Spatial Sorting and the Geography of Innovation

Santiago Truffa*

March 2026

Abstract

The scarce input in the production of new ideas is human cognitive time. Because complex problem solving depends on communication, coordination, and tacit knowledge exchange, the allocation of this time is inherently spatial — and the geographic organization of talent is therefore a first-order determinant of innovation. This paper studies how technologies that increase effective cognitive capacity, most notably artificial intelligence, reshape the spatial equilibrium of innovation. I develop and estimate a spatial equilibrium model in which heterogeneous firms and workers sort across cities while competing for scarce talent and urban space. Complementarities between worker skill and firm problem complexity generate self-reinforcing composition effects, and the framework recovers the full distributions of firm and worker types across cities, allowing technological change to affect the upper tail of economic activity rather than only average productivity. Estimated on a panel of the 50 largest U.S. metropolitan areas from 2019 to 2023, the model matches key cross-city patterns in wages, rents, and sorting. Counterfactuals show that increases in cognitive efficiency primarily reallocate innovative activity rather than scale it uniformly: top workers and complex firms concentrate further in cities with dense complementary networks. Technological progress therefore raises aggregate productivity while simultaneously intensifying the spatial concentration of frontier innovation.

JEL Codes: R12, J24, O33, D83, R31.

Keywords: Spatial equilibrium, innovation, artificial intelligence, sorting, agglomeration, superstar cities.

*Universidad Adolfo Ibáñez. This paper is based on Chapter 2 of my PhD dissertation titled "Urban Connectivity." I am grateful for helpful comments from seminar participants. All errors are my own.

1 Introduction

The scarce input in the production of new ideas is high-capacity human cognitive time. Because complex problem solving depends on communication, coordination, and tacit knowledge exchange, the allocation of this time is inherently spatial: proximity to complementary firms, collaborators, and knowledge networks raises the productivity of skilled workers. The geographic organization of talent is therefore a first-order determinant of breakthrough innovation — and a small number of metropolitan areas account for a disproportionate share of frontier activity, high-skill employment, and complex problem solving.[1]

The forces sustaining these superstar innovation hubs are qualitatively well understood, but the quantitative mechanisms through which technological change reshapes their geography are not. The arrival of general-purpose technologies has historically transformed which places lead economic activity. Widespread artificial intelligence tools — which reduce the time cost of search, drafting, coding, and information processing — constitute a new episode in this long-run relationship between technology and economic geography.

This paper studies how technologies that increase effective cognitive capacity reshape the spatial equilibrium of innovation. A central modeling choice is to interpret AI not as a simple automation or factor-augmenting shock, but as a technology that raises the effective productivity of scarce cognitive time: by reducing overhead in routine tasks, AI allows skilled workers to engage with more complex problems within a fixed time budget, increasing the effective supply of high-capacity cognitive labor without adding workers.[2] This distinction matters for spatial equilibrium. When talent remains scarce but becomes more productive, the returns to skill and to high-quality local matches can rise simultaneously, potentially reinforcing the pull of existing innovation hubs rather than allowing activity to diffuse.

To study this question, I develop a spatial equilibrium model in which heterogeneous firms and workers sort across cities while competing in two bidding markets. Firms differ in problem complexity, workers differ in skill, and cities differ in amenities, housing supply, and connectivity. The production structure draws on the knowledge-hierarchy model of Garicano (2000), in which workers solve problems up to their competence and escalate harder ones at a communication cost. This directly delivers two features the model requires: supermodularity

[1] Innovative activity is highly geographically concentrated. The productivity of top inventors rises when they locate in larger technology clusters (Moretti, 2021; Kerr, 2010), and a small number of metros account for the bulk of patents, venture capital, and high-growth startups (Kerr and Robert-Nicoud, 2020). A large literature attributes this to agglomeration forces — labor pooling, input sharing, and knowledge spillovers (Marshall, 1920; Glaeser et al., 1992; Duranton and Puga, 2004) — and to the role of localized human capital in sustaining urban innovation dynamics (Moretti, 2012).

[2] A growing empirical literature supports this interpretation. AI tools raise productivity in knowledge-intensive tasks (Brynjolfsson et al., 2023; Noy and Zhang, 2023; Peng et al., 2023), with gains concentrated among higher-skill workers and more complex task environments (Dell'Acqua et al., 2023).

in match surplus — more complex firms benefit disproportionately from high-skill workers because the problems they must escalate are harder — and communication costs as a natural spatial primitive, since cities differ in how cheaply agents can exchange tacit knowledge (Storper and Venables, 2004). The production function $V_i(m, z) = m/[h_i(1-z)]$ is a reduced form of this logic, where h_i is city-specific knowledge-transmission cost and $(1-z)^{-1}$ captures the escalation intensity of a firm at complexity z.

Scarcity operates on both sides of the market: talent is limited and urban space is limited. Firms compete for workers in labor markets; workers compete for access through housing markets. These two bidding processes generate the model's central mechanism — *reciprocal composition feedback*. Firm payoffs depend on the expected distribution of workers in a city; worker payoffs depend on the expected distribution of firms. Location choices therefore change the very distributions that determine those payoffs, making city productivity endogenous to the firms and workers who locate there. When complementarities are strong, small differences in fundamentals amplify into large differences in equilibrium outcomes. Housing markets provide a countervailing push force: as high-skill workers concentrate in a location, rising rents limit access and moderate sorting (Hsieh and Moretti, 2019; Saiz, 2010; Gyourko et al., 2013).

A key feature of the framework is that equilibrium determines the *full city-by-type distributions* of firms and workers rather than average productivity or skill levels. This matters because the geography of innovation is a tail phenomenon — it is determined by where the most skilled workers sort, where the most complex problems are solved, and where frontier activity concentrates. The model belongs to the quantitative spatial equilibrium tradition (Rosen, 1979; Roback, 1982; Allen and Arkolakis, 2022; Redding, 2022) but departs from it by centering the equilibrium problem on the production of ideas and by recovering full type distributions rather than aggregate masses.

I estimate the model on a balanced panel of the 50 largest U.S. metropolitan areas from 2019 to 2023, targeting city-year moments on wages, rents, worker and firm mass shares, type composition, and sorting patterns. Identification exploits cross-city variation in sorting and prices, differential exposure to AI-related technological change, and variation in housing supply constraints.

Three results organize the findings. First, the model reproduces key cross-city patterns in wages, rents, and population allocation. Second, strong concentration in top-tail activity emerges endogenously from bidding, scarcity, and complementarity — not from imposed heavy-tailed assumptions on city productivity. Third, technology shocks primarily *reallocate* innovative activity rather than scale it uniformly: cities with denser complementary worker–firm networks attract disproportionate re-sorting of top workers and complex firms, with

Silicon Valley capturing the largest innovation-share gains even under moderate diffusion scenarios. Technological progress therefore raises aggregate productivity while simultaneously intensifying the geographic concentration of frontier innovation.

2 Related Literature

Quantitative spatial equilibrium. The model belongs to the Rosen–Roback tradition (Rosen, 1979; Roback, 1982), extended by recent quantitative work providing tools for welfare analysis across city systems (Allen and Arkolakis, 2022; Redding, 2022). Heterogeneous-worker extensions study how sorting amplifies wage inequality across cities (Diamond, 2016; Baum-Snow and Pavan, 2012; Moretti, 2012). This paper departs from the tradition by centering the equilibrium on full type distributions — the natural object for studying innovation concentration as a tail phenomenon (Eeckhout et al., 2014; Gabaix, 1999; Eeckhout, 2004).

Agglomeration and superstar cities. Knowledge spillovers are geographically localized (Jaffe et al., 1993; Audretsch and Feldman, 1996; Lucas, 1988), and cities with initial human capital concentrations grow faster (Glaeser et al., 1992). The paper is most directly related to the superstar-city literature: cities with inelastic housing supply attract disproportionate high-skill demand (Gyourko et al., 2013); skill-biased technological change has been spatially concentrated (Moretti, 2012; Diamond, 2016); market concentration has risen alongside superstar firms (Autor et al., 2020); and top inventors are causally more productive in larger clusters (Moretti, 2021). This paper provides an equilibrium framework in which these patterns arise endogenously from composition feedback and bidding competition.

Two-sided sorting. Worker sorting and firm sorting have largely been studied in parallel. On the worker side, a substantial share of the urban wage premium reflects skill sorting rather than a pure agglomeration externality (Moretti, 2004; Diamond, 2016; Baum-Snow and Pavan, 2012). On the firm side, firm sorting amplifies city-size wage premia beyond what worker composition alone predicts (Gaubert, 2018), and matched employer-employee data confirm that both sides contribute to agglomeration premia (Combes et al., 2008). This paper embeds both into a unified equilibrium with joint composition feedbacks.

The production structure follows Garicano (2000) and yield supermodular match surplus in Garicano (2000) and its extensions. Garicano and Rossi-Hansberg (2006) showed that cheaper communication concentrates talent at the top of hierarchies a within-organization analogue of the spatial concentration result here. Antràs et al. (2006) extended this to offshoring, and Garicano and Hubbard (2007) provided empirical validation using law firm

data. Garicano and Rayo (2016) showed that the marginal product of senior problem-solvers rises with problem quality, implying that AI tools which filter routine tasks substantially raise high-skill returns. Assortative matching under supermodularity follows Becker (1973); frictions preserve the positive relationship (Shimer and Smith, 2000); and structural identification of sorting builds on Eeckhout and Kircher (2011) and Lise and Postel-Vinay (2020). Labor and housing markets are cleared via first-price Poisson auctions (Milgrom and Weber, 1982; Moen, 1997; Shimer, 2007), which yield closed-form win probabilities and link market tightness directly to informational rents.

The closest precursor to the full framework is Davis and Dingel (2019), who build a system-of-cities model in which costly idea exchange is the agglomeration force and skill sorting emerges endogenously from learning complementarities — establishing that knowledge-transmission costs, rather than amenities or input prices, can be the primitive spatial force. Gokan et al. (2019) use Garicano's communication-cost logic to study how trade costs shape firms' organizational form across space, but do not study the joint distribution of worker and firm types. No existing quantitative spatial model embeds the Garicano hierarchy to recover the full city-level distributions of heterogeneous workers *and* firms jointly.

Housing and spatial misallocation. Geographical constraints predict housing inelasticity across metros (Saiz, 2010), and restrictive zoning has substantially reduced aggregate output by blocking workers from high-productivity cities (Hsieh and Moretti, 2019; Glaeser and Gyourko, 2018). The model endogenizes who is priced out of productive cities, generating equilibria with high wages, high rents, and incomplete sorting consistent with Diamond (2016) and Gyourko et al. (2013), and capturing declining geographic mobility (Molloy et al., 2011; Kaplan and Schulhofer-Wohl, 2017).

Artificial intelligence and the organization of work. Generative AI raises productivity in knowledge-intensive tasks, with gains concentrated among higher-skill workers (Brynjolfsson et al., 2023; Noy and Zhang, 2023; Dell'Acqua et al., 2023; Peng et al., 2023). AI primarily lowers the cost of prediction, benefiting workers with complementary rather than substitutable tasks (Agrawal et al., 2018). Whether this complements or substitutes for human labor in the aggregate depends on task structure (Acemoglu, 2024; Acemoglu and Restrepo, 2018; Autor, 2022); this paper takes a complementarity view and adds a spatial dimension absent from aggregate models. AI adoption is spatially uneven and concentrated in high-skill urban areas (Bessen and Cockburn, 2024; Hampole et al., 2025), supporting the differential-exposure specification. Whether AI ultimately concentrates or disperses activity depends on whether local knowledge complementarities dominate substitution for coloca-

tion (Glaeser et al., 2014; Dingel and Neiman, 2020); the counterfactuals below show that complementarities dominate.

Contribution. Three elements distinguish this paper. First, it provides a distributional spatial equilibrium recovering full city-level type distributions for firms and workers — essential for studying innovation as a tail phenomenon. Second, it embeds AI as a spatial reallocation shock through cognitive-time efficiency, connecting micro evidence to aggregate spatial outcomes. Third, it shows that endogenous composition feedbacks generate approximately power-law innovation concentration (Gabaix, 1999; Kerr and Robert-Nicoud, 2020) without imposing heavy-tailed assumptions on city fundamentals.

The remainder of the paper is organized as follows. Section 3 presents the model. Section 4 discusses the main theoretical results. Section 5 describes estimation and data. Section 7 evaluates empirical fit. Section 9 analyzes innovation concentration. Section **??** conducts counterfactual exercises. Section 10 concludes.

3 Economic Environment

This section presents the model in the order a reader solves it conceptually: production, city primitives, bidding markets, preferences, and equilibrium closure.

3.1 Production and problem solving

The economy produces knowledge by solving heterogeneous problems. Following a market-for-skills logic, each firm faces problems of random difficulty $x \sim U[0,1]$ and has internal capability $z \in [0,1]$. A firm can solve in-house all problems with difficulty $x \leq z$. Harder problems ($x > z$) require external skill.

A worker of type $m \in [0,1]$ can solve the residual problem if $m > x$. Conditional on the firm needing help ($x > z$), the probability that the worker can solve the problem is

$$\Pr(x < m \mid x > z) = \frac{\max\{m - z, 0\}}{1 - z}.$$

This expression highlights a key asymmetry. As firm capability z rises, the remaining unsolved problems are harder on average, increasing the value of worker skill. More complex firms therefore benefit disproportionately from access to highly skilled workers.

Communication and coordination are costly and city-dependent. Let $h_i > 0$ denote the time cost of transmitting and implementing knowledge in city i. Lower h_i implies that worker cognitive time is used more effectively.

A reduced-form expression for team output is

$$V_i(m, z) = \frac{m}{h_i(1 - z)}.$$

The term $(1 - z)^{-1}$ captures that more complex firms operate closer to the technological frontier, where successful matches are more valuable. The term $1/h_i$ captures city-specific efficiency in converting cognitive time into solved problems.

This production structure is supermodular in (m, z): the marginal return to worker skill is larger in more complex firms. This complementarity provides the microfoundation for sorting in the upper tail of the skill and firm distributions (Becker, 1973; Eeckhout and Kircher, 2011).

3.2 Cities, types, and primitives

There are N cities indexed by $i \in \{1, \ldots, N\}$. Workers differ by skill $m \in [0, 1]$ and firms (equivalently, problems) differ by complexity $z \in [0, 1]$. Aggregate type densities are $g(m)$ for workers and $f(z)$ for firms.

Each city is characterized by the vector of primitives

$$\left(A_i, S_i, E_i, h_{i0}, \phi_i^F, \phi_i^W, \xi_i\right).$$

Here, A_i denotes amenities, S_i housing supply (fixed in the short run), E_i the city's exposure to AI-related technologies, and h_{i0} baseline communication frictions. The terms $(\phi_i^F, \phi_i^W, \xi_i)$ are city-specific shifters that capture persistent unobserved firms and workers attractiveness for cities.

Two primitives play a central role in the mechanism. First, h_{i0} governs how costly coordination and idea transmission are within city i prior to technological diffusion. Second, S_i disciplines spatial congestion: even if a city becomes highly attractive, limited housing supply restricts entry and generates rent pressure (Saiz, 2010; Glaeser and Gyourko, 2018).

Technology diffusion reduces coordination frictions according to

$$h_{it} = h_{i0} \exp(-\lambda_h D_t E_i), \qquad \kappa_{it} = \frac{1}{h_{it}}.$$

Here D_t is an economy-wide diffusion index, E_i captures city-specific exposure, and λ_h determines how strongly diffusion translates into lower communication costs. Diffusion therefore

affects cities heterogeneously. The semi-elasticity is

$$\frac{\partial \log h_{it}}{\partial D_t} = -\lambda_h E_i,$$

so cities with greater exposure experience larger effective shocks. This heterogeneous diffusion structure is motivated by evidence that AI adoption is spatially uneven and correlated with pre-existing high-skill labor markets (Bessen and Cockburn, 2024).

Economically, diffusion lowers the time cost of communication and problem solving. Because h_{it} enters production and bidding objects, reductions in h_{it} increase the value of high-skill matches, alter bidding intensity in labor and housing markets, and can reallocate workers and firms across cities.

3.3 Sorting shares, distributions, and ODE identities

A central feature of the model is *tacit knowledge*. Before relocating, firms and workers observe the composition of cities but do not observe the realized type of their future match. A firm learns a worker's true type only after meeting, and a worker learns a firm's effective complexity and opportunity set only after matching.

This information structure has two implications. First, location decisions depend on the *expected composition* of each city rather than on deterministic partner selection. Second, expectations must be correct in equilibrium, so the model is a rational-expectations spatial equilibrium in distributions.

Let $s_i^F(z)$ and $s_i^W(m)$ denote location shares for firms and workers of type z and m. City-level densities are therefore

$$f_i(z) = s_i^F(z) f(z), \qquad \widehat{g}_i(m) = s_i^W(m) g(m),$$

with adding-up constraints

$$\sum_i s_i^F(z) = 1, \qquad \sum_i s_i^W(m) = 1$$

for every type.

Define city masses and moments

$$M_i^F = \int_0^1 f_i(z)\, dz, \qquad \widehat{M}_i^W = \int_0^1 \widehat{g}_i(m)\, dm,$$

$$\mu_i^z = \frac{1}{M_i^F}\int_0^1 z f_i(z)\,dz, \qquad \mu_i^{m,\text{app}} = \frac{1}{\widehat{M}_i^W}\int_0^1 m\widehat{g}_i(m)\,dm,$$

and two tightness measures

$$\theta_i^{LM} = \frac{M_i^F}{\widehat{M}_i^W}, \qquad \theta_i^H = \frac{\widehat{M}_i^W}{S_i}.$$

Here θ_i^{LM} measures labor-market tightness (firm pressure per worker), while θ_i^H measures housing-market tightness (worker pressure per housing unit). These tightness measures are central to the Poisson auction structure, which maps them into bidding intensity and win probabilities in each market (Moen, 1997).

City-level cumulative distributions satisfy

$$F_i'(z) = f_i(z), \qquad \widehat{G}_i'(m) = \widehat{g}_i(m).$$

These are ODE identities: once equilibrium densities are known, city CDFs are determined by integration. The equilibrium objects of the model are therefore full type distributions rather than city averages.

3.4 Generic first-price Poisson bidding object

Both labor and housing markets are modeled as first-price auctions with random numbers of competitors. The Poisson structure provides a tractable way to map market tightness into bidding behavior (Milgrom and Weber, 1982).

For bidder type t, willingness-to-pay $X(t)$, normalized opponent CDF $\widetilde{H}(t)$, and Poisson intensity λ:

$$Q(t) = \exp\{-\lambda(1-\widetilde{H}(t))\},$$

$$b(t) = X(t) - \mathcal{R}(t), \qquad \mathcal{R}(t) = \frac{\int_{t_0}^t Q(s)\,dX(s)}{Q(t)}.$$

Here $Q(t)$ is the probability of winning the auction and $\mathcal{R}(t)$ is the informational rent generated by optimal bid shading under uncertainty about competitors.

3.5 Labor market: firms bid for workers

In city i, a firm of type z has willingness-to-pay

$$X_i^F(z) = \phi_L \frac{\mu_i^{m,\text{app}}}{h_i(1-z)}.$$

Firms therefore bid more in cities where expected worker skill is higher and communication frictions are lower.

Let $\widetilde{F}_i(z) = F_i(z)/M_i^F$ denote the normalized firm CDF. The probability that a firm wins a worker is

$$Q_i^F(z) = \exp\{-\lambda_i^F(1-\widetilde{F}_i(z))\}, \qquad \lambda_i^F = (1+\tau_L)\theta_i^{LM}.$$

Equilibrium bids are

$$b_i^F(z) = X_i^F(z) - \mathcal{R}_i^F(z),$$

and expected firm payoff is

$$\mathbb{E}[U_i^F(z)] = \left[(1-\phi_L)\frac{\mu_i^{m,\text{app}}}{h_i(1-z)} + \mathcal{R}_i^F(z)\right] Q_i^F(z).$$

The implied wage schedule for workers is

$$w_i(m) = L\frac{m\psi_i}{h_i^W(m)} - \phi_i,$$

where

$$\psi_i = \frac{1}{M_i^F}\int_0^1 \frac{1}{1-z} f_i(z)\,dz, \qquad \phi_i = \frac{1}{M_i^F}\int_0^1 \mathcal{R}_i^F(z) f_i(z)\,dz.$$

The term ψ_i captures demand pressure from complex firms, while ϕ_i reflects average informational-rent leakage in labor bidding.

3.6 Housing market: workers bid for housing

Workers compete for scarce housing using the same bidding logic. Their willingness-to-pay is a share of wages:

$$X_i^H(m) = \phi_H w_i(m), \qquad b_i^H(m) = X_i^H(m) - \mathcal{R}_i^H(m).$$

Let $\widehat{\widetilde{G}}_i(m) = \widehat{G}_i(m)/\widehat{M}_i^W$. The probability of securing housing is

$$\overline{Q}_i^W(m) = \exp\{-\lambda_i^W(1-\widehat{\widetilde{G}}_i(m))\}, \qquad \lambda_i^W = (1+\tau_H)\theta_i^H.$$

Housing closure and outside mass. Because housing supply is finite, win probabilities are scaled to satisfy

$$M_i^W = \int_0^1 q_i(m)\widehat{g}_i(m)\,dm = \min\{\widehat{M}_i^W, S_i\}.$$

Define housed and outside densities

$$g_i(m) = q_i(m)\widehat{g}_i(m), \qquad g_i^0(m) = (1 - q_i(m))\widehat{g}_i(m).$$

Therefore

$$\widehat{M}_i^W = M_i^W + M_i^0, \qquad M_i^W \le S_i.$$

Mean rent is

$$r_i = \frac{1}{M_i^W}\int_0^1 b_i^H(m)g_i(m)\,dm.$$

This closure identifies who gains access to housing and who is priced out, allowing the model to characterize the incidence of spatial competition. The mechanism is consistent with the findings of Hsieh and Moretti (2019) that housing constraints prevent workers from accessing high-productivity cities, and with the broader evidence on housing costs as a driver of declining geographic mobility (Molloy et al., 2011).

Economic interpretation. The model endogenizes push and pull forces in spatial equilibrium. Pull forces arise from high expected match values in thick labor markets. Push forces arise from tighter competition for both workers and housing. Both forces are pecuniary externalities generated by bidding for scarce talent and scarce urban space. Spatial sorting therefore emerges as the equilibrium outcome of competition and expectations under tacit knowledge.

4 Worker Preferences and Location Choice

Workers choose locations by trading off expected disposable income, access to amenities, and the probability of obtaining housing. This is where labor-market opportunities and housing-market competition jointly determine worker location decisions.

If a worker of type m is housed in city i, disposable income is

$$x_i^{in}(m) = w_i(m) - b_i^H(m) = (1 - \phi_H)w_i(m) + \mathcal{R}_i^H(m).$$

Each component has a clear interpretation:

- $w_i(m)$ is labor income determined in the talent market.
- $b_i^H(m)$ is the housing payment resulting from the housing auction.
- ϕ_H is the share of wage the worker is willing to bid for housing.
- $\mathcal{R}_i^H(m)$ is the informational rent arising from first-price bidding.

Thus $x_i^{in}(m)$ represents net disposable income conditional on living in city i.
Expected utility is

$$\mathbb{E}[U_i^W(m)] = e^{Ta_i}\left[q_i(m)x_i^{in}(m) + (1 - q_i(m))u_0\right].$$

Interpretation:

- e^{Ta_i} scales utility by city amenities. The parameter T governs how strongly workers value non-pecuniary location attributes such as climate, urban consumption, and schools, consistent with the Rosen–Roback tradition (Rosen, 1979; Roback, 1982).
- $q_i(m)$ is the probability of obtaining housing in city i.
- $q_i(m)x_i^{in}(m)$ is expected utility conditional on being housed.
- $(1 - q_i(m))u_0$ represents the outside option if the worker is not housed.

Outside-option normalization. For theoretical analysis it is convenient to normalize $u_0 = 0$ as a location-invariant baseline. This isolates cross-city utility differences to wages, rents, housing access, and amenities. In numerical implementation we replace exact zero with a small floor $\varepsilon > 0$ to avoid numerical issues. This is purely computational and does not alter the economic content.

Idiosyncratic location shocks and interior sorting. Agents receive idiosyncratic location utilities

$$U_i^F(z) = \bar{U}_i^F(z) + \varepsilon_i^F(z), \qquad U_i^W(m) = \bar{U}_i^W(m) + \varepsilon_i^W(m),$$

where ε follows a Type-I extreme value distribution. This yields logit location shares

$$s_i^F(z) = \frac{\exp\{U_i^F(z)/\sigma_F\}}{\sum_j \exp\{U_j^F(z)/\sigma_F\}}, \qquad s_i^W(m) = \frac{\exp\{U_i^W(m)/\sigma_W\}}{\sum_j \exp\{U_j^W(m)/\sigma_W\}}.$$

Economically, (σ_F, σ_W) capture unobserved heterogeneity in location preferences such as family ties, institutional fit, or moving frictions. Empirically, they prevent knife-edge full sorting and allow interior mass across cities, consistent with the empirical finding that mobility frictions are quantitatively large (Molloy et al., 2011; Kaplan and Schulhofer-Wohl, 2017). Numerically, they smooth the equilibrium mapping and improve convergence.

Consistency with distributional equilibrium. Shares represent the *choice block* of the model, while distribution identities represent the *aggregation block*. Given shares,

$$f_i(z) = s_i^F(z) f(z), \qquad \widehat{g}_i(m) = s_i^W(m) g(m),$$

and cumulative distributions follow

$$F_i'(z) = f_i(z), \qquad \widehat{G}_i'(m) = \widehat{g}_i(m).$$

Thus the model consists of a single coupled system: logit choices determine location shares, and ODE identities aggregate those shares into equilibrium distributions across cities.

Proposition 1 (Local two-sided amplification)**.** *Let x denote worker-composition states and y firm-composition states. Around an interior equilibrium,*

$$dx = A\, dy + a\, d\chi, \qquad dy = B\, dx + b\, d\chi,$$

for primitive shock $d\chi$. If $\rho(AB) < 1$,

$$dx = (I - AB)^{-1}(a + Ab) d\chi, \qquad dy = (I - BA)^{-1}(b + Ba) d\chi.$$

Hence feedback matrices (AB, BA) amplify primitive shocks through resolvent terms.

The core feedback mechanism is

$$\mathbb{E}[V_i|z] \propto \mu_i^{m,\text{app}}, \qquad \mathbb{E}[w_i|m] \propto \psi_i, \phi_i.$$

Firm values depend on worker composition, while worker values depend on firm composition. City productivity is therefore endogenous and jointly produced by both sides of the market.

Economic intuition. A positive diffusion shock ($d\chi$ lowering h_i) increases returns to high-quality matches. Improved worker composition raises firm bids and entry incentives; improved firm composition raises worker expected wages. These responses reinforce each

other. The resolvent terms $(I - AB)^{-1}$ and $(I - BA)^{-1}$ capture this amplification: stronger cross-side complementarities generate larger equilibrium reallocations from small primitive shocks. This mechanism is formally analogous to the agglomeration multipliers studied in Moretti (2010) and the composition feedbacks in Diamond (2016).

4.1 Coupled equilibrium blocks

This section formalizes equilibrium and describes computation. The key insight is that equilibrium is obtained from a single coupled system rather than separate "logit" and "distribution" models.

The model contains two interacting components:

- **Choice block:** logit shares map utilities into location probabilities by type.
- **Distribution block:** shares determine densities, and ODE identities integrate those densities into city-level distributions.

The ODE structure therefore acts as the accounting backbone that converts decentralized location choices into full distributional equilibrium objects.

Definition 1 (Spatial sorting equilibrium with housing closure)**.** *Given* (D_t, Θ)*, an equilibrium is*

$$\{s_i^F, s_i^W, f_i, \widehat{g}_i, g_i, g_i^0, F_i, \widehat{G}_i, G_i, M_i^F, \widehat{M}_i^W, M_i^W, M_i^0, \theta_i^{LM}, \theta_i^H, w_i, r_i, q_i\}_{i=1}^N$$

such that

(i) shares are logit best responses (ii) densities follow from shares and aggregate densities (iii) ODE identities and adding-up constraints hold (iv) labor and housing bids satisfy Poisson first-price formulas (v) housing closure holds city-by-city.

For a two-city exposition, indifference conditions $U_1^F(z) = U_2^F(z)$ and $U_1^W(m) = U_2^W(m)$ can be written as coupled first-order ODE systems

$$F_1'(z) = H_F(F_1, G_1; \Omega_W), \qquad G_1'(m) = H_W(G_1, F_1; \Omega_F).$$

In the general N-city case, the same economics is solved numerically as a high-dimensional fixed point in shares, with ODE accounting identities evaluated at each iteration.

Assumption 1. *(A1) primitives are bounded with* $S_i > 0$ *and* $h_{it} > 0$

(A2) utility arguments are clipped by $\varepsilon > 0$

(A3) $\sigma_F, \sigma_W > 0$

(A4) share updates preserve simplices

(A5) the housing scaling map is continuous and enforces $M_i^W = \min\{\widehat{M}_i^W, S_i\}$.

Proposition 2 (Existence)**.** *Under assumptions (A1)–(A5), the discretized update map is continuous and maps a compact convex product of simplices into itself. A fixed point therefore exists by Brouwer's theorem.*

5 Data and Measurement

This section maps raw city-year observables into model moments and clarifies what is estimated versus calibrated.

5.1 Panel and coverage

The estimation sample is a balanced panel of 50 U.S. metropolitan areas from 2019 to 2023 (250 metro-year observations). City-year variables include employment, number of firms, college share, productivity index, mean wage, mean rent, amenity index, AI exposure, and diffusion index.

Top-50 metro selection. Metros are the largest CBSAs by average employment over the sample window.

Connectivity calibration. Baseline communication friction is calibrated from base-year relative wages:

$$h_{i0} \propto \left(\frac{\overline{w}_{i,\text{base}}}{\frac{1}{N}\sum_j \overline{w}_{j,\text{base}}} \right)^{-1}.$$

This calibration pins down initial cross-city heterogeneity in h_{i0} and is not estimated.

5.2 From observables to targeted moments

Each city-year maps into six core moments:

- worker mass share,
- firm mass share,
- mean worker-type proxy,

- mean firm-type proxy,
- wage index,
- rent index.

The data mappings are

$$\text{worker share}_{it}^{data} = \frac{employment_{it}}{\sum_j employment_{jt}}, \qquad \text{firm share}_{it}^{data} = \frac{firms_{it}}{\sum_j firms_{jt}},$$

$$\text{mean worker type}_{it}^{data} = college\ share_{it}, \qquad \text{mean firm type}_{it}^{data} = rank(productivity\ index_{it}) \in [0,1],$$

$$\text{wage index}_{it}^{data} = \frac{\overline{w}_{it}}{\frac{1}{N_t}\sum_j \overline{w}_{jt}}, \qquad \text{rent index}_{it}^{data} = \frac{\overline{r}_{it}}{\frac{1}{N_t}\sum_j \overline{r}_{jt}}.$$

Model counterparts are built from equilibrium objects (f_i, g_i, w_i, r_i). The estimator also targets worker and firm decile-shape moments to discipline sorting across the distribution.

Measurement caveat. Type moments are proxies: college share is not latent skill m, and rank-normalized productivity is not structural complexity z. This measurement gap is a key reason why type-ordering is the hardest fit margin. Future work could improve on these proxies using matched employer-employee data with occupational skill scores (Acemoglu and Autor, 2011) or task-content measures (Autor et al., 2003).

5.3 Estimation objective and algorithm

The estimated parameter vector is

$$\theta = (\sigma_F, \sigma_W, \lambda_h, T_{\text{pref}}, \tau_L, \tau_H, \rho_F, \rho_W).$$

The SMM objective is

$$\mathcal{L}(\theta) = \sum_t \left[\sum_{m \in \mathcal{M}} \sum_i \left(\frac{M_{itm}^{model}(\theta) - M_{itm}^{data}}{s_m} \right)^2 + \omega_{dec} \sum_{b=1}^{10} \left(\Delta_{tb}^W(\theta)^2 + \Delta_{tb}^F(\theta)^2 \right) \right].$$

Here s_m scales each moment block by empirical dispersion and ω_{dec} is fixed ex ante to give nontrivial weight to decile-shape moments. Each objective evaluation nests a full equilibrium solve. Search uses one baseline initialization plus 16 random candidates over bounded economic ranges; the minimum-objective candidate is selected.

6 Identification: Variation, Strengths, and Weak Margins

Each parameter is tied to a distinct source of variation in cross-city levels, cross-city dispersion, or intertemporal adjustment.

Table 1: Identification map (latest build)

Parameter	Main identifying variation	Main moments affected
σ_F	Cross-city dispersion in firm location probabilities	Firm mass shares; firm decile-shape moments
σ_W	Cross-city dispersion in worker location probabilities	Worker mass shares; worker decile-shape moments
λ_h	Differential diffusion incidence through $E_i \times D_t$	Wage/rent/mass gradients vs exposure
T_{pref}	Amenity sorting conditional on prices/congestion	Worker shares and worker means
τ_L	Labor-bidding intensity $\lambda_i^F = (1+\tau_L)\theta_i^{LM}$	Wages, sorting, firm-worker composition
τ_H	Housing-bidding intensity $\lambda_i^W = (1+\tau_H)\theta_i^H$	Rents, housing pressure, worker reallocation
ρ_F	Intertemporal inertia in firm-location shares	Year-to-year firm dynamics
ρ_W	Intertemporal inertia in worker-location shares	Year-to-year worker dynamics

7 Estimation Results

7.1 Parameter estimates and objective improvement

Table 2 reports estimates.

Table 2: Estimated structural parameters

Parameter	Estimate
σ_F	1.3440
σ_W	1.5914
λ_h	0.5791
T_{pref}	0.1196
τ_L	1.3877
τ_H	2.1792
ρ_F	0.9568
ρ_W	0.4343

How do these estimates compare to the literature? A direct one-to-one comparison is not always possible, because several parameters here are model-specific equilibrium objects rather than primitive elasticities estimated in reduced-form settings. The most informative comparison is by economic margin.

First, the location-dispersion parameters (σ_F, σ_W) are positive and sizable, implying substantial unobserved location heterogeneity and interior sorting. The fact that $\sigma_W > \sigma_F$ suggests worker sorting is more diffuse than firm sorting, consistent with evidence that mobility frictions and non-wage location components are quantitatively important for workers (Diamond, 2016; Molloy et al., 2011; Kaplan and Schulhofer-Wohl, 2017). The larger dispersion for workers also aligns with the finding of Baum-Snow and Pavan (2012) that sorting across cities accounts for a substantial but incomplete share of observed skill premia.

Second, $\lambda_h = 0.5791$ implies a meaningful diffusion semi-elasticity:

$$\frac{\partial \log h_{it}}{\partial (D_t E_i)} = -\lambda_h.$$

Hence diffusion in high-exposure cities produces large effective reductions in communication frictions. This magnitude is directionally consistent with evidence that AI raises productivity in knowledge-intensive tasks and with occupation-based AI exposure measures (Felten et al., 2023; Noy and Zhang, 2023). The estimate implies that the cities with median versus top-decile AI exposure experience economically meaningful differences in friction reduction from the same aggregate diffusion shock.

Third, $T_{\text{pref}} = 0.1196$ is positive but moderate, in line with the Rosen–Roback tradition where amenities matter for sorting but do not dominate wage-rent incidence (Roback, 1982; Diamond, 2016). The magnitude is broadly consistent with amenity-sorting estimates from

reduced-form regressions in the literature.

Fourth, bidding-intensity parameters are both positive ($\tau_L = 1.3877$, $\tau_H = 2.1792$), with a stronger housing channel. This is consistent with tightness-based competition in matching environments (Petrongolo and Pissarides, 2001) and with evidence that housing supply constraints are strong and heterogeneous across metros (Saiz, 2010). The stronger housing-market parameter ($\tau_H > \tau_L$) reflects the empirical fact that housing markets in productive cities are more congested—relative to their size—than labor markets, a pattern documented by Gyourko et al. (2013).

Finally, persistence is much stronger for firms than workers ($\rho_F = 0.9568 > \rho_W = 0.4343$), matching the empirical view that firm-location structures are highly sticky while worker migration, though limited, adjusts faster on average (Molloy et al., 2011; Kaplan and Schulhofer-Wohl, 2017). The near-unit firm persistence is consistent with the finding of Gaubert (2018) that firm location choices are largely determined by sunk costs and agglomeration complementarities that are slow to unwind.

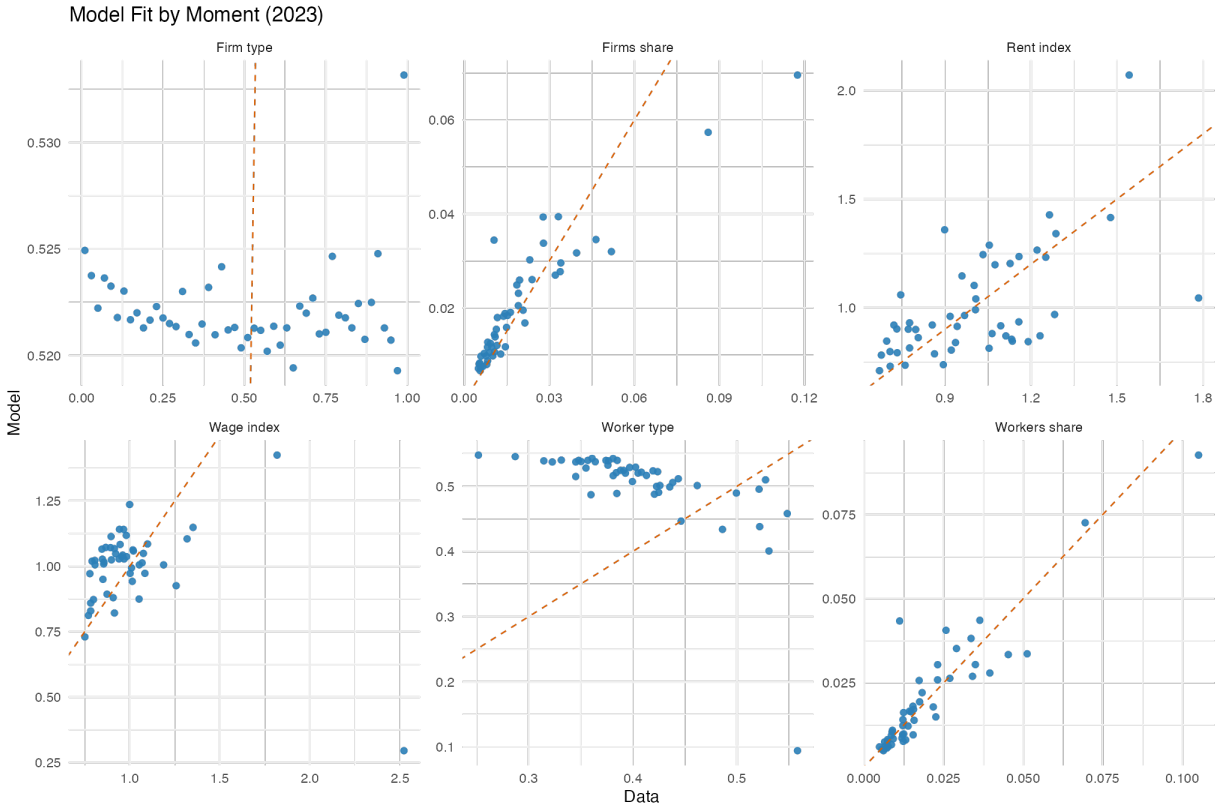


Figure 1: Model vs. data scatter (latest year): all six core moments.

Figure 1 shows strong fit for incidence outcomes—wages, rents, and mass allocation—and weaker fit for the cross-city ordering of type composition. This pattern is expected given

the economic scope of the model and the measurement strategy. The model is designed to capture the spatial organization of knowledge production and complex problem solving, whereas observed city labor markets include many occupations whose spatial allocation is primarily determined by local service demand and coverage (e.g., physicians, teachers, and other essential occupations present in most places). In addition, worker "talent" is proxied by educational attainment, which is an imperfect measure of the latent problem-solving type m in the model. These two factors naturally weaken model-data alignment in composition rankings, even when the model captures pricing and congestion incidence well.

8 Out-of-Sample Diagnostics: Within-City Inequality and Power Laws

This section evaluates non-targeted relationships between city wage levels and within-city dispersion.

8.1 Data-side proxy and model-side construction

For each year, on the data side:

$$\log(\overline{w}_{it}) = \alpha_t + \beta_t \, college \ share_{it} + u_{it},$$

and

$$Var(\log w)_{it}^{data} \approx p_{it}(1 - p_{it})\beta_t^2.$$

This yields a city-level inequality proxy from observed skill composition.

On the model side, inequality is computed directly from equilibrium wage schedules $w_i(m)$ and housed worker densities $g_i(m)$.

8.2 Results

Table 4 compares key slopes and correlations.

Table 3: Within-city inequality diagnostics

Metric	Estimate	Std. Error
Slope: model inequality on data inequality	0.7147	0.0625
Data power slope: $\log(sd)$ on $\log(mean)$	0.5938	0.0204
Model power slope: $\log(sd)$ on $\log(mean)$	0.8389	0.0598
Corr data $\log(mean)$ vs $\log(sd)$	0.8978	–
Corr model $\log(mean)$ vs $\log(sd)$	0.6075	–

The model matches signs and broad variation but overstates the elasticity of dispersion with respect to wage levels. A natural interpretation is excess convexity in the model-implied upper-tail wage schedule. The finding that larger cities exhibit more within-city inequality is consistent with the empirical literature (Eeckhout et al., 2014; Baum-Snow and Pavan, 2012; Moretti, 2012), though the model predicts a steeper relationship than is observed.

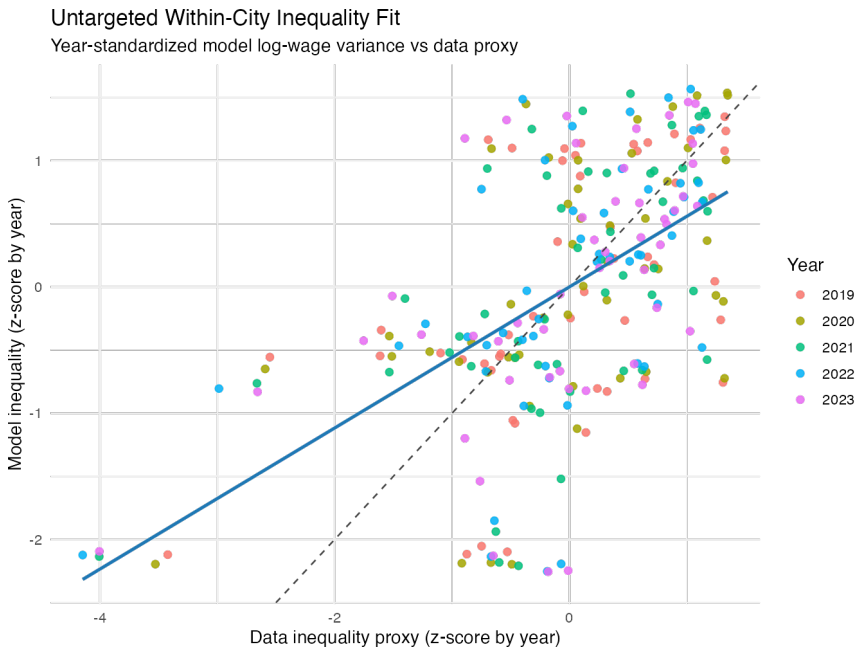


Figure 2: Model vs. data inequality proxy fit.



Figure 3: Power relation: within-city wage dispersion vs. level.

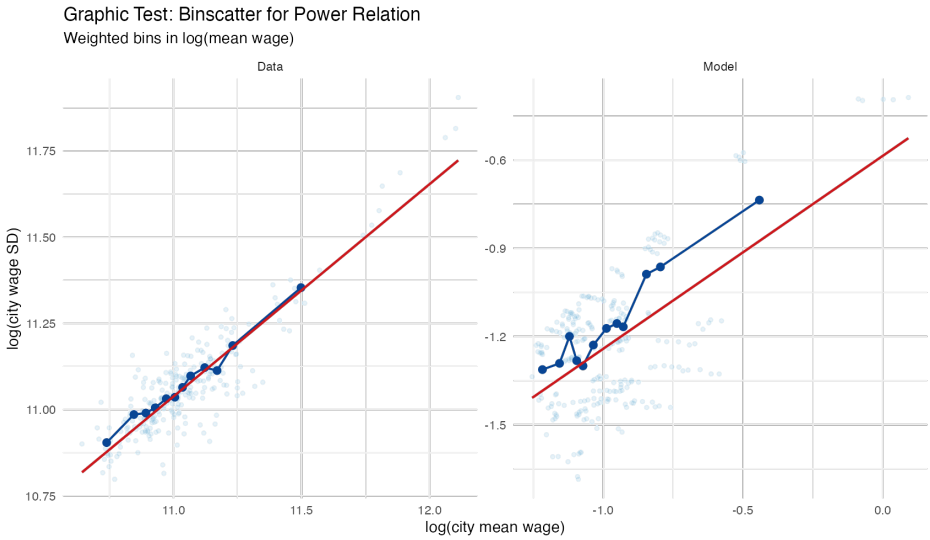


Figure 4: Binned test: mean–dispersion relation.

These moments are informative because they were not directly targeted in estimation. The model therefore passes a meaningful out-of-sample test if it reproduces the *joint* level-dispersion pattern across cities. Two findings stand out. First, the positive slope linking model inequality to data inequality (0.7147) indicates the model captures substantial cross-city rank variation in dispersion. Second, the model power slope (0.8389) exceeds the data slope (0.5938), implying the model predicts a steeper increase in within-city wage dispersion as city wage levels rise.

The core question is whether technology shocks reallocate tail activity. These diagnostics suggest the model has the right mechanism but an aggressive tail response. That makes

counterfactual concentration results directionally credible, while indicating that quantitative magnitudes in the upper tail should be interpreted with caution.

Out of sample, the model reproduces the qualitative inequality geography and a sizable share of its cross-city variation, but it over-predicts the steepness of inequality growth with city wage levels.

9 Equilibrium Distributions and Innovation Concentration

The next step is to inspect the solved equilibrium objects directly. Figures 12 and 13 report city-by-type distributions for firms and workers rather than city means.

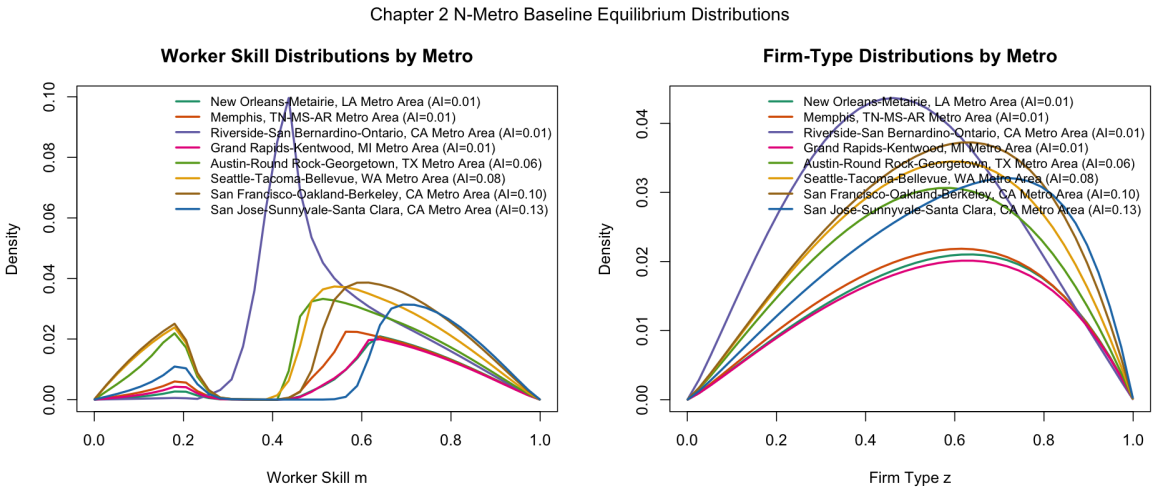


Figure 5: Estimated equilibrium type distributions across cities.

Innovation flow is constructed from upper-tail worker and firm masses:

$$M^W_{i,\text{top}} = \int_{0.9}^{1} g_i(m)\, dm, \qquad M^F_{i,\text{complex}} = \int_{0.9}^{1} f_i(z)\, dz,$$

$$\mathcal{I}_i = \kappa_i\, M^W_{i,\text{top}}\, M^F_{i,\text{complex}}.$$

This object is intentionally tail-sensitive: it captures where frontier worker-firm combinations are most likely to occur. The construction relates to reduced-form innovation measures based on patent activity and venture capital investment (Kerr and Robert-Nicoud, 2020; Moretti, 2021), but is derived endogenously from the model's sorting and bidding equilibrium.

9.1 Power-law diagnostics

Table 5 reports rank-size slopes and Pareto exponents for innovation shares.

Table 4: Innovation power-law tests

Snapshot	Full slope	Tail slope	Tail R^2	Pareto α (top 20%)
Baseline short-run	-0.6033	-0.6433	0.9214	2.8929
Baseline long-run	-0.6031	-0.6429	0.9218	2.8917
Big shock short-run (5x)	-0.7058	-0.7923	0.9203	2.4517
Big shock long-run (5x)	-0.7236	-0.8175	0.9216	2.4091

Large shocks steepen concentration slopes and lower the Pareto exponent α, indicating heavier innovation tails. The model therefore predicts that stronger diffusion can increase the spatial concentration of frontier activity even if average outcomes move modestly. The emergence of power-law concentration from the model equilibrium—without imposing Pareto primitives on city fundamentals—echoes the result of Gabaix (1999) that Zipf's law for city sizes can arise endogenously from growth dynamics rather than from parametric assumptions on fundamentals.

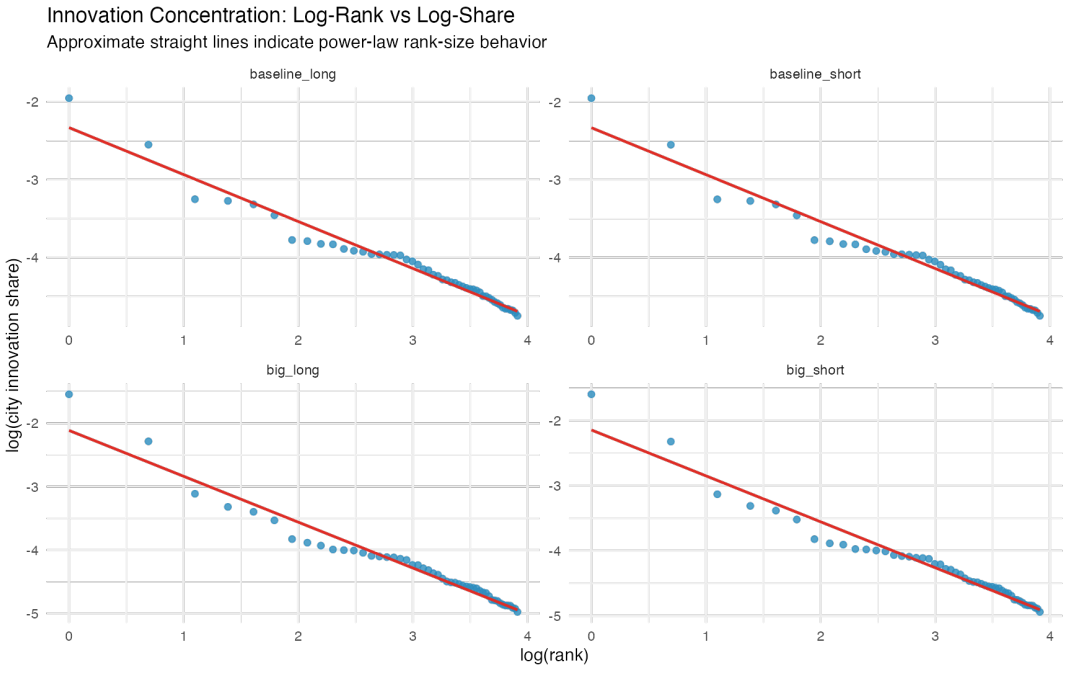


Figure 6: Innovation rank-size plot.

Figure 6 is a central result of the paper. The near-linear relationship in log-rank/log-share space shows that innovation concentration in the model is approximately power-law, even

though no Pareto law is imposed on city fundamentals. In this framework, the tail emerges endogenously from equilibrium forces: two-sided bidding, scarcity of talent and housing, and reciprocal worker-firm composition feedback. This is distinctive because concentration is not assumed; it is generated by the equilibrium mechanism. Economically, the slope summarizes how quickly innovation mass collapses into top-ranked cities. A steeper slope implies a stronger superstar-city regime and a larger gap between frontier hubs and the rest of the urban system.

9.2 Why model the technology shock as AI-driven efficiency in cognitive time

The counterfactual is motivated by a specific feature of AI technologies: they increase the effective use of skilled workers' time in knowledge-intensive tasks. In this model, production is constrained by cognitive time and coordination frictions. AI affects both margins by reducing the time cost of search, drafting, coding, communication, and problem decomposition.

Formally, we model this as a reduction in city-specific communication friction:

$$h_{it} = h_{i0} \exp(-\lambda_h D_t E_i).$$

An increase in diffusion D_t lowers h_{it} more in cities with higher AI exposure E_i. Since output and bidding objects scale with $1/h_{it}$, this is a productivity shock that raises match value per unit of worker time.

The key interpretation is not "AI replaces labor" but rather "AI changes the efficiency with which different worker types are allocated across problems and places." When h_{it} falls, higher-skill workers can solve more complex problems in the same time budget, and firms' willingness to pay for those workers rises. This shifts equilibrium sorting by type: some cities become more attractive for high-skill workers and complex firms, while others lose relative mass. This interpretation is consistent with the micro-level evidence in Noy and Zhang (2023), Dell'Acqua et al. (2023), and Brynjolfsson et al. (2023), all of which find that AI tools primarily raise output quality and breadth—not just speed—among skilled workers, consistent with an effective relaxation of cognitive time constraints.

Thus, the AI shock in this paper is a *spatial reallocation shock through efficiency*: it increases effective cognitive capacity and changes where worker-firm complementarities are most valuable in equilibrium. In contrast to accounts that emphasize AI-driven displacement (Acemoglu and Restrepo, 2018) or task substitution (Autor, 2022), the spatial-equilibrium lens reveals that even a productivity-enhancing AI shock can generate distributional consequences across cities—raising aggregate output while concentrating frontier activity.

Table 5: +15% diffusion: top city-level innovation-share gainers and losers (short run)

City	Category	Δ innovation share
San Jose-Sunnyvale-Santa Clara	Gain	+0.002070
San Francisco-Oakland-Berkeley	Gain	+0.000716
Seattle-Tacoma-Bellevue	Gain	+0.000195
Austin-Round Rock-Georgetown	Gain	+0.000017
Raleigh-Cary	Gain	+0.000001
Houston-The Woodlands-Sugar Land	Loss	-0.000135
Baltimore-Columbia-Towson	Loss	-0.000106
Chicago-Naperville-Elgin	Loss	-0.000102
Hartford-East Hartford-Middletown	Loss	-0.000102
San Diego-Chula Vista-Carlsbad	Loss	-0.000097

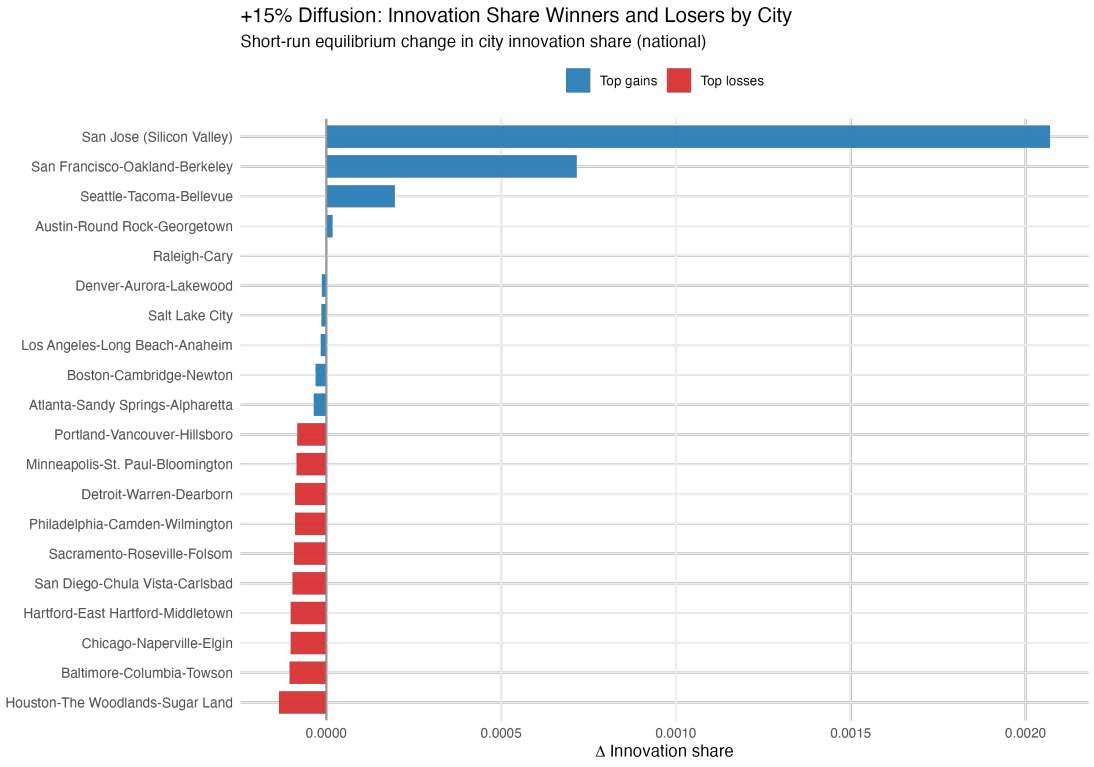


Figure 7: +15% diffusion: city innovation-share winners and losers (short run).

Table 5 and Figure 7 show that even a moderate shock generates non-trivial reallocation in innovation shares. The largest gain accrues to San Jose-Sunnyvale-Santa Clara (Silicon Valley), while several large metros lose innovation share. This is the key spatial-incidence

result: averages move little, but innovation location shifts meaningfully. The pattern mirrors the existing geography of AI exposure documented by Bessen and Cockburn (2024): cities that already have dense high-skill, technology-sector ecosystems gain disproportionately, consistent with the composition feedback mechanism in the model.

The model identifies Silicon Valley as the dominant innovation hub already in baseline, and predicts further concentration under diffusion. In the short run, the top-city innovation share rises from 0.1420 at baseline to 0.1441 under +15% diffusion (+1.46%). In the long run, it rises from 0.1418 to 0.1442 (+1.70%). Under larger shocks, this becomes first-order: the same top-city share reaches 0.2029 (short run) and 0.2136 (long run), implying increases of +42.8% and +50.7% relative to baseline.

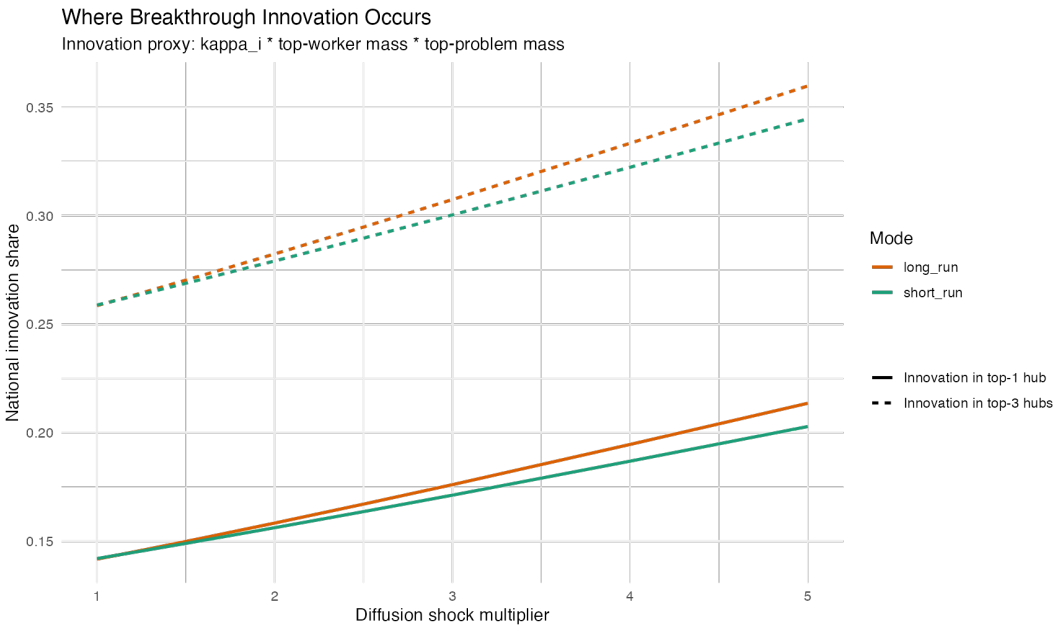


Figure 8: Innovation concentration in superstar cities across diffusion multipliers.

The +15% experiment identifies local incidence (small means, non-trivial gradients and city re-ranking). Figure 8 shows the same mechanism is nonlinear: as diffusion scales up, concentration accelerates and Silicon Valley's innovation advantage widens. This nonlinearity arises from the composition feedbacks in the model. As top workers and firms re-sort toward the leading city, the expected quality of local matches improves for remaining residents, further reinforcing the hub. The resolvent amplification captured in Proposition 1 ensures that these second-round effects can be large relative to the initial shock, generating a disproportionate response in the upper tail of the innovation distribution.

10 Conclusion

This paper studies the spatial consequences of artificial intelligence in a knowledge economy where the scarce input is cognitive time. I develop and estimate a distributional spatial-equilibrium model with two-sided heterogeneity, bidding in labor and housing markets, and diffusion-driven reductions in communication frictions. The key equilibrium objects are city-by-type distributions, which makes tail allocation, innovation concentration, and distributional incidence central outcomes of the analysis.

The central result is that modeling AI as an efficiency shock to cognitive-time allocation generates nontrivial spatial re-sorting. In the model, diffusion increases the value of high-skill matches. Because firms and workers sort across cities through reciprocal composition effects under scarcity, these gains are realized unevenly across space. Even when average effects are modest, city-level incidence can be large: some places gain innovation share, others lose it, and spatial concentration can rise endogenously without imposing heavy-tailed productivity primitives on cities.

Three conclusions follow. First, the framework fits key incidence margins—wages, rents, and mass allocation—sufficiently well to support counterfactual analysis. Second, innovation concentration emerges as an equilibrium outcome of bidding competition and two-sided sorting rather than as an imposed distributional assumption, connecting the model to the broader literature on endogenous power-law distributions in cities (Gabaix, 1999; Eeckhout, 2004). Third, technological shocks can generate a two-dimensional inequality response: partial compression within cities alongside stronger concentration across cities.

These findings speak to a broader question in economic geography: whether general-purpose technologies tend to disperse or concentrate economic activity. The present analysis suggests that the answer depends critically on whether the technology raises the returns to local complementarities—as AI appears to do for knowledge workers—or primarily reduces the cost of distance. When complementarities dominate, as they do in this model, technological progress reinforces rather than attenuates the spatial hierarchy of innovation.

Several limitations point to productive extensions. The model abstracts from endogenous knowledge capital accumulation and dynamic sorting, focusing instead on the within-period equilibrium reallocation of cognitive time. Extending the framework to allow for learning-by-doing or knowledge stocks that compound over time would generate path dependence in city trajectories and richer long-run dynamics. The model also treats AI exposure as exogenous; incorporating endogenous adoption decisions by firms and workers would allow for feedback between innovation, AI diffusion, and spatial equilibrium. Finally, the welfare implications of technology-driven spatial concentration—including the distributional consequences for

workers in cities that lose innovation share—remain important open questions for future work.

References

Acemoglu, D. (2024). The simple macroeconomics of AI. *NBER Working Paper No. 32487*. National Bureau of Economic Research.

Acemoglu, D. and Autor, D. (2011). Skills, tasks and technologies: Implications for employment and earnings. *Handbook of Labor Economics*, 4B:1043–1171.

Acemoglu, D. and Restrepo, P. (2018). The race between man and machine: Implications of technology for growth, factor shares, and employment. *American Economic Review*, 108(6):1488–1542.

Agrawal, A., Gans, J., and Goldfarb, A. (2018). *Prediction Machines: The Simple Economics of Artificial Intelligence*. Harvard Business Review Press, Boston.

Allen, T. and Arkolakis, C. (2022). The welfare effects of transportation infrastructure improvements. *Review of Economic Studies*, 89(6):2911–2957.

Antràs, P., Garicano, L., and Rossi-Hansberg, E. (2006). Offshoring in a knowledge economy. *Quarterly Journal of Economics*, 121(1):31–77.

Audretsch, D. B. and Feldman, M. P. (1996). R&D spillovers and the geography of innovation and production. *American Economic Review*, 86(3):630–640.

Autor, D. (2022). The labor market impacts of technological change: From unbridled enthusiasm to qualified optimism to vast uncertainty. *NBER Working Paper No. 30074*. National Bureau of Economic Research.

Autor, D., Dorn, D., Katz, L. F., Patterson, C., and Reenen, J. V. (2020). The fall of the labor share and the rise of superstar firms. *Quarterly Journal of Economics*, 135(2):645–709.

Autor, D. H., Levy, F., and Murnane, R. J. (2003). The skill content of recent technological change: An empirical exploration. *Quarterly Journal of Economics*, 118(4):1279–1333.

Baum-Snow, N. and Pavan, R. (2012). Understanding the city size wage gap. *Review of Economic Studies*, 79(1):88–127.

Becker, G. S. (1973). A theory of marriage: Part I. *Journal of Political Economy*, 81(4):813–846.

Bessen, J. and Cockburn, I. (2024). The spatial diffusion of artificial intelligence. *Working Paper*. Boston University School of Law.

Brynjolfsson, E., Li, D., and Raymond, L. R. (2023). Generative AI at work. *NBER Working Paper No. 31161*. National Bureau of Economic Research.

Combes, P.-P., Duranton, G., and Gobillon, L. (2008). Spatial wage disparities: Sorting matters! *Journal of Urban Economics*, 63(2):723–742.

Davis, D. R. and Dingel, J. I. (2019). A spatial knowledge economy. *American Economic Review*, 109(1):153–170.

Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., and Lakhani, K. R. (2023). Navigating the Jagged Technological Frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. *Harvard Business School Working Paper*, (24-013).

Diamond, R. (2016). The determinants and welfare implications of US workers' diverging location choices by skill: 1980–2000. *American Economic Review*, 106(3):479–524.

Dingel, J. I. and Neiman, B. (2020). How many jobs can be done at home? *Journal of Public Economics*, 189:104235.

Duranton, G. and Puga, D. (2004). Micro-foundations of urban agglomeration economies. *Handbook of Regional and Urban Economics*, 4:2063–2117.

Eeckhout, J. (2004). Gibrat's law for (all) cities. *American Economic Review*, 94(5):1429–1451.

Eeckhout, J. and Kircher, P. (2011). Identifying sorting—in theory. *Review of Economic Studies*, 78(3):872–906.

Eeckhout, J., Pinheiro, R., and Schmidheiny, K. (2014). Spatial sorting. *Journal of Political Economy*, 122(3):554–620.

Felten, E., Raj, M., and Seamans, R. (2023). Occupational, industry, and geographic exposure to artificial intelligence: A novel dataset and its potential uses. *Strategic Management Journal*, 44(9):2442–2450.

Gabaix, X. (1999). Zipf's law for cities: An explanation. *Quarterly Journal of Economics*, 114(3):739–767.

Garicano, L. (2000). Hierarchies and the organization of knowledge in production. *Journal of Political Economy*, 108(5):874–904.

Garicano, L. and Hubbard, T. N. (2007). Managerial leverage is limited by the extent of the market: Theory and evidence from the legal services industry. *Journal of Political Economy*, 115(1):1–40.

Garicano, L. and Rayo, L. (2016). Why organizations fail: Models and cases. *Journal of Economic Literature*, 54(1):137–192.

Garicano, L. and Rossi-Hansberg, E. (2006). Organization and inequality in a knowledge economy. *Quarterly Journal of Economics*, 121(4):1383–1435.

Gaubert, C. (2018). Firm sorting and agglomeration. *American Economic Review*, 108(11):3117–3153.

Glaeser, E. L. and Gyourko, J. (2018). The economic implications of housing supply. *Journal of Economic Perspectives*, 32(1):3–30.

Glaeser, E. L., Kallal, H. D., Scheinkman, J. A., and Shleifer, A. (1992). Growth in cities. *Journal of Political Economy*, 100(6):1126–1152.

Glaeser, E. L., Ponzetto, G. A., and Tobio, K. (2014). Cities, skills, and regional change. *Regional Studies*, 48(1):7–43.

Gokan, T., Kichko, S., and Thisse, J.-F. (2019). How do trade and communication costs shape the spatial organization of firms? *Journal of Urban Economics*, 113:103191.

Gyourko, J., Mayer, C., and Sinai, T. (2013). Superstar cities. *American Economic Journal: Economic Policy*, 5(4):167–199.

Hampole, M., Papanikolaou, D., Schmidt, L. D., and Seegmiller, B. (2025). Artificial intelligence and the workforce. *NBER Working Paper*. National Bureau of Economic Research.

Hsieh, C.-T. and Moretti, E. (2019). Housing constraints and spatial misallocation. *American Economic Journal: Macroeconomics*, 11(2):1–39.

Jaffe, A. B., Trajtenberg, M., and Henderson, R. (1993). Geographic localization of knowledge spillovers as evidenced by patent citations. *Quarterly Journal of Economics*, 108(3):577–598.

Kaplan, G. and Schulhofer-Wohl, S. (2017). Understanding the long-run decline in interstate migration. *International Economic Review*, 58(1):57–94.

Kerr, W. R. (2010). Breakthrough inventions and migrating clusters of innovation. *Journal of Urban Economics*, 67(1):46–60.

Kerr, W. R. and Robert-Nicoud, F. (2020). *Tech Clusters*, volume 34.

Lise, J. and Postel-Vinay, F. (2020). Multidimensional skills, sorting, and human capital accumulation. *American Economic Review*, 110(8):2328–2376.

Lucas, R. E. (1988). On the mechanics of economic development. *Journal of Monetary Economics*, 22(1):3–42.

Marshall, A. (1920). *Principles of Economics*. Macmillan, London, 8th edition.

Milgrom, P. R. and Weber, R. J. (1982). A theory of auctions and competitive bidding. *Econometrica*, 50(5):1089–1122.

Moen, E. R. (1997). Competitive search equilibrium. *Journal of Political Economy*, 105(2):385–411.

Molloy, R., Smith, C. L., and Wozniak, A. (2011). Internal migration in the United States. *Journal of Economic Perspectives*, 25(3):173–196.

Moretti, E. (2004). Human capital externalities in cities. *Handbook of Regional and Urban Economics*, 4:2243–2291.

Moretti, E. (2010). Local multipliers. *American Economic Review: Papers and Proceedings*, 100(2):373–377.

Moretti, E. (2012). *The New Geography of Jobs*. Houghton Mifflin Harcourt, Boston.

Moretti, E. (2021). The effect of high-tech clusters on the productivity of top inventors. *American Economic Review*, 111(10):3328–3375.

Noy, S. and Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654):187–192.

Peng, S., Kalliamvakou, E., Cihon, P., and Demirer, M. (2023). The impact of AI on developer productivity: Evidence from GitHub Copilot. *arXiv preprint*, arXiv:2302.06590.

Petrongolo, B. and Pissarides, C. A. (2001). Looking into the black box: A survey of the matching function. *Journal of Economic Literature*, 39(2):390–431.

Redding, S. J. (2022). Quantitative urban models: From theory to data. *Journal of Economic Perspectives*, 36(3):75–98.

Roback, J. (1982). Wages, rents, and the quality of life. *Journal of Political Economy*, 90(6):1257–1278.

Rosen, S. (1979). Wage-based indexes of urban quality of life. *Current Issues in Urban Economics*, pages 74–104.

Saiz, A. (2010). The geographic determinants of housing supply. *Quarterly Journal of Economics*, 125(3):1253–1296.

Shimer, R. (2007). Mismatch. *American Economic Review*, 97(4):1074–1101.

Shimer, R. and Smith, L. (2000). Assortative matching and search. *Econometrica*, 68(2):343–369.

Storper, M. and Venables, A. J. (2004). Buzz: Face-to-face contact and the urban economy. *Journal of Economic Geography*, 4(4):351–370.

A Appendix A: Additional Fit and Diagnostic Figures

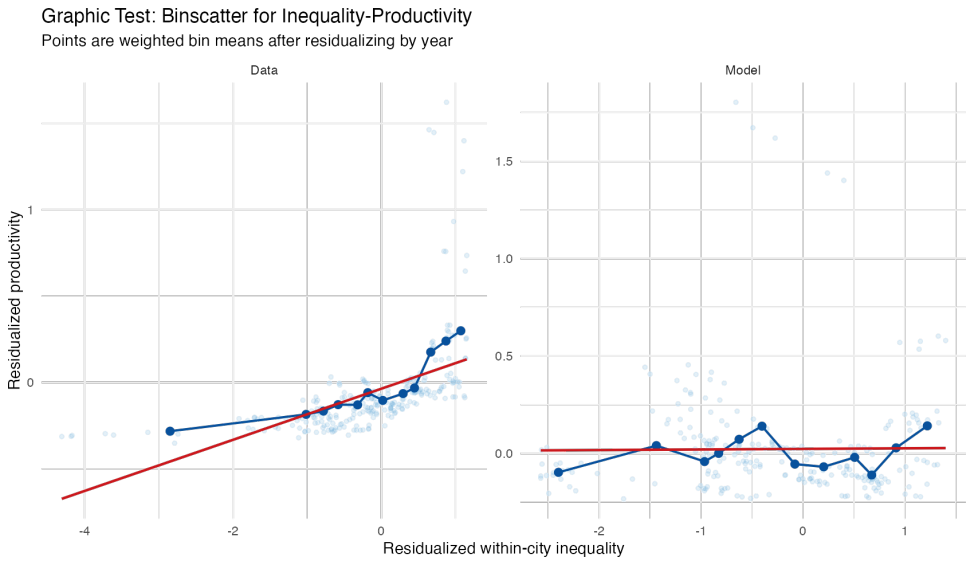


Figure 9: Residualized binscatter: productivity and inequality proxy.

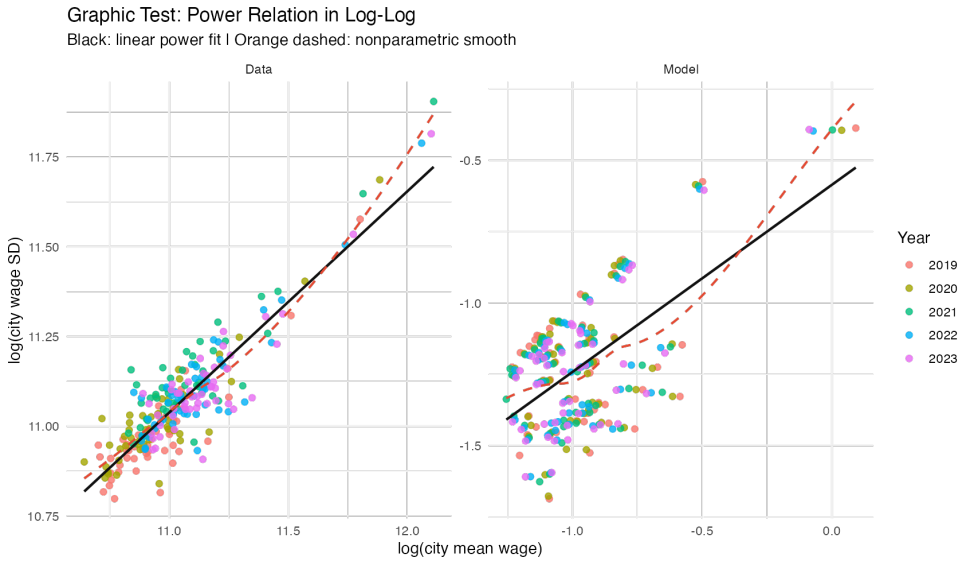


Figure 10: Log-log diagnostic: mean wage and within-city dispersion.

B Appendix B: How the Model Is Estimated

This appendix describes the numerical procedure used to solve and estimate the model. For any candidate parameter vector θ, estimation requires repeatedly solving for a rational-expectations spatial equilibrium and mapping the resulting equilibrium objects into model moments.

B.1 Equilibrium solve for a given parameter vector

For a given θ, the numerical routine iterates on firm and worker location shares until beliefs about city composition are consistent with realized composition. The algorithm proceeds as follows:

1. Initialize firm and worker share matrices on the (i, z) and (i, m) grids:

$$\{s_i^F(z),\, s_i^W(m)\}_{i=1}^N.$$

2. Construct city densities from shares:

$$f_i(z) = s_i^F(z) f(z), \qquad \widehat{g}_i(m) = s_i^W(m) g(m).$$

3. Integrate the accounting identities to recover city cumulative distributions:

$$F_i'(z) = f_i(z), \qquad \widehat{G}_i'(m) = \widehat{g}_i(m).$$

4. Compute city moments and market objects, including masses, composition moments, labor-market tightness, housing-market tightness, informational rents, wage schedules, rent schedules, and housing-access probabilities $q_i(m)$.

5. Construct firm and worker utilities implied by these objects.

6. Update location shares using damped logit in log-space:

$$\log s^{t+1} = (1 - \delta) \log s^t + \delta \log \widehat{s}(U^t).$$

7. Renormalize shares so that simplex constraints hold exactly, and repeat until convergence.

Convergence is declared when both share updates and implied moment updates are sufficiently small:

$$\|s^{t+1} - s^t\|_\infty < \texttt{tol}_s, \qquad \|M^{t+1} - M^t\|_\infty < \texttt{tol}_M.$$

This procedure is numerically stable because positivity and simplex constraints are preserved at each iteration.

B.2 Why the equilibrium is a coupled fixed point

The implemented system is neither "just logit" nor "just ODE." It is not just logit because the equilibrium objects are full functions,

$$\{f_i, g_i, F_i, G_i\}_{i=1}^{N},$$

together with endogenous wages, rents, and housing incidence. It is not just ODE because decentralized location choices are determined by random-utility best responses. Logit closes individual behavior under unobserved heterogeneity, while ODE identities aggregate those choices into city-level distributions. The equilibrium is therefore a coupled fixed point with embedded distributional accounting identities.

At convergence, beliefs about city composition coincide with realized composition. This is the rational-expectations equilibrium condition implied by tacit knowledge.

B.3 Moment construction

Once the equilibrium is solved for a candidate parameter vector θ, model-implied moments are constructed from the resulting city distributions, wages, rents, and masses. These moments include:

- worker mass shares,
- firm mass shares,
- worker composition moments,
- firm composition moments,
- wage indices,
- rent indices,
- worker and firm decile-shape moments.

The distributional structure of the model is important at this stage because it allows the researcher to recover not only city means, but also tail objects and incidence measures such as top-worker mass, top-problem mass, innovation concentration, within-city inequality, between-city inequality, and displacement under housing scarcity.

B.4 Estimation routine

Let θ denote the parameter vector to be estimated. For each candidate θ:

1. solve the equilibrium using the iterative procedure above;

2. compute model-implied moments;

3. evaluate the simulated method of moments objective function;

4. store the objective value and the associated parameter vector.

The search procedure uses one baseline starting vector plus a set of random candidate vectors drawn from economically admissible parameter ranges. Each candidate requires a full equilibrium solve. The estimator then retains the parameter vector that minimizes the objective function over the search set.

This search design is useful in a high-dimensional fixed-point environment because it is more robust than relying exclusively on local gradient-based routines.

B.5 Practical interpretation

The estimation procedure treats the model as a system of cities linked by equilibrium reallocation of firms and workers. For each candidate parameter vector, the algorithm jointly solves for location shares, city distributions, wages, rents, and housing allocation. Estimation then selects the parameter values that make these equilibrium objects match the targeted city-level moments as closely as possible.