs*, which are low-price, low-quality VMs that can be revoked (or "evicted") at short notice if needed to serve on-demand customers.[1]

[1]The price of spot VMs does not vary at a high frequency (on most major clouds, it changes at most once per day). The misleading name "spot" comes from an early design in which spot VMs were sold through auctions to the highest bidders.

In this paper, we examine the welfare effects of tiered reliability, comparing it to more traditional mechanisms for clearing markets that face congestion. Tiered reliability presents several tradeoffs. On the one hand, it allows for more efficient use of capacity by guaranteeing reliable service to customers who value it most, while still allocating the remaining capacity to others who are more price-sensitive and/or whose jobs are more amenable to being interrupted. On the other hand, it allows cloud providers to price discriminate based on product quality (i.e., reliability), with the associated effects that economists have recognized since Mussa and Rosen (1978). In theory, this form of quality differentiation may enable providers to extract a sizable surplus from customers that choose on-demand VMs (increasing profits but reducing consumer surplus) while still also extracting some surplus from spot VM users (which can improve both profits and consumer surplus). However, the presence of spot VMs may also cannibalize demand from on-demand VMs, incentivizing firms to inefficiently raise the price and degrade the quality of spot VMs.

To quantify these economic forces, we set up a structural model of the cloud computing market. On the demand side, consumers with varying needs for reliability choose between spot VMs, on-demand VMs, and an outside option. Their choices depend on the price of VMs and the perceived probability of eviction. Over time, consumers learn about the likelihood of being evicted from their own experiences with each product. On the supply side, the cloud provider owns a stock of servers that it rents out to consumers under a key technological constraint: as usage approaches the installed capacity, the provider must evict an increasing number of users, with discretion over which users to remove. Using our model, we simulate counterfactual scenarios in which the cloud provider uses different mechanisms to allocate capacity and evictions, such as tiered reliability, peak-load pricing, and rationing.

To estimate our model, we use detailed data from a major cloud provider about VM usage, prices, and evictions. There are four key elements that we need to estimate to inform our model. First, we measure users' price sensitivity by leveraging an experiment conducted by the provider. Second, to measure users' preference for reliability, we exploit the inherent randomness of evictions: when usage of a given product is high and

the provider needs to evict users, it chooses which users to evict based on technical rules that are unrelated to consumer characteristics—other than the fact that they were using the congested product. Third, we estimate a model of users' beliefs about the likelihood of eviction through our structural model by taking advantage of the fact that some customers are exposed by chance to more evictions than others. Finally, we estimate the technological relationship between capacity, usage, and evictions under the assumption that technological shocks are unrelated to demand shocks.

Our demand estimates suggest that users are sensitive to both price and evictions, and that their prior over eviction rates is large relative to the mean observed eviction rate, but very diffuse. As a result, customers are quite pessimistic about the reliability of spot VMs—over-estimating eviction rates by more than an order of magnitude—but quick to update their beliefs with experience. We find small own-price elasticities of approximately -0.5, and heterogeneous eviction responses with some users caring substantially more about reliability than others.

We present preliminary counterfactual results that compare tiered reliability to simply allowing the market to clear through congestion—similar to, for example, transportation markets. We find that tiered reliability represents a Pareto improvement relative to congestion. Cloud providers greatly benefit: clearing the market through congestion hinders their ability to extract surplus from users who place a high value on reliability, as they will inevitably be evicted from time to time. Users of all types also benefit, since they self-select according to their preferences. Those who highly value reliability have the option of choosing the (expensive) product with guaranteed reliability, whereas those who do not greatly mind evictions choose the cheaper spot product.

In future iterations, we will present additional counterfactuals analyzing other market-clearing mechanisms. We first consider peak-load pricing—clearing the market by increasing prices when demand is high. We also consider better information design, where consumers are provided with real-time information about congestion in the market.

Related Literature Some theoretical papers study dual on-demand/spot product offerings in cloud computing. These papers focus on early market designs where auctions or dynamic pricing were used for the spot market. Hoy et al. (2016) show that a dual

spot/on-demand design can yield Pareto improvements over a spot-only market when customers are risk averse. Abhishek et al. (2017) and Dierks and Seuken (2022) study under what conditions a cloud provider that offers a high-price on-demand product can benefit from also offering a spot product, without cannibalizing demand for its premium product. We make several contributions relative to these works. To the best of our knowledge, we are the first to study the interplay between on-demand and spot markets empirically, which allows us to measure the welfare of all market participants. We also study the current market design to which all major cloud providers—Amazon Web Services, Microsoft Azure, and Google Cloud—eventually converged, in which both on-demand and spot VMs are sold at a fixed price.

Our work relates to the broader literature on cloud computing. Despite the central role this market plays in today's economy, the literature on this topic remains relatively limited (see Biglaiser et al., 2024, for a review). Jin et al. (2023) measures the welfare benefits from cloud computing, with a particular focus on consumer inertia. Brand et al. (2024) study the productivity of firms when using cloud computing resources. Other works measure the effects of cloud computing on productivity (DeStefano et al., 2023) and industry dynamics (Lu et al., 2024). Kilcioglu et al. (2017) present detailed descriptive statistics about cloud computing usage patterns. Gans et al. (2023) assess the impact of proposed regulation of this market in the European Union. Hummel and Schwarz (2022) study a cloud provider's decision to allocate capacity and price computing across multiple geographic locations. We contribute to this literature by studying the pricing and market design decisions made by cloud providers and measuring the welfare effects of these decisions.

Finally, our paper contributes to a broad economics literature on mechanisms used to clear markets in the presence of supply-demand mismatches. Some classical theoretical papers study these issues in contexts such as road congestion (Vickrey, 1963), ski-lift tickets (Barro and Romer, 1987), and more general settings with capacity constraints (Williamson, 1966). More recent research has focused on empirical applications, including electricity (Joskow and Wolfram, 2012), airlines (Williams, 2022), ride-hailing platforms (Castillo, 2022), and road congestion (Kreindler, 2024). We contribute to this lit-

erature, first, by providing an empirical analysis of an innovative market clearing mechanism that might be applicable to several settings, and, second, by focusing on cloud computing—an increasingly important market in the modern economy.

2 Setting & Data

This section provides an overview of cloud computing, the data used in this paper, and some key descriptive facts that will motivate and inform our empirical model below.

2.1 Setting

Prior to the growth of cloud computing, firms that required substantial computing power would have had to acquire and maintain their own servers. Cloud computing, instead, centralizes the building and maintenance costs of this capacity, allowing individual firms to access the centralized resources via a rental market.

The majority of cloud computing services are provided in the form of virtual machines (VMs), which can be thought of as remote desktops that allow users to interact with a self-contained computer—comprising a CPU, memory (RAM), and storage (disk space)—located in a data center. VMs come in many varieties, usually defined by the number of CPU cores, the amount of memory and storage, and the manufacturer of the physical hardware (e.g., Intel or AMD). Each VM can be paired with an operating system of choice (typically some version of Linux or Windows) and can be deployed in one of the various data center regions maintained by the provider.

The standard VM products offered by cloud computing platforms are *on-demand* VMs, which customers request at a fixed listed price. Customers are then charged for every second the VM runs.[2] On-demand VMs are offered with the expectation of high reliability—indeed, they are available over 99.9% of the time. The downside of this reliability, however, is that they are offered at a relatively high price.

[2]Alternatively, customers can often make monetary commitments to the cloud platform (e.g., via "Reserved Instances" or "Savings Plans" offered by AWS and Azure or "Committed Use Discounts" on Google Cloud), which this VM usage then contributes toward.

To ensure that on-demand VMs are always available, cloud providers must ensure that they have enough capacity installed to meet peak demand. However, computing usage is cyclical throughout the day and week, frequently leaving the provider's computing capacity underutilized. All major cloud providers (including Amazon Web Services, Google Cloud, and Microsoft Azure) offer spare capacity as *spot* VMs, which provide firms with substantial discounts relative to on-demand VM prices, in exchange for granting the provider the right to terminate instances at any time in order to reallocate capacity to higher-paying customers. Spot VM markets have existed for over a decade now. Amazon Web Services was the first major cloud provider to offer spot VMs in 2009, followed by Google Cloud in 2015 and by MS Azure in 2017.

2.2 Data

We use proprietary data from a large global cloud computing provider. Our data comprises two parts. First, we observe all spot VM usage for a large sample of users from August 2020 through July 2022. These data include anonymized identifiers for users, products, geographic locations (data centers), and operating systems. For each user, we observe their total usage in hours and core-hours on every product within a large class of general-purpose VMs, which are unlikely substitutes for products outside of our sample.[3] Usage in our data has been normalized by an unknown constant for confidentiality reasons. We also observe the total number of evictions experienced by each user on each product, and whether the user initiated any new jobs on the product that day. Although we observe usage daily, for most of our analysis, we aggregate this data to the weekly level.

Second, we observe a second dataset that captures total usage, aggregated across users, for every product in our data.[4] This data incorporates the sum of all usage on a product (including spot and on-demand usage), and will be useful for estimating a supply-side model that can explain and predict eviction rates in our counterfactuals.

[3]The set of general-purpose VMs covers the vast majority of non-GPU-based VMs sold by the cloud platform but excludes some specialty products for high-performance workloads.

[4]This usage is obfuscated by a different constant than the evictable VM usage data, again for confidentiality reasons.

Finally, we observe obfuscated measures of each evictable product's price. For each product, we observe the price and the product's assigned treatment in an experiment run by the platform in 2022.

Experiment details In May 2022, a subset of spot products was randomly selected by the platform to be included in a pricing experiment. Of this subset, half were chosen to receive "shocks" which *decreased* prices relative to the standard internal price-setting process, and half were chosen to receive shocks which *increased* prices. In both cases, the largest possible shock was a 25% price change relative to the price that would have been set otherwise, and the magnitude of the shock was drawn uniformly between zero and the maximum (absolute) value. The randomized prices went into effect on June 1 2022, and prices remained fixed until July 31.

2.3 Descriptive Evidence

This section presents descriptive evidence of the key margins of response in our model. On the demand side, we present event studies that allow us to determine how users respond to eviction shocks and to price changes. On the supply side, we present evidence of congestion: eviction rates increase when usage increases. For brevity, in the sections below we refer to the combination of a *product group*, *location*, and *operating system* as a *market*, denoted as m.[5]

2.3.1 Eviction Response

Our first goal is to determine how users respond to evictions, both in terms of their continuing usage in the market where they were evicted and in alternate locations. Ex ante, because users choose to run spot VMs knowing they face a risk of eviction, they may respond to evictions by restarting their VM and continuing to run their workloads in the affected market. On the other hand, users may select spot under the belief that interruption is unlikely and thus may respond to evictions by seeking locations with lower eviction rates. This may be particularly likely if users' beliefs about eviction probabilities

[5]Our findings below validate this market definition: we find negligible substitution across markets.

are misspecified.

To this end, we conduct a matched event study of spot usage surrounding an eviction event. We focus on evictions that took place between August 2020 and May 2022. To define our treatment group, we select all user–market observations with a single eviction event, preceded by a two-week window with no other evictions and followed by a sixty-day eviction-free window. This sample selection allows us to observe a long window of usage prior to and after an eviction which is not contaminated by other evictions in the focal market. Still, this sample represents users with smaller-than-average spot usage, who are more likely to use spot products for over two months without facing any evictions. Although these customers may not be representative of the average customer, we believe that their eviction responses are informative of the average customer's preferences.

We match every treated user–market pair from this sample to one control from the set of users with no evictions in that market during the focal event-study window. Operationally, we choose the control user whose pre-eviction usage most closely aligns with that of the treated user over the fourteen-day period preceding the eviction. Formally, let t_0^i be the day of the eviction for user i. We define the dissimilarity in usage between users i and i' in market m for times between $t_0^i - 14$ (fourteen days before the eviction) and $t_0^i - 1$ (the day before the eviction) as

$$d(i, i', m) = \sqrt{\frac{1}{14} \sum_{t=t_i^0-14}^{t_i^0-1} (y_{imt} - y_{i'mt})^2} \tag{1}$$

where y_{imt} is an indicator for whether user i had positive usage in market m at time t.[6] This captures the number of days during which one user had positive usage but the other did not. We exclude users whose closest candidate has dissimilarity greater than 0.5. In practice, this procedure generates a set of matches for which nearly all pairs use the focal product on an identical set of days.

[6]For our baseline result, our outcome variable is an indicator variable for whether there was any usage. This is why our dissimilarity measure is also based on usage indicators.

We estimate an event-study regression of the form

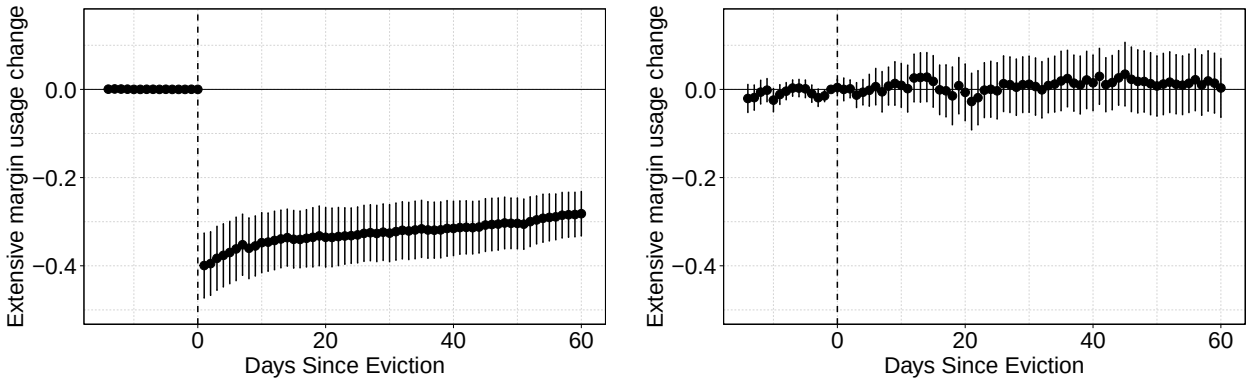
$$\Delta y_{imxt} \;=\; \sum_{\tau=-n_{pre}}^{n_{post}} \beta_\tau \, \mathbf{1}\big(t = E_{imx} + \tau\big) \;+\; \varepsilon_{imxt}, \tag{2}$$

where Δy_{imxt} denotes the difference between the usage of treated user i in market m (where she was evicted) and the usage of the control user corresponding to her x-th eviction event in that market. E_{imx} denotes the day of her x-th eviction event, so $\mathbf{1}(t = E_{imx} + \tau)$ is an indicator for day τ relative to the eviction day. We omit β_{-1} to define $\tau = -1$ as the reference day. The coefficients β_τ trace the dynamic treatment effect—how usage evolves around the time of the eviction.

Own Usage Effect and Substitution Pattern The left panel in Figure 1 reports the event-study coefficients β_τ. Immediately after the eviction, usage in the affected market falls by roughly 40% relative to the pre-eviction baseline. The magnitude of the effects slightly decline over the subsequent two months (mostly because overall usage also declines), but the point estimates remain large, negative, and statistically different from zero. The persistence of the drop suggests a durable change in the perceived attractiveness of the evicted product. One plausible mechanism is belief updating: users may revise upward the perceived likelihood of future evictions, thereby lowering their expected utility from continued usage—a hypothesis we investigate below.

The drop in usage we see on the focal product could arise from substitution towards spot products in other markets, substitution towards on-demand VMs in other markets, or simply reducing the overall usage of VMs (i.e., substitution towards the outside option). We now investigate the extent to which there is substitution towards other spot products. We estimate event study coefficients using the same matched-sample design (equation 1), where the dependent variable is now $\Delta y_{i,-m,x,t} = \sum_{k \neq m} \Delta y_{ikxt}$, the difference in total usage between the treated and control user summed across all markets other than the market m where the eviction took place.

The right panel in Figure 1 presents results from this event study. The estimated effect is small and statistically insignificant, indicating virtually no reallocation of activity to other spot products. Hence, the usage decline documented in left panel must come either



(a) Effect on usage of evicted product (b) Effect on usage of other spot products

Figure 1: Effect of eviction on usage

Note: These figures present the effect of eviction events on usage, based on an event study of the form in equation (1). The horizontal axis reports days relative to the eviction date ($t = 0$), while the vertical axis displays differences in usage between the (paired) treated and control units. Vertical bars denote 95 percent confidence intervals, based on standard errors clustered by product-location. Panel a shows effects on usage of the product the user was evicted from. To examine substitution patterns, Panel b plots effects on usage of spot VMs in every product-location combination except the one the user was evicted from. As we match on the pre-period usage pattern on the focal market, our procedure mechanically ensures no pre-trend in Panel a.

from a shift toward on-demand products or from exiting the platform altogether. The absence of cross-market substitution suggests that a product-group–location–OS tuple behaves as an economically isolated market, validating our market definition m for the remainder of the analysis.

Learning about reliability What drives the persistence of the effect that we observe in the previous event studies? One possibility is that users are learning about the likelihood of being evicted in the future. After being evicted, users infer that the likelihood of future eviction is higher and, hence, the utility from a spot product is lower. From that point on, they thus choose to make less intensive use of the spot product.[7] Under this hypothesis, users who are only starting to use a product may not have well-formed beliefs about eviction rates, making each eviction event likely to shift their beliefs about the product's reliability and thereby affect their future usage of the product. In contrast, customers who use the product intensely would be more likely to understand the characteristics of the

[7]Note that our estimation sample in the previous section focuses on customers who are not evicted for multiple weeks in a row. These customers may be particularly likely to view the focal eviction event as a "surprise" relative to the reliability they had grown to expect.

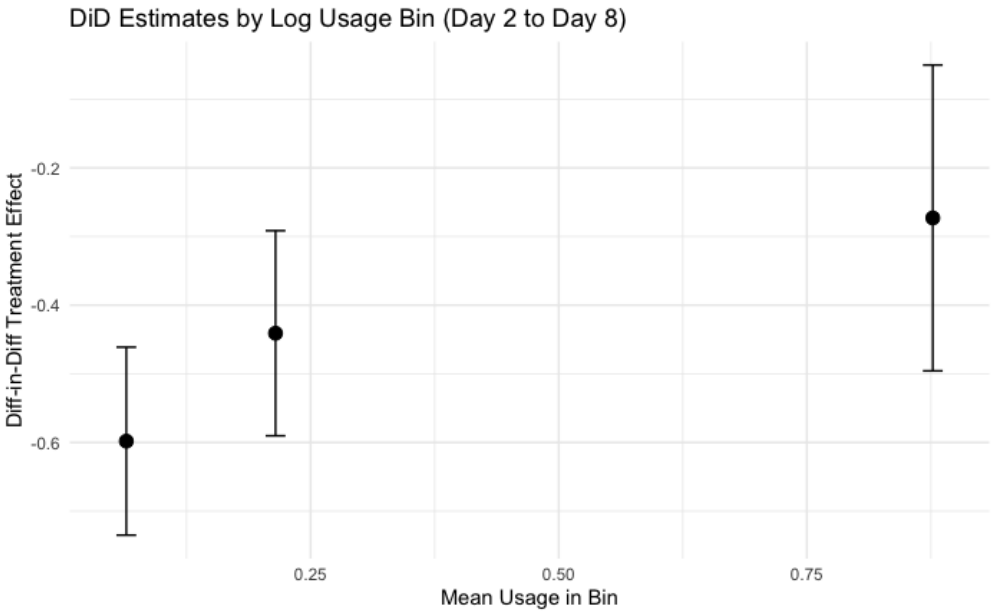


Figure 2: Demand response to evictions by past usage

Note: The figure presents differences-in-differences estimates of the average treatment effect of evictions on the extensive margin during Days $2-8$ after an eviction. We separate users by their usage before the event based on the following bins: log (usage) ≤ -2, $-2 <$ log (usage) ≤ -1, and log (usage) > -1. The horizontal axis reports the mean usage by bin. The vertical axis shows the difference-in-differences point estimate; vertical bars depict 95 percent confidence intervals based on heteroskedasticity-robust standard errors.

product well, making an additional eviction unlikely to affect expectations of reliability.

We now measure the extent of empirical evidence in our data in favor of this hypothesis. To that end, we partition users according to the intensity of their usage on the focal product before being evicted. We then estimate three specifications, each similar to equation 1, for three different samples: customers with low, medium, and high usage before the eviction event. In each specification, we report the pooled average effect of the eviction on usage for the first seven days after the eviction occurs.

Figure 2 presents our estimates. We observe that evictions result in a much stronger demand response by users with less pre-eviction usage: the lowest bin sees a 60% usage drop after an eviction, which is much less than the highest bin drop of less than 30%. This suggests that the strong pooled treatment effect may be driven by users who are still learning about the likelihood of evictions (and for whom this eviction is hence a strong signal of a high eviction rate). Still, another potential explanation is preference heterogeneity: firms that have more usage before the focal eviction are likely to have

experienced an eviction before, and if they nevertheless continued to use the product, that indicates that they may have a lower preference for reliability than other users who quit using the product immediately after their first eviction. To disentangle preference heterogeneity from learning, our structural model below exploits additional information about the market share of spot versus on-demand, especially focusing on the number of users who have ever tried spot.

2.3.2 Price Response

To measure the demand response to prices, we run a matched event study around the beginning of the experiment described above. Our estimation window spans from May 1st, 2022 (i.e., one month before the experiment) to July 31st, 2022 (when the experiment ended). Our treatment group comprises consumers who used at least one of the shocked product-group locations before or during the experiment (i.e., between May 1 and July 31), allowing us to capture effects both on users who stop using treated products following price increases and on those who begin using them following price decreases. We match treated users to control users who use the exact same products (up to the number of cores on the VM) and operating systems but in a different location that was not subject to a price change. We hence match *across* markets, minimizing the likelihood of SUTVA violations from marketplace interference effects (Blake and Coey, 2014).

Controls are chosen with replacement from the donor pool by nearest-neighbor matching on pre-event extensive margin usage paths. To be precise, let y_{imt} be an indicator for any usage by user i in market m at week t. For a treated user i and a candidate control i' in the same product-OS, we compute a pre-period dissimilarity over weeks $t \in \{t_0, \ldots, t_{-1}\}$ using the root-mean-square (RMSE) difference of the binary indicators:

$$d\left(i, i', m, m'\right) = \sqrt{\frac{1}{t_0 - t_{-1}} \sum_{t=t_{-1}}^{t_0} \left(\mathbf{1}\left\{y_{imt} > 0\right\} - \mathbf{1}\left\{y_{i'm't} > 0\right\}\right)^2}.$$

We then pair each treated unit with the control with minimal $d(\cdot)$, breaking ties at random. Similar to the eviction event study, we restrict to product-OS cells with multiple locations and keep only user-markets with full support over the event window. We

impose these restrictions so that each treated observation has a valid control from the same product–OS in a different location, and so treated and control paths are observed in every week of the window.

Let Δp_m denote the price shock for market m, which was randomly drawn for the experiment. If, in the absence of the experiment, the price in market m would have been p_{mt} during time period t, then the actual price is $(1 + \Delta p_m) \cdot p_{mt}$. Since the experiment varied prices upwards in some markets and downwards in some markets, Δp_m can be positive or negative.

For each treated-control pair x, our response variable is

$$r_{imt} = \frac{\mathbf{1}\{y_{imt} > 0\} - \mathbf{1}\{\tilde{y}_{imt} > 0\}}{|\Delta p_m|}.$$

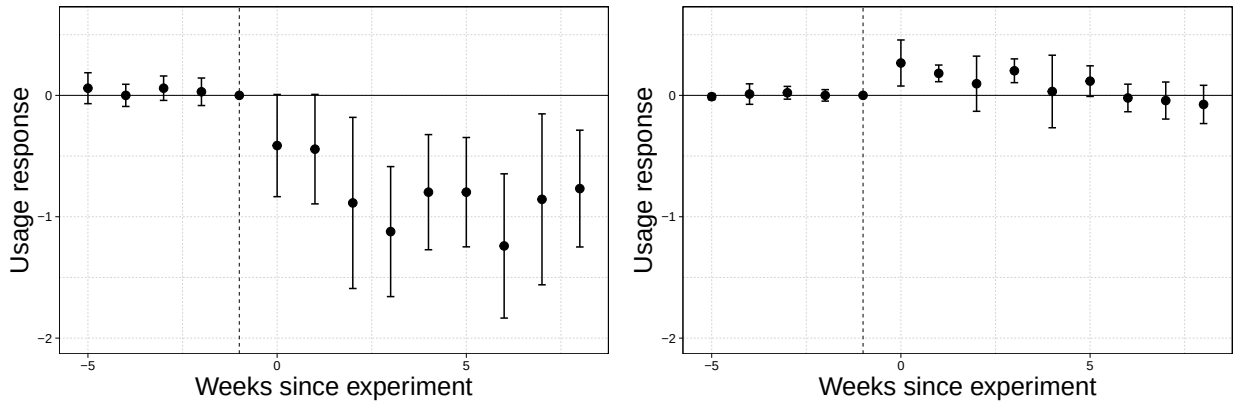
The numerator is the difference in extensive margin usage between the treated user i in market m (that is, $\mathbf{1}\{y_{imt} > 0\}$) and the usage for the corresponding control user (which we denote by $\mathbf{1}\{\tilde{y}_{imt} > 0\}$). We normalize this difference by dividing it by the absolute value of the price shock for the treated user's market, so that the treatment effects that we measure can be interpreted as elasticities. We also weight observations by $|\Delta p_m|$ so larger shocks receive more influence.

For weeks t around the experiment start t_0, we estimate

$$r_{imt} = \sum_{\tau=-n_{pre}}^{n_{post}} \beta_\tau \mathbf{1}\{t - t_0 = \tau\} + \varepsilon_{imt},$$

omitting $\tau = -1$ as the reference period, and clustering standard errors at the product-group level (i.e., the level at which treatment varies).

We present estimates separately for positive and negative shocks in Figure 3, where we see the expected effect of a decrease in demand following a price increase, and an increase in demand following a price decrease. However, our results unveil an interesting heterogeneity: while the demand response to an unexpected price increase is quite pronounced, demand is slower to adjust to an unexpected price decrease. As we are limited by our two-month experimental window, we hypothesize that the longer-run price response would likely be symmetric, suggesting the use of the estimated price elasticity



(a) Positive price shocks (b) Negative price shocks

Figure 3: Effect of price shocks on usage

Note: Both graphs display the extensive margin for usage, representing the shift from non-usage to any level of product usage. The left panel illustrates the response to positive price shocks, while the right panel demonstrates the effects of negative price shocks. Black dots indicate the estimated treatment effects, with vertical bars representing the 95% confidence intervals. The horizontal axis denotes weeks since the price shock.

from just positive price shocks (compared to the overall pooled estimate) as an important robustness check.

2.3.3 Usage and Eviction

In this section, we examine the relationship between aggregate usage (including both spot and on-demand) and eviction rates. The eviction rate is defined as the number of evictions divided by the aggregate usage measured in core-hours.

Our key identification challenge is that usage and capacity may be growing in lock-step over time, yielding a stable eviction rate in the presence of ever-increasing usage, which could lead us to falsely conclude that capacity constraints never bind. This problem emerges because capacity itself varies over long time horizons. Hence, our solution examines variation in usage over shorter time horizons when capacity is likely fixed. To this end, we aggregate to year-month by day-of-week within each product-by-location[8]. Effectively, when examining the relationship between usage and evictions separately month-by-month, this procedure controls for the year-of-the-month (and hence possible

[8]This aggregates across markets because the same underlying hardware can flexibly be deployed with any operating system; by contrast, users of VMs cannot easily change the OS they require to run their workloads.

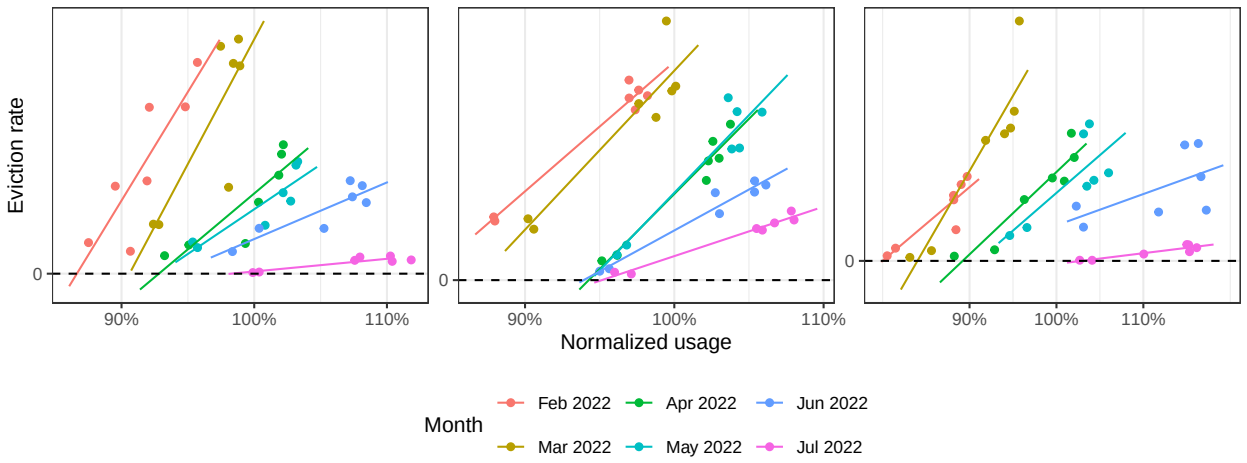


Figure 4: Eviction rates and usage

Note: This figure illustrates the relationship between the eviction rate and usage relative to the mean usage for the top three product groups across various months (February 2022 - July 2022). To avoid revealing proprietary data about the average level of eviction rates, we have removed y-axis ticks and labels.

associated capacity changes) and instruments usage with a set of dummies for the day of the week, relying on exogenous variation generated by the usual rhythm of business (e.g., the fact that employees do not work on Saturdays and Sundays).

We present a scatter plot of the eviction rate (on the vertical axis) against usage relative to a unit's average (on the horizontal axis) for the last six observed months for the top three product-groups with the most intense usage in Figure 4; the figure also exhibits a linear fit for each month. We see that within each month, there is a strictly increasing relationship between weekday average usage and the eviction rate: as the servers become more congested, additional usage increasingly becomes only possible by evicting some users. We also see a flattening of this relationship over time: comparing February to July, it is clear that additional capacity has come online, and eviction rates at the same usage level are much lower. We can also see that usage in later months is higher, as expected given the general growth of the cloud computing market during (and well beyond) our sample period.

3 Model and Estimation

This section presents our model of the cloud computing market. We first present a stylized model that highlights the key economic forces and that allows us to present some key theoretical results. We then move on to our main model, which we later estimate empirically.

3.1 Stylized Model

To fix ideas, we now present a stylized model that nevertheless captures the key features of the cloud computing market. Concretely, it involves two tiers (on-demand and spot). There is a congestion technology such that more users must be evicted the higher the utilization, and consumers receive disutility from evictions. Still, this stylized model abstracts away from some other empirically important elements that we capture in our main model below—most notably, the uncertainty about who gets evicted and consumers' beliefs about the likelihood of eviction.

Users within a market choose between an on-demand VM, a spot VM, or the outside option. They make that choice based on the prices of both types of products (p_d and p_s) and the eviction rate e for spot VMs. Hence, demand is given by:

$$(q_d, q_s, q_0) = D(p_d, p_s, e) \tag{3}$$

Demand for every product is decreasing in its own price and eviction rate, and increasing in the prices and eviction rates of other products. The overall eviction rate is given by an eviction function

$$E(q_d + q_s)$$

which is increasing in $q_d + q_s$. Only spot VMs are evicted, so their eviction rate is given by

$$e(q_d, q_s) = \frac{q_d + q_s}{q_s} E(q_d + q_s). \tag{4}$$

A market equilibrium $q^*(p_d, p_s)$ is given by a joint solution to equations 3 and 4. Note that this market does not clear through prices but through eviction rates.

Welfare-maximizing pricing Assume constant marginal costs c, and consider a social planner who chooses prices to maximize welfare:

$$\max_{p_s, p_d} U(p_d, p_s, e(q^*(p_d, p_s))) - c(q_d^*(p_d, p_s) + q_s^*(p_d, p_s)) \tag{5}$$

where $U(p_d, p_s, e(q^*(p_d, p_s)))$ denotes the gross utility function associated with demand $D(p_d, p_s, e)$.

If demand is invertible, this problem can be equivalently framed as the problem of choosing the optimal usage for each product:

$$\max_{q_s, q_d} U(q_s, q_d, e(q_s, q_d)) - c(q_s + q_d) \tag{6}$$

Taking first order conditions, one obtains[9]

$$p_s = c - q_s \bar{u}_s^e \frac{\partial e}{\partial q_s} \qquad p_d = c - q_s \bar{u}_s^e \frac{\partial e}{\partial q_d}.$$

These expressions say that the optimal prices take a Pigouvian form: marginal cost plus the marginal externality. The marginal externality is equal to the number of spot users q_s times the average marginal disutility from eviction rates on spot users $\bar{u}_s^e = 1/q_s \cdot \partial U/\partial e$ (which is negative, since it is a disutility) times the increase in eviction rates caused by an additional user $\frac{\partial e}{\partial q_d}$ or $\frac{\partial e}{\partial q_s}$.

The key difference between the two prices arises from the fact that additional on-demand usage simply congests VMs more, whereas additional spot usage has two effects: on the one hand, VMs become more congested, but on the other hand, evictions are spread out among more spot users, reducing congestion among previous spot users. Mathematically, this can be seen since:

$$\frac{\partial e}{\partial q_d} = \frac{q_s + q_d}{q_s} E'(q_s + q_d) + E\frac{1}{q_s}, \qquad \text{and} \qquad \frac{\partial e}{\partial q_s} = \frac{q_s + q_d}{q_s} E'(q_s + q_d) - E\frac{q_d}{q_s^2}. \tag{7}$$

Therefore, on-demand prices should be higher by $-\bar{u}_s^e E\frac{q_s+q_d}{q_s} = -\bar{u}_s^e e$.

These expressions give the optimal prices for tiered reliability—splitting the market

[9]Obtaining these expressions relies on the result that $\partial U/\partial q_j = p_j$. To see why, note that the usual definition of gross utility for demand of one good is $U(q) = \int_0^q p(x)dx$, where $p(x)$ is inverse demand. In the multiple-good case, gross utility $U(\mathbf{q}) = \int_{\mathbf{0}}^{\mathbf{q}} \mathbf{p}(\mathbf{r}) \cdot d\mathbf{r}$ is only well-defined when the demand function is integrable, in which case the gradient theorem gives $\partial U/\partial q_j = p_j$.

into an evictable and a non-evictable product. But is that the best design? One could have aggregated all users into one single tier. Tiered reliability is beneficial if low-eviction disutility users self-select into spot—otherwise the burden of evictions ends up being borne by people with the highest disutility. That is likely the case in real-life markets; however, one can theoretically consider forms of demand for which that is not the case. If low eviction disutility customers have a high preference shifter for on-demand, then they might actually end up preferring on-demand rather than spot. In that case it would be beneficial to merge both products together.[10]

The above result implies that, in our empirical implementation of demand, having heterogeneity in the disutility of evictions is essential to be able to assess the social benefits from segmentation: if there is no heterogeneity, then there is no benefit in terms of social welfare from segmenting the market—although there could still be some benefit in terms of profits if elasticities differ.

Profit-maximizing pricing Now consider the profit maximization problem of the cloud provider. Its objective function is

$$\max_{q_s,q_d}(p_s(q_s,q_d,e(q_s,q_d))-c)q_s+(p_d(q_s,q_d,e(q_s,q_d))-c)q_d. \tag{8}$$

The first order conditions are

$$p_s = c-\Omega_{ss}q_s-\Omega_{sd}q_d+q_s\tilde{u}_s^e\frac{\partial e}{\partial q_s} \qquad p_d = c-\Omega_{dd}q_d-\Omega_{sd}q_d+q_s\tilde{u}_s^e\frac{\partial e}{\partial q_d}, \tag{9}$$

where Ω is the inverse of the Jacobian of demand with respect to prices and $\tilde{u}_s^e$ refers to the average marginal disutility from evictions among users who are indifferent between using and not using spot. Therefore, profit maximization introduces two distortions, as usual: a markup, composed of both terms involving the inverse Jacobian, and a Spence distortion, since the provider accounts for externalities but imperfectly: it only cares about effects on users who are indifferent ($\tilde{u}_s^e$) rather than on all users ($\bar{u}_s^e$).

[10]Suppose, more generally, that the planner could choose a larger number of VM types, and that evictions could be targeted towards certain types. Then the problem becomes one of maximizing both prices and eviction fractions. Optimal prices would be given by expressions resembling 7. The planner would optimally send all evictions to the lowest eviction disutility product, thus going back to the two-tier case.

3.2 Empirical Model

We now build an empirical model that extends the stylized model in Section 3.1 by specifying how users make their choice between on-demand and spot VMs, and in particular how this choice is influenced by users' current belief about spot eviction rates. In our empirical model, in every period the following things happen:

1. Given their current beliefs about eviction rates, users choose between spot, on-demand and the outside option.

2. Given the total demand for spot and on-demand, the congestion technology implies an eviction rate.

3. Evictions are realized.

4. Users update their beliefs about eviction rates.

Crucially, our empirical model is not an equilibrium model in the sense that evictions do not contemporaneously clear the market as they only affect demand via users' beliefs about eviction rates, which adjust only in time for the next period. If no new users ever joined the market, all users would eventually learn the true eviction rates in the market, and the eviction rate and quantity of compute sold would stabilize. To the extent that there is a large influx of new users (whose beliefs initially mimic the prior), however, the average user's belief about eviction rates can remain far from the true eviction rate for an extended period.

3.2.1 Demand Model

We now present our main model of demand for VMs. The model has two parts. First, there is a task arrival process that determines users' computing needs. Second, users who have computing needs choose between an on-demand VM, a spot VM, and an outside option.

Let $\mathcal{I}_{mt}$ denote the (exogenous) set of users during month t in market m, where markets differ by compute varieties (e.g., the amount of memory per core) and location.

Each consumer $i \in \mathcal{I}_{mt}$ has a task to complete (we write $\chi_{imt} = 1$) with a probability that depends (only) on whether they did so in the previous period:

$$\mathbb{P}(\chi_{imt} = 1 | \chi_{i,m,t-1}) = \begin{cases} t_0 & \text{if } \chi_{imt} = 0, \\ t_1 & \text{if } \chi_{imt} = 1. \end{cases}$$

In periods during which the consumer has a task, she also draws a task length L from a distribution G, which we assume to be exogenous.

If consumer i has a task to complete ($\chi_{imt} = 1$), she needs to decide whether to allocate this task to an on-demand VM ($k_{imt} = d$), a spot VM ($k_{imt} = s$), or the outside option ($k_{imt} = o$). As both inside options involve purchasing services from the same cloud provider, we allow them to be more substitutable and partition the set of options into nests as $\{s, d\} \cup \{o\}$. The associated utilities that consumer i derives in choice situation (m, t) from option $k \neq o$ in nest $g(k)$ are then

$$\begin{aligned} u_{imkt} &= v_{imkt} + \zeta_{img(k)t} + (1 - \sigma)\epsilon_{imkt} \\ &= -\beta p_{mkt} - \gamma_i \times 1\{k = s\} e_{mt} - F \times 1\{k \neq k_{i,m,t-1}\} + \delta_{mkt} + \zeta_{img(k)t} + (1 - \sigma)\epsilon_{imkt}. \end{aligned} \tag{10}$$

The first term represents the disutility from paying the price p_{mkt}. The utility of choosing spot ($k = s$) is also influenced by the consumer's dislike of evictions γ_i, which varies across consumers and is distributed as $\gamma_i \sim \text{Gamma}(\alpha, \eta)$, and the rate of evictions (e_{mt}), which happen at a constant Poisson rate.

Next, consumers face a switching cost F if they choose an option that differs from that chosen in the previous month. They also face a market-level unobservable δ_{mkt} that is common to all users.

Finally, $\zeta_{img(k)t}$ is an idiosyncratic shock common to all options inside the same nest (i.e., it varies only by whether an option is the outside option) and ϵ_{imkt} is distributed i.i.d. Type-1 Extreme Value. The distribution of $\zeta_{img(k)t}$ is specified such that $\zeta_{img(k)t} + \sigma\epsilon_{imkt}$ has a Generalized Extreme Value distribution, yielding a nested-logit model (Cardell, 1997; McFadden, 1978).

We set $u_{imot} = \zeta_{img(o)t} + \sigma\epsilon_{imot}$ so that all utilities are measured relative to the mean

utility of the outside option.

The consumer does not know the actual eviction rate e_{mt}. Instead, she learns about it based on her past experience. To yield her personal posterior mean eviction rate $\tilde{\pi}_{imt}$, the consumer combines a Gamma prior with her personal eviction history (x_{imt}, n_{imt}), where x counts her prior evictions and n measures her total prior usage (the sum of all previous task lengths). Bayesian updating yields a posterior mean of

$$\tilde{\pi}_{imt} = \frac{a_0 + x_{imt}}{b_0 + n_{imt}},$$

where a_0 and b_0 are parameters measuring consumers' priors over eviction rates and are assumed to be constant across consumers. Consumers make choices based on these posteriors, which imply the following expected utilities:

$$\begin{aligned}\tilde{u}_{imkt} &= \tilde{v}_{imkt} + \zeta_{img(k)t} + (1-\sigma)\epsilon_{imkt} \\ &= -\beta p_{mkt} - \gamma_i \times 1\{k = s\}\tilde{\pi}_{imt} - F \times 1\{k \neq k_{i,m,t-1}\} + \delta_{mkt} + \zeta_{img(k)t} + (1-\sigma)\epsilon_{imkt}.\end{aligned} \tag{11}$$

Integrating over the heterogeneity in preferences as well as previous period choices, eviction histories, and task arrivals, we get the following market shares for spot and on-demand:

$$s_{smt} = \int_i \frac{\left[\exp(\frac{\tilde{v}_{ismt}}{1-\sigma}) + \exp(\frac{\tilde{v}_{idmt}}{1-\sigma})\right]^{1-\sigma}}{1 + \left[\exp(\frac{\tilde{v}_{ismt}}{1-\sigma}) + \exp(\frac{\tilde{v}_{idmt}}{1-\sigma})\right]^{1-\sigma}} \frac{\exp(\frac{\tilde{v}_{ismt}}{1-\sigma})}{\exp(\frac{\tilde{v}_{ismt}}{1-\sigma}) + \exp(\frac{\tilde{v}_{idmt}}{1-\sigma})} di.$$

$$s_{dmt} = \int_i \frac{\left[\exp(\frac{\tilde{v}_{ismt}}{1-\sigma}) + \exp(\frac{\tilde{v}_{idmt}}{1-\sigma})\right]^{1-\sigma}}{1 + \left[\exp(\frac{\tilde{v}_{ismt}}{1-\sigma}) + \exp(\frac{\tilde{v}_{idmt}}{1-\sigma})\right]^{1-\sigma}} \frac{\exp(\frac{\tilde{v}_{idmt}}{1-\sigma})}{\exp(\frac{\tilde{v}_{ismt}}{1-\sigma}) + \exp(\frac{\tilde{v}_{idmt}}{1-\sigma})} di.$$

We integrate over the joint density of $(\gamma_i, k_{im,t-1}, n_{im,t-1}, x_{im,t-1})$ as choices in the prior period are correlated with the preference parameter γ_i.

Demand for both products—that is, the usage in core-hours—is then given by

$$q_{jmt} = |\mathcal{I}_{mt}| \cdot \mathbb{E}[L] \cdot s_{jmt}, \tag{12}$$

which simply multiplies market shares by the number of users and the expected task

length by each user.

3.2.2 Capacity, congestion and evictions

We now describe how evictions are determined in our model. Consider market m, where on-demand and spot usage are given by (q_{dmt}, q_{smt}). As the same physical machine can be deployed using different operating systems, for purposes of modeling the congestion technology, we need to aggregate across markets to arrive at the total usage of a given physical machine type in a given location. Letting M index congestion-relevant partitions of markets,

$$Q_{Mt} = \sum_{m \in M} q_{dmt} + \sum_{m \in M} q_{smt}$$

Inspired by the patterns in Figure 4, we assume that the congestion technology takes the form

$$e(Q_{Mt}) = \exp\big(\gamma_{Mt} + \beta_{Mt}\, Q_{Mt}\big), \tag{13}$$

where γ_{Mt} and β_{Mt} are parameters whose value depends on how additional demand is allocated to VMs and on the underlying capacity of the system.

For instance, if capacity is K, and the allocation technology is almost perfect, very few customers need to be evicted as long as $Q_{Mt} \leq K$, but evictions rapidly rise once $Q_{Mt} > K$. The functional form in (13) can mimic this behavior by letting $\beta_{Mt} \to \infty$ and setting $\gamma_{Mt} = -\beta_{Mt} \times K + \ln(e_0)$, where e_0 is the eviction rate that obtains when the system operates at capacity (which could be arbitrarily small). When the system operates below capacity, the eviction rate is strictly smaller than e_0. As the system exceeds its capacity, the eviction rate instantaneously explodes.

Crucially, however, a smaller β_{Mt} allows (13) to reflect a 'softer' congestion, which more closely matches the patterns we observe empirically: evictions rise smoothly with aggregate usage. Such a pattern can emerge due to packing or adjacency constraints: when a new on-demand user requests an 8-core VM, the platform must find a physical server with eight available cores; it is not sufficient, for example, to find eight cores on eight different physical servers. As the system approaches capacity, this packing problem gradually becomes less likely to have an eviction-free solution.

Finally, note that we explicitly allow both γ and β to vary over time to account for the installation of additional capacity and for the possibility that the allocation technology could improve over time, allowing the system to operate closer to capacity without elevated eviction rates. In (13), such improvements would be captured by secular shifts towards higher β and lower γ.

4 Identification & Estimation

4.1 Demand

For our demand model, we need to estimate parameters corresponding to price sensitivity (β), eviction disutility (α and η), substitution patterns (σ), prior beliefs about evictions (a_0 and b_0), switching costs F, and task arrival (t_0 and t_1).

We estimate these parameters by GMM. To estimate the parameters for price sensitivity and eviction disutility—for which there are evident endogeneity concerns—we rely on indirect inference moments based on our event-study-based estimates from Section 2.3. To estimate the nest parameter σ, we match the price elasticity of on-demand VMs estimated by Jin et al. (2023). For the remaining parameters, the moments that we use are a combination of first order conditions of the likelihood function and aggregate moments.

Let the full set of parameters be $\theta = (\beta, a_0, b_0, \sigma, t_0, t_1, F, \alpha, \eta)$. We assume that $\mathcal{I}_{mt} \subset \mathcal{I}_{m,t+1}$, and choose the resulting number of new customers $|\mathcal{I}_{m,t+1} \setminus \mathcal{I}_{mt}|$ such that the total number of potential customers each period, $|\mathcal{I}_{mt}|$, equals twice the observed number.

We estimate θ by GMM using the following moments:

1. *Price experiment.* We compute the simulated spot share under observed prices, $S_{smt}(\theta_1)$, and under the *counterfactual* price path p^{cf}_{smt} that would have obtained had it not been for the experiment, giving $\widehat{\Delta}^{\text{price}}_{mt} = \left[S_{smt} - S^{\text{cf}}_{smt}\right] / S_{smt}$ for each market that was part of the experiment. The moment then matches this average percentage change in demand (weighted by inverse-variance weights $\omega_{it} \propto (p_{smt} - p^{\text{cf}}_{smt})^2$) to

the reduced-form difference-in-differences estimate of the same quantity, $\hat{b}$:

$$\sum_{jt} \omega_{it} \left(\widehat{\Delta}^{\text{price}}_{mt} - \hat{b} \right) = 0.$$

2. *Eviction experiment.* Similarly to the *Price Experiment* moment, we exploit the randomness of evictions for customers that are part of our reduced-form eviction event-study above. In particular, for control customers in this event study, we calculate a counterfactual posterior that would have obtained (under the current guess of the model parameters) had they been evicted, $\pi^{\text{cf}}_{imt} = (a_0 + x_{imt} + 1)/(b_0 + n_{imt})$. We then use this counterfactual posterior to evaluate their choices in the subsequent period, and match the resulting percentage change in market share due to the counterfactual eviction to the difference-in-differences coefficient $\hat{r}$.

3. *On-Demand price elasticity.* As we lack price variation for on-demand virtual machines, we instead exploit an external estimate of the price elasticity of on-demand compute from Jin et al. (2023), which estimates $\hat{\varepsilon}_d = -0.941$. We match the model's average individual elasticity of on–demand,

$$\varepsilon_d(\theta_1) = \beta p_{dmt} \Big[1 - \sigma S_{d|g,mt} - (1 - \sigma) S_{dmt} \Big],$$

   to this external estimate $\hat{\varepsilon}_d$.

4. *FOCs for the Gamma prior.* Treating each observed spot decision as coming from the Poisson–Gamma predictive density, the first-order conditions of the log-likelihood imply

$$\frac{1}{N} \sum_{mt} \underbrace{\left[\left(\frac{1\{k_{imt}=s\}}{S_{smt}} - \frac{1-1\{k_{imt}=s\}}{1-S_{smt}} \right) \frac{\partial S_{smt}}{\partial \pi_{imt}} \right]}_{\text{score}} \frac{\partial \pi_{imt}}{\partial a_0} = 0,$$

$$\frac{1}{N} \sum_{mt} \underbrace{\left[\left(\frac{1\{k_{imt}=s\}}{S_{smt}} - \frac{1-1\{k_{imt}=s\}}{1-S_{smt}} \right) \frac{\partial S_{smt}}{\partial \pi_{imt}} \right]}_{\text{score}} \frac{\partial \pi_{imt}}{\partial b_0} = 0.$$

   These two moments pin down (a_0, b_0).

5. *Spot popularity.* To pin down the degree of heterogeneity in the disutility of evictions—

and, hence, the parameters (α, γ)—we match the total number of consumers that ever choose spot. More heterogeneity in the utility of evictions creates persistence in individual users' preference for spot relative to on demand and, thus, a smaller number of individual users accounting for all spot usage.

6. *Spot persistence moments.* Finally, we need a set of moments that jointly identify the task arrival process as well as the switching cost F. To this end, we employ three moments:

$$P(\text{spot}_t = 1 \mid \text{spot}_{t-1} = 1)$$
$$P(\text{spot}_t = 1 \mid \text{spot}_{t-1} = 0),$$

and the ratio of SPOT $\rightarrow$ NO-SPOT $\rightarrow$ SPOT spells to all spot choices—intuitively, such back-and-forth spells are less common the higher switching costs are.

Let $G(\theta) = (G_1, \ldots, G_9)'$ collect the sample analogues of these moment conditions. Our estimate is the GMM minimiser

$$\hat{\theta} = \arg\min_{\theta} G(\theta)' W^{-1} G(\theta),$$

using an inverse variance weighting matrix.

Due to computational constraints, in practice, we estimate θ by alternating estimation of the parameters which do not require re-simulating task arrivals, and those which do. In particular, we partition θ as $\theta = (\theta_A, \theta_B)$ with $\theta_A = (\beta, a_0, b_0, \sigma)$ and $\theta_B = (t_0, t_1, F, \alpha, \eta)$ and then estimate θ by alternating estimation of θ_A while keeping θ_B fixed at $\hat{\theta}_B^{prev}$ and estimation of θ_A while keeping θ_A fixed at $\hat{\theta}_A^{prev}$.

We present our results in Table 1. As expected, we find that consumers have a distaste for price (i.e., $\hat{\beta}_p > 0$), which translates into a mean price elasticity of -0.497.[11] Furthermore, with an estimated nesting parameter of $\hat{\sigma} = 0.67$, the data strongly rejects a logit structure in favor of our nested specification of utilities.

Regarding the users' priors about eviction rates, we find that before they ever use any

[11] This elasticity, which is for all markets, slightly differs from the elasticity we estimate in Section 2.3, which is only for those users in the event studies.

spot product, users expect evictions to happen at a Poisson rate of $\frac{0.001}{0.03} \approx 0.03$ evictions per core-hour. By comparison, this is over 100 times the median observed eviction rate in our data, suggesting perhaps that customers are inherently wary of spot compute. However, this is a very diffuse prior, equivalent in precision to only 1.8 core-minutes of experience with spot products. Furthermore, our estimates lack the power to statistically distinguish the prior mean from the observed mean eviction rate.

Moving on to the consequences of evictions, the average user values one fewer eviction per core-hour about as much as 102 times the usage-weighted mean price of renting a spot product per core-hour. Crucially, our estimate of the shape parameter α suggests that there is considerable heterogeneity in this dislike of evictions, with 10% of users having a more than 17% higher dislike than the mean, and 1% of users having a more than 32% higher dislike than the mean.

While we do estimate a switching cost, it is only equivalent to 0.04 currency units or about 0.02 evictions per core-hour for the average user.

Task arrival is very persistent, with a 99% chance that a task is received this month if there was one last month, but only a 1% chance that a task arrives if there was none in the prior period. These numbers imply that the probability that a user who received a task this month receives a task in one year is still $0.99^{12} \approx 89\%$.

Task sizes have a fat right tail, with a median task size of 1.13 core-minutes but a mean task size of 35 core-hours.

Parameter	Description	Estimate	SE
β_p	Price coefficient	39.86	(9.51)
σ	Nesting parameter ($\sigma = 0$ corresponds to logit)	0.67	(0.04)
a_0	Gamma–Poisson eviction rate prior	0.001	(0.05)
b_0	Gamma–Poisson eviction rate prior	0.03	(0.01)
η	Eviction disutility distribution scale	61.47	(5.57)
α	Eviction disutility distribution shape	1.11	(0.19)
F	Switching cost	1.44	(0.20)
t_1	Task-arrival prob. if task last period	0.99	(0.08)
t_0	Task-arrival prob. if no task last period	0.01	(0.003)
μ_u	Mean of log task size	-3.97	(0.02)
σ_u	Std. dev. of log task size	3.88	(0.01)

Table 1: Parameter estimates and standard errors

Note: This table presents the results of our demand estimation, providing estimates of the parameters that determine a user's utility from choosing spot or on-demand as specified in Equation 10. Heteroskedasticity-robust standard errors in parentheses.

4.2 Supply

We now estimate the congestion technology that we introduced in Section 3.2.2. To this end, recall that the relevant usage Q_{Mt} is aggregated across both spot and on-demand markets, as well as across markets $m \in M$ that correspond to the same type of physical machine M being deployed with different operating systems.

Our attempt to estimate the relationship in (13) faces two key challenges. Firstly, as discussed above, over the long run, usage and capacity both expand, and the unobserved capacity thus confounds the relationship between usage and evictions. To address this challenge, we normalize usage within a month and isolate variation driven by the rhythm of business by restricting attention to usage differences across different days of the week. Secondly, however, this strategy throws away a lot of variation, and hence we now face an issue of limited power. To make progress, we hence (i) impose a plausible functional form on how β_{Mt} can vary over time t while leaving γ_{mt} (which determines eviction rate levels) fully flexible and (ii) employ an empirical Bayes procedure to pool information across congestion-relevant markets M.

Formally speaking, for each congestion-relevant set of markets M, calendar month t, and weekday $w \in \{1, \ldots, 7\}$, let Q_{Mtw} refer to the mean aggregate usage across both

spot and on-demand, and let e_{Mtw} refer to the corresponding mean eviction rate. Our estimating equation corresponding to (13) is then

$$\mathbb{E}[e_{Mtw}] = \exp\left(\beta_M \frac{Q_{Mtw}}{(1/7)\sum_{w'=1}^{7} Q_{Mtw'}} + \gamma_{Mt}\right). \tag{14}$$

Here, the market-specific slopes β_M are restricted to vary over time in such a way that the same percentage deviation relative to the average usage that month has the same effect on the eviction rate. Note, however, that the product-month fixed effects γ_{Mt}, which determine the eviction-rate levels, remain unrestricted. These fixed effects can hence capture capacity build-out, allowing capacity to shift over time.

Estimation is via Poisson pseudo-maximum likelihood (PPML) (Gourieroux et al., 1984), but direct PPML estimates $\widehat{\beta}_M$ are noisy for low-signal products, and a few can be negative even though eviction risk should be non-decreasing in load. To enforce positivity and share information across products, we posit a log-normal population prior

$$\beta_M \sim \text{LogNormal}(\mu, \sigma^2),$$

and exploit the asymptotic normality of maximum-likelihood estimates to conclude that $\widehat{\beta}_M \mid \beta_M \approx \mathcal{N}(\beta_M, s_M^2)$, where s_M is the PPML standard error. We estimate hyperparameters (μ, σ) by maximum likelihood (treating our estimates $\hat{\beta}_M$ as observations) and then compute the posterior mean $\widetilde{\beta}_M = \mathbb{E}[\beta_M \mid \widehat{\beta}_M, s_M; \mu, \sigma]$. Intuitively, this shrinks noisy slopes toward the cross-product mean while respecting $\beta_M > 0$. Finally, we obtain an updated estimate of γ_{Mt} by constraining $\beta_M = \tilde{\beta}_M$ (our shrunk estimate of the slope) and re-estimating Equation 14 under this constraint.

Figure 6 compares the eviction-rate fits from the baseline Poisson model to the Empirical Bayes version on a random set of product–group $\times$ location panels for the last six months of our sample. Each panel plots weekday average eviction rates $\bar{e}_{Mtw}$ against the corresponding normalized usage, with the raw PPML fit in red and the EB-shrunk fit in blue. The EB curves generally track the data more closely and eliminate counterintuitive negative slopes, indicating that shrinkage stabilizes noisy unit-level sensitivities while preserving the increasing relationship between load and evictions.

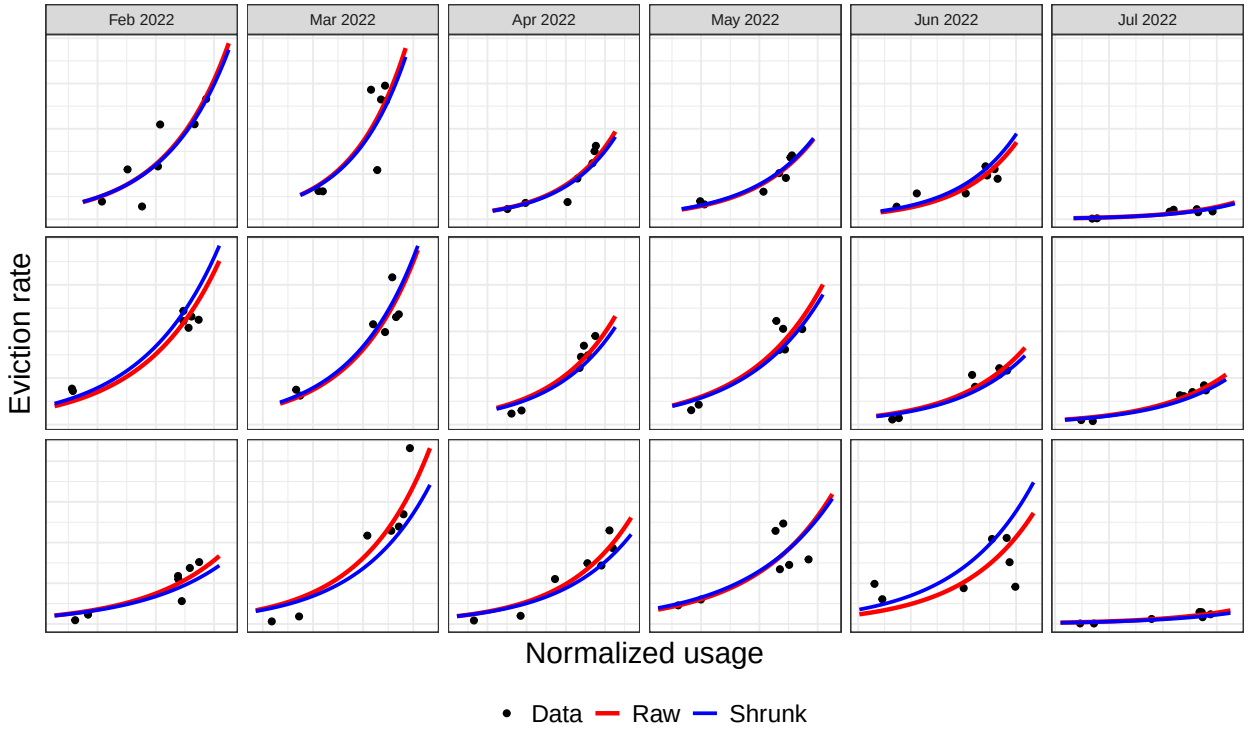


Figure 5: Model fit for top 3 product groups across locations.

Note: The figure displays the relationship between normalized usage (x-axis) and eviction rate (y-axis) from February 2022 to July 2022. The black dots represent the true eviction rates observed in the data, the red line shows the eviction rate estimates from the Poisson regression model as described in Equation (14), and the blue line illustrates the posterior estimates, refined through an Empirical Bayes approach. This approach, assuming β_M follows a log-normal distribution, was applied to address negative estimates of β_M, providing more stable estimates. A detailed illustration of this method is available in the appendix.

5 Welfare and Reliability Under Counterfactuals

We now present preliminary counterfactuals to assess the benefits of tiered reliability. We consider two mechanisms to clear the market. The first is tiered reliability, where, as in the current market, the provider offers an on-demand tier with guaranteed availability at a high price and allocates all evictions to a lower-quality spot tier. The second market clearing mechanism is congestion. The provider offers one single product that receives all evictions. The market endogenously clears through eviction rates. To provide a fair comparison between both mechanisms, we compare scenarios where prices are set optimally, both in terms of profit and welfare maximization.

Table 2 presents our counterfactual results. Tiered reliability represents a Pareto gain

Table 2: Counterfactual results

Counterfactual Objective (1)	Mechanism (2)	Δ Welfare (3)	Δ CS (4)	Δ Profit (5)	Profit (6)
Profit	Tiered	—	—	—	0.59
	Congestion	-1.22	-0.64	-0.55	0.04
Welfare	Tiered	0.17	0.37	-0.16	0.42
	Congestion	-1.02	-0.42	-0.56	0.03

Note: This table compares welfare across different counterfactuals. In each scenario, prices are set to maximize profit or welfare, as specified by column (1). Column (2) specifies the mechanism used to clear the market. Columns (3)-(5) present changes in total welfare as well as in its two components, consumer surplus and profit, measured as a fraction of total revenue in the first row (tiered reliability with profit-maximizing prices). Column (6) presents the cloud provider's profit, once again measured as a fraction of total revenue in the first row.

relative to simply allowing congestion to clear the market, regardless of whether prices are set to maximize profits or welfare. Profits are over ten times higher: when congestion clears the market, allocating evictions to the users with the highest willingness to pay for reliability greatly hinders the ability of the platform to extract surplus from them.

Importantly, switching to tiered reliability does not only benefit the provider—it increases consumer surplus by around as much as profits. Furthermore, it benefits (almost) all types of users, as shown by figure 6:[12] Users with a high willingness to pay for reliability benefit due to the availability of a product with guaranteed service (despite its high price). Users that do not mind evictions that much benefit from the existence of a low-price spot product.

References

Abhishek, Vineet, Ian A. Kash, and Peter Key, "Fixed and Market Pricing for Cloud Services," 2017.

Barro, Robert J. and Paul M. Romer, "Ski-Lift Pricing, with Applications to Labor and Other Markets," *The American Economic Review*, 1987, *77* (5), 875–890.

[12]The only exception are users with very high disutility of evictions, who are slightly worse off when prices are set to maximize profits.

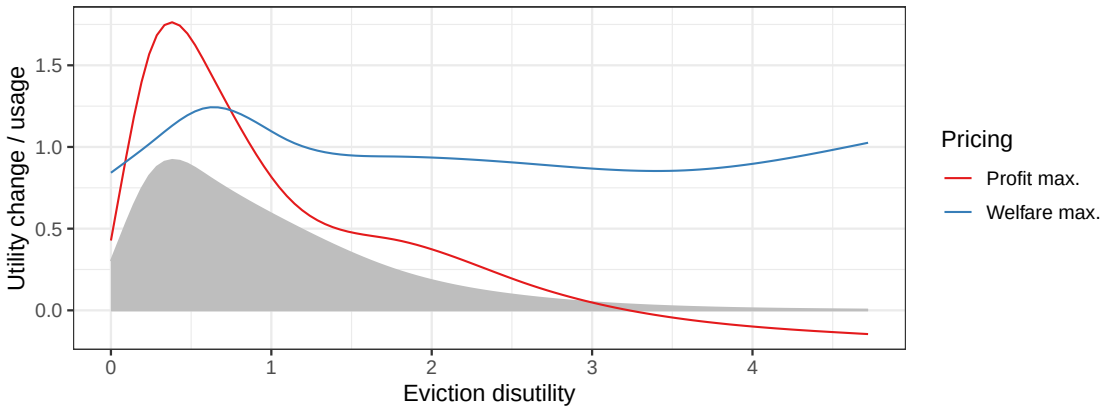


Figure 6: Heterogeneity in consumer surplus effects

Note: This figure shows consumer surplus gains of tiered reliability relative to congestion across users with higher or lower disutility of evictions. The horizontal axis represents the disutility of an eviction. It is normalized to have mean one. The vertical axis shows the change in disutility per unit of usage. It is normalized so that the average value under profit maximization is one. The color represents whether prices are set to maximize welfare or profits. The density in gray in the background represents the density of users.

Biglaiser, Gary, Jacques Crémer, and Andrea Mantovani, "The Economics of the Cloud," 2024.

Blake, Thomas and Dominic Coey, "Why marketplace experimentation is harder than it seems: the role of test-control interference," in "Proceedings of the Fifteenth ACM Conference on Economics and Computation" EC '14 Association for Computing Machinery New York, NY, USA 2014, p. 567–582.

Brand, James M, Mert Demirer, Connor Finucane, and Avner A Kreps, "Firm Productivity and Learning in the Digital Economy: Evidence from Cloud Computing," Technical Report, National Bureau of Economic Research 2024.

Cardell, Scott, "Variance Components Structures for the Extreme-Value and Logistic Distributions with Application to Models of Heterogeneity," *Econometric Theory*, 1997, pp. 185–213.

Castillo, Juan Camilo, "Who Benefits from Surge Pricing?," *Working paper*, 2022.

DeStefano, Timothy, Richard Kneller, and Jonathan Timmis, "Cloud Computing and Firm Growth," *The Review of Economics and Statistics*, 11 2023, pp. 1–47.

Dierks, Ludwig and Sven Seuken, "Cloud Pricing: The Spot Market Strikes Back," *Management Science*, 2022, *68* (1), 105–122.

Gans, Joshua, Mikaël Hervé, and Muath Masri and, "Economic analysis of proposed regulations of cloud services in Europe," *European Competition Journal*, 2023, *19* (3), 522–568.

Gourieroux, Christian, Alain Monfort, and Alain Trognon, "Pseudo Maximum Likelihood Methods: Applications to Poisson Models," *Econometrica*, 1984, *52* (3), 701–720.

Hoy, Darrell, Nicole Immorlica, and Brendan Lucier, "On-demand or spot? Selling the cloud to risk-averse customers," in "International Conference on Web and Internet Economics" Springer 2016, pp. 73–86.

Hummel, Patrick and Michael Schwarz, "Efficient Capacity Provisioning for Firms with Multiple Locations: The Case of the Public Cloud," in "Proceedings of the 23rd ACM Conference on Economics and Computation" 2022, pp. 1018–1039.

Jin, Chuqing, Sida Peng, and Peichun Wang, "Sticky Consumers and Cloud Welfare," *Working paper*, 2023.

Joskow, Paul L. and Catherine D. Wolfram, "Dynamic Pricing of Electricity," *American Economic Review*, May 2012, *102* (3), 381–85.

Kilcioglu, Cinar, Justin M Rao, Aadharsh Kannan, and R Preston McAfee, "Usage patterns and the economics of the public cloud," in "Proceedings of the 26th International Conference on World Wide Web" 2017, pp. 83–91.

Kreindler, Gabriel, "Peak-Hour Road Congestion Pricing: Experimental Evidence and Equilibrium Implications," *Econometrica*, 2024, *92* (4), 1233–1268.

Lu, Yao, Gordon M Phillips, and Jia Yang, "The impact of cloud computing and ai on industry dynamics and concentration," Technical Report, National Bureau of Economic Research 2024.

McFadden, Daniel, "Modeling the Choice of Residential Location," in Anders Karlqvist, Lars Lundqvist, Folke Snickars, and Jörgen Weibull, eds., *Spatial Interaction Theory and Planning Models*, North-Holland, 1978, pp. 75–96.

Mussa, Michael and Sherwin Rosen, "Monopoly and product quality," *Journal of Economic theory*, 1978, *18* (2), 301–317.

Vickrey, William S., "Pricing in Urban and Suburban Transport," *The American Economic Review*, 1963, *53* (2), 452–465.

White, Halbert, "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 1980, *48* (4), 817–838.

Williams, Kevin R., "The Welfare Effects of Dynamic Pricing: Evidence From Airline Markets," *Econometrica*, 2022, *90* (2), 831–858.

Williamson, Oliver E., "Peak-Load Pricing and Optimal Capacity under Indivisibility Constraints," *The American Economic Review*, 1966, *56* (4), 810–827.

A Supply Model

In this section, we detail some technical aspects of the PPML estimator and EB shrinkage, where we estimate product-specific elasticities β_M in the main text:

$$\mathbb{E}[e_{Mtw}] \;=\; \exp\Big(\beta_M \,\frac{Q_{Mtw}}{(1/7)\sum_{w'=1}^{7} Q_{Mtw'}} + \gamma_{Mt}\Big) \tag{15}$$

where M indexes product–group $\times$ location units (aggregating across operating systems), t months, and w weekdays; Q_{Mtw} is aggregate usage across both spot and on-demand. We estimate (15) by PPML with month fixed effects γ_{Mt} for each M. Identification and consistency rely on correct mean specification rather than a full Poisson law, as in Gourieroux et al. (1984), with heteroskedasticity-robust (White) covariance (White, 1980).

A.1 Setup

Let $\widehat{\beta}_M$ denote the PPML slope from (15) and s_M its standard error. We adopt the usual large-sample approximation

$$\widehat{\beta}_M \mid \beta_M \;\sim\; \mathcal{N}(\beta_M,\, s_M^2),$$

and impose a positivity-preserving log-normal prior

$$\beta_M \;\sim\; \text{LogNormal}(\mu, \sigma^2) \qquad (\beta_M > 0),$$

with density $f(\beta;\mu,\sigma) = \phi\big((\log\beta - \mu)/\sigma\big)/(\beta\sigma)$. The resulting marginal density for the summary statistic $\widehat{\beta}_M$ is the one–dimensional mixture

$$p(\widehat{\beta}_M \mid \mu,\sigma) \;=\; \int_0^\infty \phi\left(\frac{\widehat{\beta}_M - \beta}{s_M}\right)\, f(\beta;\mu,\sigma)\, d\beta. \tag{16}$$

A.2 Hyperparameter estimation: GH inside MH

To implement the EB prior and obtain shrunken slopes $\widetilde{\beta}_M$, we first estimate the hyperparameters (μ,σ) by maximizing the sample log–marginal likelihood. We there-

fore combine Gauss–Hermite (GH) and Metropolis–Hastings (MH): GH gives fast, deterministic, numerically stable evaluations of each one–dimensional normal–lognormal marginal, and MH then explores the low–dimensional hyperparameter surface using the GH–evaluated target—i.e., GH inside MH—to deliver a reliable posterior mode (and draws) for (μ, σ).

Concretely, let

$$\ell(\mu, \sigma) \;\equiv\; \sum_{M} \log p(\widehat{\beta}_M \mid \mu, \sigma), \qquad p(\widehat{\beta}_M \mid \mu, \sigma) \text{ as in (16).}$$

We evaluate each term $p(\widehat{\beta}_M \mid \mu, \sigma)$ by N-node Gauss–Hermite quadrature after the change of variables $\beta = \widehat{\beta}_M + \sqrt{2}\, s_M x$:

$$p(\widehat{\beta}_M \mid \mu, \sigma) \;\approx\; \frac{1}{\sqrt{\pi}} \sum_{i=1}^{N} W_i \, f\big(\widehat{\beta}_M + \sqrt{2}\, s_M x_i; \mu, \sigma\big),$$

accumulating the sum with a log–sum–exp routine to prevent underflow. We report the hyperparameter estimate as the posterior mode and use posterior draws for uncertainty quantification.

A.3 Posterior mean of β_M

Given (μ, σ), the posterior mean (our shrunken estimate) is

$$\widetilde{\beta}_M = \frac{\int_0^\infty \beta \; \phi\big((\widehat{\beta}_M - \beta)/s_M\big) \, f(\beta; \mu, \sigma) \, d\beta}{\int_0^\infty \; \phi\big((\widehat{\beta}_M - \beta)/s_M\big) \, f(\beta; \mu, \sigma) \, d\beta}. \tag{17}$$

To compute (17), we draw from the unit–level posterior via a Metropolis–Hastings random walk in $\eta = \log \beta$ (which enforces $\beta > 0$ automatically); the log–posterior is

$$\log \pi(\eta \mid \widehat{\beta}_M, s_M; \mu, \sigma) = -\tfrac{1}{2}\Big(\tfrac{\widehat{\beta}_M - exp(\eta)}{s_M}\Big)^2 - \tfrac{1}{2}\Big(\tfrac{\eta - \mu}{\sigma}\Big)^2 - \log(\sigma\sqrt{2\pi}).$$

A.4 Re-estimating month effects

Before using the model for fit diagnostics and counterfactuals, the month effects must be made coherent with the EB–shrunken slopes. The raw fixed effects γ_{Mt} are identified conditional on the raw $\widehat{\beta}_M$; if we were to plug $\widetilde{\beta}_M$ into the mean $\exp(\beta x + \gamma)$ while

keeping the old γ_{Mt}, the fitted curve would generally shift and no longer match the observed weekday cells. Hence, once we obtain $\widetilde{\beta}_M$, we re-estimate the month effects $\widetilde{\gamma}_{Mt}$ conditional on $\widetilde{\beta}_M$ so that the pair $(\widetilde{\beta}_M, \widetilde{\gamma}_{Mt})$ satisfies the PPML mean restriction and delivers internally consistent predictions $\widehat{e} = \exp(\widetilde{\beta}_M x + \widetilde{\gamma}_{Mt})$. Concretely, we fix $\widetilde{\beta}_M$ and form the "rate residual"

$$y_{Mtw} \equiv \frac{e_{Mtw}}{\exp(\widetilde{\beta}_M x_{Mtw})}.$$

Then $\log \mathbb{E}[y_{Mtw}] = \gamma_{Mt}$. Operationally, for each congestion-relevant market M we estimate

$$\text{PPML:} \qquad y \sim 1 \mid \text{month FE}, \quad (\text{vcov = hetero}),$$

which returns a set of month fixed effects $\{\widehat{\gamma}_{Mt}\}_{t=1}^{T}$.

A.4.1 Additional Graphs

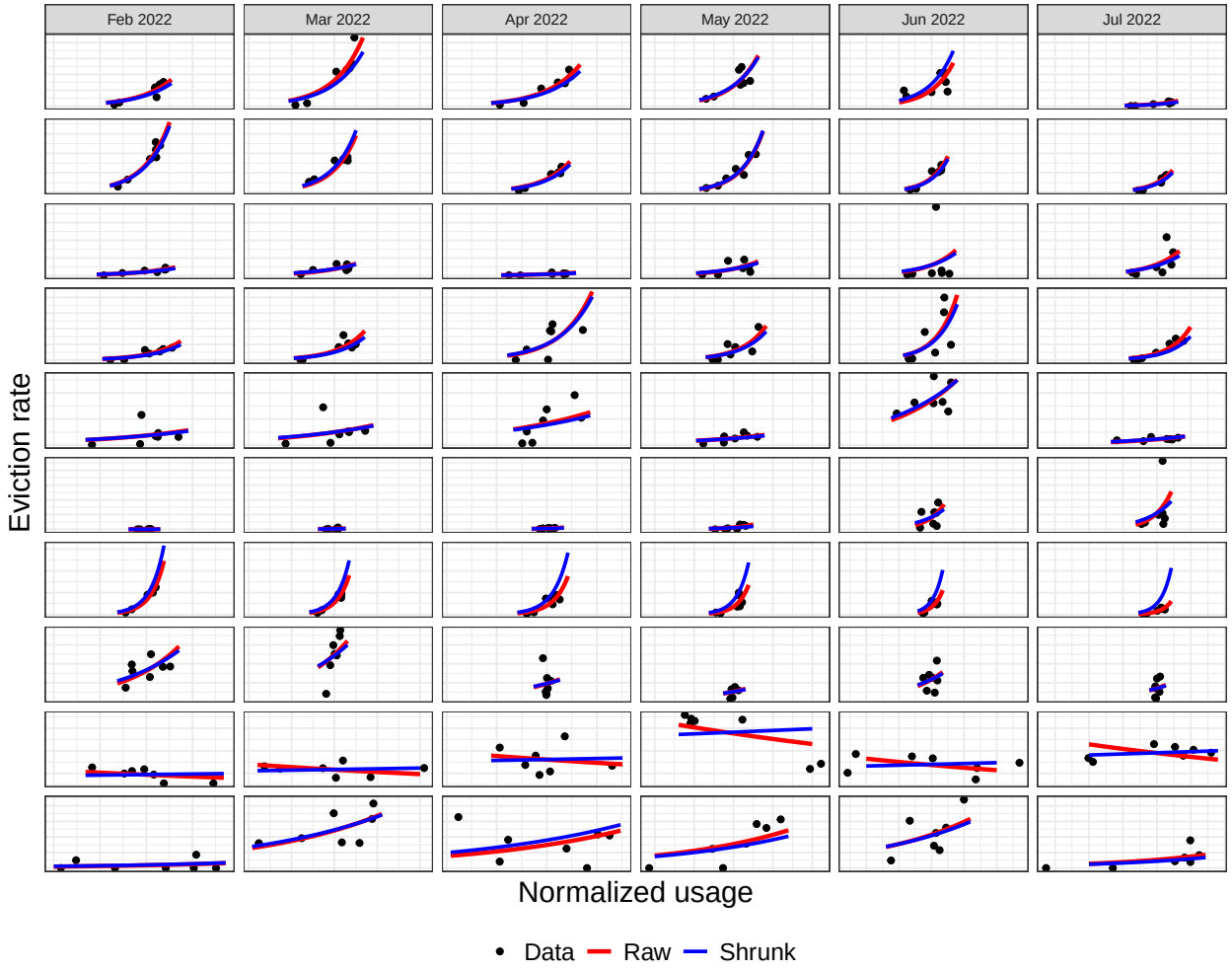


Figure 7: Model fit for random 10 product groups across locations.

Note: The figure displays the relationship between normalized usage (x-axis) and eviction rate (y-axis) from February 2022 to July 2022. The black dots represent the true eviction rates observed in the data, the red line shows the eviction rate estimates from the Poisson regression model as described in Equation (14), and the blue line illustrates the posterior estimates, refined through an Empirical Bayes approach. This approach, assuming β_M follows a log-normal distribution, was applied to address negative estimates of β_M, providing more stable estimates.