Finance and Economics Discussion Series

Federal Reserve Board, Washington, D.C.
ISSN 1936-2854 (Print)
ISSN 2767-3898 (Online)

Queuing, Service Time, and Price Dynamics in Residential Mortgage Lending

Akos Horvath, Benjamin S. Kay

2026-017

Please cite this paper as:
Horvath, Akos, and Benjamin S. Kay (2026). "Queuing, Service Time, and Price Dynamics in Residential Mortgage Lending," Finance and Economics Discussion Series 2026-017. Washington: Board of Governors of the Federal Reserve System, https://doi.org/10.17016/FEDS.2026.017.

Queuing, Service Time, and Price Dynamics in Residential Mortgage Lending*

Akos Horvath
Federal Reserve Board

Benjamin S. Kay
Federal Reserve Board

March 2026

Abstract

Building on queuing theory, we develop and empirically validate a novel theoretical model of residential mortgage supply. Our model gives insight into how the stochastic arrival and sequential servicing of loan applications affect mortgage origination. The model provides closed-form predictions for lenders' optimal response to changes in the level and price elasticity of mortgage demand. Using confidential HMDA data, we estimate that a one standard deviation increase in mortgage demand raises mortgage rate spreads by 3 to 8 basis points, loan quantities by 20 to 32 percent, and application processing times by 3 to 5 days. We also provide empirical evidence for the model prediction that a higher elasticity of mortgage demand moderates price increases due to demand shocks, which can limit lenders' exploitation of their market power.

* We are grateful to Thomas Daula, John Driscoll, James Kahn, Daniel Ringo, Florent Rouxelin, as well as the participants of the William and Mary Real Estate Research Symposium, the Boca Finance and Real Estate Conference, and the Federal Reserve Applied Economics Brown Bag seminar for their helpful feedback on this paper. The views expressed in this paper are those of the authors and do not necessarily reflect the views of the Board of Governors of the Federal Reserve System. Email: akos.horvath@frb.gov and benjamin.s.kay@frb.gov.

1 Introduction

Positive shocks to mortgage demand can cause a significant increase in lender markups because of capacity constraints (Scharfstein and Sunderam, 2016; Fuster, Lo and Willen, 2024).[1] The empirical evidence for this effect is circumstantial: markups are higher at times when the aggregate quantity of mortgage applications is higher (Fuster et al., 2021; Fuster, Lo and Willen, 2024) or in counties with more concentrated mortgage markets (Scharfstein and Sunderam, 2016). However, the finance literature has not generally studied how individual mortgage lenders change their markups when they face capacity constraints.

Our paper investigates three related questions. First, how do mortgage lenders adjust their rates when they experience a positive demand shock and their capacity is limited? Second, how does the equilibrium quantity of mortgages change because of the demand shock and the resulting lender rate adjustment? Third, how does the demand shock affect service quality, as measured by mortgage application processing times?

We make two contributions to the finance literature. First, we introduce a novel structural model of mortgage lender capacity constraints. Second, we empirically identify the model's parameters for mortgage demand and test its predictions at the individual lender level to study the effect of such capacity constraints on the U.S. residential mortgage market. Our findings speak to long-standing economic questions about the pass-through of interest rate cycles, the role of operational frictions in retail credit markets, and the micro-foundations of rate dispersion.

In our structural model, we embed a stochastic (M/M/1) queuing model in a standard profit-maximization framework to capture the effect of lender capacity constraints on mortgage rates, quantities, and processing times. In the model, queue length represents operational congestion, which directly affects borrower waiting times and indirectly affects borrower willingness to pay. To our knowledge, this is the first application of queuing theory to model residential mortgage demand. The model accommodates heterogeneity both across firms and over time while being tractable enough to yield closed-form expressions for optimal mortgage rates, quantities, and processing time as well as their sensitivities to changes in mortgage demand.

In our empirical analysis, we apply our structural model at the lender level to investigate how observed mortgage outcomes change in response to variation in mortgage demand. We use loan-level data from the confidential Home Mortgage Disclosure Act collection, which includes application and origination dates, loan amounts and rates, and borrower characteristics. The analysis proceeds in two steps. In step one, we identify lender-specific mortgage demand curves by using shifts in mortgage-backed securities yields as an instrument. In step two, we use panel regressions to estimate how changes in the levels and slopes of these lender-specific demand curves affect equilibrium interest rates, quantities, and processing times in the residential mortgage market.

[1] Such capacity markups have also been identified in other financial services, such as insurance, banking, and options markets (Gron, 1994; Aliaga-Dıaz and Olivero, 2007; Chen, Joslin and Ni, 2019). The effect is strong enough to have even macroeconomic implications (Menezes and Quiggin, 2022; Kuhn and George, 2019; Fagnart, Licandro and Sneessens, 1997).

We find that mortgage lenders respond to demand shocks consistent with the theoretical predictions of our model with capacity constraints. Specifically, we estimate that a one standard deviation increase in demand raises mortgage spreads by 3 to 8 basis points, increases loan quantities by 20 to 31 percent, and extends processing times by 3 to 5 days. These findings indicate that operational congestion, induced by demand shocks, causes lenders' pricing power and processing delays. Our results provide microeconomic evidence for the capacity constraint channel suggested by aggregate patterns in prior work, confirming that demand shocks can induce variation in mortgage rates and service quality even at the lender level.

We also find that borrower price sensitivity attenuates the increase in lender pricing power in response to demand shocks. Specifically, we estimate that a one standard deviation increase in the elasticity of mortgage demand reduces spreads by 0.7 to 2.4 basis points. Because demand shocks and elasticity are positively correlated in our analysis, this interaction moderates changes in mortgage rates. The estimated net effect of a positive demand shock, which accounts for this correlation, is a 1.3 to 7.5 basis point increase in mortgage spreads. These empirical results are consistent with market dynamics under imperfect competition, where firms face downward-sloping elastic demand curves, which gives them pricing power and the ability to adjust markups in response to both congestion and borrower behavior.

Our regression estimates are robust across specifications (over a variety of fixed effects, controls, and quadratic terms), and their signs are in line with our theoretical model predictions. Importantly, the empirical results indicate that identification is primarily driven by time-series rather than cross-sectional variation, suggesting that macroeconomic factors, such as interest rate cycles, are key determinants of mortgage demand dynamics. Indeed, once we control for time fixed effects, the estimated effects on mortgage rates and processing times become insignificant, while the effect on mortgage quantity becomes stronger. This pattern highlights the limits of lender-specific pricing power in that it suggests that lenders have more pricing power at times when common shocks drive up mortgage demand across the market and borrowers have fewer outside options.

While we focus on the residential mortgage market, our methodology can be applied more broadly to other markets in which firms face stochastic demand and limited processing capacity. Market settings such as small business lending, post-disaster reconstruction, commercial aviation, and semiconductor production often involve similar structural constraints, where service or production delays as well as prices are jointly affected by operational bottlenecks. Demonstrating how a queuing mechanism can be integrated into a structural model, we offer a flexible tool to analyze supply-side frictions in capacity-constrained industries.

The paper is structured as follows. Section 2 introduces our theoretical model. Section 3 describes the data used in the empirical analysis. Section 4 describes our empirical methodology, and Section 5 presents the results. Section 6 concludes.

2 Structural model of mortgage supply and demand

We model a mortgage lender's decision and loan application processing pipeline using the M/M/1 queuing framework (see, e.g., Erlang (1909)). In this framework, the arrival of applications follows a Poisson process with rate λ, and the processing of applications follows a Poisson process with rate μ as a "single server" (the lender) processes one application at a time on a first-come-first-served basis with an unbounded pipeline.[2] We model the stochastic mortgage applications by letting the quoted price p of mortgages determine the application arrival rate $\lambda(p)$, which in turn specifies the expected demand curve:

$$\mathbb{E}[q(p)] = \lambda(p) = \lambda_0 - \psi \cdot p, \tag{1}$$

where $\psi \geq 0$ is the price sensitivity of applicants. The parameters λ_0 and ψ provide a reduced-form representation of the determinants of demand, such as amenities offered by the lender, competition from other lenders, as well as the applicant's search and switching costs. Intuitively, a larger lender has higher arrival rate λ_0 and higher service rate μ. Furthermore, although the slope parameter ψ need not scale with lender size, similar semi-elasticities across firms are plausible.

A single firm faces independent and identically distributed negative exponential waiting times between customers. This is equivalent to a Poisson arrival process for customers, which we model as governed by the parameter λ. A single lender serves mortgage applicants one at a time from the front of the queue, according to a first-come, first-served discipline.[3] Service time is also independent and identically distributed exponentially, which we model as governed by the parameter μ, an exogenous firm characteristic. When the service is complete, the customer leaves the queue, and the number of customers in the system is reduced by one. There is no limit on the maximum number of customers a queue can contain (the buffer is infinite in size).[4]

Let N_t denote the number of applications in the system, either waiting or in service, at time t. Conditioning on the quoted price p, $\{N_t\}_{t\geq 0}$ is a birth-death Markov chain with birth rate λ and death rate μ, with transition rate matrix Q on the state space $\{0, 1, 2, \ldots\}$:

$$Q = \begin{bmatrix} -\lambda & \lambda & 0 & \cdots & 0 & 0 \\ \mu & -(\mu+\lambda) & \lambda & \cdots & 0 & 0 \\ 0 & \mu & -(\mu+\lambda) & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -(\mu+\lambda) & \lambda \\ 0 & 0 & 0 & \cdots & \mu & -\mu \end{bmatrix}. \tag{2}$$

[2] The "single server" is an abstraction of the lender's application processing work—which includes, for example, verification and underwriting activities—and does not literally represent a single employee. Alternatively, one could use an M/M/c framework, with c being the number of servers. Such an approach would mainly differ from our model in the dispersion of the average processing time defined in Equation (3).

[3] The discipline of a queue is the rule that determines the formation of a queue and the way in which a customer is selected for service from those waiting. Other possibilities include random selection for service, "last come, first served" (last in first out), and various methods of prioritizing customers (Peterson and Davie, 2024).

[4] In practice, like in the line for checkout at a store, or the queue of processes for a CPU, there is a maximum buffer, though it may not bind in an economically meaningful way.

It follows that the average processing time (APT) of an application, which includes both waiting and service times, is:

$$\text{APT} = \frac{1}{\mu - \lambda}, \tag{3}$$

which is finite if $\lambda < \mu$.[5]

To our knowledge, our model is the first in the finance literature that applies the M/M/1 queuing framework to endogenize demand within a firm-level pricing model. While queuing models have been extensively used in the field of operations research, their economic application has been limited, with only a handful of articles in finance, operations research economics, and health care economics (Azriel, Feigin and Mandelbaum, 2019; Green, 2006; Cont, Stoikov and Talreja, 2010). Naor (1969) considers the welfare implications of imposing tolls on customers arriving at service facilities with M/M/1 queues. In contrast, we focus on the producer's problem, rather than doing welfare analysis. More closely related to our work, Levy and Levy (1991) develop a model of mortgage origination to compare the revenue performance of rate lock and no-lock pricing regimes. Like our model, theirs uses the M/M/1 framework. However, their study is purely theoretical and focuses on the lender's ability to adjust prices under different contractual commitments, whereas we investigate loan pricing, service quantities, and processing times. Their model omits production costs, assumes full price discrimination, and does not model borrower behavior: customers arrive at fixed rates that do not depend on quoted prices. In contrast, in our model, borrowers respond to both prices and, indirectly, processing delays via the lender's cost function. Finally, Cowdrey et al. (2018) use the M/M/1 framework to simulate queuing strategies in a retail banking context. Their goal is to improve operational performance through alternative service scheduling rules and highlight the operational relevance of congestion, not to model loan pricing, demand, or the firm's optimization problem. Their paper treats arrival rates and service rules as exogenous and focuses on queue efficiency, not economic behavior.

2.1 The lender's profit-maximization problem

In the queuing framework described above, the lender chooses a posted price p so as to maximize its profit on the expected quantity of loans implied by the arrival rate $\lambda(p)$. Generally, p can encompass both upfront closing costs and interest rate components, such as the risk-free rate, the term premium, the prepayment premium, and the credit spread. In the empirical work in Section 5, we take p to be the primary–secondary mortgage spread, which includes the lender's origination costs and markup.[6]

[5] If average arrival times exceed average service times, the queue explodes and no stationary distribution exists. Therefore, in order to ensure that the model has a steady state, we assume that $\lambda(p) < \mu$ holds true. Appendix A presents an alternative model specification in which the lender has an objective function that endogenously leads to finite queue lengths and waiting times without this additional parametric assumption.

[6] In the lender's problem, considering the price to be the primary-secondary mortgage spread is a natural choice because this is the part of the mortgage rate that the lender can plausibly control and therefore seek to optimize. Crucially, however, this choice can be made without loss of generality, and it does not imply that mortgage demand, reflected by the arrival rate $\lambda(p)$, is unaffected by yield curve shifts. Rather, such yield curve shifts are captured by the intercept parameter λ_0 of the demand curve, specified in Equation (1).

The lender's expected total revenue is:

$$TR(p) = \mathbb{E}\left[p \cdot q(p) \mid p\right] = p \cdot \lambda(p), \tag{4}$$

and the lender's expected total cost is:

$$\begin{aligned} TC(p) &= \mathbb{E}\left[c \cdot q(p)^2 + d \cdot q(p) - e \cdot (\mu - q(p)) \mid p\right] = \\ &= c \cdot \left(\lambda(p)^2 + \lambda(p)\right) + (d+e) \cdot \lambda(p) - e \cdot \mu, \end{aligned} \tag{5}$$

where the expectation of the stochastic mortgage demand q is defined in Equation (1).[7]

We use a quadratic cost function that incorporates a linear penalty term for operating near capacity (i.e., when $q(p)$ gets close to μ) because this functional form yields closed-form model predictions, which we can test in the empirical part of our analysis. The parameters $d > 0$ and $c > 0$ capture the baseline marginal origination costs and increasing marginal costs from staffing and overtime pressures as more loans arrive, respectively. The $-e \cdot (\mu - q)$ term in the cost function captures the benefit of keeping a capacity buffer, which can be beneficial because it enables the lender to absorb positive demand shocks without delays in application processing. Such delays can be costly for the lender because they can lead to customer dissatisfaction, for example, through rate-lock violations. The penalty term can also be interpreted as a reduced-form approximation of the delay-related opportunity costs that arise in queuing environments, where expected response times increase rapidly as utilization rises.[8]

The lender's expected profit is:

$$\begin{aligned} \Pi(p) = TR(p) - TC(p) &= p \cdot \lambda(p) - c\left(\lambda(p)^2 + \lambda(p)\right) - d \cdot \lambda(p) + e\left(\mu - \lambda(p)\right) = \\ &= -c\left(\lambda_0 - \psi \cdot p\right)^2 - (c + e + d - p)\left(\lambda_0 - \psi \cdot p\right) + e \cdot \mu. \end{aligned} \tag{6}$$

As participation constraint, we impose the sign restriction:

$$\lambda_0 > \psi(c + d + e), \tag{7}$$

which rules out the corner solution in which the optimal quantity is zero.[9]

Our model permits variation in quoted prices across lenders—a feature that typically depends on product differences or search frictions. As shown by Burdett and Judd (1983), even identical firms can sustain price dispersion when some consumers face search costs. Subsequent studies, including

[7] These revenue and cost function specifications hold true even if some applications are rejected or withdrawn as long as approval and withdrawal rates are locally constant in the optimization range. If not all applications end in loan origination, then the levels of the revenue and cost functions shift, but the comparative statics remain unchanged.

[8] In Appendix A, we use an alternative cost function, replacing the linear penalty term with a "capacity wall" that generates logarithmically increasing costs as the system approaches full utilization. Under this alternative model specification, spare capacity appears explicitly in the lender's first-order condition, which endogenously creates a preference for it. Nonetheless, these different model specifications yield qualitatively similar results.

[9] This sign restriction follows from Equation (10) if $\psi > 0$ and $c, d, e > 0$ (which are natural assumptions for the demand slope and cost coefficients). The $\lambda_0 \leq \psi(c + d + e)$ case captures the lender's exit from the market.

Wolinsky (1986) and Eeckhout and Kircher (2010), show that imperfect information or sorting across consumers can lead to persistent price differences in competitive markets. Although our model abstracts from these mechanisms, quoted loan prices affect application arrival rates, allowing for price sensitivity and variation across lenders. Because a firm that deviates from the market-clearing price still retains positive demand, our model accommodates price dispersion without requiring information frictions or strategic interactions. In our model, this dispersion arises as a result of pipeline congestion, because even under identical service technologies and demand functions, differences in quoted prices lead to differences in customer flow and queue length.

2.2 First-order condition and optimal price

Differentiating the profit function in Equation (6) with respect to p and setting the derivative to zero yields the first-order condition:[10,11]

$$\frac{\partial \Pi}{\partial p} = (2c\psi + 1)\big(\lambda_0 - \psi p\big) + \psi\big(c + d + e - p\big) = 0. \tag{8}$$

Solving for the loan's optimal quoted price results in:

$$p^* = \frac{\lambda_0(1 + 2c\psi) + \psi(c + d + e)}{2\psi(1 + c\psi)}, \tag{9}$$

which implies an optimal positive price under $\lambda_0, \psi > 0$ and $c, d, e > 0$.

Substituting into Equation (1) yields the optimal average quantity:

$$\lambda(p^*) = \frac{\lambda_0 - \psi(c + d + e)}{2(1 + c\psi)}, \tag{10}$$

which is positive under the participation constraint imposed in Equation (7).

Additionally, as discussed at the beginning of Section 2, we assume that $\lambda(p^*) < \mu$ to ensure that expected waiting times are finite under optimal pricing. This assumption makes intuitive sense, as mortgage lenders typically have some spare capacity so that they can meet unanticipated demand shocks while maintaining service standards.

Substituting into Equation (4) yields the lender's optimal revenue:

$$TR(p^*) = p^* \cdot \lambda(p^*) = \frac{\big[\lambda_0(2c\psi + 1) + (c + d + e)\psi\big]\big[\lambda_0 - (c + d + e)\psi\big]}{4\,\psi\,\big(c\psi + 1\big)^2}. \tag{11}$$

[10] The profit function's second derivative is $\frac{\partial^2 \Pi}{\partial p^2} = -2\psi - 2c\psi^2 < 0$ under $\psi > 0$ and $c, d, e > 0$, which implies that the solution maximizes profit.

[11] This marginal cost function does not depend on spare capacity or include the service rate μ. Consequently, on the margin, the lender does not internalize how pricing affects waiting times. In our alternative model specification presented in Appendix A, the lender's marginal cost is a function of the average waiting time, which makes the lender endogenize this dynamic and thus also obviates the need for the assumption that $\lambda(p) < \mu$.

Finally, substituting into Equation (3) yields the optimal average processing time:

$$\mathrm{APT}(p^*) = \frac{1}{\mu - \lambda(p^*)} = \frac{2\,(c\psi + 1)}{2c\mu\psi + (c + d + e)\psi - \lambda_0 + 2\mu}. \tag{12}$$

2.3 Partial derivatives and (semi-)elasticities

The closed-form solutions presented in Section 2.2 have closed-form sensitivities with respect to the underlying parameters. In our analysis, we focus on four key derivatives:[12]

$$\frac{\partial \lambda(p^*)}{\partial \lambda_0} = \frac{1}{2\,(c\psi + 1)} > 0; \tag{13}$$

$$\frac{\partial \mathrm{APT}(p^*)}{\partial \lambda_0} = \frac{2\big(1 + c\psi\big)}{\big(2\mu - \lambda_0 + \psi(2c\mu + c + d + e)\big)^2} > 0; \tag{14}$$

$$\frac{\partial p^*}{\partial \lambda_0} = \frac{2c\psi + 1}{2\psi\,(c\psi + 1)} > 0; \tag{15}$$

$$\frac{\partial p^*}{\partial \psi} = -\frac{2c^2\lambda_0\psi^2 + c(c + d + e)\psi^2 + 2c\lambda_0\psi + \lambda_0}{2\psi^2\big(1 + c\psi\big)^2} \;<\; 0, \tag{16}$$

where the corresponding signs apply under $\lambda_0, \psi > 0$ and $c, d, e > 0$ as well as the participation constraint imposed in Equation (7).

For the purposes of our empirical analysis, we express these derivatives as (semi-)elasticities. The elasticity of the optimal flow with respect to the demand-intercept parameter λ_0 (i.e., the baseline arrival rate) is:

$$\varepsilon_{\lambda,\lambda_0} \equiv \frac{\partial \lambda}{\partial \lambda_0} \cdot \frac{\lambda_0}{\lambda} = \frac{\lambda_0}{\lambda_0 - (c + d + e)\psi} > 0.$$

The elasticity of average processing time with respect to λ_0 is:

$$\varepsilon_{\mathrm{APT},\lambda_0} \equiv \frac{\partial \mathrm{APT}}{\partial \lambda_0} \cdot \frac{\lambda_0}{\mathrm{APT}} = \frac{\lambda_0}{-\lambda_0 + 2\mu(c\psi + 1) + (c + d + e)\psi} > 0.$$

The semi-elasticity of the optimal price with respect to λ_0 is:

$$\text{semi-}\varepsilon_{p^*,\lambda_0} \equiv \frac{\partial p^*}{\partial \lambda_0} \cdot \lambda_0 = \frac{2c\psi + 1}{2\psi(c\psi + 1)} \cdot \lambda_0 > 0,$$

and the semi-elasticity of price with respect to the demand-slope parameter ψ is:

$$\text{semi-}\varepsilon_{p^*,\psi} \equiv \frac{\partial p^*}{\partial \psi} \cdot \psi = -\frac{2c^2\lambda_0\psi^2 + c(c + d + e)\psi^2 + 2c\lambda_0\psi + \lambda_0}{2\psi\,(c\psi + 1)^2} < 0,$$

[12] We report the full set of derivatives in Appendix B.

The signs of these (semi-)elasticities follows from the signs of the corresponding partial derivatives. In the rest of our paper, we empirically confirm the signs and estimate the magnitudes of these relevant elasticities, using the data in Section 3 and applying the methodology in Section 4.

3 Data

In our empirical analysis, we utilize loan-level information from the confidential Home Mortgage Disclosure Act (HMDA) collection. The HMDA mandates that all but the smallest mortgage lenders report data on residential mortgage applications and originations.[13] Importantly, in contrast to the publicly available HMDA collection, the confidential collection includes the exact dates of mortgage applications and lender decisions as well as a set of applicant and loan characteristics, such as credit scores and interest rates on originated loans. This enhanced data granularity is crucial for our empirical analysis because it enables us to measure mortgage rates and processing times as well as to control for borrower and loan characteristics in our regressions.

Our sample contains every conventional, first-lien, single-family mortgage application reported in the HMDA collection between 2018 and 2023, including approvals, withdrawals, and rejections. The sample begins in 2018 because the confidential HMDA collection does not have precise interest rate information on originated loans before that. Furthermore, we include only "purchase" and "refinance" loans and exclude "home improvement" loans because the latter are scarce and may have unique demand dynamics.[14] Altogether, our sample consists of about 41 million mortgage applications, out of which about 35 million ended in loan originations.

We augment the information in the HMDA collection in two ways. First, from the data platform of ICE Data Indices, LLC, we use the average effective yield on agency residential mortgage-backed securities to measure the level of yields in the secondary mortgage market. Specifically, we use the dollar-weighted average of the effective yields on securities included in the ICE BofA indexes MFNC and MGNC, which track outstanding current-coupon agency mortgage-backed securities with 15-year and 30-year fixed-rate residential mortgage pools.[15] Second, we add the legal names and entity types of mortgage lenders from the dataset constructed by Robert Avery.[16]

4 Empirical methodology

Our empirical analysis proceeds in two stages. First, we use variation in mortgage applications and interest rates over time to identify lender-specific mortgage demand curves (Section 4.1). Specifically, we estimate the empirical counterparts of the baseline (level) parameter λ_0 and the price sensitivity

[13] Submissions are generally mandatory for lenders that extended at least 100 closed-end mortgage loans in each of the two preceding calendar years. See the Consumer Financial Protection Bureau's flow chart for a detailed description of HMDA reporting requirements.

[14] Home improvement loans account for only a small (about 3%) share of the HMDA data in our sample period.

[15] ICE BofA indexes MFNC and MGNC are products of ICE Data Indices, LLC (ICE Data) and are used with permission. ICE is a registered trademark of ICE Data or its affiliates.

[16] The dataset constructed by Robert Avery is described and accessible at https://www.philadelphiafed.org/surveys-and-data/consumer-finance-data/home-mortgage-disclosure-act-lender-file.

(slope) parameter ψ of the mortgage demand curve in the structural model introduced in Section 2. Second, we exploit variation in these parameter estimates to study how changes in the levels and slopes of these lender-specific demand curves affect equilibrium interest rates, quantities, and processing times in the residential mortgage market (Section 4.2).

4.1 Estimating the parameters of lender-specific mortgage demand

We estimate the empirical counterpart of the expected mortgage demand curve used in the structural model introduced in Section 2. The demand curve takes the form $\lambda = \lambda_0 - \psi \cdot p$ at the individual lender level, where λ is the arrival rate of mortgage applications, p is the price of mortgage loans, and λ_0 and ψ are the level and slope parameters of the demand curve, respectively. In order to study the effect of variation in mortgage demand in the second stage of our empirical analysis, we need to let the level and slope parameters vary across lenders and over time. Accordingly, in the first stage of our analysis, we generalize the above mortgage demand curve to multiple lenders, geographical areas, and periods, and use the empirical equation:

$$q(l,s,m) = \lambda_0(l,s,m) - \psi(l,m) \cdot p(l,s,m) + \tilde{\varepsilon}(l,s,m), \tag{17}$$

where $q(l,s,m)$ is the observed number of mortgage applications submitted to lender l in state s in year-month m, $p(l,s,m)$ is the average primary-secondary mortgage rate spread on mortgage loans originated by lender l in state s in year-month m.[17] The coefficients $\lambda_0(l,s,m)$ and $\psi(l,m)$ represent the lender-state-month-specific level and lender-month-specific slope parameters of the demand curve, respectively, and $\tilde{\varepsilon}(l,s,m)$ is the error term.

We estimate the demand parameters $\lambda_0(l,s,m)$ and $\psi(l,m)$ by first standardizing the empirical demand equation and then using shifts in mortgage supply induced by weakly exogenous variation in the cost of mortgages to identify the parameters. Specifically, we apply the following affine transformation to the empirical demand equation:

$$\underbrace{\frac{q(l,s,m) - A(l,m)}{B(l,m)}}_{L(l,s,m)} = \underbrace{\frac{\lambda_0(l,s,m) - A(l,m)}{B(l,m)}}_{L_0(l,s,m)} - \frac{\psi(l,m)}{B(l,m)} \cdot p(l,s,m) + \underbrace{\frac{\tilde{\varepsilon}(l,s,m)}{B(l,m)}}_{\varepsilon(l,s,m)}, \tag{18}$$

where $A(l,m)$ and $B(l,m)$ are the sample mean and standard deviation of application volume for lender l in the specific calendar month of year-month m across all states, respectively, and L denotes standardized mortgage demand.[18]

Substituting in the definition of $p(l,s,m)$, which is the *mortgage rate spread*, further yields

$$L(l,s,m) = L_0(l,s,m) - [r(l,s,m) - \text{AMBS}(m)] \cdot \frac{\psi(l,m)}{B(l,m)} + \varepsilon(l,s,m), \tag{19}$$

[17] We construe the primary-secondary mortgage rate spread as the difference between the mortgage loan rate and the effective yield on agency mortgage-backed securities in the secondary market, defined in Section 3.

[18] By letting the mean and standard deviation of applications vary across calendar months within each lender, we allow for seasonality of mortgage applications at the individual lender level.

where $r(l,s,m)$ is the average *mortgage rate* on loans originated by lender l in state s in year-month m, and AMBS(m) is the effective yield on agency residential mortgage-backed securities in year-month m, defined in Section 3. It follows from Equation (19) that the standardized mortgage demand curve, expressed as a function of the mortgage rate, is

$$L(l,s,m) = \underbrace{\left[L_0(l,s,m) + \text{AMBS}(m) \cdot \frac{\psi(l,m)}{B(l,m)}\right]}_{\alpha(l,s,m)} - r(l,s,m) \cdot \underbrace{\frac{\psi(l,m)}{B(l,m)}}_{\beta(l,m)} + \varepsilon(l,s,m), \tag{20}$$

where α and β are the level and slope parameters of the standardized demand curve, respectively. Hence, the functional relationships between the parameters of the original and standardized demand curves are

$$\begin{aligned} \lambda_0(l,s,m) &= A(l,m) + B(l,m) \cdot [\alpha(l,s,m) - \text{AMBS}(m) \cdot \beta(l,m)] \text{ and} \\ \psi(l,m) &= B(l,m) \cdot \beta(l,m). \end{aligned} \tag{21}$$

We employ a panel instrumental variable regression model to identify the level and slope parameters of the transformed demand curve specified in Equation (20). Specifically, because the interest rate $r(l,s,m)$ and the standardized number of mortgage applications $L(l,s,m)$ are jointly determined in equilibrium, we use AMBS$(m-1)$, the one-month lag of the effective yield on agency residential mortgage-backed securities, as an instrument for $r(l,s,m)$.[19] In line with the econometric discussion in Angrist and Krueger (2001), we argue that *aggregate* shifts in secondary market yields in year-month $(m-1)$ constitute weakly exogenous (i.e., pre-determined) cost shocks for lenders in year-month m, which induce shifts in their *individual* loan supply curves. Our identification strategy uses this variation in the individual loan supply curve to identify the parameters of the individual loan demand curve.

Finally, applying Equation (21), we use $A(l,m)$ and $B(l,m)$ as well as the regression coefficient estimates $\widehat{\alpha}(l,s,m)$ and $\widehat{\beta}(l,m)$ to calculate the mortgage demand curve parameter estimates $\widehat{\lambda}_0(l,s,m)$ and $\widehat{\psi}(l,m)$ used in the next stage of the empirical analysis. Section 5 presents summary statistics for the parameter estimates.

4.2 Studying the effect of variation in lender-specific mortgage demand

We use the demand parameters estimated in Section 4 to investigate how variation in lender-specific demand affects equilibrium mortgage rate spreads, application volumes, and application processing

[19] First-stage diagnostics indicate that AMBS$(m-1)$ is a relevant instrument for $r(l,s,m)$ with and without controlling for lender-state-month fixed effects. In the first stage, the pass-through coefficient on AMBS$(m-1)$ is 0.96, with $p < 0.001$, consistent with a near one-for-one mapping from secondary-market yields to primary-market pricing. Sanderson–Windmeijer and Stock–Yogo partial-F statistics comfortably exceed conventional weak-instrument thresholds with $p < 0.001$. Additionally, the instrument also satisfies the exclusion restriction because mortgage supply–demand dynamics at the individual lender level arguably do not drive aggregate shifts, and especially lagged aggregate shifts, in secondary-market yields.

times. In particular, we consider the four partial derivatives introduced in Section 2.3:

$$\frac{\partial \lambda(p^*)}{\partial \lambda_0}, \quad \frac{\partial \text{APT}(p^*)}{\partial \lambda_0}, \quad \frac{\partial p^*}{\partial \lambda_0}, \quad \text{and} \quad \frac{\partial p^*}{\partial \psi},$$

for which our structural model yields closed-form expressions. These quantify, respectively, the sensitivity of the number of applications, application processing times, and the mortgage rate spread to the level parameter λ_0 of the mortgage demand curve as well as the sensitivity of the mortgage rate spread to the slope parameter ψ of the mortgage demand curve.

To quantify the empirical magnitudes of the above derivatives, we aggregate our data to the lender-state-month level and estimate regressions of the form:

$$\begin{aligned} y(l,s,m) &= a(l,s) + b \cdot \log(\widehat{\lambda}_0(l,s,m)) + c \cdot \log(\widehat{\lambda}_0(l,s,m))^2 + \boldsymbol{d}^\top \boldsymbol{x}(l,s,m) + \varepsilon(l,s,m), \text{ and} \\ y(l,s,m) &= a(l,s) + b \cdot \log(\widehat{\psi}(l,m)) + c \cdot \log(\widehat{\psi}(l,m))^2 + \boldsymbol{d}^\top \boldsymbol{x}(l,s,m) + \varepsilon(l,s,m), \end{aligned} \tag{22}$$

where l, s, and m index lenders, states, and year-months, respectively. The dependent variable $y(l,s,m)$ is the logarithm of the number of applications, the logarithm of the average application processing time, or the average mortgage rate spread for lender l in state s and year-month m. Depending on the specific partial derivative under examination, the regression model includes the quadratic polynomial of the logarithm of the estimated mortgage demand level parameter $\widehat{\lambda}_0(l,s,m)$ or the estimated mortgage demand slope parameter $\widehat{\psi}(l,m)$.

Control variable vector $\boldsymbol{x}$ includes loan and borrower characteristics aggregated at the lender-state-month level. Specifically, we control for the interaction of the loan purpose indicator's percent share and the jumbo loan indicator's percent share as well as the triple-interaction of the loan-to-value (LTV) ratio quartile indicator's percent share, the borrower's FICO score quartile indicator's percent share, and the borrower's debt-to-income (DTI) ratio tercile indicator's percent share.[20] Additionally, we include lender-state fixed effects, denoted by $a(l,s)$, to control for potential unobserved confounding factors across lender-states.[21]

Our panel regression model, as specified in Equation (22), captures the effect of variation in lender-specific demand curves on equilibrium interest rates, quantities, and processing times while putting equal weights on all lenders, states, and year-months. Although the resulting estimates are representative of a randomly chosen lender, state, and point in time, we recognize that residential mortgage lending activity is not equally distributed across lenders, states, and year-months. Hence, partly as a robustness test, we also apply our regression model by weighting lender-state-month cells in our aggregated dataset according to the number of mortgage applications, thereby producing estimates that are representative of a typical residential mortgage loan.

Finally, we also estimate all regression specifications by instrumenting for the logarithms of parameter estimates $\log(\widehat{\lambda}_0)$ and $\log(\widehat{\psi})$ using their one-month-lagged values to address potential

[20] In our sample, loan purpose can be "purchase" or "refinance", implying that the interaction of the loan purpose and jumbo loan indicators results in $(2 \times 2 =)$ 4 possible purpose-jumbo groups whose percent share we include in the regression. Similarly, the triple-interaction of LTV ratio quartile, FICO score quartile, and DTI ratio tercile indicators results in $(4 \times 4 \times 3 =)$ 48 possible LTV-FICO-DTI groups whose percent share we include in the regression.

[21] In alternative regression specifications, we use lender, state, and year-month fixed effects to control for unobserved confounding factors across lenders, states, and over time.

concerns about endogeneity due to backwards causation. In particular, one could argue that the outcome variables in Equation (22), such as mortgage rate spreads, influence the lender-specific mortgage demand curves, whose parameters in turn are used as explanatory variables. Therefore, we use the lagged parameter estimates as instruments, which are pre-determined and thus satisfy the exclusion restriction. These lagged instruments are also relevant and valid because of the strong autocorrelations of the structural parameter estimates.[22]

5 Empirical results

Table 1 shows summary statistics for the variables used in our empirical analysis, calculated using the dataset aggregated to the lender-state-month level. The variables are the following: the number of loan applications (q); the level (λ_0) and slope (ψ) parameters of the lender-specific demand curves estimated in the first stage of the analysis (Section 4.1); the application processing time (APT); and the mortgage rate spread (p^*) on originated loans.[23]

Table 1: Summary statistics of structural parameter estimates

	Mean	σ	$\sigma_{l,s}$	$\sigma_{l,s,t}$	p10	p25	p50	p75	p90
q	42.4	211.3			1.0	2.0	6.0	25.0	136.1
$\log(q)$	2.0	1.7			0.0	0.7	1.8	3.2	4.4
λ_0	64.4	206.4			2.6	8.0	21.5	54.3	136.1
$\log(\lambda_0)$	3.2	1.4	0.6	0.8	1.4	2.3	3.2	4.1	5.0
ψ	15.9	34.7			1.3	3.1	6.8	17.4	34.5
$\log(\psi)$	1.9	1.3	0.2	0.2	0.2	1.1	1.9	2.7	3.6
APT	50.8	32.3			27.9	34.5	43.5	57.0	77.2
log(APT)	3.8	0.5			3.3	3.5	3.8	4.0	4.3
p^*	1.4	0.7			0.8	1.1	1.4	1.7	2.0

This table reports the mean, standard deviation (σ), and selected percentiles for each variable. $\sigma_{l,s}$ is the standard deviation conditional on lender-state fixed effects. $\sigma_{l,s,t}$ is the standard deviation conditional on lender-state-month fixed effects.

We now turn to the second stage of our empirical analysis, which uses the panel regressions specified in Equation (22) to estimate the sensitivity of the number of applications, application processing times, and mortgage rate spreads to the level parameter λ_0 of the mortgage demand curve, as well as the sensitivity of the mortgage rate spread to the slope parameter ψ of the mortgage demand curve. These sensitivity estimates give us insight into the empirical magnitudes of the related partial derivatives implied by our structural model, discussed in Section 2.3.

Table 2 shows the estimates for the sensitivity of the number of applications to the level parameter λ_0 of the mortgage demand curve, which is related to the partial derivative in Equation 13. We focus on specifications (4) and (5), which do not include time fixed effects.[24] Identification

[22] The sample autocorrelations of $\log(\widehat{\lambda}_0)$ and $\log(\widehat{\psi})$ within states are 0.98 and 0.99, respectively.

[23] For notational simplicity, we suppress the "hats" on the parameter estimates in this section.

[24] Appendix C presents results from additional regression model specifications and robustness tests.

in these regressions comes from variation within lender-state cells over time. At the end of this section, we discuss specifications (8) and (9), which include lender, state, and time fixed effects. We estimate that the elasticity of the number of applications to λ_0 is 0.338–0.545 across the two specifications and two weighting methods. Table 1 shows that the sample standard deviation of $\log(\lambda_0)$ within lender-state cells is 60 percent, and thus our estimates imply that a one standard deviation log-shock to λ_0 causes a 20 to 32 percent change in the number of applications.

Table 2: Estimated effect of $\log(\lambda_0)$ on the logarithm of the number of applications

	(4)	(5)	(8)	(9)
Unweighted				
Log of λ_0	0.511***	0.460***	1.240***	1.313***
Number of Observations	909,066	745,727	910,186	748,923
Adjusted R^2	0.85	0.84	0.74	0.70
Adjusted Within R^2	0.18	0.15	0.57	0.50
Weighted				
Log of λ_0	0.338***	0.545***	1.107***	1.247***
Number of Observations	41,329,108	40,630,549	41,330,137	40,633,091
Adjusted R^2	0.96	0.78	0.93	0.63
Adjusted Within R^2	0.54	0.54	0.68	0.68
Lender Fixed Effect			x	x
State Fixed Effect			x	x
Lender × State Fixed Effect	x	x		
State × Purpose × Jumbo Controls	x	x	x	x
FICO × LTV × DTI Controls	x	x	x	x
Time (Monthly) Fixed Effect			x	x
Quadratic Specification	x		x	
Lagged IV Estimate		x		x

Standard error estimates are robust to heteroskedasticity (White, 1980).

The reason that the estimated effect on the number of applications is less than the 60 percent shock to λ_0 is likely the lender's pricing power. Indeed, in response to positive demand shocks, lenders can raise prices, thereby discouraging some borrowers from applying but raising revenues per borrower. Table 3 shows the estimates for the sensitivity of the mortgage rate spread (in percentage points) to the level parameter λ_0 of the mortgage demand curve, which is related to the partial derivative in Equation 15. We continue to focus on specifications (4) and (5), estimating that the semi-elasticity of the spread to λ_0 is 0.055 to 0.134 across the two specifications and two weighting methods. The 60 percent demand shock increases spreads by 3 to 8 basis points, which deters some of the potential borrowers from applying. This effect captures both the price sensitivity of borrowers and their option to return during less busy times as well as to forgo borrowing entirely.

Table 4 shows the estimates for the sensitivity of the application processing time to the level parameter λ_0 of the mortgage demand curve, which is related to the partial derivative in Equation 14. We continue to focus on specifications (4) and (5), estimating that the elasticity of the application

Table 3: Estimated effect of $\log(\lambda_0)$ on the interest rate spread (%)

	(4)	(5)	(8)	(9)
Unweighted				
Log of λ_0	0.107^{***}	0.127^{***}	0.006^{***}	0.007^{***}
Number of Observations	909,066	745,727	910,186	748,923
Adjusted R^2	0.56	0.59	0.59	0.62
Adjusted Within R^2	0.04	0.01	0.06	0.06
Weighted				
Log of λ_0	0.055^{***}	0.134^{***}	-0.002^{***}	0.003^{***}
Number of Observations	41,329,108	40,630,549	41,330,137	40,633,091
Adjusted R^2	0.52	0.52	0.68	0.58
Adjusted Within R^2	0.07	0.07	0.11	0.11
Lender Fixed Effect			x	x
State Fixed Effect			x	x
Lender × State Fixed Effect	x	x		
State × Purpose × Jumbo Controls	x	x	x	x
FICO × LTV × DTI Controls	x	x	x	x
Time (Monthly) Fixed Effect			x	x
Quadratic Specification	x		x	
Lagged IV Estimate		x		x

Standard error estimates are robust to heteroskedasticity (White, 1980).

processing time to λ_0 is 0.103 to 0.179. This regression estimate implies that a lender experiencing a one standard deviation log-shock to λ_0 of 60 percent raises application processing times by 6 to 10 percent. Since it takes lenders, on average, 50.8 days to reach a decision (Table 1), this effect increases processing times by about 3 to 5 days. Most residential mortgages include a rate lock that guarantees the interest rate for a fixed period: McMurray and Thomson (1997) estimate an average of about 46 days, and Consumer Financial Protection Bureau (2023) documents standard 30-, 45-, and 60-day locks. Therefore, the estimated effect on processing times is economically meaningful, yet not large enough to induce systematic rate-lock violations.

Table 5 shows the estimates for the sensitivity of the mortgage rate spread to the slope parameter ψ of the mortgage demand curve, which is related to the partial derivative in Equation 16. We continue to focus on specifications (4) and (5), estimating that the semi-elasticity of the mortgage rate spread to ψ is -0.032 to -0.107. Table 1 shows that the sample standard deviation of $\log(\psi)$ within lender-state cells is 20 percent, and thus our sensitivity estimates imply that a one standard deviation log-shock to ψ decreases spreads by 0.6 to 2.1 basis points. However, because shocks to λ_0 are correlated with shocks to ψ, the total effect of a one standard deviation log-shock to λ_0, combined with a $\rho \cdot \sigma_\psi$ log-shock to ψ, would be a $\sigma_{\lambda_0} \cdot \partial p^* / \partial \log(\lambda_0) + \rho \cdot \sigma_\psi \cdot \partial p^* / \partial \log(\psi)$ change in spreads. The sample correlation between $\log(\lambda_0)$ and $\log(\psi)$ is about 0.7, and thus the total estimated change in spreads is $(0.6 \cdot (0.055) + 0.7 \cdot 0.2 \cdot (-0.107)) \approx +1.8$ to $(0.6 \cdot (0.127) + 0.7 \cdot 0.2 \cdot (-0.050)) \approx +6.9$ basis

Table 4: Estimated effect of $\log(\lambda_0)$ on the logarithm of the average processing time

	(4)	(5)	(8)	(9)
Unweighted				
Log of λ_0	0.144^{***}	0.155^{***}	0.022^{***}	0.022^{***}
Number of Observations	908,860	745,599	909,988	748,797
Adjusted R^2	0.40	0.47	0.38	0.44
Adjusted Within R^2	0.09	0.07	0.02	0.01
Weighted				
Log of λ_0	0.103^{***}	0.179^{***}	0.021^{***}	0.005^{***}
Number of Observations	41,328,476	40,630,038	41,329,512	40,632,582
Adjusted R^2	0.70	0.37	0.71	0.40
Adjusted Within R^2	0.26	0.27	0.04	0.05
Lender Fixed Effect			x	x
State Fixed Effect			x	x
Lender × State Fixed Effect	x	x		
State × Purpose × Jumbo Controls	x	x	x	x
FICO × LTV × DTI Controls	x	x	x	x
Time (Monthly) Fixed Effect			x	x
Quadratic Specification	x		x	
Lagged IV Estimate		x		x

Standard error estimates are robust to heteroskedasticity (White, 1980).

points.[25] These estimates suggest that even a two standard deviation shock to the lender-specific demand curves would account for only a fraction of the aggregate dispersion in mortgage rate spreads estimated in the literature. For example, Bhutta, Fuster and Hizmo (2020) find that identical mortgages loans, in the same market, on the same day, have a 10th to 90th percentile rate dispersion of 54 basis points for identical loans in the same market on the same day.

Overall, the empirical results are consistent with our structural model. First, the signs of the partial derivative estimates match model predictions. Second, the magnitudes of the estimated effects are plausible. In particular, the model-implied variation in mortgage pricing is in line with the total variation in mortgage pricing implied by the empirical estimates. The estimated effect on average processing times is also reasonable. Moreover, as an indication of robustness, the instrumental variable specifications (5) and (9) produce similar, and even slightly stronger, results as the ordinary least squares specifications (4) and (8), which allays potential concerns about endogeneity bias due to backward causation in our regression model.

One notable caveat to our empirical results is that identification is mostly based on temporal, rather than cross-sectional, variation. This can be seen from regression specifications 8 and 9 in Tables 2 through 4, which also include time fixed effects. We find that once fixed effects are included, the estimated effects generally remain statistically significant and have signs consistent with the

[25] This calculation uses estimates for $\partial p^*/\partial \log(\lambda_0)$ and $\partial p^*/\partial \log(\psi)$ from corresponding regression specifications.

Table 5: Estimated effect of log of ψ on the mortgage rate spread (%)

	(4)	(5)	(8)	(9)
Unweighted				
Log of ψ	-0.032^{***}	-0.050^{***}	-0.047^{***}	-0.099^{***}
Number of Observations	956,002	776,712	956,789	779,923
Adjusted R^2	0.54	0.58	0.58	0.61
Adjusted Within R^2	0.03	−0.00	0.06	0.06
Weighted				
Log of ψ	-0.107^{***}	-0.095^{***}	-0.081^{***}	-0.149^{***}
Number of Observations	41,540,204	40,832,047	41,540,913	40,834,613
Adjusted R^2	0.52	0.50	0.67	0.57
Adjusted Within R^2	0.07	0.08	0.11	0.11
Lender Fixed Effect			x	x
State Fixed Effect			x	x
Lender × State Fixed Effect	x	x		
State × Purpose × Jumbo Controls	x	x	x	x
FICO × LTV × DTI Controls	x	x	x	x
Time (Monthly) Fixed Effect			x	x
Quadratic Specification	x		x	
Lagged IV Estimate		x		x

Standard error estimates are robust to heteroskedasticity (White, 1980).

structural model, but their economic significance decreases meaningfully. This pattern suggests that our identification mostly exploits common, rather than idiosyncratic, shocks to lender-specific mortgage demand curves, which induce variation *within* lenders over time. This finding is not surprising because lender-specific demand curves have a strong common component, driven by monetary policy and other macroeconomic factors, while the effects of idiosyncratic shocks are likely attenuated by the mechanism that borrowers can avoid higher interest rates and longer waiting times by shopping across and switching between lenders.

6 Conclusion

In this paper, we introduce and empirically test a novel structural model of the individual mortgage lender with pricing power in which mortgage applications arrive and are processed at stochastic time intervals on a first-come-first-served basis. Leveraging loan-level information in the confidential HMDA dataset, we investigate how application volumes, mortgage rate spreads, and processing times change in response to fluctuations in lender-specific mortgage demand.

Our empirical results confirm several key predictions of our structural model. First, increases in demand significantly raise mortgage rate spreads, application volumes, and application processing times. A one standard deviation increase in mortgage demand leads to 3 to 8 basis points higher spreads, approximately 20 to 32 percent higher application volumes, and 3 to 5 days longer processing

times. These results highlight the trade-offs faced by mortgage lenders in periods of elevated demand, in which lenders balance mortgage rate increases, higher application volumes, and the risk of borrower dissatisfaction due to prolonged delays.

Additionally, our analysis highlights the importance of the interest rate sensitivity of mortgage demand, which mitigates the first-order effect of demand shocks on mortgage rates. We observe a positive empirical correlation between the level and interest rate sensitivity of mortgage demand. Because higher interest rate sensitivity puts downward pressure on mortgage rates, it partially offsets the upward pressure on mortgage rates from positive demand shocks.

Combining queuing theory, structural modeling, and econometric analysis, our work gives novel theoretical and empirical insights into equilibrium mortgage market dynamics as well as the pricing decisions of individual mortgage lenders in response to demand shocks. These insights not only deepen the economic comprehension of capacity constraints but also have practical implications for lenders and policymakers seeking to improve market efficiency and consumer outcomes in the residential mortgage market.

References

Aliaga-Dıaz, Roger, and Marıa Pıa Olivero. 2007. "Macroeconomic Implications of Market Power in Banking." Citeseer.

Angrist, Joshua D, and Alan B Krueger. 2001. "Instrumental variables and the search for identification: From supply and demand to natural experiments." *Journal of Economic perspectives*, 15(4): 69–85.

Azriel, David, Paul D Feigin, and Avishai Mandelbaum. 2019. "Erlang-S: A data-based model of servers in queueing networks." *Management Science*, 65(10): 4607–4635.

Bhutta, Neil, Andreas Fuster, and Aurel Hizmo. 2020. "Paying too much? Price dispersion in the US mortgage market." Federal Reserve Board.

Burdett, Kenneth, and Kenneth L Judd. 1983. "Equilibrium price dispersion." *Econometrica: Journal of the Econometric Society*, 955–969.

Chen, Hui, Scott Joslin, and Sophie Xiaoyan Ni. 2019. "Demand for crash insurance, intermediary constraints, and risk premia in financial markets." *The Review of Financial Studies*, 32(1): 228–265.

Consumer Financial Protection Bureau. 2023. "What's a lock-in or a rate lock on a mortgage?" Accessed: March 13, 2025.

Cont, Rama, Sasha Stoikov, and Ram Talreja. 2010. "A Stochastic Model for Order Book Dynamics." *Operations Research*, 58(3): 549–563.

Cowdrey, Kevin WG, Jaco de Lange, Reza Malekian, Johan Wanneburg, and Arun Cyril Jose. 2018. "Applying queueing theory for the optimization of a banking model." *Journal of Internet Technology*, 19(2): 381–389.

Eeckhout, Jan, and Philipp Kircher. 2010. "Sorting and decentralized price competition." *Econometrica*, 78(2): 539–574.

Erlang, Agner Krarup. 1909. "The Theory of Probabilities and Telephone Conversations." *Nyt Tidsskrift for Matematik*, 20: 33–39.

Fagnart, Jean-François, Omar Licandro, and Henri R Sneessens. 1997. "Capacity utilization and market power." *Journal of Economic Dynamics and Control*, 22(1): 123–140.

Fuster, Andreas, Aurel Hizmo, Lauren Lambie-Hanson, James Vickery, and Paul S Willen. 2021. "How resilient is mortgage credit supply? Evidence from the COVID-19 pandemic." National Bureau of Economic Research.

Fuster, Andreas, Stephanie H Lo, and Paul S Willen. 2024. "The time-varying price of financial intermediation in the mortgage market." *The Journal of Finance*, 79(4): 2553–2602.

Green, Lawrence V. 2006. "Queueing Analysis in Healthcare." In *Patient Flow: Reducing Delay in Healthcare Delivery*. 281–307. Springer.

Gron, Anne. 1994. "Capacity constraints and cycles in property-casualty insurance markets." *The RAND Journal of Economics*, 110–127.

Kuhn, Florian, and Chacko George. 2019. "Business cycle implications of capacity constraints under demand shocks." *Review of Economic Dynamics*, 32: 94–121.

Levy, Anat, and Hanoch Levy. 1991. "Lock and no-lock mortgage plans: is it only a matter of risk shifting?" *Operations research letters*, 10(4): 233–240.

McMurray, John, and Thomas Thomson. 1997. "Determinants of the closing probability of residential mortgage applications." *Journal of Real Estate Research*, 14(1): 55–64.

Menezes, Flavio M., and John Quiggin. 2022. "Market power amplifies the price effects of demand shocks." *Economics Letters*, 221: 110908.

Naor, Pinhas. 1969. "The regulation of queue size by levying tolls." *Econometrica: journal of the Econometric Society*, 15–24.

Peterson, Larry, and Bruce Davie. 2024. "Queuing Disciplines." In *Computer Networks: A Systems Approach*. Chapter 6.2. No Publisher. Accessed March 24, 2025.

Scharfstein, David, and Adi Sunderam. 2016. "Market power in mortgage lending and the transmission of monetary policy." *Unpublished working paper. Harvard University*, 2.

White, Halbert. 1980. "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity." *Econometrica: journal of the Econometric Society*, 817–838.

Wolinsky, Asher. 1986. "True monopolistic competition as a result of imperfect information." *The Quarterly Journal of Economics*, 101(3): 493–511.

A Alternative cost function specification

We explore how a variation on the lender's total cost function affects the baseline results. Specifically, keeping the baseline model in Section 2 unchanged, we use an alternative (logarithmic, rather than quadratic) cost function specification, which endogenously generates a preference for capacity slack.

A.1 Alternative cost function

We model the lender's cost function over the planning horizon as depending on the expected arrival rate λ relative to service capacity μ. We specify the expected total cost function such that it captures two essential features of queuing dynamics: **convexity of costs** in volume and an **asymptotic cost penalty** as utilization approaches capacity:

$$TC(\lambda, \mu) = d\,\lambda - c\ln\big(\mu - \lambda\big), \tag{23}$$

where $d > 0$ is the baseline marginal cost of mortgage origination, and $c > 0$ is the sensitivity of costs to capacity constraints. As λ approaches μ, the term $\ln(\mu - \lambda)$ diverges to negative infinity, so $TC(\lambda, \mu)$ imposes an infinite penalty on a lender operating with no slack. This captures the severe opportunity costs, such as overtime wages, error correction, and reputation loss, incurred when operating arbitrarily close to capacity and where customers face extremely long expected response times.

Similar to the quadratic cost function used in Equation (5), the logarithmic cost function implies convex, increasing marginal costs in λ. Specifically, the marginal cost of production is

$$\frac{\partial TC}{\partial \lambda} = MC(\lambda) = d + \frac{c}{\mu - \lambda}, \tag{24}$$

and the second derivative is

$$\frac{\partial^2 TC}{\partial \lambda^2} = \frac{c}{(\mu - \lambda)^2}, \tag{25}$$

which is positive for $0 < \lambda < \mu$.

This marginal cost function is closely linked to system congestion, which can be seen by substituting for average processing time (APT), defined earlier in Equation (3). This yields

$$MC(\lambda) = d + c \cdot \text{APT}, \tag{26}$$

which implies a direct link between the lender's pricing problem and average processing time, the canonical measure of congestion in the queuing literature.

A.2 The lender's profit maximization under the alternative cost function

We solve the lender's profit maximization problem described in Section 2.1 under the alternative cost function specified in Equation (23). Substituting the logarithmic cost function yields the profit function:

$$\Pi(p) = p \cdot \lambda(p) - \Big[d \cdot \lambda(p) - c\ln\big(\mu - \lambda(p)\big)\Big].$$

Using this profit function and the notation introduced in Section 2, we derive the closed-form solution as follows:

$$\text{(0) Arrival rate:} \qquad \lambda(p) = \lambda_0 - \psi p, \tag{27}$$

$$\text{(0') Response time:} \qquad \mathrm{APT}(p) = \frac{1}{\mu - \lambda(p)}, \tag{28}$$

$$\text{(1) Total cost function:} \qquad TC(\lambda) = d\,\lambda - c\ln\big(\mu - \lambda\big), \tag{29}$$

$$\text{(1') Production domain:} \qquad 0 < \lambda < \mu, \tag{30}$$

$$C(p) = d\,\lambda(p) - c\ln\big(\mu - \lambda(p)\big), \tag{31}$$

$$\text{(2) Profit function:} \qquad \Pi(p) = p\,\lambda(p) - C\big(\lambda(p)\big) \tag{32}$$

$$= c\ln\big(\mu - \lambda(p)\big) - \big(d - p\big)\,\lambda(p), \tag{33}$$

$$\text{(3) First-order condition:} \qquad \frac{\partial \Pi}{\partial p} = \frac{c\psi}{\mu - \lambda(p)} + \lambda_0 + \psi d - 2\psi p, \tag{34}$$

$$\frac{\partial \Pi}{\partial p} = 0 \iff 2\psi^2 p^2 - \psi(\lambda_0 + \psi d - 2(\mu - \lambda_0))\,p \tag{35}$$

$$- \Big[(\mu - \lambda_0)\big(\lambda_0 + \psi d\big) + c\psi\Big] = 0, \tag{36}$$

$$\text{(4) Optimal price } p^* = p^*_+ : \qquad p^* = \frac{3\lambda_0 - 2\mu + \psi d}{4\psi} \tag{37}$$

$$+ \frac{\sqrt{\big(3\lambda_0 - 2\mu + \psi d\big)^2 + 8\Big[(\mu - \lambda_0)\big(\lambda_0 + \psi d\big) + c\psi\Big]}}{4\psi} \tag{38}$$

$$\text{(5) Optimal arrival rate:} \qquad \lambda(p^*) = \lambda_0 - \psi p^*_+, \tag{39}$$

$$\text{(6) Restriction for } \lambda(p^*) > 0 : \qquad \lambda_0 > \psi\left(d + \frac{c}{\mu}\right) \tag{40}$$

$$\implies \lambda(p^*_+) > 0, \tag{41}$$

$$\text{(7) Optimal response time and finiteness:} \qquad 0 < \lambda(p^*_+) < \lambda_0 < \mu, \tag{42}$$

$$\mu - \lambda(p^*_+) > \mu - \lambda_0 > 0, \tag{43}$$

$$\mathrm{APT}(p^*_+) = \frac{1}{\mu - \lambda(p^*_+)} \;\in\; \left(0,\; \frac{1}{\mu - \lambda_0}\right) \;<\; \infty. \tag{44}$$

Lemma 1 implies that the optimal arrival rate $\lambda(p^*)$ is strictly positive if and only if the parameter restriction in Equation (41) holds. Notably, like the restriction $\lambda(p^*) > 0 \iff \lambda_0 \geq \psi(c + d + e)$ in the body of the paper, the restriction $\lambda_0 > \psi\left(d + \frac{c}{\mu}\right)$ is an assumption about the marginal cost at zero volume being less than the price at zero volume. Failing to satisfy this condition in either model implies that it is not profitable to produce the first unit, and the lender exits the mortgage market.

Lemma 1. *Suppose $\psi > 0$ and $\lambda_0 < \mu$. Let p^* be the larger root of the first–order condition* (35)*, given by* (38)*, and define $\lambda(p^*) = \lambda_0 - \psi p^*$. Then*

$$\lambda(p^*) > 0 \quad \Longleftrightarrow \quad \lambda_0 > \psi\left(d + \frac{c}{\mu}\right),$$

with $\lambda(p^*) = 0$ exactly at $\lambda_0 = \psi\left(d + \frac{c}{\mu}\right)$, and $\lambda(p^*) < 0$ for $\lambda_0 < \psi\left(d + \frac{c}{\mu}\right)$.

Proof. By construction $\lambda(p) = \lambda_0 - \psi p$, so

$$\lambda(p^*) = \lambda_0 - \psi p^*.$$

Solving the first–order condition for p and taking the larger root gives

$$p^* = \frac{3\lambda_0 - 2\mu + \psi d}{4\psi} + \frac{\sqrt{\Delta}}{4\psi},$$

where

$$\Delta = \big(3\lambda_0 - 2\mu + \psi d\big)^2 + 8\Big[(\mu - \lambda_0)\big(\lambda_0 + \psi d\big) + c\psi\Big].$$

Substituting this expression into $\lambda(p^*) = \lambda_0 - \psi p^*$ and simplifying yields

$$\lambda(p^*) = \frac{\lambda_0 + 2\mu - d\psi - \sqrt{\Delta}}{4}. \tag{45}$$

To analyze the sign of $\lambda(p^*)$, define the deviation of λ_0 from the candidate threshold as

$$k \;\equiv\; \lambda_0 - \psi\left(d + \frac{c}{\mu}\right),$$

so that

$$\lambda_0 = \psi\left(d + \frac{c}{\mu}\right) + k, \qquad k \in \mathbb{R}.$$

This is just a reparameterization of λ_0; at this stage we do not assume any sign for k.

Substituting this expression for λ_0 into Equation!(45) and simplifying gives

$$\lambda(p^*) = \frac{S - \sqrt{T}}{4\mu},$$

where

$$S \equiv c\psi + k\mu + 2\mu^2,$$

$$T \equiv \Delta\big(\lambda_0 = \psi(d + \tfrac{c}{\mu}) + k\big) = c^2\psi^2 + 2ck\mu\psi + 4c\mu^2\psi + k^2\mu^2 - 4k\mu^3 + 4\mu^4.$$

By expanding and simplifying, we obtain that

$$S^2 - T = 8k\mu^3.$$[26]

[26] Expanding $S = c\psi + k\mu + 2\mu^2$ gives $S^2 = c^2\psi^2 + k^2\mu^2 + 4\mu^4 + 2ck\mu\psi + 4c\mu^2\psi + 4k\mu^3$. Subtracting $T = c^2\psi^2 + 2ck\mu\psi + 4c\mu^2\psi + k^2\mu^2 - 4k\mu^3 + 4\mu^4$ term by term cancels all but the $k\mu^3$ terms, leaving $S^2 - T = 8k\mu^3$.

Because $\mu > 0$, the sign of $S^2 - T$ is exactly the sign of k:

$$S^2 > T \iff k > 0, \qquad S^2 = T \iff k = 0, \qquad S^2 < T \iff k < 0.$$

Now note that $\sqrt{T} \geq 0$ by definition. Consider three cases:

(i) $k = 0$. Then $S^2 = T$, so $\sqrt{T} = |S|$. In this case $\lambda(p^*) = (S - \sqrt{T})/(4\mu) = 0$.

(ii) $k > 0$. Then $S^2 > T$, so $|S| > \sqrt{T}$. In the admissible parameter region $\mu > 0$, $c > 0$, and $\lambda_0 < \mu$, one can check that $S > 0$, so $S > \sqrt{T}$ and hence $S - \sqrt{T} > 0$. Because $4\mu > 0$, it follows that $\lambda(p^*) > 0$.

(iii) $k < 0$. Then $S^2 < T$, so $|S| < \sqrt{T}$. In all cases this implies $S - \sqrt{T} < 0$, and therefore $\lambda(p^*) < 0$.

Putting these cases together,

$$\lambda(p^*) > 0 \iff k > 0, \qquad \lambda(p^*) = 0 \iff k = 0, \qquad \lambda(p^*) < 0 \iff k < 0.$$

Recalling that $k = \lambda_0 - \psi\big(d + c/\mu\big)$, we obtain

$$\lambda(p^*) > 0 \quad \iff \quad \lambda_0 > \psi\left(d + \frac{c}{\mu}\right),$$

with equality and strict negativity corresponding to $\lambda_0 = \psi\left(d + \frac{c}{\mu}\right)$ and $\lambda_0 < \psi\left(d + \frac{c}{\mu}\right)$, respectively. This proves the lemma. □

B Complete list of partial derivatives

This section provides a complete list of partial derivatives implied by the structural model introduced in Section 2.

B.1 Price sensitivity to parameters

$$\frac{\partial p^*}{\partial \lambda_0} = \frac{2c\psi + 1}{2\psi\left(c\psi + 1\right)} \tag{46}$$

The sign of $\frac{\partial p^*}{\partial \lambda_0}$ is positive.

$$\frac{\partial p^*}{\partial d} = \frac{1}{2\left(c\psi + 1\right)} \tag{47}$$

The sign of $\frac{\partial p^*}{\partial d}$ is positive.

$$\frac{\partial p^*}{\partial c} = \frac{\lambda_0 + 1 - \psi(d+e)}{2\left(c\psi + 1\right)^2} \tag{48}$$

Under the stability restriction $\lambda_0 \geq \psi(c+d+e)$, the sign of $\frac{\partial p^*}{\partial c}$ is positive.

$$\frac{\partial p^*}{\partial e} = \frac{1}{2\left(c\psi + 1\right)} \tag{49}$$

The sign of $\frac{\partial p^*}{\partial e}$ is positive.

$$\frac{\partial p^*}{\partial \psi} = -\frac{2c^2\lambda_0\psi^2 + c^2\psi^2 + cd\psi^2 + ce\psi^2 + 2c\lambda_0\psi + \lambda_0}{2\psi^2\left(c\psi + 1\right)^2} \tag{50}$$

The sign of $\frac{\partial p^*}{\partial \psi}$ is negative.

B.2 Expected quantity sensitivity to structural parameters

Because $\rho = \lambda/\mu$, it follows that $\rho(p^*) = \lambda(p^*)/\mu$, and therefore the signs of the derivatives of ρ^* are the same as those of $\lambda(p^*)$.

$$\frac{\partial \lambda(p^*)}{\partial \lambda_0} = \frac{1}{2\left(c\psi + 1\right)} \tag{51}$$

The sign of $\frac{\partial \lambda(p^*)}{\partial \lambda_0}$ is positive.

$$\frac{\partial \lambda(p^*)}{\partial d} = -\frac{\psi}{2\left(c\psi + 1\right)} \tag{52}$$

The sign of $\frac{\partial \lambda(p^*)}{\partial d}$ is negative.

$$\frac{\partial \lambda(p^*)}{\partial c} = -\frac{\psi\left(\lambda_0 + 1 - \psi(d+e)\right)}{2\left(c\psi + 1\right)^2} \tag{53}$$

Under $\lambda_0 \geq \psi(c+d+e)$, the sign of $\frac{\partial \lambda(p^*)}{\partial c}$ is negative.

$$\frac{\partial \lambda(p^*)}{\partial e} = -\frac{\psi}{2\left(c\psi + 1\right)} \tag{54}$$

The sign of $\frac{\partial \lambda(p^*)}{\partial e}$ is negative.

$$\frac{\partial \lambda(p^*)}{\partial \psi} = -\frac{c\lambda_0 + c + d + e}{2\left(c\psi + 1\right)^2} \tag{55}$$

The sign of $\frac{\partial \lambda(p^*)}{\partial \psi}$ is negative.

B.3 Average processing time sensitivity to structural parameters

$$\frac{\partial \mathrm{APT}(p^*)}{\partial \lambda_0} = \frac{2\left(c\psi + 1\right)}{\left(2c\mu\psi + (c + d + e)\psi - \lambda_0 + 2\mu\right)^2} \tag{56}$$

The sign of $\frac{\partial \mathrm{APT}(p^*)}{\partial \lambda_0}$ is positive.

$$\frac{\partial \mathrm{APT}(p^*)}{\partial d} = -\frac{2\psi\left(c\psi + 1\right)}{\left(2c\mu\psi + (c + d + e)\psi - \lambda_0 + 2\mu\right)^2} \tag{57}$$

The sign of $\frac{\partial \mathrm{APT}(p^*)}{\partial d}$ is negative.

$$\frac{\partial \mathrm{APT}(p^*)}{\partial c} = -\frac{2\psi\left(\lambda_0 + 1 - \psi(d + e)\right)}{\left(2c\mu\psi + (c + d + e)\psi - \lambda_0 + 2\mu\right)^2} \tag{58}$$

Under $\lambda_0 \geq \psi(c + d + e)$, the sign of $\frac{\partial \mathrm{APT}(p^*)}{\partial c}$ is negative.

$$\frac{\partial \mathrm{APT}(p^*)}{\partial e} = -\frac{2\psi\left(c\psi + 1\right)}{\left(2c\mu\psi + (c + d + e)\psi - \lambda_0 + 2\mu\right)^2} \tag{59}$$

The sign of $\frac{\partial \mathrm{APT}(p^*)}{\partial e}$ is negative.

$$\frac{\partial \mathrm{APT}(p^*)}{\partial \psi} = -\frac{2\left(c\lambda_0 + c + d + e\right)}{\left(2c\mu\psi + (c + d + e)\psi - \lambda_0 + 2\mu\right)^2} \tag{60}$$

The sign of $\frac{\partial \mathrm{APT}(p^*)}{\partial \psi}$ is negative.

C Complete set of regression estimates

Table C.1: Estimated effect of $\log(\lambda_0)$ on the logarithm of the number of applications

	(1)	(2)	(3)	(6)	(7)
Unweighted					
Log of λ_0	0.917^{***}	0.387^{***}	0.317^{***}	1.161^{***}	1.158^{***}
Number of Observations	933,315	932,678	909,066	933,307	910,186
Adjusted R^2	0.57	0.83	0.84	0.69	0.69
Adjusted Within R^2	0.57	0.10	0.14	0.48	0.48
Weighted					
Log of λ_0	1.063^{***}	0.949^{***}	0.482^{***}	1.215^{***}	1.204^{***}
Number of Observations	41,405,633	41,405,052	41,329,108	41,405,630	41,330,137
Adjusted R^2	0.88	0.93	0.95	0.92	0.93
Adjusted Within R^2	0.88	0.26	0.52	0.65	0.67
Lender Fixed Effect				x	x
State Fixed Effect				x	x
Lender × State Fixed Effect		x	x		
State × Purpose × Jumbo Controls			x		x
FICO × LTV × DTI Controls			x		x
Time (Monthly) Fixed Effect				x	x
Quadratic Specification			x		
Lagged IV Estimate				x	

Standard error estimates are robust to heteroskedasticity (White, 1980).

Table C.2: Estimated effect of $\log(\lambda_0)$ on the mortgage rate spread (%)

	(1)	(2)	(3)	(6)	(7)
Unweighted					
Log of λ_0	0.008^{***}	0.074^{***}	0.087^{***}	0.009^{***}	0.007^{***}
Number of Observations	933,315	932,678	909,066	933,307	910,186
Adjusted R^2	0.00	0.56	0.56	0.57	0.59
Adjusted Within R^2	0.00	0.01	0.04	0.00	0.06
Weighted					
Log of λ_0	-0.024^{***}	0.119^{***}	0.059^{***}	-0.003^{***}	0.000
Number of Observations	41,405,633	41,405,052	41,329,108	41,405,630	41,330,137
Adjusted R^2	0.01	0.50	0.52	0.64	0.68
Adjusted Within R^2	0.01	0.02	0.07	0.00	0.11
Lender Fixed Effect				x	x
State Fixed Effect				x	x
Lender × State Fixed Effect		x	x		
State × Purpose × Jumbo Controls			x		x
FICO × LTV × DTI Controls			x		x
Time (Monthly) Fixed Effect				x	x
Quadratic Specification			x		
Lagged IV Estimate				x	

Standard error estimates are robust to heteroskedasticity.

Table C.3: Estimated effect of $\log(\lambda_0)$ on the logarithm of average processing time

	(1)	(2)	(3)	(6)	(7)
Unweighted					
Log of λ_0	0.034^{***}	0.147^{***}	0.108^{***}	0.022^{***}	0.022^{***}
Number of Observations	933,093	932,444	908,860	933,085	909,988
Adjusted R^2	0.01	0.36	0.39	0.36	0.38
Adjusted Within R^2	0.01	0.05	0.09	0.00	0.02
Weighted					
Log of λ_0	-0.020^{***}	0.242^{***}	0.114^{***}	0.004^{***}	0.008^{***}
Number of Observations	41,404,976	41,404,385	41,328,476	41,404,973	41,329,512
Adjusted R^2	0.01	0.64	0.70	0.70	0.71
Adjusted Within R^2	0.01	0.11	0.26	0.00	0.04
Lender Fixed Effect				x	x
State Fixed Effect				x	x
Lender × State Fixed Effect		x	x		
State × Purpose × Jumbo Controls			x		x
FICO × LTV × DTI Controls			x		x
Time (Monthly) Fixed Effect				x	x
Quadratic Specification			x		
Lagged IV Estimate				x	

Standard error estimates are robust to heteroskedasticity.

Table C.4: Estimated effect of $\log(\psi)$ on the mortgage rate spread (%)

	(1)	(2)	(3)	(6)	(7)
Unweighted					
Log of ψ	-0.039^{***}	-0.037^{***}	-0.032^{***}	-0.053^{***}	-0.047^{***}
Number of Observations	982,212	981,926	956,002	982,204	956,789
Adjusted R^2	0.01	0.54	0.54	0.56	0.58
Adjusted Within R^2	0.01	0.00	0.03	0.00	0.06
Weighted					
Log of ψ	-0.036^{***}	-0.088^{***}	-0.113^{***}	-0.111^{***}	-0.085^{***}
Number of Observations	41,619,150	41,618,908	41,540,204	41,619,147	41,540,913
Adjusted R^2	0.02	0.49	0.52	0.64	0.67
Adjusted Within R^2	0.02	0.00	0.07	0.00	0.11
Lender Fixed Effect				x	x
State Fixed Effect				x	x
Lender × State Fixed Effect		x	x		
State × Purpose × Jumbo Controls			x		x
FICO × LTV × DTI Controls			x		x
Time (Monthly) Fixed Effect				x	x
Quadratic Specification			x		
Lagged IV Estimate				x	

Standard error estimates are robust to heteroskedasticity.