Output Market Power and Spatial Misallocation*

Santiago Franco
Boston University

September 2026

Abstract

Most product industries are local. In the U.S., firms selling goods and services to local consumers account for half of the economy's sales and generate sixty percent of jobs. Competition in these industries occurs in local product markets. I propose a theory of such competition in which heterogeneous firms have output market power and choose where to operate. Firm sorting generates endogenous differences in local competition. Variable markups induce more productive firms to over-concentrate in larger markets, leading to spatial misallocation. I use U.S. Census microdata to estimate markups in local industries and find significant heterogeneity across U.S. markets. Producers in markets in the top decile of the market-size distribution have markups approximately thirty percent lower than those in the bottom decile. I then estimate the model and quantify the general-equilibrium effects of place-based policies. A quasi-optimal policy eliminates markups and reallocates productive firms toward smaller markets. It generates large local productivity gains and raises aggregate welfare. In contrast, a wage-bill subsidy targeted to smaller markets produces qualitatively similar effects, but with substantially smaller magnitudes.

*Email: safranco@bu.edu. This article is a revised version of my Ph.D. dissertation at the University of Chicago. I am grateful to my advisors, Esteban Rossi-Hansberg, Ufuk Akcigit, Chang-Tai Hsieh, and Erik Hurst, for their guidance. I thank Lukas Nord for discussing the paper. I also thank Rodrigo Adao, Fernando Alvarez, Adrien Bilal, Olivia Bordeu, Jonathan Dingel, Agustin Gutierrez, Greg Kaplan, Benny Kleinman, Santiago Lacouture, Marco Loseto, Sreyas Mahadevan, Kiminori Matsuyama, Simon Mongey, Pauline Mourot, Jose M. Quintero, Pascual Restrepo, Marcos Sora, and Chad Syverson for insightful comments. Any views expressed are mine and not those of the U.S. Census Bureau. The Census Bureau has reviewed this data product to ensure appropriate access, use, and disclosure avoidance protection of the confidential source data used to produce this product. This research was performed at a Federal Statistical Research Data Center under FSRDC Project Number 2441 (CBDRB-FY23-P2441-R10613).

1 Introduction

Most product industries are local. In the U.S., firms selling goods and services to local consumers account for 50% of sales and generate 60% of jobs.[1] Competition in these industries is inherently local. For example, restaurants or retail stores in New York do not compete with similar establishments in Chicago or Seattle. As a result, the location choices of firms operating in these industries are a key determinant of the strength of local competition. Some places attract productive firms that set low prices, while others attract few producers and face weaker competition. When firms exert output market power, these differences in local competition can have important consequences for local welfare and resource allocation. What determines firms' location decisions in imperfectly competitive markets, and what are the consequences for local competition? What are the general equilibrium effects of altering firm location choices through place-based policies?

In this paper, I answer these questions with a quantitative spatial model of firm location and endogenous variable markups. In the framework, spatial differences in local competition arise endogenously from the location choices of heterogeneous firms. The equilibrium allocation is inefficient because more productive firms over-concentrate in larger cities. I then use U.S. Census micro-data to estimate markups for establishments in local industries. I document substantial heterogeneity in markups across U.S. cities, with cities in the top decile of the city-size distribution exhibiting aggregate markups approximately 30% lower than those in the bottom decile. Finally, I estimate the model and use it to quantify the general-equilibrium effects of place-based policies. I contrast the optimal policy chosen by a utilitarian social planner with policies that attract firms to less-developed locations, such as the U.S. Empowerment Zones program.

In the baseline framework, production takes place across locations that I call cities. Workers are freely mobile and consume local goods, housing, and a traded good. Cities differ ex ante in traded-good productivity and have a fixed supply of housing. The traded good is homogeneous, produced competitively in every city, and freely traded across locations. Local goods, in contrast, are produced by heterogeneous firms that differ in productivity and choose where to locate.

The location choices of heterogeneous local-goods producers is at the core of the framework. Potential entrants pay an entry cost to learn their productivity and then choose a city in which to operate. Once located, they produce a differentiated variety using labor and compete monopolistically with other local producers. Workers have Kimball preferences (Kimball, 1995) over local varieties, which make each firm's demand elasticity depend on its position along the residual demand curve and generate heterogeneous markups. In equilibrium, more productive firms set lower prices, face less elastic demand, and charge higher markups.

Local producers choose where to locate based on two endogenous city characteristics. The first is market size, which is determined by workers' income and governs potential sales. The second is local competition, which depends on competitors' prices and local wages and determines firms' markups conditional on their productivity. In equilibrium, larger cities are more competitive, so firms face a trade-off between higher sales and lower markups when choosing where to locate. Since more productive firms charge higher markups, they benefit relatively more from the additional sales

[1] See Delgado, Porter, and Stern (2015) and my calculations in Table 1.

available in larger markets. Hence, there is positive assortative matching, with more productive firms sorting into larger, more competitive cities.

Two opposing forces determine spatial differences in market power. First, local competition restrains markups across locations. Because larger markets are more competitive, this competition force lowers markups in such places. Second, markets that attract more productive firms tend to exhibit higher markups, reflecting the positive relationship between firm productivity and markups. Since larger markets attract more productive firms, this selection force pushes markups up in such places. Hence, the distribution of markups across cities depends on which of these two forces dominates. Intuitively, when selection into larger markets is weak, firms in larger cities are relatively similar and stronger competition leads to lower markups. In contrast, when selection is strong, a few productive firms dominate larger markets, raising aggregate markups.

In the model, local producers markups generate multiple inefficiencies. I characterize the necessary conditions that must hold in any efficient allocation and compare them with the conditions in the decentralized equilibrium. This comparison reveals three margins of inefficiency. First, and central to my framework, firm location decisions are inefficient, leading to an over-concentration of productive firms in larger markets. Second, aggregate entry is inefficient. Third, the local-goods sector is too small because markups distort consumption away from local goods and toward housing and the traded good. Worker allocation is efficient conditional on firm location because the framework features no explicit worker-location externalities. Finally, I use these efficiency conditions to characterize the optimal transfers that implement the efficient allocation.

The spatial misallocation arises because entry generates two opposing externalities. First, consumers value greater variety. Therefore, a firm entering a market creates a positive externality by lowering the local price index and increasing aggregate expenditure on all local goods. Because entrants do not internalize the resulting increase in demand for incumbents, this force leads to insufficient entry. Second, because local varieties are imperfect substitutes, entrants reduce the profits of incumbents without internalizing this effect. Entry therefore creates a negative business-stealing externality, which pushes toward excessive entry. The balance between these two forces determines whether entry in a given location is excessive, insufficient, or efficient.[2] I show that the strength of these externalities varies systematically with local competition. Where competition is strong, the business-stealing externality is larger. In contrast, where competition is weak, the gains from additional variety are greater. Because larger cities are more competitive in equilibrium, there is excessive entry in larger cities and insufficient entry in smaller ones. This imbalance therefore generates spatial misallocation, with productive firms over-concentrating in larger markets.

I extend the baseline model for the quantitative analysis along five dimensions. First, the productivity of local-goods producers is determined in two stages. Firms choose locations based on their initial productivity and, after locating, draw a final productivity that is correlated with their initial draw. This extension generates within-city markup heterogeneity while preserving positive sorting. As a result, markups create an additional inefficiency: within-city misallocation of production inputs. Second, cities also differ in local amenities that capture non-monetary compen-

[2] Dixit and Stiglitz (1977) show that in single-market models with monopolistic competition and CES preferences, these externalities are constant and exactly offset each other, yielding efficient entry.

sating differentials, such as public goods and weather. Third, workers have idiosyncratic location preferences, which generate a well-defined migration elasticity with respect to real wages. Fourth, both traded-good and local-good producers use housing as an input in production. Fifth, housing supply is elastic in each city. I show that the main results from the baseline model carry over to this extended framework.

In the second part of the paper, I examine the model's positive implications. I estimate markups for U.S. local producers by combining workers' demand with firms' production decisions, in the spirit of De Loecker (2011). I start from the firm's first-order condition for labor, which links the labor cost share of sales to the markup and the labor output elasticity. Consumer preferences imply that markups are also a function of local sales shares and local price indices. I use this relationship to construct a nonparametric control function for markups and combine it with the labor first-order condition to jointly estimate markups and the labor output elasticity. Within-city variation in sales shares and labor cost shares identifies these two objects.

I implement the empirical strategy using U.S. Census microdata. I use the Longitudinal Business Database (LBD) to construct establishments' labor cost shares of sales and the Economic Censuses (EC) to construct their local sales shares. Using 2017 as the baseline year, I estimate markups for five million local establishments. I then aggregate these estimates to the city level using the aggregation implied by the model.

I find substantial heterogeneity in markups across U.S. cities. Cities in the top decile of the city-size distribution have aggregate markups approximately 30% lower than those in the bottom decile. This finding implies that, in U.S. local industries, the competition force dominates the selection force. This relationship is stable across census years and remains robust when defining cities as either commuting zones or counties.

The cross-sectional markup patterns are robust across alternative specifications. First, I extend the baseline estimation to allow for labor market power and sectoral heterogeneity in technologies and preferences. Second, I show that the cross-city patterns remain when restricting the sample to firms operating in a single location. Finally, using the Census of Manufactures, I consider alternative markup estimates that use cost shares to recover output elasticities, allow for establishment-specific output elasticities, and treat materials or energy as the flexible input. Across these exercises, the negative relationship between aggregate markups and market size remains stable.

In the final section, I estimate the parameters of the model to study its normative implications. First, I use the establishment-level markup estimates to identify the Kimball demand parameters. The model delivers estimating equations that link these parameters to firm markups and sales shares. Second, I use a Simulated Method of Moments (SMM) approach to estimate the parameters governing local producers' productivity distribution and the aggregate entry cost. I identify the productivity distribution by targeting average establishment employment across different quantiles of the market-size distribution, while the entry cost is disciplined by the economy-wide aggregate markup. Finally, I infer city-specific productivities and amenities using data on population and average wages in each city. The estimated model reproduces the negative relationship between aggregate markups and city size.

I use the estimated model to quantify the general-equilibrium effects of two place-based policies. I first study a policy that eliminates markups for all firms in all locations. I refer to it as the quasi-optimal policy because it combines a location-specific output subsidy from the optimal policy, which is more generous in larger markets, with a proportional wage tax on workers.[3] The policy corrects all inefficiencies associated with markups and, in particular, the spatial misallocation of local producers. As a result, productive firms relocate toward smaller markets, increasing competition and generating large local productivity gains. TFP increases across all cities, but the gains are substantially larger in smaller markets, where productivity rises by as much as 60%. Larger markets also experience productivity gains from the elimination of within-city markup distortions, which raises allocative efficiency. Nevertheless, this channel is quantitatively less important. In addition, higher aggregate entry raises productivity across all locations. In terms of welfare, the policy generates aggregate welfare gains of 3.1%, with significantly larger gains in smaller markets.

I contrast the quasi-optimal policy with a policy inspired by the U.S. Empowerment Zones (EZ) program. I introduce a 20% wage-bill subsidy in the smallest commuting zones, which together account for 2.3% of the U.S. population. The EZ program operates through a spatial mechanism similar to that of the quasi-optimal policy: it attracts firms to targeted locations, increasing competition and productivity. Its quantitative effects, however, are substantially smaller. TFP increases at most by 2% in treated locations, while non-treated markets experience small productivity losses as firms relocate away. Welfare increases by up to 2% in targeted locations but falls slightly elsewhere, in part because workers outside targeted markets help finance the program without receiving direct local benefits. Overall, the EZ program raises aggregate welfare by 0.03%. Despite its smaller effects, the exercise shows that a relatively simple place-based subsidy can partially reproduce the spatial reallocation generated by the quasi-optimal policy.

Related Literature. This paper contributes to five strands of the literature. The first studies competition in local output markets. Hsieh and Rossi-Hansberg (2023), Oberfield et al. (2023), and Kleinman (2025) analyze competition by multi-establishment firms across regions under CES preferences, which imply constant markups. I contribute to this literature by introducing a framework with endogenous variable markups. At the cost of abstracting from the multi-location expansion problem, the model allows me to study richer local pricing dynamics and introduces an additional force shaping firm location decisions. In related empirical work, Rossi-Hansberg, Sarte, and Trachter (2020) and Autor, Patterson, and Van Reenen (2023) document diverging trends in market concentration at the national and local levels. My paper complements this work by providing direct estimates of output market power across local markets.

Relatedly, the second strand studies firm sorting across space. Gaubert (2018), Bilal (2023), Oh (2023), Lindenlaub, Oh, and Peters (2024), and Lhuillier (2025) examine sorting driven by agglomeration forces, monopsony power, and labor-market frictions. I depart from these studies by emphasizing sorting through competitive price pressures. Nocke (2006), Combes et al. (2012), and Matsuyama and Ushchev (2022) also study sorting through competitive price pressures, but

[3] The policy does not attain the first best because the resulting allocation of workers across locations is inefficient. The optimal policy instead requires lump-sum taxes on workers.

take market size as exogenous. In my model, market size is endogenously determined, allowing me to study additional general-equilibrium effects of firms' location choices. Finally, Behrens, Duranton, and Robert-Nicoud (2014) study the productivity advantages of large cities arising from agglomeration, firm sorting, and firm selection. I complement their analysis by quantifying the role of allocative efficiency in explaining cross-sectional productivity differences.

Spatial inefficiencies connect the paper to a third strand studying optimal spatial policies. Fajgelbaum and Gaubert (2020, 2025), Rossi-Hansberg, Sarte, and Schwartzman (2026), and Donald, Fukui, and Miyauchi (2026) characterize spatial policies in settings where the allocation of workers across space is inefficient. I contribute to this literature by focusing on the allocation of firms and characterizing optimal policies in a setting where spatial inefficiencies arise from firms' market power. I show that restoring efficiency requires policies that target firms directly through output subsidies that vary across markets. In this sense, the optimal policy is consistent with Bilal (2023), whose optimal policy also reallocates firms toward smaller markets. My characterization also highlights that spatial misallocation arises from endogenous pricing wedges, in contrast to the exogenous constraints studied by Hsieh and Moretti (2019), leading to different efficient spatial allocations. Finally, the optimal policy eliminates within-city markup variation, thereby alleviating resource misallocation across firms in the sense of Restuccia and Rogerson (2008) and Hsieh and Klenow (2009).

The role of markups as distortions also connects the paper to a fourth strand studying their aggregate implications. Edmond, Midrigan, and Xu (2023) quantify the aggregate welfare costs of markups in models of monopolistic competition with Kimball preferences and oligopolistic competition with nested CES demand. I adopt their monopolistic competition structure and extend it to a spatial setting with many local output markets. My main departure is to quantify the welfare costs of markups through an additional channel: the inefficient spatial allocation of firms. Other studies, including Peters (2020), De Loecker, Eeckhout, and Mongey (2022), Aghion et al. (2023), and Akcigit and Ates (2023), study the implications of markups for business dynamism. In contrast, I focus on how markups shape regional outcomes, including productivity and prices.

Finally, my empirical analysis relates to a literature studying markups across space. Hottman (2021), Anderson, Rebelo, and Wong (2023), and Sangani (2025) examine markup differences across cities within the retail sector. I depart from these studies in two ways. First, I estimate markups across all local industries rather than focusing only on retail. Second, I embed these estimates in a general-equilibrium model to evaluate place-based policies. I also find that retail markups are lower in larger cities, consistent with Hottman (2021). Anderson, Rebelo, and Wong (2023) and Sangani (2025) instead find higher markups in higher-income places. Their results therefore speak to a different source of cross-city variation than the market-size differences I study.

The rest of the paper is organized as follows. Section 2 develops the baseline theoretical framework and studies its efficiency properties, with a focus on the spatial misallocation of firms across cities. Section 3 extends the baseline framework for the quantitative analysis. Section 4 presents the empirical analysis, estimating markups across cities and testing the model's predictions. Section 5 presents the quantitative analysis, where I estimate the model and quantify the welfare effects of place-based policies. Finally, Section 6 concludes.

2 Baseline Model

This section develops a theory of spatial differentials in market power, where production takes place in locations I call cities. The theory abstracts from dynamics and describes a long-run steady state of the economy. I extend the model along several dimensions in Section 3.

2.1 Environment

Geography. There is a continuum of ex-ante heterogeneous cities, each endowed with one unit of housing. Cities differ in their local productivity, a, which is distributed according to the cumulative distribution function $F(a)$ with density $f(a)$ over the connected support $[\underline{a}, \overline{a}]$. I index cities by their productivity a rather than by their particular name.

Workers. There is a unit mass of freely mobile, identical workers. Each worker has one unit of labor, which is supplied inelastically. In every city a, workers derive utility from consuming a bundle of local varieties $Y(a)$, housing $H(a)$, and a freely traded good $Q(a)$, according to

$$U(a) = \left(\frac{Y(a)}{\eta}\right)^{\eta} \left(\frac{H(a)}{\alpha}\right)^{\alpha} \left(\frac{Q(a)}{1-\eta-\alpha}\right)^{1-\eta-\alpha}, \tag{1}$$

where η and α are the expenditure shares on local goods and housing. Workers have symmetric Kimball preferences (Kimball, 1995) over local varieties. These preferences belong to the Homothetic with Direct Implicit Additivity (HDIA) family, as defined by Matsuyama and Ushchev (2017). The per-capita consumption of the local goods bundle, $Y(a)$, is implicitly given by

$$\int_z \Upsilon\left(\frac{y(z,a)}{Y(a)}\right) dG_a(z) = 1, \tag{2}$$

where $y(z,a)$ is the per-capita consumption of a local variety produced by a firm with productivity z, $G_a(\cdot)$ is the endogenous local producer productivity distribution, and $\Upsilon(\cdot)$ is a strictly increasing and concave function satisfying $\Upsilon(0) = 0$.[4] CES preferences are a special case when $\Upsilon(q) = q^{\frac{\sigma-1}{\sigma}}$.

Kimball preferences offer three key advantages. First, they generate cross-sectional variation in markups, the central object of this paper. Second, they are homothetic, allowing the analysis to focus on markup differences across cities driven by price competition rather than income effects.[5] Third, despite their flexibility, Kimball preferences remain tractable enough to ensure a unique characterization of the equilibrium. Oligopolistic models with nested CES demand, such as Atkeson and Burstein (2008) and Edmond, Midrigan, and Xu (2015), provide an alternative framework for generating variable markups. However, characterizing entry in these models is challenging because they often feature multiple equilibria. In my setting, this problem is particularly difficult because firms make market-specific entry decisions.

[4] This notation anticipates that, in equilibrium, two firms with the same productivity in the same city make identical pricing decisions. Hence, workers consume the same amount of each variety produced by each firm.

[5] Income effects may also shape spatial differences in market power, but I leave them for future research. See Combes et al. (2012) for a spatial model with endogenous markups generated by non-homothetic preferences.

Local Goods. A mass of M_e potential entrants pays an entry cost c_e to learn their productivity z. This productivity is drawn from a common distribution with cumulative distribution function $G(\cdot)$, density function $g(\cdot)$, and connected support $[\underline{z}, \overline{z}]$. After learning their productivity, firms choose a city in which to produce and sell. This location decision determines the set of locally available varieties. Within a city, local producers compete monopolistically and produce according to the linear production function

$$y(z, a) = z\, l(z, a), \tag{3}$$

where $l(z, a)$ denotes labor. Local labor markets are competitive, and firms pay a constant wage $W(a)$. I assume free entry, so firms enter until aggregate profits are exhausted.

Traded Good. A perfectly competitive representative firm produces a homogeneous good in every location. This good is freely traded and used as the *numeraire*. Similar to local varieties, the traded good is produced with a linear production function:

$$Q^T(a) = a\, L^T(a), \tag{4}$$

where $Q^T(a)$ and $L^T(a)$ denote total traded good production and total employment in city a, respectively. Traded good producers compete with local goods producers in the same local labor market, paying the wage $W(a)$.

Housing. A representative mutual fund owns housing and rents it to workers at equilibrium rents $R(a)$. Profits from the mutual fund are rebated back to workers as a lump-sum transfer.

2.2 Consumption and Location Decisions of Workers

We start by characterizing workers' optimal consumption and location decisions. Workers face the budget constraint

$$\mathbb{P}(a)Y(a) + R(a)H(a) + Q(a) = W(a) + s, \tag{5}$$

where $\mathbb{P}(a)$ is the price of the bundle of local varieties, $R(a)$ is the housing price, s are the transfers from the mutual fund, and we have normalized the traded good price to one.[6] The homotheticity of Kimball preferences guarantees the existence of a price index for the bundle of local varieties. Therefore, workers maximize (1) subject to (2) and (5). The per capita consumption of local varieties, housing, and the traded good resulting from this maximization is given by

$$Y(a) = \frac{\eta(W(a)+s)}{\mathbb{P}(a)}, \qquad H(a) = \frac{\alpha(W(a)+s)}{R(a)}, \qquad \text{and} \qquad Q(a) = (1-\eta-\alpha)(W(a)+s). \tag{6}$$

Appendix A.1 shows that the per-capita consumption of an individual variety $y(z, a)$ is given by

[6] Throughout the text, I use blackboard bold font for indices. There are three: the ideal price index $\mathbb{P}(a)$, the price competition index $\mathbb{D}(a)$, and the competition index $\mathbb{C}(a)$.

$$\frac{y(z,a)}{Y(a)} = \varphi\left(\frac{p(z,a)}{\mathbb{D}(a)}\right), \tag{7}$$

where $p(z,a)$ is the price of a variety produced by a firm with productivity z in city a, $\varphi(\cdot) \equiv (\Upsilon')^{-1}(\cdot)$, and $\mathbb{D}(a)$ is a price index implicitly defined by

$$\int_z \Upsilon\left(\varphi\left(\frac{p(z,a)}{\mathbb{D}(a)}\right)\right) dG_a(z) = 1. \tag{8}$$

The expression in (7) represents the residual demand curve faced by local variety producers. For an individual firm, changes in other firms' prices are summarized by the price index $\mathbb{D}(a)$. Firms in every location compete against the price index $\mathbb{D}(a)$ when choosing their optimal price. Therefore, $\mathbb{D}(a)$ captures the degree of local competition, and I refer to it as the *competition price index*.

In contrast, the *ideal price index*, $\mathbb{P}(a)$, determines the price of the bundle of local varieties

$$\mathbb{P}(a) = \int_z p(z,a)\varphi\left(\frac{p(z,a)}{\mathbb{D}(a)}\right) dG_a(z). \tag{9}$$

While the competition price index $\mathbb{D}(a)$ governs the relative consumption between local varieties, the ideal price index $\mathbb{P}(a)$ determines the consumption of the bundle $Y(a)$ relative to housing and the traded good. These two price indices are proportional when $\Upsilon(\cdot)$ takes the CES form.

Workers' consumption decisions (6) imply that the indirect utility of workers in city a equals the real wage in that location. Therefore, the free mobility condition implies that

$$\frac{W(a)+s}{\mathbb{P}(a)^{\eta} R(a)^{\alpha}} = \overline{U}, \tag{10}$$

for all cities and some common utility level $\overline{U}$. Moreover, the local housing market equilibrium implies that $R(a) = \alpha L(a)(W(a)+s)$. Substituting this expression into (10) and solving for $L(a)$ yields the labor supply for city a:

$$L(a) = \kappa \frac{(W(a)+s)^{\frac{1-\alpha}{\alpha}}}{\mathbb{P}(a)^{\frac{\eta}{\alpha}}}, \qquad \kappa \equiv \left(\alpha \overline{U}^{\frac{1}{\alpha}}\right)^{-1}. \tag{11}$$

All else equal, cities with higher wages or lower local goods prices attract more workers.

2.3 Pricing and Location Decisions of Local Goods Producers

Firms producing local goods pay the entry cost, c_e, to learn their productivity. Then, they choose a city to operate and the price that maximizes profits. We solve this problem backwards.

2.3.1 Pricing Decision

The production function in (3) implies that a firm with productivity z located in city a chooses $p(z,a)$ to maximize

$$\pi(z,a) = \max_{p(z,a)} \left[p(z,a)y(z,a) - \frac{W(a)}{z} y(z,a) \right] L(a) \qquad \text{s.t.} \quad (7). \tag{12}$$

The first-order condition of this problem is

$$\frac{p(z,a)}{\mathbb{D}(a)} \left[1 - \frac{1}{\sigma\left(\frac{p(z,a)}{\mathbb{D}(a)}\right)} \right] = \frac{\mathbb{C}(a)}{z}, \tag{13}$$

where $\sigma(\cdot)$ is the price elasticity of demand implied by the residual demand curve (7),

$$\sigma\left(\frac{p(z,a)}{\mathbb{D}(a)}\right) \equiv -\frac{\partial \log y(z,a)}{\partial \log p(z,a)} = \frac{-\frac{p(z,a)}{\mathbb{D}(a)} \varphi'\left(\frac{p(z,a)}{\mathbb{D}(a)}\right)}{\varphi\left(\frac{p(z,a)}{\mathbb{D}(a)}\right)}, \tag{14}$$

and $\mathbb{C}(a)$ is a competition index summarizing local competitive pressures:

$$\mathbb{C}(a) \equiv \frac{W(a)}{\mathbb{D}(a)}. \tag{15}$$

The competition index $\mathbb{C}(a)$ accounts for competition in both local input and output markets. When wages increase, raising production costs, $\mathbb{C}(a)$ rises. Similarly, when other local producers lower their prices, the competition price index $\mathbb{D}(a)$ decreases, leading to an increase in $\mathbb{C}(a)$.

Under the concavity assumption of the aggregator $\Upsilon(\cdot)$, the first-order condition (13) defines a strictly increasing function $\psi(\cdot)$ that determines the optimal relative price $p(z,a)/\mathbb{D}(a)$:

$$\frac{p(z,a)}{\mathbb{D}(a)} = \psi\left(\frac{\mathbb{C}(a)}{z}\right). \tag{16}$$

The relative price (16) reveals two forces that determine prices within and across cities. First, within cities, more productive firms charge lower prices. Second, firms charge higher relative prices in more competitive cities, i.e. those cities with higher $\mathbb{C}(a)$.

The profit-maximizing markup is given by the usual Lerner formula that combines (14) and (16),

$$\mu\left(\frac{\mathbb{C}(a)}{z}\right) = \frac{\sigma\left(\psi\left(\frac{\mathbb{C}(a)}{z}\right)\right)}{\sigma\left(\psi\left(\frac{\mathbb{C}(a)}{z}\right)\right) - 1}. \tag{17}$$

The extent to which markups vary with relative prices is determined by the curvature of the

elasticity function (14). When $\sigma(\cdot)$ is an increasing function, this property is often referred to as Marshall's second law of demand (MSLD). In this case, firms charging higher relative prices face more elastic demand and set lower markups. When preferences are CES, $\sigma(\cdot)$ is constant, meaning all firms charge the same markup in all cities, regardless of local competition or their productivity.

We can combine (7) and (16) to solve for the optimal relative quantities and labor demand:

$$\frac{y(z,a)}{Y(a)} = \varphi\left(\psi\left(\frac{\mathbb{C}(a)}{z}\right)\right), \qquad \text{and} \qquad l(z,a) = \frac{\varphi\left(\psi\left(\frac{\mathbb{C}(a)}{z}\right)\right) Y(a)L(a)}{z}. \tag{18}$$

Finally, it is useful to define the consumer surplus ratio, $\delta(\cdot)$, which we will use later when studying the framework's normative implications. Following Baqaee, Farhi, and Sangani (2023), we define $\delta(\cdot)$ as the ratio of the area under the demand curve to firms' sales:

$$\delta\left(\frac{\mathbb{C}(a)}{z}\right) = \frac{\Upsilon\left(\varphi\left(\psi\left(\frac{\mathbb{C}(a)}{z}\right)\right)\right)}{\varphi\left(\psi\left(\frac{\mathbb{C}(a)}{z}\right)\right)\psi\left(\frac{\mathbb{C}(a)}{z}\right)} \geq 1. \tag{19}$$

2.3.2 Location Decision

We turn to analyze the location decisions of local goods producers. A producer with productivity z evaluates potential profits in every city and chooses the location that yields the highest profits.

Let $M(a)$ denote the total expenditure on local varieties in city a:

$$M(a) \equiv \mathbb{P}(a)Y(a)L(a). \tag{20}$$

We refer to $M(a)$ as the *market size* of city a, as it represents the potential revenue of local producers in that location.

Using (16), we can write the location problem a firm with productivity z as:

$$a^*(z) = \underset{a}{\operatorname{argmax}}\, \pi(z,a) = \underbrace{M(a)}_{\text{Market size}} \times \underbrace{\psi\left(\frac{\mathbb{C}(a)}{z}\right)\varphi\left(\psi\left(\frac{\mathbb{C}(a)}{z}\right)\right)\frac{\mathbb{D}(a)}{\mathbb{P}(a)}}_{\text{Market share}} \times \underbrace{\frac{\mu\left(\psi\left(\frac{\mathbb{C}(a)}{z}\right)\right) - 1}{\mu\left(\psi\left(\frac{\mathbb{C}(a)}{z}\right)\right)}}_{\text{Profit share}}. \tag{21}$$

Equation (21) reveals three forces shaping the location decisions of local producers. The first is market size, which reflects total expenditure on local varieties in city a. All producers prefer larger markets because they offer greater potential sales.

The second force is market share, which determines how much of total local expenditure each firm captures. The market share of firm z depends on its relative price, $\psi\left(\mathbb{C}(a)/z\right)$, and on the ratio $\mathbb{D}(a)/\mathbb{P}(a)$, which summarizes competitors' sales.[7] Firms that charge lower relative prices capture

[7] Indeed, note that (9) and (16) imply that: $\mathbb{P}(a)/\mathbb{D}(a) = \int_z \psi\left(\frac{\mathbb{C}(a)}{z}\right)\varphi\left(\psi\left(\frac{\mathbb{C}(a)}{z}\right)\right) dG_a(z)$.

larger market shares and therefore greater sales.

The final force is the share of sales that accrues as profits. Under CES demand, this share is constant, so profits are proportional to sales. Outside CES, demand elasticity varies across firms, and so does the fraction of sales converted into profits. Firms facing less elastic demand retain a larger share of their revenue as profits and therefore, all else equal, place greater value on locations where they can charge higher markups.

2.4 Traded Good Producers

We now characterize the problem of traded good producers. These producers are immobile and make only production decisions.

The production function (4) implies that the marginal cost of perfectly competitive producers is $W(a)/a$. Thus, the zero-profit condition in every city requires

$$W(a) = a, \tag{22}$$

implying that, in equilibrium, wages are higher in more productive cities. Finally, (4) determines total labor demand from traded good producers in city a:

$$L^T(a) = \frac{Q^T(a)}{W(a)}. \tag{23}$$

2.5 Equilibrium Definition and Characterization

Having established how workers and firms make consumption, production, and location decisions, I now define and characterize the conditions for a decentralized equilibrium.

First, define $L^N(a) \equiv \int_z l(z,a) dG_a(z)$, with $l(z,a)$ given by (18). Then, local housing and labor markets must clear in each city a:

$$R(a) = \alpha L(a)(W(a) + s), \qquad L(a) = L^N(a) + L^T(a), \tag{24}$$

where $L^T(a)$ is as in (23). The traded good market clears:

$$\int_a Q^T(a) dF(a) = \int_a L(a) Q(a) dF(a) + c_e M_e. \tag{25}$$

Free entry ensures that local producers' expected profits equal the entry cost, and the aggregate labor market clears:

$$\int_z \pi(z, a^*(z)) dG(z) = c_e, \qquad \int_a L(a) dF(a) = 1. \tag{26}$$

Finally, the mutual fund budget constraint is:

$$\int_a \alpha L(a)(W(a)+s)dF(a) = s. \tag{27}$$

A decentralized equilibrium consists of a mass of entering local producers M_e, a utility level $\overline{U}$, transfers s, an optimal price function $\psi(\cdot)$, a location choice function $a^*(z)$, a local productivity distribution $G_a(\cdot)$, a wage function $W(a)$, housing prices $R(a)$, a competition index function $\mathbb{C}(a)$, and population distribution $L(a)$ such that:

1. Workers make consumption and location decisions given wages and prices, as in (6) and (11)
2. Local producers maximize profits given location aggregates, as in (16).
3. Local producers optimally choose their location, as in (21), and the local productivity distribution $G_a(\cdot)$ is consistent with these choices.
4. Local labor and housing markets clear, as in (24).
5. The traded good market clears, as in (25).
6. Local producers make zero expected profits, and aggregate labor market clears, as in (26).
7. Mutual fund transfers are given by (27).

I next characterize the existence, uniqueness, and properties of the equilibrium.

Firstly, taking logs of (21) and deriving the first-order condition reveals the complementarities between firm productivity and local competition:

$$\underbrace{\frac{M'(a)}{M(a)} + \frac{\mathbb{D}'(a)}{\mathbb{D}(a)} - \frac{\mathbb{P}'(a)}{\mathbb{P}(a)}}_{\substack{\text{Marginal benefit from locating} \\ \text{in more productive cities}}} = \underbrace{\frac{\mathbb{C}'(a)}{\mathbb{C}(a)} \times \frac{1}{\mu\left(\frac{\mathbb{C}(a)}{z}\right) - 1}}_{\substack{\text{Marginal cost from locating} \\ \text{in more productive cities}}}. \tag{28}$$

The left-hand side of (28) captures the benefits of locating in more productive cities. The term $M(a)$ captures changes in market size, while $\mathbb{D}(a)$ and $\mathbb{P}(a)$ reflect changes in market share.[8] Together, these terms determine the marginal revenue generated by moving toward more productive cities. Importantly, this component is identical across firms, irrespective of their productivity.

The right-hand side captures the cost of operating in more competitive cities. The first term reflects changes in the competition index, $\mathbb{C}(a)$, while the second determines how these changes affect firms with different productivity. In particular, stronger competition reduces profits in proportion to $1/(\mu(\cdot)-1)$. Under MSLD, more productive firms charge higher markups and are therefore less affected by stronger competition. I refer to this complementarity between productivity and markups as *pricing complementarities*. As a result, more productive firms benefit relatively more from the additional sales available in larger markets.

[8] The re-optimization of prices across cities has no first-order effect on profits, as shown in Appendix B.1.

In summary, firms face a trade-off when choosing where to locate. Larger markets offer greater sales but also tougher competition and lower markups, while smaller markets offer lower sales but weaker competition and higher markups. Pricing complementarities imply that this trade-off affect firms differentially as more productive firms charge higher markups and are therefore less affected by stronger competition. Market size and local competition are endogenous, so the strength of these forces is determined in equilibrium. Proposition 1 shows that, under MSLD, more productive firms self-select into more productive cities, which are larger and endogenously more competitive.

To formally characterize the problem in (21), I restrict attention to parameterizations of $\Upsilon(\cdot)$ that satisfy MSLD. Empirical evidence supports this feature of consumer preferences, as shown by De Loecker and Goldberg (2014) and Amiti, Itskhoki, and Konings (2019).[9] For clarity, I adopt the functional form proposed by Klenow and Willis (2016), which satisfies MSLD. Appendix B provides conditions under which general Kimball aggregators satisfying MSLD ensure the main theoretical predictions of the model hold.

Assumption 1 (Parametric form of Kimball aggregator)**.** *The Kimball aggregator $\Upsilon(\cdot)$ is given by the Klenow and Willis (2016) functional form:*

$$\Upsilon(q) = (\overline{\sigma} - 1) \exp\left(\frac{1}{\varepsilon}\right) \varepsilon^{\frac{\overline{\sigma}}{\varepsilon}-1} \left[\Gamma\left(\frac{\overline{\sigma}}{\varepsilon}, 0\right) - \Gamma\left(\frac{\overline{\sigma}}{\varepsilon}, \frac{q^{\varepsilon/\overline{\sigma}}}{\varepsilon}\right)\right], \tag{29}$$

where $\overline{\sigma} > 1$, $\varepsilon > 0$, and $\Gamma(\cdot, \cdot)$ is the Upper incomplete Gamma function,

$$\Gamma(x_1, x_2) = \int_{x_2}^{\infty} t^{x_1 - 1} e^{-t} dt.$$

Appendix A.2 derives expressions for the price elasticity $\sigma(\cdot)$, relative price $\psi(\cdot)$, relative quantities $\varphi(\cdot)$, and markup $\mu(\cdot)$ under the functional form (29) for the Kimball aggregator.

I define an *assignment pair* as a pair of functions $a \to (z(a), M(a))$, where $z(a)$ is the assignment function mapping local producers to cities, and is the inverse of $a^*(z)$.[10] The function $M(a)$ represents the equilibrium market size that sustains this location choice.

Proposition 1 (Sorting)**.** *Suppose Assumption 1 holds and that $\alpha > \eta/(\overline{\sigma} - 1)$. Then, there exists a threshold $\underline{\varepsilon}$ such that for all $\varepsilon \in (0, \underline{\varepsilon}]$, every solution to (21) exhibits positive assortative matching. The functions $z(a)$, $M(a)$, and $\mathbb{C}(a)$ are strictly increasing.*

Proof. See Appendix B.1. □

Proposition 1 establishes the existence and uniqueness of the assignment between city productivity a and local producer productivity z, featuring positive assortative matching (PAM): more productive firms locate in more productive cities, $z'(a) > 0$, which are larger, $M'(a) > 0$, and more competitive, $\mathbb{C}'(a) > 0$. Intuitively, higher wages in more productive cities attract more

[9] Melitz (2018) provides a detailed discussion on the implications of violating MSLD.

[10] As Proposition 1 shows, $z(a)$ is strictly increasing, ensuring that its inverse is well defined.

workers, expanding market size. This, in turn, incentivizes local producers to enter, intensifying competition. Due to pricing complementarities, only the most productive firms can withstand the heightened competition, while lower-productivity firms settle in smaller cities where competition is weaker. This proposition extends the sorting results of Nocke (2006) and Matsuyama and Ushchev (2022) to a setting with mobile workers, where local population is determined in equilibrium.

The parameter restrictions in Proposition 1 ensure the existence of a unique assignment. The term α captures the only congestion force in this framework: housing. As workers concentrate in a city, housing costs rise, limiting further in-migration. Conversely, when ε is sufficiently small, workers' valuation of an additional local variety is $\eta/(\overline{\sigma}-1)$, representing the main agglomeration force in this setting. Thus, the condition $\alpha > \eta/(\overline{\sigma}-1)$ mirrors the standard requirement for uniqueness of equilibrium in spatial models: congestion forces must outweigh agglomeration forces (Redding & Rossi-Hansberg, 2017). Proposition 2 formalizes this by proving that, under the assumptions of Proposition 1, a unique equilibrium exists.

Proposition 2 (Existence and Uniqueness)**.** *Under the assumptions of Proposition 1, a decentralized equilibrium with positive assortative matching exists. Furthermore, if the supports of $G(\cdot)$ and $F(\cdot)$ are sufficiently small, the decentralized equilibrium is unique.*

Proof. See Appendix B.2. □

It is possible to gain additional insight into how differences in local competition depend on the sorting of local producers by considering a particular limiting equilibrium. Suppose that ex-ante differences across cities become arbitrarily small, meaning all cities have nearly identical productivity. In this case, any ex-post differences across cities are solely determined by differences in the pool of local producers. Corollary 1 below shows that spatial differentials in local competition can arise even in the absence of ex-ante heterogeneity between locations.

Corollary 1 (Equilibrium with Ex-Ante Identical Locations)**.** *Suppose the assumptions of Proposition 1 hold. Then, as exogenous differences in productivity a vanish, there exists a limiting decentralized equilibrium in which differences in market size and local competition persist, and more productive firms locate in larger, more competitive cities.*

Proof. See Appendix B.3. □

The results in Corollary 1 highlight that pricing complementarities drive spatial differences in local competition. In the limiting economy, cities are ex-ante identical, but ex-post differences arise endogenously due to the spatial sorting of local producers. On the one hand, pricing complementarities allow more productive firms to better adapt to environments with intense competition, leading them to locate in larger cities where competition is endogenously tougher. On the other hand, the least productive firms find it optimal to locate in smaller cities where competition is weaker. Since locations do not differ ex-ante along any dimension, city size and local competition levels are determined entirely by the location choices of local producers.

Corollary 1 is a natural result since there are no technological complementarities between local producer productivity and city productivity. Local producers value city productivity only through its effect on city size. When cities have identical productivity, firms still sort across space due to endogenous differences in city size and local competition. Naturally, city size and local competition are determined in equilibrium by the location choices of local producers. Corollary 1 highlights that in the limiting economy, the joint determination of these three factors leads to a non-degenerate equilibrium in which larger cities are more competitive and attract more productive firms.

2.6 Spatial Misallocation

Having established the positive implications of the framework, I now turn to its normative properties. The exposition closely follows Fajgelbaum and Gaubert (2020) to highlight the inefficiencies in firms' location decisions that arise from markups.

I characterize the optimal allocation using a utilitarian social planner who maximizes workers' utility. The planner chooses the distribution of workers across cities, $L(a)$, and their consumption of local goods, $Y(a)$, the traded good, $Q(a)$, and housing, $H(a)$, subject to free mobility and local housing constraints. The planner also allocates labor between traded- and local-good production and determines the location of local-good producers, $z(a)$. The formal definition of the planning problem is relegated to Appendix C, while the main text focuses on characterizing the solution.

Inefficiencies arise when decentralized equilibrium outcomes differ with the ones from the planning problem. Let $E(a)$ denote the per-capita expenditure of workers in location a. In the decentralized equilibrium, this object is given by the left-hand-side of (5). The following results highlights the inefficiencies in the distribution of workers and firms across locations.

Proposition 3 (Efficient Allocation)**.** *If a competitive equilibrium is efficient, then*

$$W(a) = E(a) + \varkappa, \qquad \frac{M'(a)}{M(a)} + \frac{\mathbb{D}'(a)}{\mathbb{D}(a)} - \frac{\mathbb{P}'(a)}{\mathbb{P}(a)} = \frac{\mathbb{C}'(a)}{\mathbb{C}(a)} \times \frac{1}{\delta\left(\frac{\mathbb{C}(a)}{z(a)}\right) - 1}, \tag{30}$$

for some constant $\varkappa$. Additionally, the value of local output equals the local wage bill, $M(a) = W(a)L^N(a)$, and the mass of entrants satisfies $c_e M_e = \int M(a)\left[\delta\left(\mathbb{C}(a)/z(a)\right) - 1\right] f(a)\, da$.

Proof. See Appendix C.2. □

Proposition 3 highlights the sources of inefficiency in the decentralized equilibrium. The first condition in (30) is a special case of Proposition 1 in Fajgelbaum and Gaubert (2020) and equates the planner's marginal benefit of locating a worker in location a to its marginal cost. In my setting, there are no explicit worker agglomeration or congestion externalities, so the marginal benefit is the marginal product of labor, $W(a)$. The marginal cost is the per-capita expenditure in that location, together with $\varkappa$, the Lagrange multiplier on the planner's national labor constraint, expressed in units of the traded good. It captures the opportunity cost of assigning a worker to location a.

The second condition in (30) concerns firms' location decisions and is the novel part of my framework. It is the planner's counterpart to (28) and states that, in the efficient allocation, the social

marginal benefit of locating a firm in a more productive location equals the social marginal cost. Importantly, the social marginal benefit is still determined by changes in market size and market share, as in (28). The social marginal cost, however, is proportional to $1/(\delta(\cdot)-1)$ rather than $1/(\mu(\cdot)-1)$, where $\delta(\cdot)$ is the consumer-surplus ratio defined in (19). This difference reflects the entry externalities described below and shows that efficiency requires firms' location decisions to be governed by the "social markup", $\delta(\cdot)$, rather than by private markups. This result extends the intuition from previous studies of firm entry and market power to a spatial setting (Zhelobodko et al., 2012; Bilbiie, Ghironi, & Melitz, 2019; Dhingra & Morrow, 2019; Baqaee, Farhi, & Sangani, 2023; Edmond, Midrigan, & Xu, 2023).

The remaining conditions capture additional margins of inefficiency. Markups distort the relative consumption of local goods, housing, and the traded good. The condition $M(a) = W(a)L^N(a)$ implies that, in an efficient allocation, firms price at marginal cost, eliminating this distortion. Finally, markups also distort the aggregate entry margin. As in the firm-location condition, the efficient free-entry condition uses $\delta(\cdot)$ to compute social profits, rather than $\mu(\cdot)$, which determines private profits, as shown in equation (21). Because firms sort perfectly across cities, there is no within-city markup variation and therefore there is no misallocation of production inputs. I introduce this margin in the extended model.

Proposition 3 shows that the decentralized equilibrium is inefficient because of firms' markups. First, equation (30) shows that the allocation of workers is efficient whenever the transfers in the decentralized equilibrium satisfy $s = -\varkappa$. All the inefficiencies in the economy therefore arise from firms' market power rather than from workers' externalities. Second, when consumer demand is not CES, the location decision of local producers is inefficient, since social returns, $\delta(\cdot)$, differ from private returns, $\mu(\cdot)$. In addition, the aggregate entry margin and the relative consumption of goods are also inefficient. Third, in the limiting case of CES preferences, the location decision of local producers is efficient given the allocation of workers, since $\delta = \mu$. Nevertheless, the aggregate entry margin and the relative consumption of goods remain inefficient under constant markups. In this case, the allocation of workers is also inefficient because s does not coincide with $-\varkappa$.

Hence, firm spatial misallocation is inherently a product of firms' endogenous markups. This distinction separates the place-based component of the optimal policy from its uniform component.

Proposition 4 (Optimal Policy)**.** *The efficient allocation is attained by:*

1. *A location-specific output subsidy increasing with market size and firm relative output:*

$$T(y(z,a),a) = \tau\left(\frac{y(z,a)}{Y(a)}\right)\iota(a)M(a), \tag{31}$$

$$\tau\left(\frac{y(z,a)}{Y(a)}\right) \equiv \Upsilon\left(\frac{y(z,a)}{Y(a)}\right) - \Upsilon'\left(\frac{y(z,a)}{Y(a)}\right)\frac{y(z,a)}{Y(a)}, \qquad \tau'\left(\frac{y(z,a)}{Y(a)}\right) > 0, \qquad \iota(a) \equiv \frac{\mathbb{D}(a)}{\mathbb{P}(a)}.$$

2. *A lump-sum tax to workers, $t = -(s + \varkappa)$.*

Proof. See Appendix C.3. □

Proposition 4 shows that the first-best allocation can be implemented through a location-specific output subsidy for firms and a lump-sum tax on workers. The subsidy to firms eliminates markups, thereby correcting the three margins of inefficiency created by market power. The tax finances the subsidy and, at the same time, corrects the allocation of workers. In the decentralized equilibrium, firms pay the entry cost out of markup profits. Once the subsidy eliminates markups, entry is financed with a lump-sum tax on workers. This tax equals the amount by which workers' expenditure exceeds its efficient level, so it also restores the efficient worker allocation. The two instruments therefore mirror the two location margins. The tax on workers is uniform across space because, conditional on the location of firms, worker allocation is efficient. The subsidy to firms is location-specific because markups vary across space in the decentralized equilibrium.

The transfers that implement the efficient allocation are connected to previous work. The lump-sum tax on workers is a special case of the transfers in Proposition 2 in Fajgelbaum and Gaubert (2020). The firm transfers in (31) extend the optimal subsidies of Edmond, Midrigan, and Xu (2023) under their Kimball specification to a multi-market setting. These transfers contain a negative component proportional to $\Upsilon'\left(y(z,a)/Y(a)\right)\left(y(z,a)/Y(a)\right)$ and a positive component proportional to $\Upsilon\left(y(z,a)/Y(a)\right)$. The first effectively taxes the sales generated under the firm's revenue curve, while the second compensates the firm for the consumer surplus generated by its production. Together, these two components induce firms to produce at marginal cost and set markups equal to one in all locations. They also replace the markup governing firms' private returns with the consumer-surplus ratio governing social returns, thereby correcting the location and aggregate entry margins. Importantly, the subsidy scales with market size, $M(a)$, giving the policy a place-based component. Intuitively, holding relative output fixed, firms in larger markets receive larger transfers because they generate more consumer surplus through greater sales.

Why does output market power lead to spatial misallocation? The answer lies in two externalities associated with local entry. First, when firms enter a market, they create a positive externality, known as the *variety-gains externality*. Because consumers value variety, the entry of a new firm lowers the ideal price index, increasing aggregate expenditure on local goods.[11] Second, because local varieties are imperfect substitutes, new producers reduce demand for existing varieties, lowering the profits of incumbent firms. This negative externality is referred to as the *business-stealing externality*. These externalities are not unique to my framework, as they arise in standard models of firm entry and differentiated varieties (Dixit & Stiglitz, 1977). With constant markups, however, the two externalities are constant and offset each other. Once markups vary across firms and markets, the two externalities no longer offset each other, creating spatial distortions in firm entry.

Whether there is too much or too little entry in a city depends on the relative strength of the two local externalities. When the variety-gains externality dominates business stealing, there is insufficient entry, while the opposite leads to excessive entry. The balance between the two externalities can be understood by comparing (28) and (30). To ease notation, denote $\delta(a) \equiv \delta\left(\mathbb{C}(a)/z(a)\right)$ and $\mu(a) \equiv \mu\left(\mathbb{C}(a)/z(a)\right)$. Holding everything else fixed, when $\delta(a) > \mu(a)$, the variety-gains externality dominates: locating in more productive cities is relatively more costly for

[11] Mankiw and Whinston (1986) provide an alternative interpretation of the variety-gains externality. Entry increases consumer surplus, but firms capture only part of that surplus in their profits.

firms in the decentralized equilibrium than in the efficient allocation, leading to insufficient entry in these locations. Conversely, when $\mu(a) > \delta(a)$, the business-stealing externality dominates: locating in more productive cities is relatively less costly in the decentralized equilibrium, leading to excessive entry in these locations.

Whether these externalities lead to excessive or insufficient concentration of firms in larger markets cannot be determined at this level of generality. First, as highlighted by Dhingra and Morrow (2019), the relationship between $\delta(a)$ and $\mu(a)$ depends on the exact parametrization of demand. Second, the differences between (28) and (30) extend beyond the cost of locating in more productive markets. Aggregate objects generally differ between the decentralized and efficient allocations, so the benefits of locating in larger markets may differ as well. Finally, transfers differ across the two allocations, changing the spatial distribution of market size and therefore its gradient, $M'(a)/M(a)$.

To make progress, I compare the decentralized equilibrium with an allocation that preserves the same structure of worker transfers while correcting firms' market power. Consider a version of the economy in which profits from the national portfolio are rebated to workers through a uniform wage subsidy, so that income in location a is $W(a)(1+s)$. Starting from this allocation, I implement the output subsidy in Proposition 4 and finance it with a uniform tax on labor income, so that workers' income becomes $W(a)(1+t)$. I refer to the resulting allocation as the *quasi-optimal policy*.

The quasi-optimal policy eliminates markups and therefore corrects firms' location decisions, the relative consumption of goods, and aggregate entry. It differs from the first best only in the treatment of workers. Rather than using a lump-sum tax to set the level of workers' expenditure, it preserves income proportional to local wages. Because income is proportional to wages in both the decentralized and quasi-optimal allocations, the comparison preserves the same worker-transfer structure and isolates the distortions in firms' location decisions. This is also the policy convention I adopt in the extended model and in the quantitative analysis. Let $z^*(a)$ denote the productivity of local producers in location a under the quasi-optimal policy.

Proposition 5 (Spatial Misallocation)**.** *Under Assumption 1 and the assumptions of Proposition 1, there exists $\widetilde{\varepsilon} > 0$ such that, for $\varepsilon \in (0, \widetilde{\varepsilon})$,*

$$z(a) < z^*(a), \qquad \text{for all} \qquad a \in (\underline{a}, \overline{a}). \tag{32}$$

Proof. See Appendix C.2. □

Proposition 5 shows that productive producers are over-concentrated in larger markets. The quasi-optimal policy reallocates them toward smaller cities, so that interior city hosts more productive firms. Under Assumption 1 and when ε is not too large, $\mu(a) > \delta(a)$, so the private cost of locating in more competitive markets is lower than the social cost. The business-stealing externality therefore dominates, leading productive firms to locate excessively in larger markets. This result is similar to Dhingra and Morrow (2019) and Matsuyama and Ushchev (2021), who show that entry is excessive in single-market models with homogeneous firms under MSLD. The proportional structure of worker transfers is key for this comparison. The elasticity of market size with respect

to productivity does not depend on the level of transfers, so the policy's financing does not affect the spatial gradient of market size or firms' marginal returns to locating in larger markets.

This section has outlined a general equilibrium model in which spatial differences in markups arise endogenously from the location choices of heterogeneous local producers. The framework shows how markups distort the allocation of firms across cities, leading to an overconcentration of the most productive firms in the largest markets. The baseline model remains intentionally parsimonious in its treatment of spatial heterogeneity, workers' idiosyncratic location preferences, housing supply, and within-city variation in markups across firms. In the next section, I extend the framework along these dimensions. I then take the extended model to the data in Section 5 and use it to quantify the welfare effects of alternative place-based policies.

3 Extended Model

In this section, I first describe the extensions of the baseline model. Then, I demonstrate how the results from Section 2 carry over to this richer environment.

3.1 Environment

The extended model is as in Section 2, with the following additions:

Productivity Realization. Local producers' productivity is now determined in two stages. As in the baseline framework, firms first pay the entry cost c_e and draw an initial productivity z, based on which they choose where to locate. After locating, firms draw their final productivity ω from a conditional distribution with CDF $G(\omega|z)$ and PDF $g(\omega|z)$. Firms then observe ω and make their pricing and production decisions. This extension generates an ex-post, non-degenerate productivity distribution among local producers within each city while preserving positive sorting across cities based on firms' initial productivity draw, z. It also introduces an additional margin of inefficiency: within-city markup variation leads to misallocation of production inputs.

Local Amenities. Cities now differ in both traded-good productivity, a, and local residential amenities, b. I index locations by the productivity-amenity pair $c = (a, b)$, which follows a cumulative distribution function $F_c(a, b)$ with density $f_c(a, b)$. Residential amenities capture attributes that attract workers to particular locations, such as climate or local public goods. This additional dimension of city heterogeneity allows the model to jointly match the distributions of population and wages across cities.

Workers' Idiosyncratic Location Preferences. Workers are now heterogeneous in their valuations of specific locations, and I index them by i. Worker i observes a set of idiosyncratic location-specific preference shocks, $\varsigma_i(c)$, and chooses where to reside. If worker i locates in c, utility is given by

$$U_i(c) = b(c) \left(\frac{Y(c)}{\eta}\right)^{\eta} \left(\frac{H(c)}{\alpha}\right)^{\alpha} \left(\frac{Q(c)}{1-\eta-\alpha}\right)^{1-\eta-\alpha} \varsigma_i(c). \tag{33}$$

The idiosyncratic preference shock $\varsigma_i(c)$ is independently and identically distributed across workers and locations and follows a Fréchet distribution with shape parameter θ. Under the preferences in (33), real income no longer equalizes across cities in equilibrium. Consequently, θ determines the migration elasticity with respect to real income in the quantitative model.

Housing in Production. Local and traded goods producers now use housing in production, which I refer to as commercial structures. A local goods producer in city c with final productivity ω produces according to a Cobb-Douglas production function:

$$y(\omega, c) = \omega \, l(\omega, c)^{\beta} h(\omega, c)^{1-\beta}, \tag{34}$$

where $l(\omega, c)$ is labor input, $h(\omega, c)$ is the demand for commercial structures (buildings), and $\beta \in (0, 1)$ represents the labor elasticity of output.

Similarly, traded goods producers in city c produce according to:

$$Q^T(c) = a(c) \, \left(L^T(c)\right)^{\gamma} \left(H^T(c)\right)^{1-\gamma}, \tag{35}$$

where $L^T(c)$ is employment, $H^T(c)$ is commercial structures, and $\gamma \in (0, 1)$ is the labor elasticity of output. By incorporating housing into production, the quantitative model captures an additional local congestion force, as a larger local population drives up production costs.

Housing Supply. Perfectly competitive land developers use the traded good to produce housing and commercial structures in every city, according to the isoelastic production function:

$$\overline{H}(c) = \left(\frac{1+\phi}{\phi} \, \overline{Q}(c) \right)^{\frac{\phi}{1+\phi}}, \tag{36}$$

where $\overline{H}(c)$ is the total supply of buildings in city c (housing and commercial structures) and $\overline{Q}(c)$ is the amount of traded good used for buildings.

Ownership. The representative mutual fund aggregates profits from land developers and rebates them to workers through a flat wage subsidy.

3.2 Characterization

Appendix D illustrates how the quantitative extensions preserve the fundamental structure of local producers' production and location decisions. On the one hand, the firms' pricing problem in Section 2.3.1 continues to hold, with two key modifications: (i) the competition index now accounts for local structures prices, and (ii) firms make pricing decisions based on their final productivity.

On the other hand, the location choices of local producers in Section 2.3.2 follow the same structure, with two modifications. First, firms choose where to locate based on a composite index, x, which

I refer to as a city's *appeal*. Although a city is characterized by a traded-good productivity and local amenities pair, (a, b), Appendix D shows that firms value these characteristics only through:

$$x \equiv x(a,b) \;=\; \left(a^{\frac{1+\theta(1-\eta\beta)}{\gamma}}\, b^{\theta}\right)^{\frac{1}{\theta\xi}}, \tag{37}$$

where ξ is a positive constant defined in Proposition 6 below. Intuitively, traded-good productivity shapes wages, while amenities influence population. Hence, city size, measured by total expenditure on local goods, depends on a combination of these two characteristics, and firms therefore sort based on the appeal index x rather than the pair (a, b).[12]

Second, given the two-stage realization of productivity, firms make location choices based on expected profits. Indexing cities by their appeal, x, the location problem (21) becomes

$$x^*(z) = \underset{x}{\operatorname{argmax}}\, \Pi(z,x) \equiv \mathbb{E}_{\omega}[\pi(\omega,x)] = \int_{\omega} \frac{M(x)\psi\left(\frac{\mathbb{C}(x)}{\omega}\right)\varphi\left(\psi\left(\frac{\mathbb{C}(x)}{\omega}\right)\right)\overline{\delta}(x)}{\sigma\left(\psi\left(\frac{\mathbb{C}(x)}{\omega}\right)\right)} dG(\omega|z), \tag{38}$$

where $\overline{\delta}(x)$ is the sales-weighted mean consumer surplus ratio, which is equal to the ratio of the local price indices.

The location problem (38) is a standard stochastic optimization problem. Because final productivity is not observed at the time of location choice, firms choose locations based on expected profits. Using standard tools of monotone comparative statics under uncertainty (Athey, 2002), Proposition 6 extends the baseline sorting and equilibrium results under the following assumption on the destination productivity distribution.

Assumption 2 (Log-Supermodularity in Destination Productivity)**.** *The PDF of the destination productivity distribution for local producers, $g(\omega|z)$, is strictly log-supermodular in (ω, z).*

Proposition 6 (Sorting, Existence, and Uniqueness of Equilibrium in the Extended Model)**.** *Suppose Assumptions 1 and 2 hold. Moreover, assume that $\xi > \eta/(\overline{\sigma} - 1)$, where*

$$\xi \equiv \frac{\gamma\Big(1+\phi+\theta(\alpha+\eta(1-\beta))\Big)+(1-\gamma)\Big(1+\theta(1-\eta\beta)\Big)}{\theta\gamma(1+\phi)}.$$

Then the results of Propositions 1 and 2 and Corollary 1 carry over to the extended model with three modifications. First, cities are indexed by their appeal, x, defined in (37). Second, market size $M(x)$, firm location $z(x)$, and local housing-structure rents $R(x)$ depend only on city appeal, whereas the price indices $\mathbb{D}(x,a)$ and $\mathbb{P}(x,a)$, wages $W(x,a)$, and population $L(x,a)$ depend on both city appeal and local productivity. Third, the competition index is now defined as $\mathbb{C}(x) \equiv \nu W(x,a)^{\beta} R(x)^{1-\beta}/\mathbb{D}(x,a)$, where $\nu \equiv 1/[\beta^{\beta}(1-\beta)^{1-\beta}]$, and depends only on x.

Proof. See Appendix D. □

[12] Firms are indifferent between locating in cities $c = (a, b)$ and $c' = (a', b')$ whenever $x(a, b) = x(a', b')$.

It is worth highlighting how the additional model ingredients affect the sorting and equilibrium results. First, under Assumption 1, the profit function in (21) is log-supermodular in x and z. Assumption 2 ensures that expected profits in (38) preserve this property, so the PAM result carries over to the extended framework.

Second, the parameter ξ generalizes the congestion force in the baseline model. In addition to the housing expenditure share α, it incorporates the housing supply elasticity ϕ, the dispersion of idiosyncratic location preferences θ, and the input elasticities β and γ, since production now also uses housing. Together, these parameters determine how the costs of locating in a city change with its size. The condition $\xi > \eta/(\overline{\sigma} - 1)$ ensures the existence of the PAM equilibrium in the extended model when ε is not too large. When $\beta = \gamma = 1$, $\phi = 0$, and $\theta \to \infty$, ξ converges to α, and the condition reduces to that in Proposition 1.

Finally, price indices, wages, and population can no longer be summarized solely by city appeal. Wages are pinned down by the zero-profit condition of traded-good producers and therefore depend directly on local productivity a. Workers also value amenities directly, whereas local producers value them only through their effects on market size and structure rents.[13]

Aggregate Objects. Two aggregate objects determine misallocation and local welfare in the extended model. The first is the city-level aggregate markup, which influences the location decisions of local producers. The second is local TFP, which captures misallocation within and across cities induced by markups. Using standard aggregation results (Edmond, Midrigan, & Xu, 2023), these objects are given by

$$\mathcal{M}(x) \equiv \left[\int_{\omega} \frac{s^{y}(\omega, x)}{\mu\left(\frac{\mathbb{C}(x)}{\omega}\right)} dG_{x}(\omega)\right]^{-1}, \qquad \mathcal{Z}(x) \equiv \left[\int_{\omega} \frac{\varphi\left(\frac{\mathbb{C}(x)}{\omega}\right)}{\omega} dG_{x}(\omega)\right]^{-1}, \tag{39}$$

where $s^{y}(\omega, x) \equiv \psi\left(\frac{\mathbb{C}(x)}{\omega}\right) \varphi\left(\frac{\mathbb{C}(x)}{\omega}\right) \overline{\delta}(x)$ denotes the sales share of firm ω in city x. The distribution $G_{x}(\cdot)$ denotes the productivity distribution of local producers in city x, with density $g(\omega|z(x))$, while $\varphi(\mathbb{C}/\omega)$ and $\mu(\mathbb{C}/\omega)$ are shorthand for $\varphi(\psi(\mathbb{C}/\omega))$ and $\mu(\psi(\mathbb{C}/\omega))$, respectively. The aggregate markup is a sales-weighted harmonic mean of firm-level markups, while aggregate productivity is a firm-size-weighted harmonic mean of firm-level productivity.

Spatial Misallocation in the Extended Model. In principle, Propositions 3 and 4 can be extended to the richer environment. If cities differ only in local productivity, abstracting from amenities, the propositions continue to apply with some modifications. In particular, the aggregate markup and aggregate consumer surplus enter conditions (28) and (30) instead of their firm-level counterparts. Moreover, worker transfers include a labor tax equal to $1/(1 + \theta)$, as in Section III.E of Fajgelbaum and Gaubert (2020). When cities differ in both productivity and amenities, however, characterizing the efficient allocation becomes substantially more difficult. As in the baseline model, efficiency requires lump-sum transfers to workers. These transfers imply that x is no longer a sufficient statistic for firms' location choices, which instead depend separately on

[13] Local population initially depends directly on local amenities. For consistency with the notation above, I express amenities as a function of x and a using (37).

local productivity and amenities. Characterizing the efficient allocation therefore requires solving a multidimensional assignment problem, which has proved particularly difficult.[14] For this reason, in the extended model I rebate transfers proportionally to wages rather than as lump-sum payments.

In Section 5, I implement the optimal location-specific output subsidy for firms from Proposition 4 within the extended model. As in the baseline model, this policy eliminates markups for all firms, thereby alleviating the main source of inefficiency. Importantly, relative to the baseline model, the policy also eliminates within-city markup variation, addressing an additional source of production misallocation generated by the richer productivity process for local producers. I avoid lump-sum transfers and instead finance the output subsidy with a flat proportional tax on labor income. Workers receive the profits of land developers through a flat wage subsidy, together with transfers from the national portfolio net of taxes. Because this transfer structure may not achieve the first best, I refer to this allocation as the quasi-optimal policy, as in Section 2.6.

This section extended the baseline model along several dimensions to better capture spatial heterogeneity across cities. Relative to the baseline, the extended model features both the misallocation of firms across cities and the misallocation of production within cities. I study the quantitative importance of this additional margin in Section 5. Before turning to the model quantification, the next section introduces the data and estimates markups for local producers across U.S. cities.

4 Empirical Analysis

In this section, I estimate firm-level markups across U.S. cities. I begin by describing the data and defining local industries. I then present evidence that, consistent with the model, firms in larger markets appear to be more productive than those in smaller markets. Next, I estimate markups for firms in local industries. These firm-level estimates constitute the key empirical input used to estimate the parameters that discipline the quantitative analysis. Finally, I examine how aggregate markups vary across cities. This cross-sectional analysis provides evidence on the mechanisms highlighted by the model, but it is not required for the subsequent quantification.

4.1 Data

My primary dataset is the microdata from the U.S. Census Bureau's Longitudinal Business Database (LBD). The LBD is constructed from administrative employment records and covers nearly every nonfarm private-sector establishment in the United States. For my analysis, I use information on employment, wage bills, geographic location at the county and commuting-zone levels, six-digit NAICS industries, and establishment identifiers.

I supplement the LBD with sales data from the Economic Censuses (EC), which are conducted every five years. I use microdata from the 2002, 2007, 2012, and 2017 censuses of Construction Industries, Manufacturing, Retail Trade, Services, Wholesale Trade, Finance, Insurance and Real Estate, and Transportation, Communications and Utilities. I link establishments in the Economic

[14] See Boerma, Ottolini, and Tsyvinski (2025) for a discussion of such problems and a characterization in the context of workers and firms with general output functions and distributions.

Censuses to those in the LBD using establishment identifiers. The resulting sample consists of LBD establishments matched to sales data from the Economic Censuses.[15] I use 2017 as the baseline year and the earlier census years for robustness exercises.

I use the Census of Manufactures to conduct additional markup estimation exercises. It contains detailed establishment-level information on expenditures on intermediate inputs, capital (equipment and structures), and energy. Comparable information is not available in the other Economic Censuses. I construct measures of capital, materials, and labor following standard procedures in the literature (Foster, Grim, & Haltiwanger, 2016).

I map cities in the model to commuting zones in the data, following Autor and Dorn (2013). My sample covers 722 commuting zones in the continental United States. Under this mapping, each commuting zone corresponds to a collection of nearby model locations. Formally, a commuting zone represents an interval $[x, x + dx]$ in the model's location space. In some robustness exercises, I instead define cities at the county level.

4.2 Local Industries

I use the classification of Delgado, Porter, and Stern (2015) to distinguish between local and traded establishments in the LBD. Broadly, their approach classifies six-digit NAICS industries as either "Local" or "Traded" based on employment specialization, geographic concentration, and proximity to final consumers. Local industries are present in most geographic areas and primarily serve local consumers. By contrast, traded industries serve customers in other regions and may be geographically concentrated. The classification assigns 310 six-digit NAICS industries to the local category and 778 to the traded category.[16] Using the industry codes in the LBD, I classify an establishment as a local producer if it belongs to one of the 310 local industries.[17]

Table 1 displays summary statistics for the baseline sample. The sample includes 85% of all LBD establishments in 2017. Establishments in local industries are smaller, on average, than those in traded industries in terms of employment, sales, and wage bill. Nevertheless, there are almost three times as many local establishments as traded establishments. As a result, local industries account for more than half of U.S. economic activity, employing 67% of the labor force and generating 52% of total sales and 59% of total labor income.

To remain consistent with the framework, I treat all establishments belonging to the same firm within a city as a single establishment. The model abstracts from both within-city geography and the possibility that a firm operates multiple establishments within the same city. I follow the same convention in the data by aggregating all establishments of a given firm within a city. This aggregation is also motivated by the possibility that firms make pricing and production decisions jointly across their establishments. The results are robust to instead treating these establishments as separate units or excluding multi-establishment firms from the sample.

[15] The resulting sample is similar to the one constructed by Hsieh and Rossi-Hansberg (2023).

[16] The full list of 310 local NAICS industries is available through the Cluster Mapping Project.

[17] This classification is also used by Berger, Herkenhoff, and Mongey (2022) in their study of labor market power.

Table 1: Summary statistics baseline 2017 sample

	All Industries (1)	Local Industries (2)	Traded Industries (3)
Number of establishments	6,655,000	5,075,000	1,579,000
Avg. Employment (# of workers)	17.82	15.70	24.61
Avg. Sales (thousands)	3,463	2,376	6,955
Avg. Wage bill (thousands)	627.4	488.1	1,075
Agg. employment share	...	0.67	0.33
Agg. sales share	...	0.52	0.48
Agg. wage bill share	...	0.59	0.41

Notes: Table 1 displays summary statistics for the 2017 LBD-EC matched sample. The traded-local industries classification is based on Delgado, Porter, & Stern (2015).

4.3 Model Validation

Before studying markups across U.S. cities, I provide empirical support for the model's sorting prediction. Proposition 1 in the baseline model and Proposition 6 in the extended model imply that local producers are more productive in larger cities. However, the LBD-EC data do not allow me to estimate total factor productivity outside the manufacturing sector. I therefore examine how two proxies for productivity relate to city size.

Sales per worker and city size. My first proxy for productivity is sales per worker. Ideally, one would use physical output per worker as a measure of labor productivity. However, because I observe only establishment-level sales, I cannot separately identify prices and quantities. I therefore use sales per worker as my establishment-level measure of labor productivity.

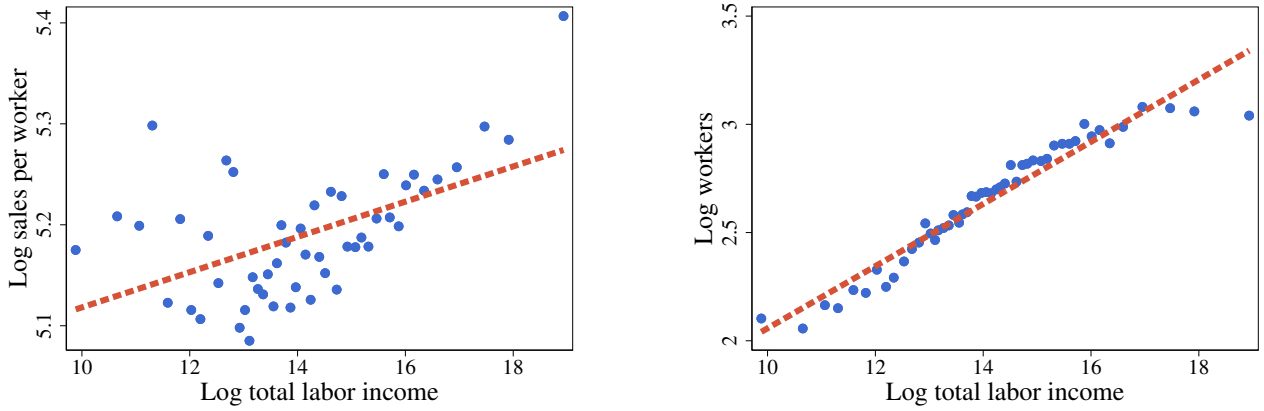
I next aggregate this measure to the commuting-zone level. Using each establishment's location and local-traded classification, I compute sales per worker for establishments in local industries and average this measure across establishments within each commuting zone. This yields my commuting-zone measure of labor productivity. Guided by the framework, I define commuting-zone size as the total income of workers residing in the commuting zone.[18]

Figure 1(a) presents a bin scatter of log sales per worker against log commuting-zone size. Consistent with the model's predictions, the relationship is positive. A 10% increase in commuting-zone size is associated with an approximately 0.3% increase in sales per worker. Moreover, commuting zones in the top decile of the size distribution have sales per worker about 10% higher than those in the bottom decile.

Establishment size and city size. My second proxy for firm productivity is establishment size, measured by employment. For each commuting zone, I compute average employment per

[18]Consistent with the model, this measure includes all workers, regardless of whether they are employed in local or traded industries.

Figure 1: Local industries revenue per worker and employment across commuting zones



(a) Sales per worker

(b) Size

Notes: Figure 1(a) shows a bin-scatter of commuting zone log average sales per worker and log total labor income. Figure 1(b) shows a bin-scatter of commuting zone log average establishment employment and log total labor income. Both scatters consider 50 equally sized commuting zone bins according to their total labor income, and bin averages are computed among establishments in local industries. Dashed line is a linear fit.

establishment among firms in local industries. I then divide commuting zones into 50 equally sized bins based on total labor income and compute average establishment size within each bin, following the same procedure used for labor productivity.

Figure 1(b) presents the corresponding bin scatter for average establishment employment. The relationship is considerably steeper and tighter than that for sales per worker. A 10% increase in commuting-zone size is associated with an approximately 1.4% increase in average establishment employment. Establishments in commuting zones in the top decile of the size distribution employ about two and a half times as many workers as those in the bottom decile. Thus, average establishment size more than doubles across the commuting-zone size distribution.

One concern with both proxies is that sales and employment at establishments belonging to firms that operate in multiple commuting zones may reflect within-firm linkages across locations. In particular, outcomes at an establishment in one commuting zone may depend on the firm's distribution of establishments across other locations, as in frameworks such as Kleinman (2025). Because my model abstracts from this margin, I assess the robustness of the results using a sample that more closely matches the model's environment. Appendix Figure F1 reproduces both panels after restricting the sample to firms operating in a single commuting zone. Both relationships remain, and, if anything, the labor-productivity gradient is tighter in this restricted sample.

4.4 Markup Estimation

This section outlines the empirical strategy I use to estimate markups for establishments in local industries. I combine the consumer demand and firm labor demand implied by the model to derive an estimating equation from which markups can be recovered, building on the approach in De Loecker (2011). Before presenting the method, I discuss why existing approaches are difficult

to implement in the matched LBD–Economic Census sample.

4.4.1 Existing Methods

One approach to markup estimation is commonly referred to as the demand approach. This methodology relies on data on prices and product characteristics, together with information on quantities sold or market shares, to estimate a system of own- and cross-price elasticities across products. Given these estimates, markups can be recovered from the pricing first-order conditions after specifying a model of firm conduct (Ackerberg et al., 2007; Berry, Gaynor, & Scott Morton, 2019). Unfortunately, I do not observe prices or product characteristics in the LBD-EC data, so I cannot implement the demand approach.

Instead, my method builds on the production approach to markup estimation. Originally developed by Hall (1988) and extended by De Loecker and Warzynski (2012), this approach combines data on sales and variable input expenditures with estimates of output elasticities to recover markups.

To introduce this method, I index establishments by j and cities by c.[19] Under standard regularity conditions for the firm's cost-minimization problem, the markup of establishment j in city c can be expressed as the ratio of the output elasticity of a flexible input to that input's expenditure share in sales.[20] The conditions required for this ratio estimator hold in my quantitative setting, and the labor demand function derived in Section D.1 implies that

$$\mu_{jc} = \frac{\beta}{\alpha^l_{jc}} e^{-\nu_{jc}}, \tag{40}$$

where $\alpha^l_{jc} = (W_c l_{jc})/(p_{jc} y_{jc})$ is the labor expenditure share in total sales, which is observed in the data. The term ν_{jc} captures either an ex-post production shock that firms do not observe when choosing their inputs or measurement error in quantities.[21] Equation (40), known as the *ratio estimator*, implies that markups can be recovered with estimates of the labor output elasticity and the ex-post shock. The main econometric challenge is therefore to estimate these two objects.

A common alternative for estimating output elasticities is the production function approach. Under this approach, researchers estimate a production function by regressing output on inputs while controlling for, or otherwise removing, unobserved productivity. Estimation is typically carried out using control function methods, as in Ackerberg, Caves, and Frazer (2015) and Gandhi, Navarro, and Rivers (2020), or dynamic panel methods, as in Arellano and Bover (1995) and Blundell and Bond (1998, 2000). Each approach has its own costs and benefits.[22] However, these approaches generally require data on physical output quantities. When sales are used instead, unobserved differences in output prices confound the identification of production function parameters (Bond et al., 2021; DeRidder, Grassi, & Morzenti, 2024). Data on physical quantities for establishments

[19] In the empirical application, cities are defined as either commuting zones or counties.

[20] A flexible input satisfies: 1) it can be freely adjusted, and 2) establishments take its price as given.

[21] Specifically, incorporating this term into (34), the production function takes the form $y_{jc} = \omega_{jc} e^{\nu_{jc}} l^{\beta}_{jc} h^{1-\beta}_{jc}$.

[22] De Loecker and Syverson (2021) provide an extensive review of control function and dynamic panel methods.

across different sectors of the U.S. economy are generally unavailable.[23] Hence, I cannot implement the production function approach to construct the ratio estimator.

A second alternative for estimating production elasticities is the cost-share approach.[24] Under cost minimization, an input's output elasticity equals its share of total costs multiplied by the production function's returns to scale, which some applications normalize to one by assuming constant returns to scale. Unlike the production function approach, this method does not require data on physical quantities. It does, however, require data on an establishment's total costs. As highlighted by De Loecker and Syverson (2021), such data are rare, with capital costs being particularly difficult to observe. Indeed, data on total establishment costs in the United States are generally unavailable outside publicly traded firms and manufacturing establishments. Thus, I cannot implement the cost-share approach to construct the ratio estimator.

As data limitations prevent me from estimating output elasticities, I adopt an alternative approach to recover markups using the ratio estimator. Following the insight in De Loecker (2011), I use the demand structure of the model to address the identification problem. De Loecker (2011) uses a CES demand system to control for unobserved firm prices when estimating a production function. I instead use the Kimball demand structure to construct a control function for markups. The two approaches differ along two main dimensions. First, I use the demand structure to estimate markups directly rather than to recover production elasticities. Second, the Kimball demand system allows markups to vary across firms, whereas CES preferences imply constant markups. In this respect, my approach is closest to Blum et al. (2024), who construct a proxy for markups in the context of productivity estimation while also allowing for firm-specific markup variation.

4.4.2 Alternative Procedure

To avoid tying the estimates to a particular functional form, I consider a general parameterization of $\Upsilon(\cdot)$ and impose only the existence of a choke price.

Assumption 3 (Kimball Aggregator for Markup Estimation)**.** *Assume* $\Upsilon(\cdot)$ *is strictly increasing and concave and satisfies* $\Upsilon(0) = 0$*. Moreover, assume that there exists* $\overline{p} < \infty$ *such that*

$$\left(\Upsilon'\right)^{-1}(\overline{p}) = 0.$$

Assumption 3 implies that there is a finite price at which workers demand zero quantities. This condition is satisfied by the Klenow and Willis (2016) parameterization used in Sections 2 and 3.

To illustrate my approach, I begin by rearranging (40) and taking logs:

$$\log \alpha_{jc}^{l} = \beta - \log \mu_{jc} - \nu_{jc}. \tag{41}$$

An initial reading of (41) might suggest that markups could be estimated from the residual of a

[23] Exceptions include manufacturing subsamples in Foster, Haltiwanger, and Syverson (2008) and Atalay (2014).

[24] De Loecker, Eeckhout, and Unger (2020) and Edmond, Midrigan, and Xu (2023) use this approach to estimate markups for publicly traded firms and manufacturing establishments, respectively.

regression of the labor share on a constant capturing the labor output elasticity. This procedure, however, faces two important challenges. First, the measurement error ν_{jc} is absorbed by the residual, thereby confounding markup estimates.[25] Second, as Appendix E.5 shows, if the production function is generalized so that output elasticities vary with input usage, the resulting correlation between markups and inputs invalidates the identification of markups from the residual in (41). To overcome these challenges, I use the demand system in Assumption 3 to construct a control function for markups.

Appendix E.1 shows that the markup of establishment j in city c can be expressed as a function of the firm's sales share, s^y_{jc}, and the local price indices $\mathbb{P}_c$ and $\mathbb{D}_c$:

$$\mu_{jc} = \mu\left(s^y_{jc}\frac{\mathbb{P}_c}{\mathbb{D}_c}\right). \tag{42}$$

The exact functional form of $\mu(\cdot)$ depends on the parameterization of $\Upsilon(\cdot)$. To keep the estimation parsimonious, I approximate the markup function semiparametrically using a sieve series estimator, following Hansen (2014) and its application to production function estimation by Gandhi, Navarro, and Rivers (2020). In the baseline specification, I approximate the log markup function with a third-degree polynomial in $s^y_{jc}(\mathbb{P}_c/\mathbb{D}_c)$.[26] Under this approximation, the resulting estimating equation takes the form of a heterogeneous-slopes model:

$$\log \alpha^l_{jc} = \log\beta + \varsigma_{1,c}s^y_{jc} + \varsigma_{2,c}(s^y_{jc})^2 + \varsigma_{3,c}(s^y_{jc})^3 + \upsilon_{jc}, \tag{43}$$

where $\varsigma_{1,c}$, $\varsigma_{2,c}$, and $\varsigma_{3,c}$ are city-specific coefficients that absorb the variation in $\mathbb{P}_c/\mathbb{D}_c$, and υ_{jc} is a composite error term containing both the measurement error ν_{jc} and the approximation error. Importantly, Assumption 3 implies that the approximation to the markup function has no constant term. This restriction allows me to separately identify markup levels and the output elasticity β. The resulting markup estimates are

$$\hat{\mu}_{jc} = \exp(-\hat{\varsigma}_{1,c}s^y_{jc} - \hat{\varsigma}_{2,c}(s^y_{jc})^2 - \hat{\varsigma}_{3,c}(s^y_{jc})^3). \tag{44}$$

Equation (43) clarifies the variation that identifies markups. Conditional on the output elasticity, β, identification comes from within-city variation in sales shares and labor cost shares. Within a city, establishments with higher sales shares and lower labor expenditure shares have higher markups. Because the measurement error ν_{jc} may be correlated with contemporaneous sales shares s^y_{jc}, I also report robustness exercises that instrument contemporaneous sales shares with their lagged values in the estimation of (43).

Appendix Table F1 presents summary statistics for the estimated markups. I report results for the baseline markups obtained by estimating (43) by ordinary least squares (OLS), as well as

[25] Peters (2020) follows a related strategy but corrects for measurement error using data on all inputs, which allow him to estimate a nonparametric production function. This correction is not feasible with my data.

[26] The baseline results are virtually unchanged when higher-order polynomials are used.

for markups estimated by instrumenting the contemporaneous sales-share terms with their lags. Overall, the two estimates are very similar. Establishments in U.S. local industries charge an average markup of 5.6%, while those near the bottom of the markup distribution have markups close to zero. Consistent with Assumption 3, establishments with very small local sales shares have markups close to zero. In contrast, establishments near the top of the distribution charge markups of approximately 10%. These estimates are comparable to those reported by DeRidder, Grassi, and Morzenti (2024) for French manufacturing firms, but lower than those reported by De Loecker, Eeckhout, and Unger (2020) for publicly traded U.S. firms.

I defer one technical discussion and two estimation extensions to the appendix. Appendix E.2 derives conditions on $\Upsilon(\cdot)$ under which the regularity conditions in Newey (1997) hold, ensuring consistency of the sieve estimator in (43). Appendix E.3 then extends the estimation framework to allow for multiple sectors. The estimating equation (43) carries over with two modifications. First, the city-specific slopes become city-sector specific because local price indices also vary across sectors. Second, establishment market shares are defined at the city-sector level. Finally, Appendix E.4 extends the baseline approach to account for potential labor market power. Intuitively, markdowns can be recovered using a control function based on local wage-bill shares.

4.5 Markups Across Cities

I now examine the cross-sectional patterns of markups across cities. I estimate markups for establishments in local industries using (44).[27] I then compute the city-level aggregate markup as the sales-weighted harmonic mean of establishment-level markups. Formally, I define $\mathcal{M}_c$, the empirical counterpart of (39), as

$$\mathcal{M}_c = \left(\sum_{j \in c} \frac{s^y_{jc}}{\mu_{jc}} \right)^{-1}, \tag{45}$$

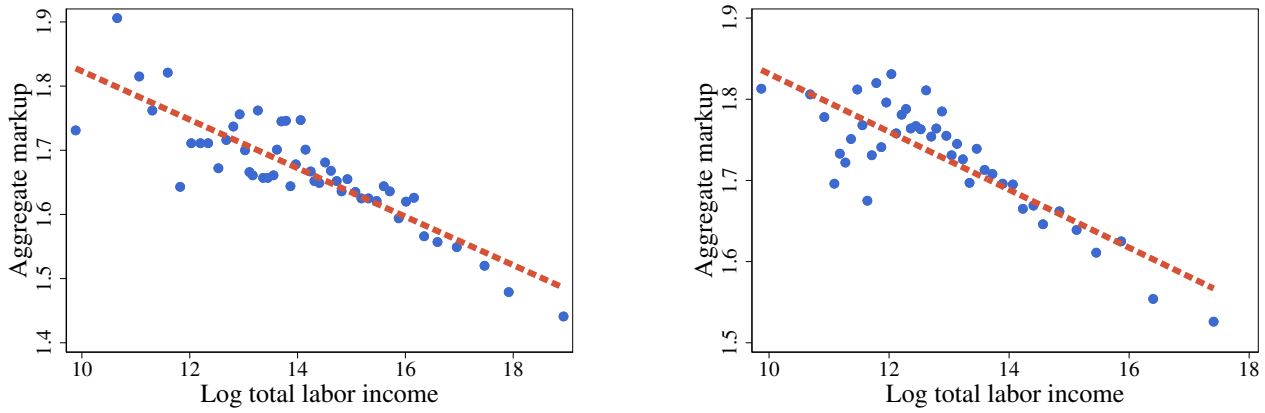
where the sum is taken over all establishments located in city c.

Figure 2 shows the relationship between city-level aggregate markups and market size, measured by total labor income, for both commuting zones and counties. The results reveal a clear empirical pattern. Larger markets have significantly lower aggregate markups than smaller ones. Figure 2(a) shows that commuting zones in the top decile of the size distribution, including those that encompass major urban areas such as New York and Chicago, have aggregate markups approximately 20% lower than those in the bottom decile. Across the size distribution, a 1% increase in commuting-zone size is associated with an approximately 0.02% decrease in the aggregate markup. Figure 2(b) documents the same pattern across counties.

The relationship between aggregate markups and market size is stable over time. Table 2 reports the elasticity of aggregate markups with respect to market size for commuting zones and counties

[27] In principle, (44) could also be used to estimate markups for establishments in traded industries. Doing so, however, would require taking a stand on the geographic markets these establishments serve and accounting for imports when computing their market shares.

Figure 2: Aggregate markup and market size



(a) Commuting zones

(b) Counties

Notes: Figure 2(a) presents a bin scatter of commuting-zone aggregate markups against log total labor income. Figure 2(b) presents the corresponding bin scatter for counties. Both panels divide cities into 50 equally sized bins based on total labor income. The city-level aggregate markup is the sales-weighted harmonic mean of local establishments' markups, as defined in (45). The dashed line shows a linear fit across bins.

in each census year from 2002 to 2017, extending the baseline analysis for 2017 presented in the preceding figures. The negative relationship is present in every specification. For commuting zones, the estimated elasticity remains relatively stable, ranging from -2% to -3.2%. For counties, the estimates are of similar magnitude across years, except in 2017, when the elasticity is smaller in magnitude at -1.4%.

The negative relationship between aggregate markups and market size is robust to alternative estimation approaches. I first address the concern that contemporaneous sales shares may be correlated with the ex post shock, ν_{jc}. Appendix Figure F2 presents a bin scatter of aggregate markups against commuting-zone size when establishment markups are estimated using the instrumental-variables approach discussed above. Under this alternative specification, aggregate markups levels slightly decline, but the strong negative relationship documented in Figure 2(a) remains.

I next examine whether the baseline results are affected by potential labor market power. Firms may exercise market power in local labor markets, which could affect the markups inferred from (43). To address this possibility, I re-estimate markups using the methodology developed in Appendix E.4, which allows for firm-specific markdowns. Figure F3 reproduces Figure 2 using these alternative estimates. The implied city-level aggregate markups increase substantially, but the negative relationship between market size and aggregate markups becomes even more pronounced. Furthermore, Table F2 shows that the elasticity of aggregate markups with respect to market size remains negative across census years for both counties and commuting zones.

As a third robustness exercise, I consider a simpler measure of markups based solely on the inverse of the labor share. As equation (40) shows, the labor output elasticity is constant across establishments in my sample. Therefore, in the absence of ex post production shocks, cross-sectional variation in establishment markups would be driven entirely by variation in the inverse

Table 2: Average City Markup elasticity with respect to City Size

	Dep. var.: Log aggregate markup			
	2002	2007	2012	2017
	(1)	(2)	(3)	(4)
	Panel A: Commuting Zones			
Log total labor income	-0.0209***	-0.0317***	-0.0327***	-0.0207***
	(0.0020)	(0.0026)	(0.0029)	(0.0027)
Observations	700	700	700	700
R-squared	0.184	0.285	0.243	0.141
	Panel B: Counties			
Log total labor income	-0.0222***	-0.0322***	-0.0289***	-0.0147***
	(0.0016)	(0.0018)	(0.0020)	(0.0017)
Observations	3200	3200	3200	3200
R-squared	0.076	0.126	0.081	0.029

Notes: Table 2 displays the average elasticity of commuting zone aggregate markup and city size. City aggregate markups is defined as (45) and city size is defined as total labor income. Panel A shows the mean elasticity defining cities as commuting zones. Panel B shows the mean elasticity defining cities as counties. Robust standard errors in parenthesis. *10% level, **5% level, ***1% level.

labor share.[28] Motivated by this observation, I construct an alternative markup measure that abstracts from ex post shocks and normalizes the labor output elasticity to one. I then aggregate these establishment-level measures to the city level as before. Because this procedure identifies only relative markup differences across locations, I normalize each city's aggregate markup by that of the largest city. Figure F4 shows that the strong negative relationship between aggregate markups and market size persists. It also implies markup gaps of a similar magnitude across the size distribution, with smaller markets having aggregate markups approximately 25% to 30% higher than larger markets. Table F3 further shows that this negative relationship remains stable over time for both counties and commuting zones.

Having established that the empirical pattern is robust to alternative markup estimates and across census years, I finally examine whether multi-location firms complicate its interpretation as evidence of local competition. These firms pose two potential challenges. First, retail chains set nearly uniform prices across their stores (DellaVigna & Gentzkow, 2019). Combined with higher labor costs in larger markets, uniform pricing could mechanically lower measured establishment markups in large cities, even in the absence of differences in local market power. Second, because multi-establishment firms typically originate in large markets and expand into smaller ones (Hsieh & Rossi-Hansberg, 2023), their prevalence varies systematically with market size. The aggregate relationship could therefore partly reflect differences in the composition of firms across locations rather than differences in local competition. Both concerns are reduced when the sample is re-

[28] This builds on the insight in DeRidder, Grassi, and Morzenti (2024) that the ratio estimator remains informative about cross-sectional and time-series variation in markups even when output elasticities are measured imprecisely.

stricted to firms operating in a single commuting zone, whose pricing decisions are more likely to reflect local market conditions. Figure F5 replicates Figure 2(a) using this restricted sample. The negative relationship between aggregate markups and market size is, if anything, steeper.

4.5.1 Markups Across Cities and Sectors

Although the primary goal of this section is to present the markup estimates used in the quantitative analysis, I now turn to a sector-specific analysis of markups across cities. The purpose of this exercise is to illustrate more clearly the variation in the data that identifies markups and to show that the negative relationship between city-level markups and market size is not mechanical.

I estimate markups for the same sample of local establishments using the multi-sector approach described in Appendix E.3. Under this extension, conditional on a sector-specific output elasticity, markups are identified from variation in establishment labor expenditure shares relative to establishment market shares within each city-sector. I then construct city-sector aggregate markups analogously to (45). I implement this procedure for fifteen two-digit NAICS industries.

Tables F4 and F5 report the elasticity of city-sector aggregate markups with respect to market size for counties and commuting zones, respectively. The results reveal substantial heterogeneity across sectors, with some displaying positive elasticities and others negative ones. Importantly, two of the three sectors with the largest shares of local employment, Retail and Accommodation and Food Services, display a negative elasticity under both city definitions. These sectors therefore contribute largely to the negative aggregate relationship documented in Figure 2.

The sector-level results also show that the negative relationship between aggregate markups and market size is not mechanical. In particular, because the estimation uses establishment market shares, one might be concerned that markups decline with market size simply because establishments tend to have smaller market shares in larger markets with more competitors. If this mechanism mechanically generated the aggregate relationship, however, the same negative relationship should emerge across sectors. Instead, the sector-level estimates reveal substantial heterogeneity, with several sectors displaying positive elasticities. Thus, the relationship between markups and market size depends on the joint variation in establishment market shares and labor expenditure shares within each city-sector, rather than on either variable in isolation.

4.6 Robustness Using the Census of Manufactures

I use the manufacturing sector to conduct additional robustness exercises that provide further empirical support for my markup estimates. Unlike the full sample, the Census of Manufactures contains detailed information on production inputs, including materials, energy, and capital expenditures. These data allow me to examine the variation in markups across cities using three alternative markup estimates. Appendix F.4 provides the details of these exercises.

I address two initial concerns simultaneously. First, one potential concern with my markup estimation procedure is that labor may not be fully flexible. Second, the Cobb-Douglas assumption may be too restrictive because establishments can differ in the technologies they operate. I ad-

dress both concerns in Appendix E.5 by extending the baseline estimating equation (43) to use materials and energy, which are generally viewed as more flexible and less subject to adjustment costs than labor. The extension also allows for a general Hicks-neutral production function with establishment-specific output elasticities, which I approximate using a flexible polynomial in labor, materials, energy, and capital, following Gandhi, Navarro, and Rivers (2020).

Table F6 reports the robustness results. It presents regressions of the baseline markup estimates, which use labor as the flexible input, on alternative markup estimates that use materials or energy and allow for a general production function. The estimated coefficients are close to one for both counties and commuting zones, and the regressions display high R^2 values, indicating a very tight relationship across the alternative markup measures.

Finally, I perform a third robustness exercise by computing markups for local manufacturing producers using a cost-share approach. Under this methodology, one recovers an estimate of the labor output elasticity by exploiting the firm's first-order conditions from cost-minimization and by assuming a level of the scale elasticity. Similarly to De Loecker, Eeckhout, and Unger (2020) and Edmond, Midrigan, and Xu (2023), I assume constant returns to scale and compute the labor elasticity at the 2-digit and 4-digit NAICS industry levels. Then, I obtain markup estimates by plugging back these estimates into (40). Figures F8 and F9 show that these markup estimates exhibit virtually the same levels and cross-sectional variation as my baseline markup estimates.

This section outlined the empirical approach to estimating markups using U.S. microdata and presented the resulting estimates. Markups vary substantially across local producers and across cities, with larger cities exhibiting lower aggregate markups than smaller ones. I now turn to the quantitative analysis, where I use the firm-level markup estimates to recover key elasticities of the model. I then use the estimated model to quantify the welfare effects of place-based policies.

5 Quantitative Analysis

In this section, I estimate the extended model and use it to quantify the general equilibrium effects of place-based policies. The quantitative analysis focuses on the continental U.S. and defines cities as commuting zones, for a total of 722 locations.

5.1 Model Estimation

To estimate the extended model, I first specify the initial productivity distribution $G_0(\cdot)$ and the ex post productivity distribution $G(\cdot|z)$. For the former, I assume a uniform distribution on $[\underline{z}, \overline{z}]$, with $\overline{z} < 1$. For the latter, I assume a Pareto distribution with shape parameter $1/z$ and location parameter normalized to one. This specification satisfies Assumption 2.[29]

Under these functional-form assumptions, the extended model has 11 parameters. I divide them into three groups according to the estimation approach. First, I calibrate three parameters exter-

[29] The conditional density is $g(\omega|z) = \frac{\omega^{-\left(\frac{1}{z}+1\right)}}{z}$. Then $\frac{\partial^2 \log g(\omega|z)}{\partial \omega \partial z} = \frac{1}{z^2 \omega} > 0$, so $g(\omega \mid z)$ is strictly log-supermodular in (ω, z).

Table 3: First Group: External Calibration

Parameter	Description	Source	Value
α	Housing expenditure share	Davis & Ortalo-Magne (2011)	0.24
ϕ	Housing supply elasticity	Baum-Snow & Han (2024)	0.29
θ	Dispersion location preferences	Monte, Redding, & Rossi-Hansberg (2018)	3.3

Notes: Table 3 displays parameter values for the externally calibrated parameters.

nally using standard values from the literature. Second, the model delivers estimating equations for five additional parameters, which I estimate using Minimum Distance Estimation (MDE) and the Generalized Method of Moments (GMM). Finally, I estimate the remaining three parameters using a Simulated Method of Moments (SMM) routine. These parameters are the lower and upper bounds of the initial local-goods productivity distribution and the entry cost. I infer them by targeting average firm size across quintiles of the market-size distribution, together with the economy-wide aggregate markup.

In addition to these common parameters, I must estimate the city-specific traded-good productivities and amenities. I recover these fundamentals nonparametrically by exactly matching employment and average wages in each commuting zone.

Externally Calibrated (3 parameters). This group comprises the housing expenditure share, α, the housing supply elasticity, ϕ, and the dispersion of idiosyncratic location preferences, θ.

Table 3 summarizes the calibrated values. I set the housing expenditure share to $\alpha = 0.24$, following Davis and Ortalo-Magne (2011). I set the housing supply elasticity to $\phi = 0.29$, the mean estimate in Baum-Snow and Han (2024). Finally, I set the dispersion of location preferences to $\theta = 3.3$, the baseline estimate of Monte, Redding, and Rossi-Hansberg (2018) for U.S. commuting zones.[30]

MDE and GMM Estimates (5 parameters). This group contains five parameters: the local-goods expenditure share, η; the Kimball demand parameters, $\overline{\sigma}$ and ε; and the output elasticities of local- and traded-good producers, β and γ. Table 4 summarizes the estimates.

I estimate the local-goods expenditure share, η, using the local-goods sales share reported in Table 1. Given the calibrated housing expenditure share, $\alpha = 0.24$, an aggregate local-goods sales share of 52% implies $\eta = 0.39$.

I estimate the Kimball demand parameters in two steps. Here, I summarize the variation that identifies these parameters, while Appendix G.1 provides the estimation details. First, as shown by Edmond, Midrigan, and Xu (2023), the Klenow-Willis specification for household preferences implies a log-linear relationship between markups and sales shares. Using this relationship, together with my markup estimates and data on sales shares, I recover an estimate of the ratio $\varepsilon/\overline{\sigma}$.[31]

Equipped with an estimate of $\varepsilon/\overline{\sigma}$, I then use an iterative GMM procedure to estimate $\overline{\sigma}$. Under the Klenow-Willis demand system, relative prices and the ratio of the price indices, $\mathbb{P}_c/\mathbb{D}_c$, can be expressed as functions of sales shares and known parameters. Given a guess for $\overline{\sigma}$, I solve this

[30] Fajgelbaum et al. (2018) estimate $\theta = 1.73$ using U.S. states as the geographic unit.

[31] My estimating equation therefore extends Edmond, Midrigan, and Xu (2023) to a multi-location setting.

Table 4: Second Group: GMM

Parameter	Description	Moment and Method	Estimate
η	Local goods expenditure share	Local sales share (MDE)	0.39
$\varepsilon/\overline{\sigma}$	Demand super-elasticity	Markup and sales share (OLS)	0.11
$\overline{\sigma}$	Demand elasticity	Markup levels (GMM)	2.26
β	Labor output elasticity (local)	Labor demand, local (OLS)	0.25
γ	Labor output elasticity (traded)	Labor demand, traded (MDE)	0.29

Notes: Table 4 summarizes the GMM estimation results for the second block of parameters. MDE: Minimum Distance Estimator, OLS: Ordinary Least Squares and GMM: Generalized Methond of Moments.

system to obtain relative prices and then use (16) and (42) to compute implied markups. I estimate $\overline{\sigma}$ by minimizing the distance between these implied markups and the markup estimates from (43). Intuitively, within-city variation in markups and sales shares identifies the super-elasticity $\varepsilon/\overline{\sigma}$, while the level of markups pins down $\overline{\sigma}$.

Finally, I estimate the output elasticities for labor. For establishments in local industries, I estimate β from the markup estimating equation (43). For establishments in traded industries, I estimate γ using (23), which implies that γ equals the labor cost share of sales.

SMM Estimates (3 parameters). The parameters in the last group are the ones governing the local producers' productivity distribution, $\underline{z}$ and $\overline{z}$, in addition to the entry cost c_e. Table 5 displays the estimated parameters and the goodness of fit.

I estimate three parameters by SMM. Within-city productivity follows a Pareto distribution with shape parameter z^{-1}, so $\underline{z}^{-1}$ and $\overline{z}^{-1}$ govern productivity dispersion in the smallest and largest commuting zones by market size. I use two sources of variation to identify these parameters. First, I target the elasticity of the aggregate markup with respect to market size reported in column 5 of Panel A in Table 2. Second, I use the distribution of firm employment across markets of different sizes. I divide commuting zones into quintiles by market size and compute average firm employment within each quintile. I then target average firm size in the second, third, and fourth quintiles relative to the first. I omit the fifth quintile because it is dominated by extremely large firms in small commuting zones with high wages. To pin down the entry cost c_e, I target the economy-wide aggregate markup for local industries, defined as the population-weighted average of commuting-zone-level markups. I construct this moment using the markup estimates from Section 4. Overall, the model is overidentified, with five moments for three parameters.

Formally, let $\Theta = (\underline{z}, \overline{z}, c_e)$ denote the vector of parameters to estimate. I implement the SMM by choosing Θ to minimize the sum of squared percentage deviations between the model-simulated moments, $\Psi^m(\Theta)$, and their empirical counterparts, Ψ^d. I use the TikTak algorithm of Arnoud, Guvenen, and Kleineberg (2019) to search globally over the parameter space.[32] At each iteration of the optimization routine, I invert the equilibrium mapping to nonparametrically recover a_c and b_c from the labor-supply and wage conditions. Appendix G.2 details the inversion procedure.

Table 5 reports the parameter estimates and moment fit. The estimates imply that productivity

[32] I use 3000 starting points and a simplex search method for local optimization.

Table 5: Third Group: SMM Estimates and Moment Fit

Panel A: Parameter Estimates		
Parameter	Description	Estimate
$\underline{z}$	Prod. Dispersion, smallest CZ	0.264
$\overline{z}$	Prod. Dispersion, largest CZ	0.529
c_e	Entry Cost	4.9×10^{-3}
Panel B: Targeted Moments		
Moment	Data	Model
Aggregate Markup	1.51	1.45
Markup elast. to Market Size	-0.021	-0.021
Avg. Firm Emp., Q2/Q1	1.29	1.27
Avg. Firm Emp., Q3/Q1	1.52	1.47
Avg. Firm Emp., Q4/Q1	1.81	1.89

Notes: Table 5 presents the SMM estimates and goodness of fit for the third block of parameters.

dispersion increases with market size. In particular, the productivity distribution in larger markets has a thicker upper tail, resulting in mean firm productivity that is 56% higher than in the smallest markets. The model closely matches all five targeted moments. It also performs reasonably well on untargeted moments. For example, the sales share of the top 5% of firms is 0.65 in the largest commuting zones compared with 0.76 in the data.

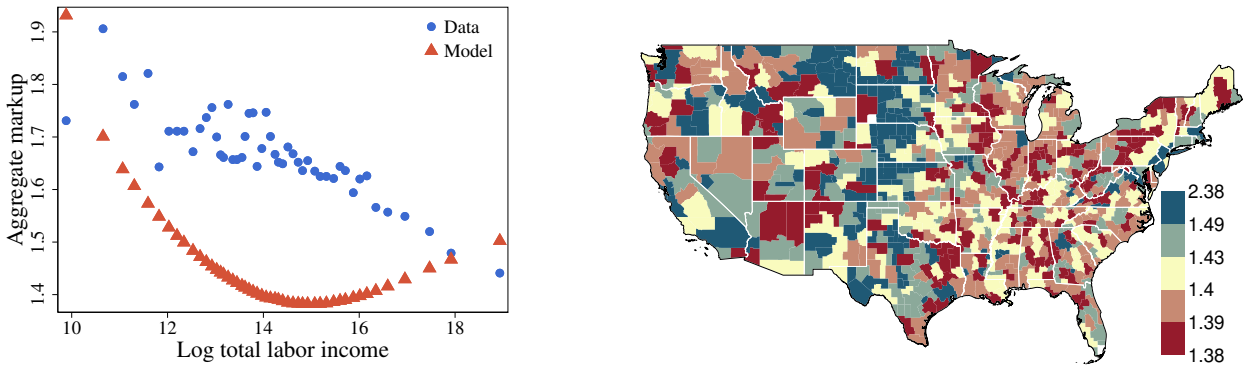
5.2 The Estimated Decentralized Equilibrium

In this section, I solve for the decentralized equilibrium using the estimated parameters and assess how well the model reproduces key spatial patterns in the data.

First, Figure 3 presents the model-implied markups. Figure 3(a) computes markups for each commuting zone and reports a bin scatter, which I compare with its empirical counterpart in Figure 2. The model closely matches the level of markups in the smallest and largest commuting zones. In particular, it reproduces the fact that commuting zones in the top decile of the size distribution have markups close to 30% lower than those in the bottom decile. Nevertheless, the model predicts slightly lower markups than observed in the data for mid-sized cities. Figure 3(b) shows the geography of these markups across U.S. commuting zones. Small commuting zones in the central United States exhibit markups above 2. In contrast, midsize and larger commuting zones, such as the one containing Houston, exhibit markups between 1.4 and 1.5. Markup variation across these locations is also substantially smaller than across smaller markets.

Figure 4 plots the main equilibrium objects in the decentralized equilibrium and shows that the model reproduces several additional empirical regularities across cities. First, the top-right panel shows the cross-sectional variation in markups across the support of the appeal distribution, consistent with the reduced-form evidence in Section 4. Second, larger cities are more competitive, as measured by the competition index, and more productive. This productivity advantage reflects

Figure 3: Markups in the Decentralized Equilibrium



(a) Aggregate markup and market size

(b) Markups across commuting zones

Notes: Figure 3(a) plots data and model aggregate markups across the 50 commuting zone bins of Figure 2. Figure 3(b) shows model markups across commuting zones.

both the higher mean productivity of firms located in larger cities and the contribution of firm mass and allocative efficiency captured by the TFP term in (39). Third, the bottom-left panel reports the price index for local varieties, $\mathbb{P}(a)$, which is higher in larger cities, in line with Diamond and Moretti (2026), who document that the prices of non-housing non-tradables are systematically higher in such locations. Finally, the model also reproduces the facts that larger commuting zones have higher housing rents and higher wages.[33]

Appendix Figure F6 displays the city characteristics recovered from the model inversion. Local traded-good productivity is shown in the left panel, while local amenities are shown in the right panel. The two characteristics are highly correlated across commuting zones. Because the inversion exactly matches wages and population, the model rationalizes high-wage commuting zones as having high productivity and more populous commuting zones as having higher amenities.

Finally, Appendix Figure F7 provides an over-identification exercise by comparing sales per worker in the model and the data. The estimation routine does not target differences in sales per worker across locations. Nevertheless, the model closely reproduces the empirical pattern documented in Figure 1(a).

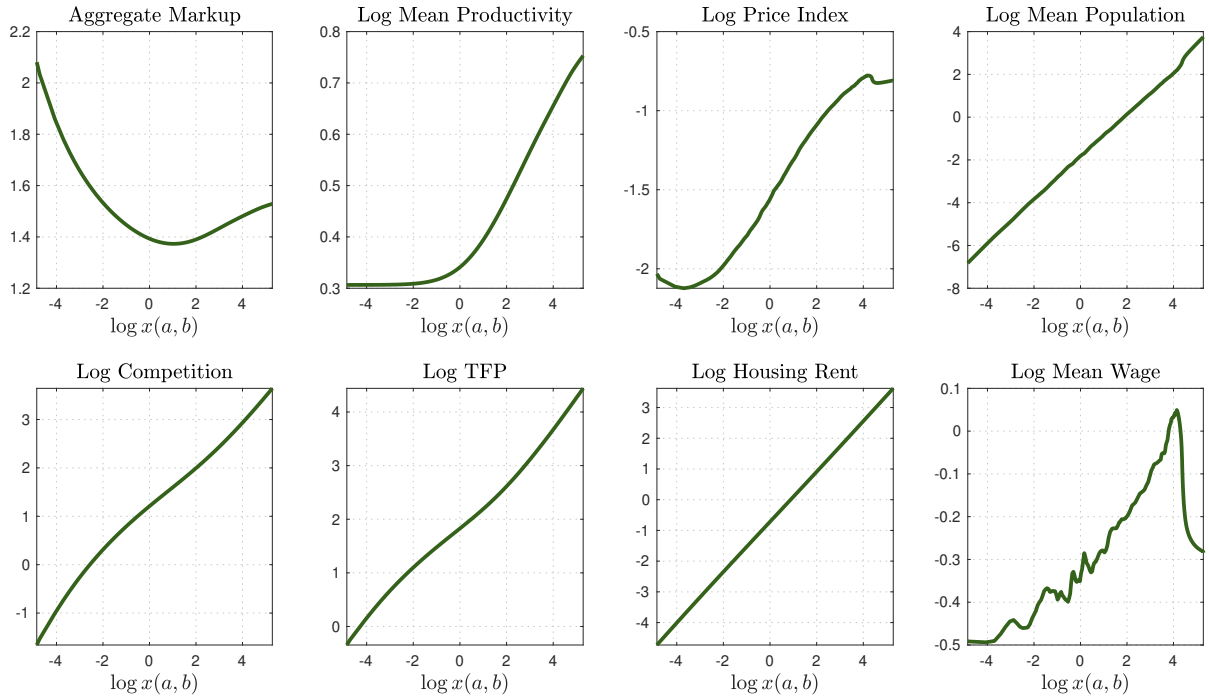
5.3 Place-based Policy Counterfactuals

This final section studies policy counterfactuals. I quantify the aggregate effects of implementing the quasi-optimal policy characterized in Proposition 5 and contrast them with policies designed to attract firms to smaller places, in the spirit of the U.S. Empowerment Zones program.

The main policy I study is the location-specific output subsidy in (31), which eliminates markups. This policy is financed by a flat wage tax on workers rather than by a lump-sum tax. As a result, it does not attain the first best, and I therefore refer to it as the quasi-optimal policy.

[33]The exception is at the very top of the appeal distribution. Wages peak in high-productivity locations such as San Jose and then decline for the highest-appeal cities, most notably Los Angeles.

Figure 4: Cross Sectional Patters in the Decentralized Equilibrium



Notes: Figure 4 displays the extended model's decentralized equilibrium. The x-axis represents the city's estimated appeal, $x(a,b)$.

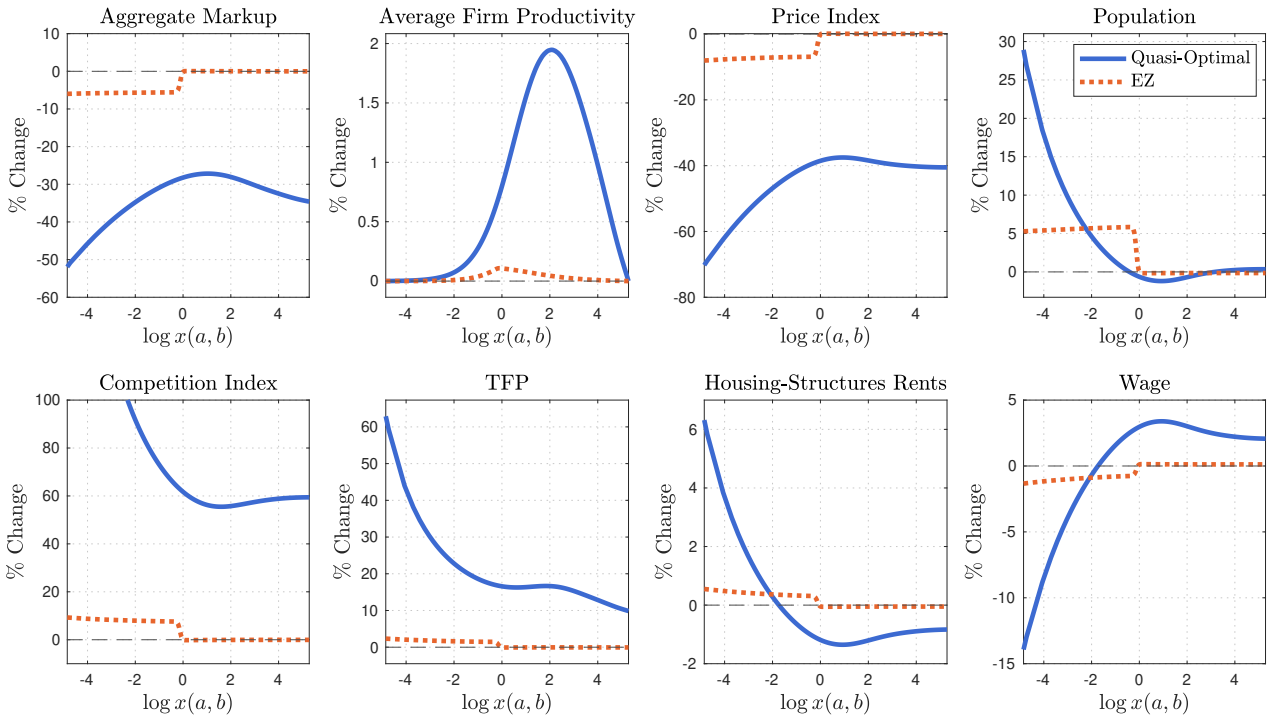
Nevertheless, the policy addresses the three inefficiencies associated with markups in the baseline model: i) spatial misallocation of firms, ii) inefficient aggregate entry, and iii) inefficient relative consumption of local goods, housing, and the traded good. In addition, it corrects a fourth inefficiency introduced in the extended model: iv) within-city misallocation of production inputs arising from markup variation.

I contrast the effects of the quasi-optimal policy with a wage-bill subsidy targeted to smaller commuting zones, consistent with the U.S. Empowerment Zones (EZ) program. Implemented in the 1990s, the EZ program is one of the largest federal place-based policies. Designed to spur economic activity in distressed urban and rural communities, it combines place-based tax incentives with federal grants tied to local development plans.[34] A central component of the program is the EZ Employment Credit, which offers a tax credit of up to 20% on wages paid to EZ residents. I model the EZ program by implementing this wage-bill subsidy in the smallest commuting zones that collectively account for 2.3% of the U.S. population. For tractability, I abstract from other EZ components, such as grants and investment incentives.

Figure 5 compares cross-sectional changes in equilibrium outcomes under the quasi-optimal policy and the EZ program.[35] Consistent with Proposition 4, the quasi-optimal policy eliminates markups in every location. Aggregate markups fall by more than 50% in the smallest locations, where baseline markups are highest, and by roughly 30% elsewhere. Competition increases everywhere, with substantially larger effects in smaller locations. In larger locations, this increase mainly

[34] Glaeser and Gottlieb (2008) and Busso, Gregory, and Kline (2013) provide overviews of the EZ program.
[35] Figure G1 in Appendix G.3 displays the equilibrium outcomes in levels rather than in changes.

Figure 5: Changes in equilibrium objects under the Quasi-Optimal Policy and EZ Program



Notes: Figure 5 shows percent changes in equilibrium under the quasi-optimal policy and the EZ program, relative to the decentralized equilibrium. The x-axis represents the city's estimated appeal $x(a,b)$.

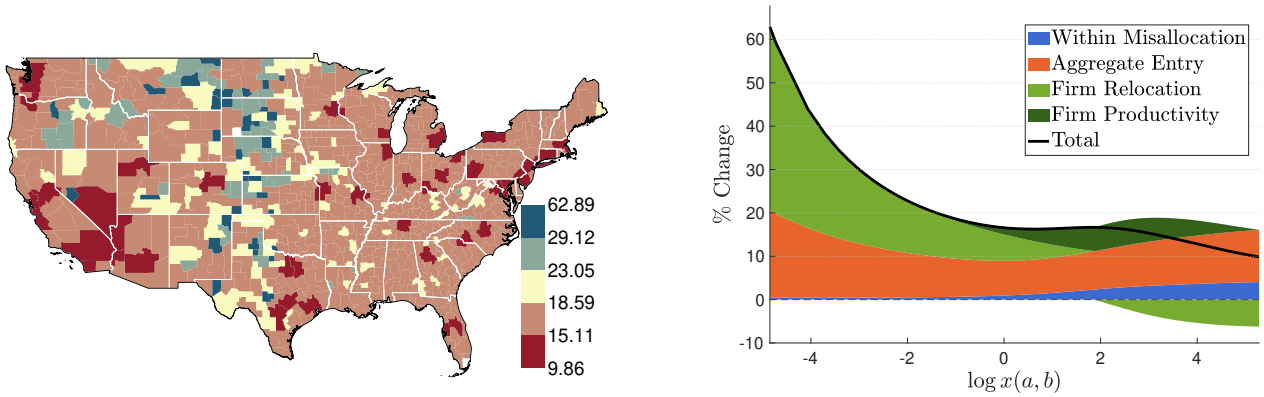
reflects markup removal, which boosts production. In smaller locations, the policy also induces productive firms to relocate toward these markets. This reallocation sharply increases the mass of firms in the smallest locations and raises mean firm productivity by about 2% in mid-sized locations. As a result, local TFP, which captures both firm productivity and the mass of firms, increases by more than 60% in the smallest locations.

The EZ program generates qualitatively similar effects in the targeted locations. The subsidy induces incumbent firms to expand production and attracts additional firms to these markets, increasing competition and reducing aggregate markups. Quantitatively, however, the effects are more modest than under the quasi-optimal policy. Aggregate markups fall by about 6%, while the increases in average firm productivity and local TFP are roughly an order of magnitude smaller.

Smaller cities expand under both policies. The top-right panel of Figure 5 shows that population increases by almost 30% under the quasi-optimal policy and by about 5% under the EZ program. Under the quasi-optimal policy, this expansion is accompanied by a decline in population in mid-sized cities. By contrast, the EZ program has little effect on population in non-targeted locations.

Prices move consistently with this population reallocation. Under the quasi-optimal policy, the local price index falls in all locations as markups are eliminated, with substantially larger declines of about 70% in smaller cities. Under the EZ program, local prices fall by about 7% in targeted smaller locations and remain nearly unchanged in larger ones. The reallocation of firms and workers also affects wages and housing rents. Under the quasi-optimal policy, wages rise by roughly 3% in larger markets and fall by more than 10% in smaller locations. The EZ program has little effect on

Figure 6: Changes in local TFP under the Quasi-Optimal Policy



(a) Geography (b) Decomposition

Notes: Figure 6(a) displays the changes in local TFP, as defined in (39), across commuting zones under the quasi-optimal policy. Figure 6(b) decomposes these changes a function of city appeal, $x(a,b)$.

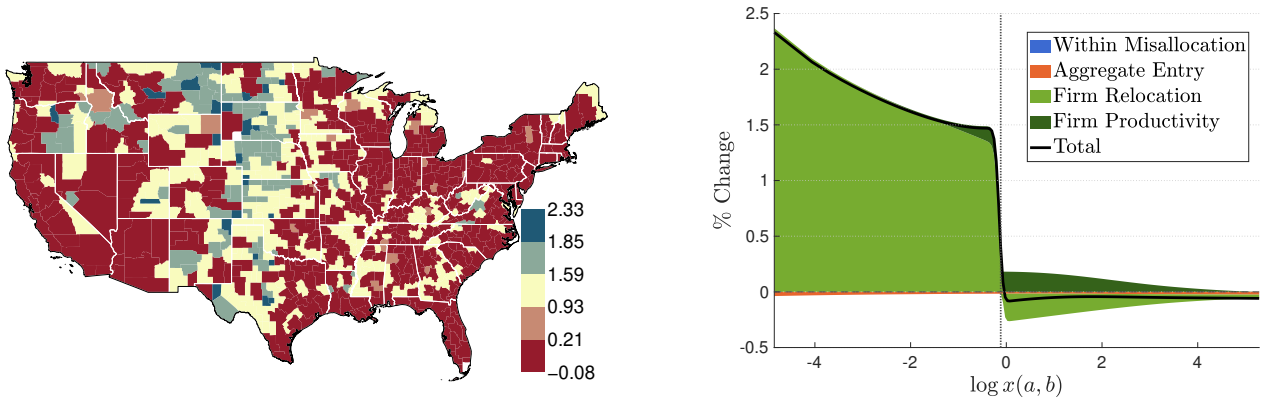
wages in larger places and reduces them by less than 1% in smaller cities. Housing rents move in the opposite direction. Under the quasi-optimal policy, rents decline by around 1% in larger cities and increase by about 6% in smaller ones. The EZ program similarly raises rents in smaller cities by about 1% while leaving rents in larger places nearly unchanged. Workers nonetheless move into these locations because the decline in the local price index more than offsets the lower nominal wages and higher housing rents, resulting in higher real income.

I now turn to decomposing changes in local TFP across space and across channels. Figure 6 presents this decomposition for the quasi-optimal policy. First, Figure 6(a) illustrates the geography of TFP changes across commuting zones. The figure reveals substantial heterogeneity. Commuting zones in the central part of the country, from central Texas to North Dakota and Montana, experience the largest TFP gains, ranging from 20% to 60%. In contrast, major commuting zones, such as those containing Houston and Boston, experience smaller but still substantial gains of around 10% to 15%. The remaining locations fall in between, with TFP increases of about 15% to 20%.

Figure 6(b) decomposes these TFP changes into four channels. Firm relocation accounts for most of the variation. By correcting spatial misallocation, the policy induces productive firms to relocate toward smaller markets. This reallocation generates TFP gains of up to 40% in smaller locations but lowers TFP in larger markets as productive firms move away. The gains from this channel decrease with market size and turn negative in the largest locations, where TFP falls by 5%.

The remaining channels also contribute to the TFP response in different ways. Aggregate entry is the second main driver. The quasi-optimal policy increases aggregate entry by almost 20%, generating relatively uniform TFP gains of around 10% to 20% across locations. Differences in average firm productivity provide an additional source of gains, contributing around 10% to TFP in mid-sized to large locations. Finally, the within-city misallocation channel, which captures the allocative gains from eliminating markup distortions, is quantitatively less important than the other channels. It generates TFP gains of less than 5% in larger cities and becomes almost negligible in smaller locations. This pattern is consistent with Assumption 2, which implies greater

Figure 7: Changes in local TFP under the EZ Program



(a) Geography (b) Decomposition

Notes: Figure 7(a) displays the changes in local TFP, as defined in (39), across commuting zones under the EZ program. Figure 7(b) decomposes these changes a function of city appeal, $x(a, b)$. Locations to the left of the vertical dotted line are those treated by the policy.

markup dispersion and therefore greater scope for within-city allocative gains in larger markets.

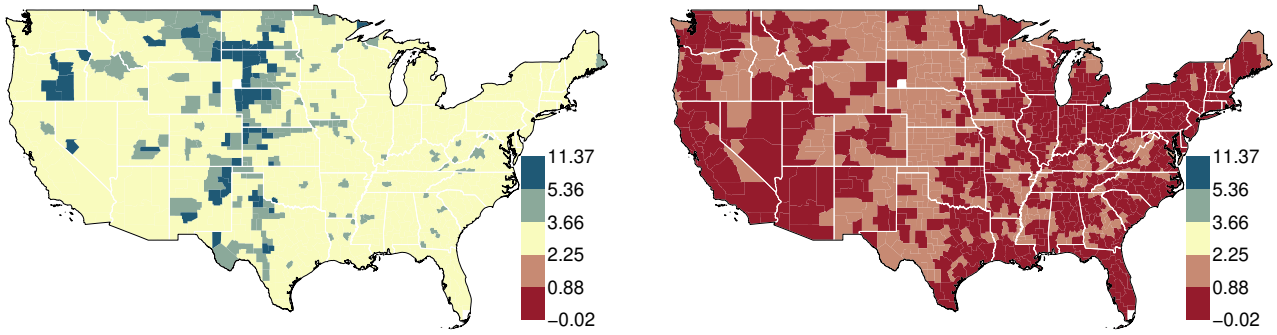
Figure 7 presents the decomposition of TFP changes across space and channels under the EZ program. Figure 7(a) reveals sharper spatial differences in TFP changes, consistent with the more targeted nature of the EZ program. Most treated commuting zones are located in the central part of the country, where TFP gains are the largest, ranging from 1.5% to 2%. Some adjacent commuting zones also benefit indirectly from the policy, as discussed below, experiencing gains of 1% to 1.5%. In contrast, most non-treated commuting zones experience small TFP losses. This group includes major commuting zones, and the losses remain below 0.1%.

Figure 7(b) decomposes the TFP response across the different channels. Locations to the left of the vertical dotted line are those treated by the policy. As under the quasi-optimal policy, firm relocation is the main driver, accounting for most of the gains in treated commuting zones. The policy also generates spillovers in non-treated locations, as marginal firms relocate toward smaller markets, increasing selection among the firms that remain and raising average firm productivity. As a result, firm relocation lowers TFP in non-treated places, but the increase in average firm productivity nearly offsets these losses, leaving only small net declines. These patterns show that the EZ program qualitatively replicates the intended effects of the quasi-optimal policy through a simpler instrument, but with smaller quantitative effects.

The other channels differ more sharply from those under the quasi-optimal policy. Aggregate entry plays a negligible role, as the mass of entrants barely changes, declining by only 0.04%. The within-city misallocation channel is also muted because the EZ program does not eliminate variation in markups within cities, even in treated locations. As a result, allocative efficiency remains essentially unchanged.

I now close the quantitative analysis by comparing the welfare effects of the two policies. Figure 8 illustrates the geography of changes in local mean utility under the quasi-optimal policy and the EZ program. Even though workers are freely mobile across locations, local mean utility provides a

Figure 8: Changes in mean Utility under the Quasi-Optimal and EZ Policies



(a) Quasi-Optimal Policy

(b) EZ Policy

Notes: Figures 8(a) and 8(b) show the percentage change in mean utility under the quasi-optimal policy and the EZ program, respectively, relative to the decentralized equilibrium. Mean utility is defined as real income scaled by local amenities.

useful measure of local welfare for workers who do not change location in response to the policy.[36]

First, Figure 8(a) shows that the welfare gains from the quasi-optimal policy are somewhat more localized than the productivity gains. Less populated commuting zones in the central part of the country benefit the most, with increases in mean utility ranging from 3.5% to 12%. The remaining locations also experience substantial welfare gains, ranging from 2% to 3.5%.

Second, Figure 8(b) displays the welfare effects of the EZ program across commuting zones. Overall, the EZ program delivers much more muted welfare gains than the quasi-optimal policy. Utility in targeted locations increases by no more than 2%. In contrast, the remaining locations experience small welfare losses of about 0.02%. This illustrates that, even though the policy raises welfare in targeted locations, these gains come at the expense of welfare losses elsewhere. These losses reflect both the decline in productivity in non-treated locations and the fact that workers in these locations help finance the program without receiving direct local benefits.

I aggregate local welfare changes to compute the welfare gains from the quasi-optimal policy and the EZ program. The quasi-optimal policy delivers an aggregate welfare gain of 3.1%. This effect is consistent with the patterns in Figure 8(a), where most locations experience welfare gains of this order and only a small number of commuting zones achieve substantially larger gains. The EZ program increases aggregate welfare more modestly, by only 0.03%. Nevertheless, it is worth highlighting that this much smaller program still generates positive aggregate welfare gains, primarily by correcting spatial misallocation in a way that is qualitatively consistent with the rationale of the quasi-optimal policy.

The two policies also differ sharply in their fiscal requirements. Financing the quasi-optimal policy requires a proportional tax equivalent to 17.2% of workers' income, whereas the EZ program requires only 0.13%. For each policy, the ratio of the aggregate welfare gain to the tax rate required to finance is therefore similar: 0.18 for the quasi-optimal policy and 0.23 for the EZ program. This

[36] Mean utility is given by real income scaled by local amenities.

comparison, however, reflects average rather than marginal returns. The quasi-optimal policy expands until the relevant markup distortions are corrected, which lowers its average return. In contrast, the EZ program is evaluated at one particular scale, and its return may change as the program expands.

The aggregate welfare gains from both policies are in line with those reported in the literature. Compared with previous studies of firm misallocation across space, the effects are of similar magnitude to those found by Gaubert (2018) and Bilal (2023). The welfare gains from eliminating markups under the quasi-optimal policy, however, are considerably smaller than those reported by Edmond, Midrigan, and Xu (2023). Their framework differs from mine along several dimensions. In particular, it does not feature a spatial component but allows for elastic labor supply. Introducing an elastic labor-supply margin into the current framework could substantially amplify the welfare gains from reducing markups, since the aggregate markup effectively acts as a tax on output and therefore also distorts labor supply.

6 Conclusion

This paper has developed a quantitative theory of competition in local markets in which heterogeneous producers have output market power and choose where to operate. The model shows how market size and local competition jointly shape firm location decisions and markups across space. Entry externalities associated with variable markups lead productive firms to over-concentrate in larger markets, generating spatial misallocation. I document substantial spatial heterogeneity in markups using U.S. Census microdata. Local producers in larger markets charge significantly lower markups than those in smaller markets. I then use the framework to quantify the effects of place-based policies. A quasi-optimal policy that eliminates markups reallocates productive firms toward smaller markets, generates large local productivity gains, and raises aggregate welfare. A wage-bill subsidy targeted to smaller locations produces qualitatively similar but substantially smaller effects. A central message of the paper is therefore that local market power generates spatial misallocation through firms' location choices, creating a rationale for place-based policies that redirect productive firms toward smaller markets.

The framework can be extended in two promising directions. First, the current analysis abstracts from studying specific sectors. Yet, I document that sectors exhibit distinct patterns in markups across cities, suggesting that the forces shaping firm location and local competition may vary across industries. The setting could therefore be used to study the joint design of industrial and spatial policies. Second, the model can be extended to study the spatial expansion of multi-location firms. Their presence affects local competition and markups across markets. However, it is not clear whether firms' expansion decisions are efficient and whether there is scope for policy to improve the spatial configuration of their establishments.

References

Ackerberg, Daniel et al. (2007). "Econometric Tools for Analyzing Market Outcomes". In: *Handbook of Econometrics*. Ed. by J.J. Heckman and E.E. Leamer. 1st ed. Vol. 6A. Elsevier. Chap. 63.

Ackerberg, Daniel A., Kevin Caves, and Garth Frazer (2015). "Identification Properties of Recent Product Function Estimators". In: *Econometrica* 83.6, pp. 2411–2451.

Aghion, Philippe et al. (Feb. 2023). "A Theory of Falling Growth and Rising Rents". In: *The Review of Economic Studies*, rdad016.

Akcigit, Ufuk and Sina T. Ates (2023). "What Happened to US Business Dynamism?" In: *Journal of Political Economy* 131.8, pp. 2059–2124.

Amiti, Mary, Oleg Itskhoki, and Jozef Konings (Feb. 2019). "International Shocks, Variable Markups, and Domestic Prices". In: *The Review of Economic Studies* 86.6, pp. 2356–2402.

Anderson, Eric, Sergio Rebelo, and Arlene Wong (Mar. 2023). *Markups Across Space and Time*. Working Paper 24434. National Bureau of Economic Research.

Arellano, Manuel and Olympia Bover (1995). "Another Look at the Instrumental Variable Estimation of Error-Components Models". In: *Journal of Econometrics* 68.1, pp. 29–51.

Arnoud, Antoine, Fatih Guvenen, and Tatjana Kleineberg (Oct. 2019). *Benchmarking Global Optimizers*. Working Paper 26340. National Bureau of Economic Research.

Atalay, Enghin (2014). "MATERIALS PRICES AND PRODUCTIVITY". In: *Journal of the European Economic Association* 12.3, pp. 575–611.

Athey, Susan (Feb. 2002). "Monotone Comparative Statics under Uncertainty". In: *The Quarterly Journal of Economics* 117.1, pp. 187–223. ISSN: 0033-5533.

Atkeson, Andrew and Ariel Burstein (Dec. 2008). "Pricing-to-Market, Trade Costs, and International Relative Prices". In: *American Economic Review* 98.5, pp. 1998–2031.

Autor, David, Christina Patterson, and John Van Reenen (Apr. 2023). *Local and National Concentration Trends in Jobs and Sales: The Role of Structural Transformation*. Working Paper 31130. National Bureau of Economic Research.

Autor, David H. and David Dorn (Aug. 2013). "The Growth of Low-Skill Service Jobs and the Polarization of the US Labor Market". In: *American Economic Review* 103.5, pp. 1553–97.

Baqaee, David Rezza, Emmanuel Farhi, and Kunal Sangani (June 2023). "The Darwinian Returns to Scale". In: *The Review of Economic Studies*, rdad061.

Baum-Snow, Nathaniel and Lu Han (2024). "The microgeography of housing supply". In: *Journal of Political Economy* 132.6, pp. 1897–1946.

Behrens, Kristian, Gilles Duranton, and Frédéric Robert-Nicoud (2014). "Productive Cities: Sorting, Selection, and Agglomeration". In: *Journal of Political Economy* 122.3, pp. 507–553.

Berger, David, Kyle Herkenhoff, and Simon Mongey (Apr. 2022). "Labor Market Power". In: *American Economic Review* 112.4, pp. 1147–93.

Berry, Steven, Martin Gaynor, and Fiona Scott Morton (Aug. 2019). "Do Increasing Markups Matter? Lessons from Empirical Industrial Organization". In: *Journal of Economic Perspectives* 33.3, pp. 44–68.

Bilal, Adrien (Mar. 2023). "The Geography of Unemployment". In: *The Quarterly Journal of Economics* 138.3, pp. 1507–1576. ISSN: 0033-5533.

Bilbiie, Florin O., Fabio Ghironi, and Marc J. Melitz (Oct. 2019). "Monopoly Power and Endogenous Product Variety: Distortions and Remedies". In: *American Economic Journal: Macroeconomics* 11.4, pp. 140–74.

Blum, Bernardo S. et al. (2024). "The ABCs of Firm Heterogeneity When Firms Sort into Markets: The Case of Exporters". In: *Journal of Political Economy* 132.4, pp. 1162–1208.

Blundell, Richard and Stephen Bond (1998). "Initial Conditions and Moment Restrictions in Dynamic Panel Data Models". In: *Journal of Econometrics* 87.1, pp. 115–143.

– (2000). "GMM Estimation with Persistent Panel Data: An Application to Production Functions". In: *Econometric Reviews* 19.3, pp. 321–340.

Boerma, Job, Andrea Ottolini, and Aleh Tsyvinski (Dec. 2025). *Multidimensional Sorting: Comparative Statics*. Working Paper 34591. National Bureau of Economic Research.

Bond, Steve et al. (2021). "Some unpleasant markup arithmetic: Production function elasticities and their estimation from production data". In: *Journal of Monetary Economics* 121, pp. 1–14.

Busso, Matias, Jesse Gregory, and Patrick Kline (Apr. 2013). "Assessing the Incidence and Efficiency of a Prominent Place Based Policy". In: *American Economic Review* 103.2, pp. 897–947.

Combes, Pierre-Philippe et al. (2012). "The Productivity Advantages of Large Cities: Distinguishing Agglomeration From Firm Selection". In: *Econometrica* 80.6, pp. 2543–2594.

Costinot, Arnaud and Jonathan Vogel (2010). "Matching and Inequality in the World Economy". In: *Journal of Political Economy* 118.4, pp. 747–786.

Davis, Morris A. and Francois Ortalo-Magne (Apr. 2011). "Household Expenditures, Wages, Rents". In: *Review of Economic Dynamics* 14.2, pp. 248–261.

De Loecker, Jan (2011). "Product Differentiation, Multiproduct Firms, and Estimating the Impact of Trade Liberalization on Productivity". In: *Econometrica* 79.5, pp. 1407–1451.

De Loecker, Jan, Jan Eeckhout, and Simon Mongey (2022). "Quantifying market power and business dynamism in the macroeconomy". In.

De Loecker, Jan, Jan Eeckhout, and Gabriel Unger (Jan. 2020). "The Rise of Market Power and the Macroeconomic Implications". In: *The Quarterly Journal of Economics*.

De Loecker, Jan and Pinelopi Koujianou Goldberg (2014). "Firm Performance in a Global Market". In: *Annual Review of Economics* 6.1, pp. 201–227.

De Loecker, Jan and Chad Syverson (2021). "An Industrial Organization Perspective on Productivity". In: *Working Paper*.

De Loecker, Jan and Frederic Warzynski (May 2012). "Markups and Firm-Level Export Status". In: *American Economic Review* 102.6, pp. 2437–71.

Delgado, Mercedes, Michael E. Porter, and Scott Stern (June 2015). "Defining clusters of related industries". In: *Journal of Economic Geography* 16.1, pp. 1–38.

DellaVigna, Stefano and Matthew Gentzkow (June 2019). "Uniform Pricing in U.S. Retail Chains". In: *The Quarterly Journal of Economics* 134.4, pp. 2011–2084.

DeRidder, Martin, Basile Grassi, and Giovanni Morzenti (2024). "The Hitchhiker's Guide to Markup Estimation: Assessing Estimates from Financial Data". In.

Dhingra, Swati and John Morrow (2019). "Monopolistic Competition and Optimum Product Diversity under Firm Heterogeneity". In: *Journal of Political Economy* 127.1, pp. 196–232.

Diamond, Rebecca and Enrico Moretti (Jan. 2026). *Where is Standard of Living the Highest? Local Prices and the Geography of Consumption*. Working Paper 29533. National Bureau of Economic Research.

Dixit, Avinash K. and Joseph E. Stiglitz (1977). "Monopolistic Competition and Optimum Product Diversity". In: *The American Economic Review* 67.3, pp. 297–308.

Donald, Eric, Masao Fukui, and Yuhei Miyauchi (2026). "Unpacking Aggregate Welfare in a Spatial Economy". In: *The Review of Economic Studies*. Forthcoming.

Edmond, Chris, Virgiliu Midrigan, and Daniel Yi Xu (Oct. 2015). "Competition, Markups, and the Gains from International Trade". In: *American Economic Review* 105.10, pp. 3183–3221.

– (2023). "How Costly Are Markups?" In: *Journal of Political Economy* 131.7, pp. 1619–1675.

Fajgelbaum, Pablo D and Cecile Gaubert (Jan. 2020). "Optimal Spatial Policies, Geography, and Sorting". In: *The Quarterly Journal of Economics* 135.2, pp. 959–1036.

– (2025). "Optimal Spatial Policies". In: *Handbook of Regional and Urban Economics, Volume 6*. Forthcoming. Elsevier.

Fajgelbaum, Pablo D et al. (Sept. 2018). "State Taxes and Spatial Misallocation". In: *The Review of Economic Studies* 86.1, pp. 333–376.

Foster, Lucia, Cheryl Grim, and John Haltiwanger (2016). "Reallocation in the Great Recession: Cleansing or Not?" In: *Journal of Labor Economics* 34.S1, S293–S331.

Foster, Lucia S, John C Haltiwanger, and Chad A Syverson (Mar. 2008). "Reallocation, Firm Turnover, and Efficiency: Selection on Productivity or Profitability?" In: *American Economic Review* 98.1, pp. 394–425.

Foster, Lucia S, John C Haltiwanger, and Cody Tuttle (Sept. 2022). *Rising Markups or Changing Technology?* Working Paper 30491. National Bureau of Economic Research.

Galichon, Alfred (2016). *Optimal transport methods in economics*. Princeton: Princeton University Press.

Gandhi, Amit, Salvador Navarro, and David Rivers (2020). "On the Identification of Gross Output Production Functions". In: *Journal of Political Economy* 128.8, pp. 2973–3016.

Gaubert, Cecile (2018). "Firm Sorting and Agglomeration". In: *American Economic Review* 108.11, pp. 3117–53.

Glaeser, Edward L and Joshua D Gottlieb (2008). *The economics of place-making policies*. Tech. rep. National Bureau of Economic Research.

Hall, Robert E (1988). "The Relation between Price and Marginal Cost in U.S. Industry". In: *Journal of Political Economy* 96.5, pp. 921–47.

Hansen, Bruce E. (Feb. 2014). "214215Nonparametric Sieve Regression: Least Squares, Averaging Least Squares, and Cross-Validation". In: *The Oxford Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics*. Oxford University Press.

Hottman, Colin (2021). *Retail Markups, Misallocation, and Store Variety across U.S. Cities*. Tech. rep.

Hsieh, Chang-Tai and Peter J. Klenow (Nov. 2009). "Misallocation and Manufacturing TFP in China and India". In: *The Quarterly Journal of Economics* 124.4, pp. 1403–1448.

Hsieh, Chang-Tai and Enrico Moretti (Apr. 2019). "Housing Constraints and Spatial Misallocation". In: *American Economic Journal: Macroeconomics* 11.2, pp. 1–39.

Hsieh, Chang-Tai and Esteban Rossi-Hansberg (2023). "The Industrial Revolution in Services". In: *Journal of Political Economy Macroeconomics* 1.1, pp. 3–42.

Kimball, Miles S. (1995). "The Quantitative Analytics of the Basic Neomonetarist Model". In: *Journal of Money, Credit and Banking* 27.4, pp. 1241–1277.

Kleinman, Benny (2025). "Wage Inequality and the Spatial Expansion of Firms".

Klenow, Peter J. and Jonathan L. Willis (2016). "Real Rigidities and Nominal Price Changes". In: *Economica* 83.331, pp. 443–472.

Lhuillier, Hugo (2025). "The local root of wage inequality". In: *University of Chicago, Becker Friedman Institute for Economics Working Paper* 2025-78.

Lindenlaub, Ilse, Ryungha Oh, and Michael Peters (2024). "Spatial Firm Sorting and Local Monopsony Power". In.

Mankiw, N. Gregory and Michael D. Whinston (1986). "Free Entry and Social Inefficiency". In: *The RAND Journal of Economics* 17.1, pp. 48–58.

Matsuyama, Kiminori and Phillip Ushchev (2017). "Beyond CES: Three Alternative Classes of Flexible Homothetic Demand Systems". In.

– (2021). "When Does Procompetitive Entry Imply Excessive Entry". In.

– (2022). "Selection and Sorting of Heterogeneous Firms Through Competitive Pressures". In.

Melitz, Marc J. (2018). "Competitive effects of trade: theory and measurement". In: *Review of World Economics* 154.1.

Monte, Ferdinando, Stephen J. Redding, and Esteban Rossi-Hansberg (Dec. 2018). "Commuting, Migration, and Local Employment Elasticities". In: *American Economic Review* 108.12, pp. 3855–90.

Newey, Whitney K. (1997). "Convergence rates and asymptotic normality for series estimators". In: *Journal of Econometrics* 79.1, pp. 147–168.

Nocke, Volker (2006). "A GAP FOR ME: ENTREPRENEURS AND ENTRY". In: *Journal of the European Economic Association* 4.5, pp. 929–956.

Oberfield, Ezra et al. (2023). "Plants in Space". In: *Journal of Political Economy* 0.ja, null.

Oh, Ryungha (2023). "Spatial sorting of workers and firms". In: *Yale University, Job Market Paper*.

Owen, Guillermo (1977). "Values of Games with a Priori Unions". In: *Mathematical Economics and Game Theory: Essays in Honor of Oskar Morgenstern*. Ed. by Rudolf Henn and Otto Moeschlin. Berlin: Springer, pp. 76–88.

Peters, Michael (2020). "Heterogeneous Markups, Growth, and Endogenous Misallocation". In: *Econometrica* 88.5, pp. 2037–2073.

Redding, Stephen J. and Esteban Rossi-Hansberg (2017). "Quantitative Spatial Economics". In: *Annual Review of Economics* 9.1, pp. 21–58.

Restuccia, Diego and Richard Rogerson (2008). "Policy distortions and aggregate productivity with heterogeneous establishments". In: *Review of Economic Dynamics* 11.4, pp. 707–720.

Rossi-Hansberg, Esteban, Pierre-Daniel Sarte, and Felipe Schwartzman (2026). "Cognitive Hubs and Spatial Redistribution". In: *American Economic Journal: Macroeconomics* 18.2, pp. 72–111.

Rossi-Hansberg, Esteban, Pierre-Daniel Sarte, and Nicholas Trachter (2020). "Diverging Trends in National and Local Concentration". In.

Sangani, Kunal (2025). "Markups Across the Income Distribution: Measurement and Implications". In: *American Economic Journal: Macroeconomics*. Conditionally accepted.

Shapley, Lloyd S. (1953). "A Value for n-Person Games". In: *Contributions to the Theory of Games, Volume II*. Ed. by Harold W. Kuhn and Albert W. Tucker. Princeton: Princeton University Press, pp. 307–317.

Yeh, Chen, Claudia Macaluso, and Brad Hershbein (July 2022). "Monopsony in the US Labor Market". In: *American Economic Review* 112.7, pp. 2099–2138.

Zhelobodko, Evgeny et al. (2012). "Monopolistic Competition: Beyond the Constant Elasticity of Substitution". In: *Econometrica* 80.6, pp. 2765–2784.

Appendix

A Derivations **51**
- A.1 Local Goods Demand . . . 51
- A.2 Klenow-Willis Derivations . . . 51

B Proofs of Baseline Model **53**
- B.1 Proof of Proposition 1 . . . 53
- B.2 Proof of Proposition 2 . . . 55
- B.3 Proof of Corollary 1 . . . 59

C Efficiency **60**
- C.1 Social Planner's Problem . . . 60
- C.2 Proof of Proposition 5 . . . 65
- C.3 First-best Implementation . . . 68

D Proofs of the Extended Model **71**
- D.1 Relative Price and Input Demand Functions . . . 71
- D.2 Extension of Proposition 1 . . . 71
- D.3 Extension of Proposition 2 . . . 75
- D.4 Extension of Corollary 1 . . . 79
- D.5 Equilibrium Conditions Quasi-Optimal Policy . . . 79
- D.6 Equilibrium Conditions EZ Policy . . . 81
- D.7 Decomposition of Local TFP Changes . . . 83

E Markup Estimation **86**
- E.1 Derivations for Estimating Equation . . . 86
- E.2 Consistency of Sieve Estimator . . . 87
- E.3 Multi-Sector Estimation . . . 88
- E.4 Labor Market Power . . . 88
- E.5 General Production Function . . . 90

F Empirical Appendix **91**
- F.1 Additional Figures . . . 91
- F.2 Additional Tables . . . 95
- F.3 Markups Across Cities for additional Sectors . . . 96
- F.4 Robustness Using the Census of Manufactures . . . 100
  - F.4.1 General Production Function . . . 100
  - F.4.2 Cost-Shares Approach . . . 102

G Quantitative Appendix **104**
- G.1 GMM Estimation . . . 104
- G.2 Estimation of City Fundamentals . . . 105
- G.3 Additional Figures . . . 106

A Derivations

A.1 Local Goods Demand

For a given $Y(a)$, consumers minimize total expenditure on local goods subject to the utility constraint (2). The associated Lagrangian is

$$\mathcal{L}^W(a) = \int_z p(z,a)y(z,a)dG_a(z) + \lambda^W(a)\left[1 - \int_z \Upsilon\left(\frac{y(z,a)}{Y(a)}\right)dG_a(z)\right],$$

where $\lambda^W(a)$ is the Lagrange multiplier. The first-order condition with respect to $y(z,a)$ is

$$p(z,a) = \frac{\lambda^W(a)}{Y(a)}\Upsilon'\left(\frac{y(z,a)}{Y(a)}\right).$$

Defining the competition price index as

$$\mathbb{D}(a) \equiv \frac{\lambda^W(a)}{Y(a)},$$

we can express the inverse demand for a single variety as

$$\frac{p(z,a)}{\mathbb{D}(a)} = \Upsilon'\left(\frac{y(z,a)}{Y(a)}\right).$$

Similarly, defining $\varphi(\cdot) \equiv (\Upsilon')^{-1}(\cdot)$, the demand function for a single variety is

$$\frac{y(z,a)}{Y(a)} = \varphi\left(\frac{p(z,a)}{\mathbb{D}(a)}\right).$$

Using these expressions, the competition price index $\mathbb{D}(a)$ satisfies

$$\int_z \Upsilon\left(\varphi\left(\frac{p(z,a)}{\mathbb{D}(a)}\right)\right)dG_a(z) = 1.$$

Finally, given $\mathbb{D}(a)$ and firm-level prices, the ideal price index $\mathbb{P}(a)$ is

$$\mathbb{P}(a) = \mathbb{D}(a)\int_z \frac{p(z,a)}{\mathbb{D}(a)}\varphi\left(\frac{p(z,a)}{\mathbb{D}(a)}\right)dG_a(z). \tag{A1}$$

A.2 Klenow-Willis Derivations

The functional form (29) implies that the price elasticity $\sigma(\cdot)$ takes the form

$$\sigma\left(\frac{p(z,a)}{\mathbb{D}(a)}\right) = \frac{\overline{\sigma}}{1+\varepsilon\log\frac{\overline{\sigma}-1}{\overline{\sigma}}-\varepsilon\log\frac{p(z,a)}{\mathbb{D}(a)}}.$$

Moreover, the optimal relative price function is

$$\frac{p(z,a)}{\mathbb{D}(a)} = \frac{\overline{\sigma}}{\varepsilon}\frac{\mathbb{C}(a)}{z}\frac{1}{\Omega\left(\overline{\lambda}\frac{\mathbb{C}(a)}{z}\right)},$$

where $\overline{\lambda}$ is a constant, and $\Omega(\cdot)$ is the main branch of the Lambert-W function.[37]

The optimal relative quantity is

$$\log\frac{y(z,a)}{Y(a)} = \frac{\overline{\sigma}}{\varepsilon}\log\left(1+\varepsilon\log\frac{\overline{\sigma}-1}{\overline{\sigma}}-\varepsilon\log\frac{p(z,a)}{\mathbb{D}(a)}\right).$$

Since local producers operate under monopolistic competition, they never produce in the inelastic region of the demand curve. Formally, producers set $p(z,a)/\mathbb{D}(a)$ such that

$$\sigma\left(\frac{p(z,a)}{\mathbb{D}(a)}\right) > 1 \iff \frac{p(z,a)}{\mathbb{D}(a)} > \frac{\overline{\sigma}-1}{\overline{\sigma}}\exp\left(-\frac{\overline{\sigma}-1}{\varepsilon}\right).$$

Similarly, Klenow and Willis (2016) define a choke price, ensuring that optimal relative quantities remain positive. From the expression for $y(z,a)/Y(a)$, the condition on the relative price is

$$\frac{p(z,a)}{\mathbb{D}(a)} < \frac{\overline{\sigma}-1}{\overline{\sigma}}\exp\left(\frac{1}{\varepsilon}\right).$$

[37] $\overline{\lambda} \equiv \frac{\overline{\sigma}^2}{(\overline{\sigma}-1)\varepsilon}\exp\left(\frac{\overline{\sigma}-1}{\varepsilon}\right)$, and $\Omega(x)$ is implicitly defined by $x = \Omega(x)\exp(\Omega(x))$.

B Proofs of Baseline Model

B.1 Proof of Proposition 1

The proof of Proposition 1 consists of three steps. First, we establish positive assortative matching conditional on the competition index $\mathbb{C}(a)$. Second, we characterize general equilibrium objects when firms sort based on competition. Third, we show that positive assortative matching holds when the competition index is determined in general equilibrium.

Step 1: Sorting conditional on the competition index. The optimal relative price and markup depend on a firm's location only through the aggregate variable $\mathbb{C}(a)$. Following the insights of Bilal (2023), we index locations by their competition index $\mathbb{C}$ rather than by their productivity a, temporarily considering the inverse function $a(\mathbb{C})$.[38] Under this formulation, the profit function (12) can be rewritten as:

$$\log \Pi(z, \mathbb{C}) = \log M(\mathbb{C}) + \log \psi\left(\frac{\mathbb{C}}{z}\right) + \log \varphi\left(\psi\left(\frac{\mathbb{C}}{z}\right)\right) - \log \frac{\mathbb{P}(\mathbb{C})}{\mathbb{D}(\mathbb{C})} + \log\left(1 - \frac{\mathbb{C}}{z}\frac{1}{\psi\left(\frac{\mathbb{C}}{z}\right)}\right), \quad \text{(B1)}$$

where we have substituted the optimal relative prices and quantities from the profit maximization problem. In this alternative formulation, firms choose their location based on the competition index $\mathbb{C}$ rather than directly on city productivity a.

To prove the existence of a strictly increasing assignment between $\mathbb{C}$ and z, we apply the standard methods from the assignment literature (e.g., Galichon (2016)). By the envelope theorem,

$$\frac{\partial \log \Pi(z, \mathbb{C})}{\partial \mathbb{C}} = \frac{M'(\mathbb{C})}{M(\mathbb{C})} - \frac{\mathbb{P}'(\mathbb{C})}{\mathbb{P}(\mathbb{C})} + \frac{\mathbb{D}'(\mathbb{C})}{\mathbb{D}(\mathbb{C})} - \frac{1}{z\psi\left(\frac{\mathbb{C}}{z}\right) - \mathbb{C}}.$$

Substituting the first-order condition (13), we obtain

$$\frac{\partial \log \Pi(z, \mathbb{C})}{\partial \mathbb{C}} = \frac{M'(\mathbb{C})}{M(\mathbb{C})} - \frac{\mathbb{P}'(\mathbb{C})}{\mathbb{P}(\mathbb{C})} + \frac{\mathbb{D}'(\mathbb{C})}{\mathbb{D}(\mathbb{C})} - \frac{\sigma\left(\psi\left(\frac{\mathbb{C}}{z}\right)\right) - 1}{\mathbb{C}}.$$

Differentiating again, we obtain

$$\frac{\partial^2 \log \Pi(z, \mathbb{C})}{\partial z \partial \mathbb{C}} = \frac{1}{z^2}\sigma'\left(\psi\left(\frac{\mathbb{C}}{z}\right)\right)\underbrace{\psi'\left(\frac{\mathbb{C}}{z}\right)}_{>0} > 0 \iff \sigma'(\cdot) > 0.$$

Thus, the profit function is log-supermodular if and only if the elasticity of demand is increasing in the relative price. Since Marshall's second law holds under the specification of Klenow and Willis (2016), there exists a strictly increasing assignment function between z and $\mathbb{C}$, denoted by $z(\mathbb{C})$.

[38] See Bilal (2023), Appendix B.3.3.

Step 2: General equilibrium objects. We now derive expressions for the general equilibrium objects under the positive assortative matching (PAM) between z and $\mathbb{C}$. First, given the strictly increasing assignment $z(\mathbb{C})$, the definition of the competition price index (8) implies:[39]

$$z'(\mathbb{C})\frac{M_e g(z(\mathbb{C}))}{f_{\mathbb{C}}(\mathbb{C})} = \frac{1}{\Upsilon\left(\varphi\left(\psi\left(\frac{\mathbb{C}}{z(\mathbb{C})}\right)\right)\right)}, \tag{B2}$$

where $f_{\mathbb{C}}(\mathbb{C})$ is the equilibrium density of $\mathbb{C}$. Using this expression, the ideal price index in (A1) simplifies to

$$\mathbb{P}(\mathbb{C}) = \mathbb{D}(\mathbb{C})\frac{\psi\left(\frac{\mathbb{C}}{z(\mathbb{C})}\right)\varphi\left(\psi\left(\frac{\mathbb{C}}{z(\mathbb{C})}\right)\right)}{\Upsilon\left(\varphi\left(\psi\left(\frac{\mathbb{C}}{z(\mathbb{C})}\right)\right)\right)}.$$

Define $\delta(\mathbb{C}/z(\mathbb{C}))$ as the ratio $\mathbb{D}(\mathbb{C})/\mathbb{P}(\mathbb{C})$:

$$\delta\left(\frac{\mathbb{C}}{z(\mathbb{C})}\right) \equiv \frac{\mathbb{D}(\mathbb{C})}{\mathbb{P}(\mathbb{C})} = \frac{\Upsilon\left(\varphi\left(\psi\left(\frac{\mathbb{C}}{z(\mathbb{C})}\right)\right)\right)}{\psi\left(\frac{\mathbb{C}}{z(\mathbb{C})}\right)\varphi\left(\psi\left(\frac{\mathbb{C}}{z(\mathbb{C})}\right)\right)}.$$

Then, the ideal price index $\mathbb{P}(\mathbb{C})$ is given by:

$$\mathbb{P}(\mathbb{C}) = \frac{W(\mathbb{C})}{\mathbb{C}\delta\left(\frac{\mathbb{C}}{z(\mathbb{C})}\right)}. \tag{B3}$$

On the other hand, household consumption and labor decisions, (6) and (11), imply that the market size $M(\mathbb{C})$ is expressed as:

$$M(\mathbb{C}) = \eta\kappa\frac{(W(\mathbb{C})+s)^{\frac{1}{\alpha}}}{\mathbb{P}(\mathbb{C})^{\frac{\eta}{\alpha}}}. \tag{B4}$$

Substituting (B3) into (B4) and using the equilibrium wage condition (22), we obtain:

$$M(\mathbb{C}) = \eta\kappa\mathbb{C}^{\frac{\eta}{\alpha}}\delta\left(\frac{\mathbb{C}}{z(\mathbb{C})}\right)^{\frac{\eta}{\alpha}}\frac{(a(\mathbb{C})+s)^{\frac{1}{\alpha}}}{a(\mathbb{C})^{\frac{\eta}{\alpha}}}. \tag{B5}$$

Step 3: Sorting in general equilibrium. In this final step, we determine under what conditions more productive cities are also more competitive. Equivalently, we characterize when the function $a(\mathbb{C})$ is increasing in general equilibrium. The first-order condition of (B1) is

[39] See Costinot and Vogel (2010) for a detailed exposition of derivatives of assignment functions in monotone assignments.

$$\mathcal{E}_M(\mathbb{C}) + \mathcal{E}_\delta(\mathbb{C}) = \sigma\left(\psi\left(\frac{\mathbb{C}}{z(\mathbb{C})}\right)\right) - 1,$$

where $\mathcal{E}_f(x)$ denotes the elasticity of any function $f(\cdot)$ with respect to x, defined as $\mathcal{E}_f(x) \equiv \frac{f'(x)x}{f(x)}$. Substituting (B5) into this condition yields

$$\eta + (\alpha + \eta)\mathcal{E}_\delta(\mathbb{C}) + (\widetilde{a}(\mathbb{C}) - \eta)\mathcal{E}_a(\mathbb{C}) = \alpha\left[\sigma\left(\psi\left(\frac{\mathbb{C}}{z(\mathbb{C})}\right)\right) - 1\right],$$

where $\widetilde{a}(\mathbb{C}) \equiv \frac{a(\mathbb{C})}{a(\mathbb{C})+s}$. Under the Klenow and Willis (2016) specification, when $\varepsilon = 0$, the functions $\sigma(\cdot)$ and $\delta(\cdot)$ are constant, taking values $\overline{\sigma}$ and $\frac{\overline{\sigma}}{\overline{\sigma}-1}$, respectively. In this case, the expression simplifies to

$$(\widetilde{a}(\mathbb{C}) - \eta)\mathcal{E}_a(\mathbb{C}) = \alpha\left(\overline{\sigma} - 1\right) - \eta.$$

To be fully rigorous, we also assume that $\underline{a}/\overline{a} > \frac{\alpha\eta}{(1-\alpha)(1-\eta)}$. The mutual fund budget constraint (27) implies that $s \leq \frac{\alpha}{1-\alpha}\overline{a}$. Then, this condition guarantees that $\widetilde{a}(\mathbb{C}) - \eta > 0$ for all cities. It follows that $\mathcal{E}_a(\mathbb{C}) > 0$ if and only if $\alpha\left(\overline{\sigma} - 1\right) > \eta$. This condition can be written equivalently as

$$\alpha > \frac{\eta}{\overline{\sigma} - 1}.$$

This result implies that there exists a region of the parameter space such that, for sufficiently small ε, PAM holds in general equilibrium. In particular, both $z(\mathbb{C})$ and $a(\mathbb{C})$ are strictly increasing, ensuring that more productive firms locate in more productive cities.

B.2 Proof of Proposition 2

The proof of Proposition 2 consists of three steps. First, we derive the system of ODEs that characterizes the equilibrium for the local goods sector. Second, we prove the existence of a solution to this system, conditional on the general equilibrium aggregates M_e and s. Third, we establish the existence and uniqueness of general equilibrium objects.

Step 1: ODE system for local goods producers. Imposing Assumption 1 and the assumptions of Proposition 1, positive assortative matching (PAM) between firm and location productivity holds. The definition of the competition price index (8) then implies:

$$z'(a) = \frac{f(a)}{g(z(a))M_e}\frac{1}{\Upsilon\left(\varphi\left(\psi\left(\frac{\mathbb{C}(a)}{z(a)}\right)\right)\right)}. \tag{B6}$$

Furthermore, the first-order condition of the location problem (21) is:

$$\mathcal{E}_M(a) + \mathcal{E}_\delta(a) = \frac{1}{\mu\left(\psi\left(\frac{\mathbb{C}(a)}{z(a)}\right)\right) - 1}\mathcal{E}_\mathbb{C}(a), \tag{B7}$$

and the equilibrium expression for market size (B5) simplifies to:

$$M(a) = \eta\kappa\mathbb{C}(a)^{\frac{\eta}{\alpha}}\delta\left(\frac{\mathbb{C}(a)}{z(a)}\right)^{\frac{\eta}{\alpha}}\frac{(a+s)^{\frac{1}{\alpha}}}{a^{\frac{\eta}{\alpha}}}.$$

Taking logs and differentiating, we obtain:

$$\mathcal{E}_M(a) = \frac{\eta}{\alpha}\mathcal{E}_\mathbb{C}(a) + \frac{\eta}{\alpha}\mathcal{E}_\delta(a) + \frac{\widetilde{a} - \eta}{\alpha},$$

with $\widetilde{a} \equiv a/(a+s)$. To derive an expression for $\mathcal{E}_\delta(a)$, note that under PAM, equations (A1) and (B6) yield:

$$\delta\left(\frac{\mathbb{C}(a)}{z(a)}\right) = \frac{\Upsilon\left(\varphi\left(\psi\left(\frac{\mathbb{C}(a)}{z(a)}\right)\right)\right)}{\psi\left(\frac{\mathbb{C}(a)}{z(a)}\right)\varphi\left(\psi\left(\frac{\mathbb{C}(a)}{z(a)}\right)\right)}. \tag{B8}$$

Differentiating, we obtain:

$$\mathcal{E}_\delta(a) = \Theta\left(\frac{\mathbb{C}(a)}{z(a)}\right)\mathcal{E}_z(a) - \Theta\left(\frac{\mathbb{C}(a)}{z(a)}\right)\mathcal{E}_\mathbb{C}(a),$$

where

$$\Theta\left(\frac{\mathbb{C}(a)}{z(a)}\right) = \rho\left(\frac{\mathbb{C}(a)}{z(a)}\right)\sigma\left(\frac{\mathbb{C}(a)}{z(a)}\right)\left[\frac{1}{\delta\left(\frac{\mathbb{C}(a)}{z(a)}\right)} - \frac{1}{\mu\left(\frac{\mathbb{C}(a)}{z(a)}\right)}\right]. \tag{B9}$$

For simplicity, we omit the dependence of $\sigma(\cdot)$ and $\mu(\cdot)$ on physical quantities in the notation:

$$\sigma\left(\frac{\mathbb{C}(a)}{z(a)}\right) \equiv \sigma\left(\psi\left(\frac{\mathbb{C}(a)}{z(a)}\right)\right), \qquad \mu\left(\frac{\mathbb{C}(a)}{z(a)}\right) \equiv \mu\left(\psi\left(\frac{\mathbb{C}(a)}{z(a)}\right)\right) = \frac{\sigma\left(\psi\left(\frac{\mathbb{C}(a)}{z(a)}\right)\right)}{\sigma\left(\psi\left(\frac{\mathbb{C}(a)}{z(a)}\right)\right) - 1},$$

and $\rho(\cdot)$ denotes the pass-through rate:

$$\rho\left(\frac{\mathbb{C}(a)}{z(a)}\right) \equiv \frac{1}{1 + \frac{\psi\left(\frac{\mathbb{C}(a)}{z(a)}\right)\sigma'\left(\psi\left(\frac{\mathbb{C}(a)}{z(a)}\right)\right)}{\sigma\left(\frac{\mathbb{C}(a)}{z(a)}\right)\left[\sigma\left(\frac{\mathbb{C}(a)}{z(a)}\right) - 1\right]}},$$

where $\sigma'(\cdot)$ denotes the derivative of the price elasticity with respect to the relative price.

Substituting the expressions for $\mathcal{E}_M(a)$ and $\mathcal{E}_\delta(a)$ into (B7), we obtain:

$$(\alpha+\eta)\Theta\left(\frac{\mathbb{C}(a)}{z(a)}\right)\frac{z'(a)a}{z(a)}+(\widetilde{a}-\eta)=\frac{\mathbb{C}'(a)a}{\mathbb{C}(a)}\left[\frac{\alpha}{\mu\left(\frac{\mathbb{C}(a)}{z(a)}\right)-1}+(\alpha+\eta)\Theta\left(\frac{\mathbb{C}(a)}{z(a)}\right)-\eta\right], \tag{B10}$$

where we explicitly express the elasticities $\mathcal{E}_z(a)$ and $\mathcal{E}_\mathbb{C}(a)$. Given M_e, equations (B6) and (B10) define a coupled system of ODEs with boundary conditions $z(\overline{a})=\overline{z}$ and $z(\underline{a})=\underline{z}$. The first boundary condition states that the most productive firms locate in the most productive city, while the second ensures that the least productive firms settle in the least productive city. The solution to this system determines the assignment function $z(a)$ and the competition index function $\mathbb{C}(a)$.

Step 2: Existence and uniqueness of a solution to the ODE system given M_e and s. The system (B6) - (B10) is Lipschitz continuous and, therefore, has a unique solution if two initial conditions are provided (by the Picard–Lindelöf Theorem). However, this system constitutes a two-point boundary value problem, as it has one initial and one terminal condition. To address this, we transform the problem into an initial value problem and apply a shooting method.

Define $\overline{\mathbb{C}}\equiv\mathbb{C}(\overline{a})$. Given M_e and the terminal condition $z(\overline{a})=\overline{z}$, we show that there exists $\overline{\mathbb{C}}$ such that $z(\underline{a})=\underline{z}$. By Lipschitz continuity, $z(\underline{a})$ is a continuous function of $\overline{\mathbb{C}}$. Since both $z(a)$ and $\mathbb{C}(a)$ are strictly increasing, we obtain that as $\overline{\mathbb{C}}\to\infty$, $z(\underline{a})\to 0$, whereas when $\overline{\mathbb{C}}\to 0$, we have $z(\underline{a})\to\overline{z}>\underline{z}$.

By the intermediate value theorem, there exists at least one $\overline{\mathbb{C}}$ such that $z(\underline{a})=\underline{z}$. Moreover, when the supports of $F(\cdot)$ and $G(\cdot)$ are sufficiently small, $\overline{\mathbb{C}}$ is unique. Up to first order, $z'(a)$ is constant and strictly increasing in $\overline{\mathbb{C}}$, since $\Upsilon\left(\varphi\left(\psi\left(\cdot\right)\right)\right)$ is strictly decreasing. Hence $z(\underline{a})$ is strictly decreasing in $\overline{\mathbb{C}}$. Given the unique $\overline{\mathbb{C}}$, the Picard–Lindelöf Theorem delivers a unique solution to the ODE system.

Step 3: Existence and uniqueness of M_e, s and $\overline{U}$. Denote $\mathbb{C}(a;M_e,s)$ and $z(a;M_e,s)$ the solution to the system (B6) - (B10). First, the aggregate labor adding-up constraint uniquely determines $\overline{U}$:

$$\overline{U}^{\frac{1}{\alpha}}=\frac{1}{\alpha}\int_a\frac{(a+s)^{\frac{1-\alpha}{\alpha}}}{\mathbb{P}(a;M_e,s)^{\frac{\eta}{\alpha}}}f(a)da,$$

where $\mathbb{P}(a;M_e,s)=\frac{a}{\mathbb{C}(a;M_e,s)\delta\left(\frac{\mathbb{C}(a;M_e,s)}{z(a;M_e,s)}\right)}$. Now, we establish the existence and uniqueness of M_e, given s. We examine the free entry condition (26):

$$c_eM_e=\int_a\frac{M(a;M_e,s)}{\sigma\left(\frac{\mathbb{C}(a;M_e,s)}{z(a;M_e,s)}\right)}f(a)da. \tag{B11}$$

For existence, note first that $\Upsilon(\cdot)$ is bounded above by a constant $\overline{\Upsilon}$ under Assumption 1. Then,

(B6) implies that any solution requires $M_e \geq M_e^{min} \equiv \frac{(\overline{a}-\underline{a})\underline{f}}{(\overline{z}-\underline{z})\overline{g}\,\overline{\Upsilon}}$, where $\underline{f}$ and $\overline{g}$ bound the densities $f(\cdot)$ and $g(\cdot)$. The left-hand side (LHS) of (B11) is increasing in M_e. As $M_e \to \infty$, (B6) implies that $\mathbb{C}(a) \to \infty$ for all a. Otherwise, $z'(a) \to 0$ for all a, violating the initial condition $z(\underline{a}) = \underline{z}$. Since profits in all locations decrease with local competition, the right-hand side (RHS) of (B11) tends to zero as $M_e \to \infty$. At $M_e = M_e^{min}$, competition is at its weakest and the RHS is bounded away from zero. Provided the entry cost is not too large, the LHS at M_e^{min} lies below the RHS. By the intermediate value theorem, there exists at least one $M_e \in [M_e^{min}, \infty)$ satisfying (B11), for a given s.

The proof of uniqueness of M_e is more challenging. Due to the highly nonlinear nature of the ODE system (B6) - (B10), characterizing the monotonicity of the RHS of (B11) is difficult without additional structure. To address this, we assume that the supports of $F(\cdot)$ and $G(\cdot)$ are sufficiently small. This assumption allows us to use a first-order approximation of the equilibrium conditions.

Up to first order, (B6) can be written as

$$M_e z'(a) \approx \frac{f(\underline{a})}{g(\underline{z})} \frac{1}{\Upsilon\left(\varphi\left(\psi\left(\frac{\underline{\mathbb{C}}}{\underline{z}}\right)\right)\right)},$$

where $\underline{\mathbb{C}} \equiv \mathbb{C}(\underline{a})$. Integrating this expression yields

$$M_e = \frac{\overline{a}-\underline{a}}{\overline{z}-\underline{z}} \frac{f(\underline{a})}{g(\underline{z})} \frac{1}{\Upsilon\left(\varphi\left(\psi\left(\frac{\underline{\mathbb{C}}}{\underline{z}}\right)\right)\right)}. \tag{B12}$$

Similarly, the aggregate labor adding-up constraint implies that $L(a) \approx 1$ up to first order. Then, market size satisfies $M(\underline{a}) \approx \eta(\underline{a}+s)$, and the free entry condition (B11) simplifies to

$$c_e M_e = \frac{\eta(\underline{a}+s)}{\sigma\left(\frac{\underline{\mathbb{C}}}{\underline{z}}\right)} f(\underline{a})(\overline{a}-\underline{a}). \tag{B13}$$

Substituting M_e from (B12) into (B13) gives

$$\frac{1}{\Upsilon\left(\varphi\left(\psi\left(\frac{\underline{\mathbb{C}}}{\underline{z}}\right)\right)\right) g(\underline{z})(\overline{z}-\underline{z})} = \frac{\eta(\underline{a}+s)}{\sigma\left(\frac{\underline{\mathbb{C}}}{\underline{z}}\right) c_e}. \tag{B14}$$

The LHS of (B14) is strictly increasing in $\underline{\mathbb{C}}$, since $\Upsilon\left(\varphi\left(\psi\left(\cdot\right)\right)\right)$ is strictly decreasing. The RHS is strictly decreasing under MSLD, since the price elasticity rises with local competition. Thus, (B14) uniquely determines $\underline{\mathbb{C}}$. Given $\underline{\mathbb{C}}$, either (B12) or (B13) uniquely determine M_e, given s.

Finally, we show existence and uniqueness of s. First, noting that $\int L(a) f(a)\,\mathrm{d}a = 1$, the mutual-fund budget constraint is

$$s = \frac{\alpha}{1-\alpha}\int_{\underline{a}}^{\overline{a}} a\, L\Big(a; M_e(s), s, \overline{U}(M_e(s), s)\Big) f(a)\, da \equiv \mathcal{T}(s) \tag{B15}$$

Then transfers satisfy the fixed-point equation $s = \mathcal{T}(s)$. To show existence, note that because $L(\cdot)$ is a density over cities, we have $\underline{a} \leq \int aL(a)f(a)\,\mathrm{d}a \leq \overline{a}$. Therefore, $\mathcal{T}(s) \in \left[\frac{\alpha}{1-\alpha}\underline{a}, \frac{\alpha}{1-\alpha}\overline{a}\right] \equiv \mathcal{I}$ for all s. Hence, $\mathcal{T}$ is a continuous map from the compact interval $\mathcal{I}$ into itself. Brouwer's fixed point theorem implies existence of a fixed point $s^* \in \mathcal{I}$.

For uniqueness, we use the first-order approximation to the system. When supports are sufficiently small, city-level objects are nearly constant at first order. Then

$$\int a\, L(a) f(a)\, da \approx \underline{a}.$$

Plugging into (B15) yields $s^* = (\alpha\underline{a})/(1-\alpha)$.

B.3 Proof of Corollary 1

The limiting case economy considers a scenario where the support of $G(\cdot)$ is wide, but the support of $F(\cdot)$ is small. In this case, we index locations by their competition index, $\mathbb{C}$, rather than by their productivity a. The first-order condition (B7) now takes the form:

$$\mathcal{E}_M(\mathbb{C}) + \mathcal{E}_\delta(\mathbb{C}) = \sigma\left(\psi\left(\frac{\mathbb{C}}{z(\mathbb{C})}\right)\right) - 1.$$

Solving for the elasticities $\mathcal{E}_M(\mathbb{C})$ and $\mathcal{E}_\delta(\mathbb{C})$ and substituting them into the above expression yields

$$(\alpha+\eta)\Theta\left(\frac{\mathbb{C}}{z(\mathbb{C})}\right)\frac{z'(\mathbb{C})\mathbb{C}}{z(\mathbb{C})} + (\tilde{a}(\mathbb{C}) - \eta)\frac{a'(\mathbb{C})\mathbb{C}}{a(\mathbb{C})} = \frac{\alpha}{\mu\left(\frac{\mathbb{C}}{z(\mathbb{C})}\right) - 1} + (\alpha+\eta)\Theta\left(\frac{\mathbb{C}}{z(\mathbb{C})}\right) - \eta.$$

In the limiting economy, shrinking the support of $F(\cdot)$ implies that $a'(\mathbb{C}) \to 0$. Consequently, the first-order condition of the local producers defines a non-degenerate assignment $z(\mathbb{C})$:

$$(\alpha+\eta)\Theta\left(\frac{\mathbb{C}}{z(\mathbb{C})}\right)\frac{z'(\mathbb{C})\mathbb{C}}{z(\mathbb{C})} = \frac{\alpha}{\mu\left(\frac{\mathbb{C}}{z(\mathbb{C})}\right) - 1} + (\alpha+\eta)\Theta\left(\frac{\mathbb{C}}{z(\mathbb{C})}\right) - \eta.$$

The supermodularity properties of the profit function $\Pi(z, \mathbb{C})$ derived in Appendix B.1 remain valid in this limiting economy. Hence, there is positive assortative matching, $z'(\mathbb{C}) > 0$. Moreover, the first-order condition implies that $(\alpha+\eta)\Theta\left(\mathbb{C}/z(\mathbb{C})\right)\left[\mathcal{E}_z(\mathbb{C}) - 1\right] = \alpha\left[\sigma\left(\mathbb{C}/z(\mathbb{C})\right) - 1\right] - \eta$. Then, the market size elasticity is $\mathcal{E}_M(\mathbb{C}) = \eta/\alpha + (\eta/\alpha)\Theta\left(\mathbb{C}/z(\mathbb{C})\right)\left[\mathcal{E}_z(\mathbb{C}) - 1\right] = \eta\sigma\left(\mathbb{C}/z(\mathbb{C})\right)/(\alpha+\eta) > 0$. Therefore, more competitive cities are larger, $M'(\mathbb{C}) > 0$.

C Efficiency

C.1 Social Planner's Problem

A utilitarian planner seeks to maximize workers' utility, $\overline{U}$. The planner selects the consumption of local goods, $Y(a)$, housing, $H(a)$, and traded goods, $Q(a)$, in each location, subject to local housing market clearing. On the production side, the planner determines the allocation of workers between the traded and local goods sectors, $L^T(a)$ and $L^N(a)$, subject to the aggregate population constraint. For the location of local goods producers, I anticipate that the planner will choose PAM. Thus, the planner selects the matching function $z(a)$, its slope, and the mass of entrants, M_e. To simplify the derivations, I use an alternative formulation in which, instead of directly choosing the slope of the assignment function, the planner selects the mass of active firms in city a, $\zeta(a) \equiv \frac{z'(a)M_e g(z(a))}{f(a)}$.[40] Thus, the planner maximizes $\overline{U}$ subject to:

$$\begin{aligned}
\int_{\underline{a}}^{\overline{a}} \left(L^T(a) + L^N(a)\right) f(a)da &= 1, \\
c_e M_e + \int_{\underline{a}}^{\overline{a}} Q(a) \left(L^T(a) + L^N(a)\right) f(a)da &= \int_{\underline{a}}^{\overline{a}} aL^T(a)f(a)da, \\
H(a) \left(L^T(a) + L^N(a)\right) &= 1 \; \forall \, a, \\
\Upsilon \left(\frac{z(a)L^N(a)}{\zeta(a) \left(L^T(a) + L^N(a)\right) Y(a)} \right) \zeta(a) &= 1 \; \forall \, a, \\
\int_{\underline{a}}^{a} \zeta(r)f(r)dr &= M_e G(z(a)) \; \forall \, a, \\
\left(\frac{Y(a)}{\eta}\right)^{\eta} \left(\frac{H(a)}{\alpha}\right)^{\alpha} \left(\frac{Q(a)}{1-\eta-\alpha}\right)^{1-\eta-\alpha} &= \overline{U} \; \forall \, a.
\end{aligned}$$

The first constraint ensures aggregate labor market clearing. The second constraint represents the traded goods aggregate resource constraint. The third constraint ensures that housing markets clear in every location. The fourth constraint corresponds to the local resource constraint, derived from the Kimball preferences for local goods. The fifth constraint is an adding-up condition based on the definition of $\zeta(a)$. Finally, the last constraint enforces the free mobility condition.

The planner's Lagrangian is given by:

[40] Given $z(a)$ and $\zeta(a)$, one can recover the slope of the assignment function as $z'(a) = \zeta(a)f(a)/(M_e g(z(a)))$.

$$
\begin{aligned}
\mathcal{L}^P &= \overline{U} \\
&+\varkappa_1 \left[1 - \int_{\underline{a}}^{\overline{a}} \left(L^T(a) + L^N(a)\right) f(a) da\right] \\
&+\varkappa_2 \left[\int_{\underline{a}}^{\overline{a}} a L^T(a) f(a) da - c_e M_e - \int_{\underline{a}}^{\overline{a}} Q(a) \left(L^T(a) + L^N(a)\right) f(a) da\right] \\
&+ \int_{\underline{a}}^{\overline{a}} \varsigma_1(a) \left[1 - H(a) \left(L^T(a) + L^N(a)\right)\right] f(a) da \\
&+ \int_{\underline{a}}^{\overline{a}} \varsigma_2(a) \left[\Upsilon\left(\frac{z(a) L^N(a)}{\zeta(a) \left(L^T(a) + L^N(a)\right) Y(a)}\right) \zeta(a) - 1\right] f(a) da \\
&- \int_{\underline{a}}^{\overline{a}} \varsigma_3(a) M_e G(z(a)) f(a) da + \int_{\underline{a}}^{\overline{a}} (\overline{\varsigma}_3 - \vartheta(a)) \zeta(a) f(a) da \\
&+ \int_{\underline{a}}^{\overline{a}} \varsigma_4(a) \left[\left(\frac{Y(a)}{\eta}\right)^{\eta} \left(\frac{H(a)}{\alpha}\right)^{\alpha} \left(\frac{Q(a)}{1-\eta-\alpha}\right)^{1-\eta-\alpha} - \overline{U}\right] f(a) da,
\end{aligned}
$$

where $\varkappa_1$, $\varkappa_2$, $\varsigma_1(a)$, $\varsigma_2(a)$, $\varsigma_3(a)$, and $\varsigma_4(a)$ are Lagrange multipliers. Moreover, I integrate by parts the constraint on $\zeta(a)$ with multiplier $\varsigma_3(a)$ and define $\overline{\varsigma}_3 \equiv \int_{\underline{a}}^{\overline{a}} \varsigma_3(a) f(a) da$ and $\vartheta(a) \equiv \int_{\underline{a}}^{a} \varsigma_3(r) f(r) dr$.

Consumption. The first-order conditions with respect to $Y(a)$, $H(a)$, and $Q(a)$ are given by:

$$
\begin{aligned}
\eta \varsigma_4(a) \overline{U} &= \varkappa_2 \mathbb{P}^*(a) Y(a) L(a), \\
\alpha \varsigma_4(a) \overline{U} &= \varkappa_2 R^*(a) H(a) L(a), \\
(1-\eta-\alpha) \varsigma_4(a) \overline{U} &= \varkappa_2 Q(a) L(a),
\end{aligned}
\tag{C1}
$$

where

$$
\begin{aligned}
\mathbb{P}^*(a) &\equiv \frac{\mathbb{D}^*(a)}{\delta(q(a))}, & \mathbb{D}^*(a) &\equiv \frac{\varsigma_2(a)}{\varkappa_2 Y(a) L(a)}, \\
R^*(a) &\equiv \frac{\varsigma_1(a)}{\varkappa_2}, & q(a) &\equiv \frac{z(a) L^N(a)}{\zeta(a) L(a) Y(a)},
\end{aligned}
$$

with $L(a) \equiv L^T(a) + L^N(a)$. Note that $\mathbb{P}^*(a)$, $\mathbb{D}^*(a)$, and $R^*(a)$ correspond to the planner's ideal price index, the planner's competition price index, and the planner's land shadow price, respectively, all in units of the traded good. Moreover, the first-order condition with respect to $\overline{U}$ implies:

$$1 = \int_{\underline{a}}^{\overline{a}} \varsigma_4(a) f(a) da. \tag{C2}$$

Summing the three consumption conditions defines expenditure per worker in units of the traded good, $x(a) \equiv \mathbb{P}^*(a)Y(a) + R^*(a)H(a) + Q(a) = \varsigma_4(a)\overline{U}/(\varkappa_2 L(a))$. Integrating over a and using (C2) gives $\varkappa_2 \int_{\underline{a}}^{\overline{a}} x(a)L(a)f(a)da = \overline{U}$. This condition determines the level of $\varkappa_2$.

Labor. The planner's shadow wage $W^*(a)$, in units of the traded good, is defined as:

$$W^*(a) = \frac{1}{\varkappa_2}\left[\frac{\varsigma_4(a)\overline{U}}{L(a)} + \varkappa_1\right] = x(a) + \frac{\varkappa_1}{\varkappa_2}. \tag{C3}$$

With this definition, the first-order conditions with respect to $L^T(a)$ and $L^N(a)$ are:

$$\frac{Q^T(a)}{L^T(a)} = W^*(a), \tag{C4}$$

$$\frac{\mathbb{P}^*(a)Y(a)L(a)}{L^N(a)} = W^*(a). \tag{C5}$$

Given the traded goods production technology, (C4) implies that the planner's wage is:

$$W^*(a) = a. \tag{C6}$$

Similarly, given the local goods production technology, (C5) determines the planner's optimality condition for $q(a)$:

$$\Upsilon'(q(a)) = \frac{\mathbb{C}^*(a)}{z(a)}, \qquad \text{where } \mathbb{C}^*(a) \equiv \frac{W^*(a)}{\mathbb{D}^*(a)}. \tag{C7}$$

Condition (C7) is the planner's counterpart to (13). Importantly, this condition shows that the planner's optimal relative quantities, $q(a)$, are determined without any wedge between $\Upsilon'(\cdot)$ and $\mathbb{C}^*(a)/z(a)$. In other words, the planner chooses markups equal to one for all firms in all locations.

Local goods producers' spatial allocation. The first-order conditions with respect to $z(a)$ and $\zeta(a)$ are:

$$\varsigma_2(a)\frac{\Upsilon'(q(a))q(a)\zeta(a)}{z(a)} = \varsigma_3(a)M_e g(z(a)),$$

$$\varsigma_2(a)\left[\Upsilon(q(a)) - \Upsilon'(q(a))q(a)\right] = \vartheta(a) - \overline{\varsigma}_3.$$

Define $J(a) \equiv \vartheta(a) - \bar{\varsigma}_3$. Then, since $J'(a) = \varsigma_3(a) f(a)$, we can rewrite the first-order condition with respect to $z(a)$ as:

$$\varsigma_2(a) \frac{\Upsilon'(q(a)) q(a) \zeta(a)}{z(a)} = \frac{J'(a)}{J(a)} J(a) \frac{M_e g(z(a))}{f(a)}.$$

Combining this equation with the first-order condition for $\zeta(a)$, we obtain:

$$\mathcal{E}_J(a) = \frac{\mathcal{E}_z(a)}{\delta(q(a)) - 1}, \tag{C8}$$

where we use the definitions of $\zeta(a)$ and $\delta(q(a))$. Furthermore, rewriting the first-order condition for $\zeta(a)$ yields:

$$J(a) = \varkappa_2 M^*(a) \Upsilon(q(a)) \left[\delta(q(a)) - 1\right], \qquad \text{where } \; M^*(a) \equiv \mathbb{P}^*(a) L(a) Y(a), \tag{C9}$$

using $\varsigma_2(a) = \varkappa_2 \mathbb{D}^*(a) Y(a) L(a)$. Since $\varkappa_2$ is a constant,

$$\mathcal{E}_J(a) = \mathcal{E}_{M^*}(a) + \mathcal{E}_q(a) \left[\frac{1}{\delta(q(a))} + \mathcal{E}_\delta(q(a)) \frac{\delta(q(a))}{\delta(q(a)) - 1} \right].$$

The elasticities $\mathcal{E}_\delta(q(a))$ and $\mathcal{E}_q(a)$ are given by:

$$\mathcal{E}_\delta(q(a)) = \frac{1}{\delta(q(a))} + \frac{1}{\sigma(q(a))} - 1, \tag{C10}$$

$$\mathcal{E}_q(a) = \sigma(q(a)) \left(\mathcal{E}_z(a) - \mathcal{E}_{\mathbb{C}^*}(a)\right), \tag{C11}$$

where we use the definition of $\delta(q(a))$ and (C7). Define $\Theta^*(q(a))$ as:

$$\Theta^*(q(a)) \equiv \sigma(q(a)) \left(\frac{1}{\delta(q(a))} - \frac{1}{\mu(q(a))} \right). \tag{C12}$$

where $\mu(q(a)) \equiv \sigma(q(a))/(\sigma(q(a)) - 1)$. With this notation:

$$\mathcal{E}_\delta(a) = \mathcal{E}_\delta(q(a)) \mathcal{E}_q(a) = \Theta^*(q(a)) \left(\mathcal{E}_z(a) - \mathcal{E}_{\mathbb{C}^*}(a)\right).$$

Combining these expressions gives:

$$\mathcal{E}_J(a) = \mathcal{E}_{M^*}(a) + \mathcal{E}_\delta(a) + \frac{\mathcal{E}_z(a)}{\delta(q(a)) - 1} - \frac{\mathcal{E}_{\mathbb{C}^*}(a)}{\delta(q(a)) - 1}. \tag{C13}$$

Equating (C8) and (C13) implies:

$$\mathcal{E}_{M^*}(a) + \mathcal{E}_{\delta}(a) = \frac{\mathcal{E}_{\mathbb{C}^*}(a)}{\delta(q(a)) - 1}. \tag{C14}$$

Note that equation (C14) is the planner's counterpart of (B7) in the decentralized equilibrium.

We follow the derivations of Appendix B.1 to derive an expression for the planner's market size, $M^*(a)$. The free mobility condition, local housing market clearing, and consumption allocation (C1) imply:

$$M^*(a) = \eta\kappa\mathbb{C}^*(a)^{\frac{\eta}{\alpha}}\delta(q(a))^{\frac{\eta}{\alpha}}\frac{(W^*(a) - \varkappa_1/\varkappa_2)^{\frac{1}{\alpha}}}{W^*(a)^{\frac{\eta}{\alpha}}}, \tag{C15}$$

where $W^*(a) - \varkappa_1/\varkappa_2 = x(a)$ is expenditure per worker and $\kappa \equiv (\alpha\overline{U}^{1/\alpha})^{-1}$. Using (C12) and (C15), we rewrite (C14) as:

$$(\alpha+\eta)\Theta^*\left(\frac{\mathbb{C}^*(a)}{z(a)}\right)\frac{z'(a)a}{z(a)} + (\widetilde{a} - \eta) = \frac{\mathbb{C}^{*\prime}(a)a}{\mathbb{C}^*(a)}\left[\frac{\alpha}{\delta\left(\frac{\mathbb{C}^*(a)}{z(a)}\right) - 1} + (\alpha+\eta)\Theta^*\left(\frac{\mathbb{C}^*(a)}{z(a)}\right) - \eta\right], \tag{C16}$$

where $\widetilde{a} \equiv a/(a - \varkappa_1/\varkappa_2)$. We also use (C7) to express $q(a)$ as a function of $\mathbb{C}^*(a)/z(a)$.

Finally, the local goods resource constraint implies:

$$z'(a) = \frac{f(a)}{g(z(a))M_e}\frac{1}{\Upsilon\left(q\left(\frac{\mathbb{C}^*(a)}{z(a)}\right)\right)}. \tag{C17}$$

The system (C16) - (C17), together with the boundary conditions $z(\overline{a}) = \overline{z}$ and $z(\underline{a}) = \underline{z}$, uniquely determines $\mathbb{C}^*(a)$ and $z(a)$.

Local Goods Producers Aggregate Entry. The first-order condition with respect to the mass of entrants M_e is:

$$\varkappa_2 c_e = -\int_{\underline{a}}^{\overline{a}} \varsigma_3(a)G(z(a))f(a)da.$$

Integrating by parts the right-hand side of the equation above yields:

$$\varkappa_2 c_e M_e = \int_{\underline{a}}^{\overline{a}} \left(\vartheta(a) - \overline{\varsigma}_3\right)\zeta(a)f(a)da.$$

Using (C9) and the local resource constraint $\Upsilon(q(a))\zeta(a) = 1$, the multiplier $\varkappa_2$ cancels and we obtain the planner's free entry condition:

$$c_e M_e = \int_{\underline{a}}^{\overline{a}} M^*(a) \left[\delta \left(q \left(\frac{\mathbb{C}^*(a)}{z(a)} \right) \right) - 1 \right] f(a) da. \tag{C18}$$

Closing the Planner's Problem. The solution to the planner's ODE system determines $\mathbb{C}^*(a)$ and $z(a)$. Given these equilibrium objects, the planner's optimality conditions allow us to recover the location-specific equilibrium variables, up to the aggregate objects $\overline{U}$, $\varkappa_1/\varkappa_2$, and M_e.

To solve for these aggregate objects, we first substitute (C15) into (C18):

$$c_e M_e = \int_{\underline{a}}^{\overline{a}} \eta \kappa \mathbb{C}^*(a)^{\frac{\eta}{\alpha}} \delta \left(q \left(\frac{\mathbb{C}^*(a)}{z(a)} \right) \right)^{\frac{\eta}{\alpha}} \frac{(W^*(a) - \varkappa_1/\varkappa_2)^{\frac{1}{\alpha}}}{W^*(a)^{\frac{\eta}{\alpha}}} \left[\delta \left(q \left(\frac{\mathbb{C}^*(a)}{z(a)} \right) \right) - 1 \right] f(a) da. \tag{C19}$$

Second, the population adding-up constraint implies:

$$1 = \int_{\underline{a}}^{\overline{a}} \kappa \mathbb{C}^*(a)^{\frac{\eta}{\alpha}} \delta \left(q \left(\frac{\mathbb{C}^*(a)}{z(a)} \right) \right)^{\frac{\eta}{\alpha}} \frac{(W^*(a) - \varkappa_1/\varkappa_2)^{\frac{1-\alpha}{\alpha}}}{W^*(a)^{\frac{\eta}{\alpha}}} f(a) da. \tag{C20}$$

Finally, the traded goods resource constraint closes the system. The consumption conditions give $Q(a) = (1-\eta-\alpha)x(a)$, and (C5) together with (C1) gives $L^N(a) = \eta x(a) L(a)/W^*(a)$. Substituting into the resource constraint with $W^*(a) = a$, $x(a) = a - \varkappa_1/\varkappa_2$, and (C20):

$$(1-\alpha) \left(\int_{\underline{a}}^{\overline{a}} a L(a) f(a) da - \frac{\varkappa_1}{\varkappa_2} \right) = \int_{\underline{a}}^{\overline{a}} a L(a) f(a) da - c_e M_e, \tag{C21}$$

$$L(a) = \kappa \mathbb{C}^*(a)^{\frac{\eta}{\alpha}} \delta \left(q \left(\frac{\mathbb{C}^*(a)}{z(a)} \right) \right)^{\frac{\eta}{\alpha}} \frac{(a - \varkappa_1/\varkappa_2)^{\frac{1-\alpha}{\alpha}}}{a^{\frac{\eta}{\alpha}}}.$$

Using the relationship $\frac{1}{\kappa} = \alpha \overline{U}^{\frac{1}{\alpha}}$, equations (C19) - (C21) jointly determine the values of $\overline{U}$, $\varkappa_1/\varkappa_2$, and M_e.

C.2 Proof of Proposition 5

We prove Proposition 5 under the small-ε condition of Proposition 1 and the small-support assumption of Proposition 2. The proof proceeds in four steps. First, we establish two inequalities between the decentralized equilibrium and the social planner allocation, which determine the strength of the business-stealing and variety-gain externalities. Second, we show that general equilibrium aggregate objects coincide across allocations up to first order. Third, we show that the competition index $\mathbb{C}(a)$ increases more steeply in the decentralized equilibrium than in the social planner

allocation. Finally, we derive the implications for the assignment function $z(a)$.

Step 1: Sign of Externalities. Under the CES benchmark, $\varepsilon = 0$, $\delta(q) = \mu(q) = \overline{\sigma}/(\overline{\sigma} - 1)$. Expanding to first order in ε around the CES benchmark we get:

$$\mu(q) \approx \frac{\overline{\sigma}}{\overline{\sigma} - 1} + \frac{\varepsilon}{(\overline{\sigma} - 1)^2} \log q, \qquad \text{and} \qquad \delta(q) \approx \frac{\overline{\sigma}}{\overline{\sigma} - 1} + \frac{\varepsilon}{(\overline{\sigma} - 1)^2} \log q - \varepsilon \frac{\overline{\sigma}}{(\overline{\sigma} - 1)^3}$$

This implies that

$$\frac{1}{\mu(q) - 1} < \frac{1}{\delta(q) - 1}. \tag{C22}$$

Equation (C22) illustrates the business-stealing externality. The left-hand side of this equation is the private cost of entering a market, as can be seen from (B7), while the right-hand side is the social cost, as shown in (C14). Equation (C22) then implies that the social cost of a firm entering a particular city is larger than the private cost.

Second, the super-elasticities defined in (B9) and (C12) satisfy $\Theta(q) = \rho(q)\Theta^*(q)$. Since $0 < \rho(q) < 1$ for $\varepsilon > 0$:

$$\Theta(q) < \Theta^*(q). \tag{C23}$$

These super-elasticities capture the change in consumer surplus ratio generated by an additional variety in a given city. Expression (C23) therefore implies that the private benefit of locating in a given city is smaller than the social benefit: variety gains externality.

Step 2: Aggregate objects. Under the small-support assumption, let $\Delta_a \equiv \overline{a} - \underline{a}$ and $\Delta_z \equiv \overline{z} - \underline{z}$. Define $\underline{q} \equiv \underline{\mathbb{C}}/\underline{z}$. (B6) integrates to first order to:

$$M_e = \frac{\Delta_a}{\Delta_z} \frac{f(\underline{a})}{g(\underline{z})} \frac{1}{\Upsilon\left(\varphi\left(\psi\left(\underline{q}\right)\right)\right)}.$$

The planner's z-ODE (C17) yields an analogous expression for M_e^*. At $\varepsilon = 0$, the planner's optimality condition (C7) implies that the relative quantities $q(a) = \varphi(\psi(\mathbb{C}(a)/z(a)))$ coincide with those in the decentralized equilibrium, so for a small ε, $M_e \approx M_e^*$. Similarly, combining (B12)–(B14) with a first-order approximation to their planner counterparts (C19)–(C20), the competition index at the lower boundary satisfies $\underline{\mathbb{C}} \approx \underline{\mathbb{C}}^*$.

Step 3: Comparison of $\mathcal{E}_{\mathbb{C}}(a)$. We compare the ODE governing $\mathbb{C}(a)$ across allocations. From (B10), the decentralized allocation satisfies:

$$\mathcal{E}_{\mathbb{C}}(a) = \frac{(\alpha + \eta)\,\Theta(q(a))\,\mathcal{E}_z(a) + (\widetilde{a} - \eta)}{\frac{\alpha}{\mu(q(a)) - 1} + (\alpha + \eta)\,\Theta(q(a)) - \eta},$$

while from (C16), the planner's allocation satisfies:

$$\mathcal{E}_{\mathbb{C}^*}(a) = \frac{(\alpha+\eta)\,\Theta^*(q^*(a))\,\mathcal{E}_{z^*}(a) + (\widetilde{a}-\eta)}{\frac{\alpha}{\delta(q^*(a))-1} + (\alpha+\eta)\,\Theta^*(q^*(a)) - \eta}.$$

Under the small-support assumption, we evaluate both expressions at $\underline{q}$ and define $\kappa_{\mathbb{C}} \equiv \mathcal{E}_{\mathbb{C}}(\underline{a})$ and $\kappa_{\mathbb{C}^*} \equiv \mathcal{E}_{\mathbb{C}^*}(\underline{a})$. Using $\mathcal{E}_z(\underline{a}) = \mathcal{E}_{z^*}(\underline{a}) \equiv \kappa_z$ from Step 2 and the identity $\Theta(\underline{q}) = \rho(\underline{q})\,\Theta^*(\underline{q})$, we get that:

$$\begin{aligned}
\kappa_{\mathbb{C}} > \kappa_{\mathbb{C}^*} \iff\ & (\alpha+\eta)\,\Theta^*(\underline{q})\,\kappa_z\,\alpha\left[\frac{\rho(\underline{q})}{\delta(\underline{q})-1} - \frac{1}{\mu(\underline{q})-1}\right] \\
& + \eta\,(\alpha+\eta)\,\Theta^*(\underline{q})\,\kappa_z\,(1-\rho(\underline{q})) \\
& + (\widetilde{a}_0-\eta)\left[\alpha\left(\frac{1}{\delta(\underline{q})-1} - \frac{1}{\mu(\underline{q})-1}\right) + (\alpha+\eta)\,\Theta^*(\underline{q})\,(1-\rho(\underline{q}))\right] \\
& > 0.
\end{aligned}$$

By (C22) and $\rho(\underline{q}) < 1$, the second and third lines are strictly positive. The first line is non-negative whenever

$$\frac{1}{\delta(\underline{q})-1} \geq \frac{1}{\rho(\underline{q})}\,\frac{1}{\mu(\underline{q})-1}. \tag{C24}$$

Condition (C24) requires the business-stealing externality to be sufficiently strong relative to the pass-through rate. At $\varepsilon = 0$, $\delta(\underline{q}) = \mu(\underline{q})$ and $\rho(\underline{q}) = 1$, so (C24) holds with equality. Since $1/(\delta-1)$ and $1/(\rho(\mu-1))$ are continuous in ε and coincide at $\varepsilon = 0$, there exists $\widetilde{\varepsilon} > 0$ such that (C24) is satisfied for all $\varepsilon \in (0, \widetilde{\varepsilon})$. Therefore $\kappa_{\mathbb{C}} > \kappa_{\mathbb{C}^*}$.

Step 4: Comparison of assignment functions. Since $\mathcal{E}_{\mathbb{C}}(a)$ is approximately constant to first order in Δ_a, the competition index admits the power-law form:

$$\mathbb{C}(a) \approx \underline{\mathbb{C}}\left(\frac{a}{\underline{a}}\right)^{\kappa_{\mathbb{C}}}, \qquad \mathbb{C}^*(a) \approx \underline{\mathbb{C}}\left(\frac{a}{\underline{a}}\right)^{\kappa_{\mathbb{C}^*}}.$$

Substituting these into the z-ODEs (B6) and (C17), and using the shared initial condition $z(\underline{a}) = z^*(\underline{a}) = \underline{z}$, define $x \equiv a/\underline{a}$ and the normalized assignment:

$$h(x;\kappa) \equiv \frac{z(a(x)) - \underline{z}}{\overline{z} - \underline{z}}.$$

To first order, the linearized z-ODE with power-law competition integrates to:

$$h(x;\nu) = \frac{x^\nu - 1}{\overline{x}^\nu - 1}, \tag{C25}$$

where $\overline{x} = \overline{a}/\underline{a}$ and ν is an increasing function of $\kappa_{\mathbb{C}}$, depending on other parameters. Since both h and h^* satisfy the boundary conditions $h(1) = 0$ and $h(\overline{x}) = 1$, and $\kappa_{\mathbb{C}} > \kappa_{\mathbb{C}^*}$ from Step 3, we

obtain $\nu > \nu^*$. It remains to show that for $\nu > \nu^*$ and $x \in (1, \overline{x})$, $h(x;\nu) < h(x;\nu^*)$. From (C25):

$$\frac{\partial h}{\partial \nu} = \frac{x^\nu \log x \left(\overline{x}^\nu - 1\right) - \overline{x}^\nu \log \overline{x} \left(x^\nu - 1\right)}{\left(\overline{x}^\nu - 1\right)^2}.$$

Setting $u = x^\nu$ and $\overline{u} = \overline{x}^\nu$, the sign of the numerator is that of $\Phi(u) - \Phi(\overline{u})$, where $\Phi(u) \equiv u \log u/(u-1)$. Since $\Phi'(u) > 0$ for $u > 1$ and $1 < u < \overline{u}$, we have $\Phi(u) < \Phi(\overline{u})$, which gives $\partial h/\partial \nu < 0$. Since $\nu > \nu^*$, we get that $h(x;\nu) < h(x;\nu^*)$ for all $x \in (1, \overline{x})$, which further implies:

$$z(a) < z^*(a) \qquad \text{for all} \qquad a \in (\underline{a}, \overline{a}).$$

At the boundaries, $z(\underline{a}) = z^*(\underline{a}) = \underline{z}$ and $z(\overline{a}) = z^*(\overline{a}) = \overline{z}$ by construction.

C.3 First-best Implementation

In this section, I explain how the firm subsidy (31) and the worker transfers $-(s + \varkappa_1/\varkappa_2)$ implement the first-best allocation in three steps. First, I establish that the subsidy eliminates markups. Second, I verify that PAM remains under the optimal policy. Third, I show that the policy induces firms to choose locations efficiently and that the policy also corrects the aggregate entry margin.

Step 1: Markup removal. First, (12) and (31) imply that net profits after transfers are given by:

$$\widehat{\Pi}(z,a) = \left[\Upsilon\left(\frac{y(z,a)}{Y(a)}\right) - \frac{\mathbb{C}(a)}{z}\frac{y(z,a)}{Y(a)}\right]\frac{\mathbb{D}(a)}{\mathbb{P}(a)}M(a). \tag{C26}$$

The first-order condition of this problem determines the optimal relative quantity $y(z,a)/Y(a)$:

$$\Upsilon'\left(\frac{y(z,a)}{Y(a)}\right) = \frac{\mathbb{C}(a)}{z}. \tag{C27}$$

This expression coincides with the social planner's first-order condition (C7), showing that firms set a markup equal to one.

Step 2: PAM. Using the envelope theorem, (C26) implies that:

$$\frac{\partial \log \widehat{\Pi}(z,a)}{\partial \mathbb{C}(a)} = -\frac{1}{\mathbb{C}(a)}\frac{1}{\delta\left(\frac{y(z,a)}{Y(a)}\right) - 1},$$

where I use (C27) and the definition of $\delta(\cdot)$. Differentiating once more, we obtain:

$$\frac{\partial^2 \log \widehat{\Pi}(z,a)}{\partial z \partial \mathbb{C}(a)} = \frac{1}{\mathbb{C}(a)\left[\delta\left(\frac{y(z,a)}{Y(a)}\right) - 1\right]^2} \times \underbrace{\frac{\partial\left(\frac{y(z,a)}{Y(a)}\right)}{\partial z}}_{>0} \times \delta'\left(\frac{y(z,a)}{Y(a)}\right),$$

where the second term on the right-hand side is positive by the first-order condition (C27). Therefore, $\widehat{\Pi}(z,a)$ is log-supermodular if and only if $\delta'(\cdot) > 0$, which holds for any $\Upsilon(\cdot)$ satisfying MSLD. This establishes PAM between z and $\mathbb{C}(a)$.

I now provide conditions under which $\mathbb{C}(a)$ is increasing in general equilibrium. The first-order condition of (C26) with respect to a is:

$$\mathcal{E}_\delta(a) + \mathcal{E}_M(a) = \mathcal{E}_\mathbb{C}(a) \frac{1}{\delta\left(\frac{y(z,a)}{Y(a)}\right) - 1}.$$

The term $\mathcal{E}_M(a)$ is the same as before, except that worker transfers are now given by the planner's lump-sum levy, $\varkappa_1/\varkappa_2$. Hence,

$$\mathcal{E}_M(a) = \frac{\eta}{\alpha}\mathcal{E}_\mathbb{C}(a) + \frac{\eta}{\alpha}\mathcal{E}_\delta(a) + \frac{\widetilde{a} - \eta}{\alpha},$$

with $\widetilde{a} \equiv a/(a - \varkappa_1/\varkappa_2)$. Moreover, (C27) implies that:

$$\mathcal{E}_\delta(a) = \Theta^*\left(\frac{\mathbb{C}(a)}{z(a)}\right)\mathcal{E}_z(a) - \Theta^*\left(\frac{\mathbb{C}(a)}{z(a)}\right)\mathcal{E}_\mathbb{C}(a),$$

with $\Theta^*(\cdot)$ defined as in (C12). Putting everything together, we obtain

$$(\alpha+\eta)\Theta^*\left(\frac{\mathbb{C}(a)}{z(a)}\right)\frac{z'(a)a}{z(a)} + (\widetilde{a} - \eta) = \frac{\mathbb{C}'(a)a}{\mathbb{C}(a)}\left[\frac{\alpha}{\delta\left(\frac{\mathbb{C}(a)}{z(a)}\right) - 1} + (\alpha+\eta)\Theta^*\left(\frac{\mathbb{C}(a)}{z(a)}\right) - \eta\right]. \quad \text{(C28)}$$

When $\Upsilon(\cdot)$ takes the Klenow and Willis (2016) specification and $\varepsilon \to 0$, (C28) collapses to

$$\mathcal{E}_\mathbb{C}(a)\left[\alpha(\overline{\sigma} - 1) - \eta\right] = (\widetilde{a} - \eta).$$

Note that, given an appropriate normalization of a, $\tilde{a} - \eta > 0$ and the conditions of Proposition 1 coincide with the condition above for $\mathcal{E}_\mathbb{C}(a)$ to be strictly increasing in general equilibrium.

Step 3: Optimal Firm Location and Aggregate Entry. Under the optimal policy, the competition price index still satisfies (B6). Hence, (B6) together with (C28) determines local producers' location decisions under the optimal policy. Note that the resulting system of ODEs coincides with the planner's system, (C16)–(C17), provided that the mass of entrants M_e under the optimal policy coincides with that in the planner's problem.

It remains to show that the optimal policy also generates an efficient economy-wide entry rate. Combining (C26), (C27), and (B6), we can write the free-entry condition under the policy as

$$\int_a \left[\delta\left(\frac{y(z(a),a)}{Y(a)}\right) - 1\right] M(a) f(a) da = c_e M_e,$$

which coincides with the social planner's entry condition (C18).

D Proofs of the Extended Model

This section proves Proposition 6 by showing how Propositions 1 and 2 and Corollary 1 carry over to the extended model. The proofs follow the structure of the baseline model, so I omit some details to keep the appendix concise. Before proceeding to extend the proofs, I derive input demand and optimal pricing functions in the extended model.

D.1 Relative Price and Input Demand Functions

First, indexing cities by c, workers' demand for local varieties retains the same expressions as in Appendix A.1. Second, input demands follow from the firms' cost minimization problem with associated Lagrangian:

$$\mathcal{L}^F(\omega, c) = W(c)l(\omega, c) + R(c)h(\omega, c) + \lambda^F(\omega, c)\left[y(\omega, c) - \omega l(\omega, c)^{\beta} h(\omega, c)^{1-\beta}\right],$$

where $\lambda^F(\omega, c)$ is a Lagrange multiplier that equals the marginal cost of production and we index firms by their final productivity ω. Taking first-order conditions and recognizing that $\mu(\omega, c) \equiv p(\omega, c)/\lambda^F(\omega, c)$ we obtain:

$$l(\omega, c)W(c) = \beta \frac{p(\omega, c)y(\omega, c)}{\mu(\omega, c)}, \qquad h(\omega, c)R(c) = (1-\beta)\frac{p(\omega, c)y(\omega, c)}{\mu(\omega, c)}.$$

Moreover, the marginal cost is given by $\nu W(c)^{\beta} R(c)^{1-\beta}/\omega$, with $\nu \equiv 1/(\beta^{\beta}(1-\beta)^{1-\beta})$. Then, the profit maximization problem becomes:

$$\pi(\omega, c) = \max_{p(\omega,c)} \left[p(\omega, c)y(\omega, c) - \nu \frac{W(c)^{\beta} R(c)^{1-\beta}}{\omega} y(\omega, c)\right] L(c) \qquad \text{s.t.} \quad \frac{y(\omega, c)}{Y(c)} = \varphi\left(\frac{p(\omega, c)}{\mathbb{D}(c)}\right).$$

As in the main text, this problem implies an optimal relative price function:

$$\frac{p(\omega, c)}{\mathbb{D}(c)} = \psi\left(\frac{\mathbb{C}(c)}{\omega}\right), \qquad \text{with} \qquad \mathbb{C}(c) \equiv \nu \frac{W(c)^{\beta} R(c)^{1-\beta}}{\mathbb{D}(c)}. \tag{D1}$$

D.2 Extension of Proposition 1

We extend the results of Proposition 1 in three steps. First, we show that there is positive assortative matching conditional on a location's competition index $\mathbb{C}(c)$. Second, we characterize general equilibrium objects when firms sort based on competition. Third, we construct a single index of locations that is a sufficient statistic for firms' location decisions. We then show that there is positive assortative matching when the location's competition index is determined in general equilibrium.

Step 1: sorting conditional on competition index. As in Section B.1, we index locations by their competition index $\mathbb{C}$ rather than by their productivity–amenities pair c. We obtain the same functional form as in the baseline model for the profit function conditional on final productivity ω, denoted $\pi(\omega, \mathbb{C})$. Hence, under Assumption 1, the profit function $\pi(\omega, \mathbb{C})$ is log-supermodular. Firms make their location decision based on expected profits:

$$\Pi(z, \mathbb{C}) \;=\; \mathbb{E}_\omega[\pi(\omega, \mathbb{C})] \;=\; \int_\omega \pi(\omega, \mathbb{C})\, g(\omega|z)\, d\omega.$$

Under Assumption 2, Theorem 1 in Athey (2002) implies that $\mathbb{C}^*(z)$ is strictly increasing, where $\mathbb{C}^*(z) = \arg\max_{\mathbb{C}} \Pi(z, \mathbb{C})$. Hence, more productive firms ex-ante locate in more competitive cities.

Step 2: general equilibrium objects. We now derive expressions for the general equilibrium objects under z–$\mathbb{C}$ PAM. To ease notation, we drop intermediate arguments and write

$$\Upsilon\left(\frac{\mathbb{C}}{\omega}\right) \equiv \Upsilon\left(\varphi\left(\psi\left(\frac{\mathbb{C}}{\omega}\right)\right)\right), \qquad \varphi\left(\frac{\mathbb{C}}{\omega}\right) \equiv \varphi\left(\psi\left(\frac{\mathbb{C}}{\omega}\right)\right).$$

For any function $d(u/\omega)$, define the conditional mean and the conditional sales-weighted mean as

$$\begin{aligned}
\mathbb{E}_\omega\left[d\left(\frac{u}{\omega}\right)\right] &\equiv \mathbb{E}_\omega\left[d\left(\frac{u}{\omega}\right) \mid z(u)\right] = \int d\left(\frac{u}{\omega}\right)\, g(\omega \mid z(u))\, d\omega, \\
\mathbb{E}_\lambda\left[d\left(\frac{u}{\omega}\right)\right] &\equiv \mathbb{E}_\lambda\left[d\left(\frac{u}{\omega}\right) \mid z(u)\right] = \frac{\int d\left(\frac{u}{\omega}\right)\, \psi\left(\frac{u}{\omega}\right) \varphi\left(\frac{u}{\omega}\right)\, g(\omega \mid z(u))\, d\omega}{\int \psi\left(\frac{u}{\omega}\right) \varphi\left(\frac{u}{\omega}\right)\, g(\omega \mid z(u))\, d\omega}.
\end{aligned}$$

Then the generalized version of (B2) is

$$z'(\mathbb{C})\, \frac{M_e\, g(z(\mathbb{C}))}{f_{\mathbb{C}}(\mathbb{C})} = \frac{1}{\mathbb{E}_\omega\left[\Upsilon\left(\frac{\mathbb{C}}{\omega}\right)\right]}, \tag{D2}$$

where Step 1 allows us to invert $\mathbb{C}^*(z)$ and obtain $z(\mathbb{C})$. Using this, the ratio $\mathbb{D}(\mathbb{C})/\mathbb{P}(\mathbb{C})$, which I denote by $\overline{\delta}(\mathbb{C})$, is

$$\overline{\delta}(\mathbb{C}) \equiv \frac{\mathbb{D}(\mathbb{C})}{\mathbb{P}(\mathbb{C})} = \mathbb{E}_\lambda\left[\delta\left(\frac{\mathbb{C}}{\omega}\right)\right],$$

where $\delta(\mathbb{C}/\omega)$ is defined in (B8). This and the definition of $\mathbb{C}$ in (D1) imply that the ideal price index is

$$\mathbb{P}(\mathbb{C}) = \nu\, \frac{W(\mathbb{C})^\beta R(\mathbb{C})^{1-\beta}}{\mathbb{C}\, \overline{\delta}(\mathbb{C})}. \tag{D3}$$

Fréchet-distributed idiosyncratic location tastes imply that the number of workers residing in city $\mathbb{C}$ is

$$L(\mathbb{C}) = \frac{1}{\overline{U}^\theta} \left(b(\mathbb{C}) \frac{W(\mathbb{C})(1+s)}{\mathbb{P}(\mathbb{C})^\eta R(\mathbb{C})^\alpha}\right)^\theta, \tag{D4}$$

where $\overline{U}$ is the average ex-ante utility before idiosyncratic tastes are realized. Hence the market

size is

$$M(\mathbb{C}) = \frac{\eta(1+s)^{1+\theta}}{\overline{U}^{\theta}} \frac{b(\mathbb{C})^{\theta} W(\mathbb{C})^{1+\theta}}{\mathbb{P}(\mathbb{C})^{\eta\theta} R(\mathbb{C})^{\alpha\theta}}, \tag{D5}$$

where s denotes transfers from the national portfolio.

Housing-land market clearing yields

$$R(\mathbb{C})^{\phi} = \frac{\alpha}{\eta} \frac{M(\mathbb{C})}{R(\mathbb{C})} + (1-\gamma) \frac{Q^T(\mathbb{C})}{R(\mathbb{C})} + \frac{1-\beta}{\overline{\mu}(\mathbb{C})} \frac{M(\mathbb{C})}{R(\mathbb{C})}, \tag{D6}$$

where the three terms on the right-hand side are total housing consumption of workers, traded-good producers, and local producers, respectively, and $\overline{\mu}(\mathbb{C})$ is the sales-weighted harmonic mean of markups:

$$\overline{\mu}(\mathbb{C}) \equiv \left(\mathbb{E}_{\lambda} \left[\frac{1}{\mu(\mathbb{C}/\omega)} \right] \right)^{-1}.$$

Local labor-market clearing implies

$$L(\mathbb{C}) = \beta \frac{M(\mathbb{C})}{W(\mathbb{C})\,\overline{\mu}(\mathbb{C})} + \gamma \frac{Q^T(\mathbb{C})}{W(\mathbb{C})}.$$

Using $M(\mathbb{C})$ and solving for traded-good output gives

$$Q^T(\mathbb{C}) = \frac{M(\mathbb{C})}{\gamma} \left[\frac{1}{\eta(1+s)} - \frac{\beta}{\overline{\mu}(\mathbb{C})} \right]. \tag{D7}$$

Substituting (D7) into (D6) and solving for $R(\mathbb{C})$ yields equilibrium land rents:

$$R(\mathbb{C}) = M(\mathbb{C})^{\frac{1}{1+\phi}}\, \chi(\mathbb{C})^{\frac{1}{1+\phi}}, \tag{D8}$$

with

$$\chi(\mathbb{C}) \equiv \left[\frac{\alpha}{\eta} + \frac{1-\gamma}{\gamma} \frac{1}{\eta(1+s)} + \frac{(1-\beta)\gamma - \beta(1-\gamma)}{\gamma\,\overline{\mu}(\mathbb{C})} \right].$$

The traded-good producers' zero-profit condition implies equilibrium wages:

$$W(\mathbb{C}) = \left(\frac{a(\mathbb{C})}{R(\mathbb{C})^{1-\gamma}\, \varrho} \right)^{\frac{1}{\gamma}}, \tag{D9}$$

with $\varrho \equiv \gamma^{-\gamma}(1-\gamma)^{\gamma-1}$. Substituting (D3), (D8), and (D9) into (D5) and solving for $M(\mathbb{C})$ gives

$$M(\mathbb{C}) = \left[\frac{\eta\,(1+s)^{1+\theta}}{\overline{U}^{\theta}\, \nu^{\eta\theta}\, \varrho^{\frac{1+\theta(1-\eta\beta)}{\gamma}}} \frac{\mathbb{C}^{\eta\theta}\, \overline{\delta}(\mathbb{C})^{\eta\theta}}{\chi(\mathbb{C})^{\theta\xi-1}}\, a(\mathbb{C})^{\frac{1+\theta(1-\eta\beta)}{\gamma}}\, b(\mathbb{C})^{\theta} \right]^{\frac{1}{\theta\xi}}, \tag{D10}$$

where $\xi \equiv \frac{\gamma(1+\phi+\theta(\alpha+\eta(1-\beta)))+(1-\gamma)(1+\theta(1-\eta\beta))}{\theta\gamma(1+\phi)}$, as in Proposition 6.[41]

Step 3: single index property. The assignment function $z(\mathbb{C})$ and the competition index $\mathbb{C}$ are jointly determined by a coupled ODE system:

$$\begin{aligned} z'(\mathbb{C})\frac{M_e g(z(\mathbb{C}))}{f_{\mathbb{C}}(\mathbb{C})} &= \frac{1}{\mathbb{E}_\omega\left[\Upsilon\left(\frac{\mathbb{C}}{\omega}\right)\right]}, \\ \frac{1}{\overline{\mu}(\mathbb{C})-1} &= \mathcal{E}_{\overline{\delta}}(\mathbb{C}) + \mathcal{E}_M(\mathbb{C}), \end{aligned} \tag{D11}$$

where the first equation comes from (D2) and the second from the first-order condition of the firm's location problem (38). The first equation only depends on demand parameters, densities, $z(\mathbb{C})$ and $\mathbb{C}$. On the other hand, to inspect the second equation, we look at (D10). The term on the left-hand side and the first term on the right-hand side depend on the equilibrium objects, $z(\mathbb{C})$ and $\mathbb{C}$. On the contrary, the last term on the right-hand side depends on the equilibrium objects and the combined object, $a(\mathbb{C})^{\frac{1+\theta(1-\eta\beta)}{\gamma}} b(\mathbb{C})^{\theta}$. Therefore, the equilibrium objects $z(\mathbb{C})$ and $\mathbb{C}$ depend only on the characteristics of a location through a combined index with specific weights on productivity and amenities. This result implies that local goods producers make their sorting decisions based on a uni-dimensional index $x(c)$ rather than considering the two dimensions of location heterogeneity separately:

$$x(c) \equiv a(c)^{\frac{1+\theta(1-\eta\beta)}{\theta\gamma\xi}} b(c)^{\frac{1}{\xi}}.$$

Step 4: sorting in general equilibrium. In the last step we characterize under which conditions more appealing locations have higher competition: i.e. the function $x(\mathbb{C})$ is increasing in general equilibrium. We can use (D10) to write the first-order condition (D11) as:

$$\begin{aligned} \frac{1}{\overline{\mu}(\mathbb{C})-1} &= \mathcal{E}_{\overline{\delta}}(\mathbb{C}) + \frac{1}{\theta\xi}\left[\eta\theta + \eta\theta\mathcal{E}_{\overline{\delta}}(\mathbb{C}) - (\theta\xi-1)\mathcal{E}_\chi(\mathbb{C})\right] + \mathcal{E}_x(\mathbb{C}), \\ &= \frac{\eta}{\xi} + \frac{\xi+\eta}{\xi}\mathcal{E}_{\overline{\delta}}(\mathbb{C}) + \mathcal{E}_x(\mathbb{C}) - \frac{\theta\xi-1}{\theta\xi}\mathcal{E}_\chi(\mathbb{C}) \end{aligned}$$

Under the Klenow and Willis (2016) specification, when $\varepsilon = 0$, $\overline{\delta}(\cdot)$, $\chi(\cdot)$, and $\overline{\mu}(\cdot)$ become constants and hence the above equation collapses to:

$$\overline{\sigma} - 1 = \frac{\eta}{\xi} + \mathcal{E}_x(\mathbb{C}),$$

[41] Note that $\theta\xi - 1 = \frac{\gamma\theta(\alpha+\eta(1-\beta))+(1-\gamma)(1+\theta(1-\eta\beta))}{\gamma(1+\phi)} > 0$.

which implies that $\mathcal{E}_x(\mathbb{C}) > 0$ if $(\xi/\eta)(\overline{\sigma}-1) > 1$. Therefore, when

$$\frac{1}{\overline{\sigma}-1} < \frac{\xi}{\eta},$$

there exists a region of the parameter space where ε is small and PAM is obtained in general equilibrium: both $z(\mathbb{C})$ and $x(\mathbb{C})$ are strictly increasing, and therefore more productive firms locate in more appealing cities.

D.3 Extension of Proposition 2

The extension of Proposition 2 is structured in three steps. First, derive the system of ODEs that determine the equilibrium for the local good sector. Second, show existence of a solution to these systems conditional on general equilibrium aggregates, M_e, $\overline{U}$, and s. Third, show existence and uniqueness of general equilibrium objects.

Step 1: ODE system for local goods producers. Impose Assumption 1, Assumption 2 and the assumptions of Proposition 6. Then, PAM between firm and location productivity is obtained. Indexing locations by x, (D11) becomes:

$$\begin{aligned} z'(x)\frac{M_e g(z(x))}{f_x(x)} &= \frac{1}{\mathbb{E}_\omega\left[\Upsilon\left(\frac{\mathbb{C}(x)}{\omega}\right)\right]}, \\ \frac{\mathcal{E}_{\mathbb{C}}(x)}{\overline{\mu}(x)-1} &= \mathcal{E}_{\overline{\delta}}(x) + \mathcal{E}_M(x), \end{aligned} \tag{D12}$$

where $f_x(\cdot)$ is the density of the appeal index, x. Market size is given by:

$$M(x) = \left[\frac{\eta(1+s)^{1+\theta}}{\overline{U}^{\theta}\nu^{\eta\theta}\varrho^{\frac{1+\theta(1-\eta\beta)}{\gamma}}}\frac{\mathbb{C}(x)^{\eta\theta}\overline{\delta}(x)^{\eta\theta}}{\chi(x)^{\theta\xi-1}}\right]^{\frac{1}{\theta\xi}} x,$$

which further implies that

$$\mathcal{E}_M(x) = \frac{\eta}{\xi}\mathcal{E}_{\mathbb{C}}(x) + \frac{\eta}{\xi}\mathcal{E}_{\overline{\delta}}(x) + 1 - \frac{\theta\xi-1}{\theta\xi}\mathcal{E}_{\chi}(x).$$

We start deriving expressions for each of the terms in the expressions above. First:

$$\mathcal{E}_{\overline{\delta}}(x) = \Theta_1(x)\,\mathcal{E}_z(x) - \Theta_2(x)\,\mathcal{E}_{\mathbb{C}}(x),$$

with

$$\Theta_1(x) = \mathbb{E}_\lambda \left[\mathcal{E}_g(\omega, z(x)) \frac{\left[\delta\left(\frac{\mathbb{C}(x)}{\omega}\right) - \overline{\delta}(x)\right]}{\overline{\delta}(x)} \right],$$
$$\Theta_2(x) = \mathbb{E}_\lambda \left[\rho\left(\frac{\mathbb{C}(x)}{\omega}\right) \sigma\left(\frac{\mathbb{C}(x)}{\omega}\right) \left(\frac{1}{\overline{\delta}(x)} - \frac{1}{\mu\left(\frac{\mathbb{C}(x)}{\omega}\right)} \right) \right]$$

with $\rho(\cdot)$ defined as in the proof of Proposition 2, and $\mathcal{E}_g(\omega, z(x))$ the elasticity of the conditional PDF $g(\omega|z(x))$ with respect to $z(x)$. Second,

$$\mathcal{E}_\chi(x) = -\Lambda_1(x)\,\mathcal{E}_z(x) + \Lambda_2(x)\,\mathcal{E}_\mathbb{C}(x),$$

with

$$\Lambda_1(x) = \frac{(1-\gamma)\beta - (1-\beta)\gamma}{\gamma} \frac{\mathbb{E}_\lambda\left[\mathcal{E}_g(\omega, z(x))\left(\frac{1}{\mu\left(\frac{\mathbb{C}(x)}{\omega}\right)} - \frac{1}{\overline{\mu}(x)}\right)\right]}{\chi(x)},$$
$$\Lambda_2(x) = \frac{(1-\gamma)\beta - (1-\beta)\gamma}{\gamma} \frac{\mathbb{E}_\lambda\left[\frac{\rho\left(\frac{\mathbb{C}(x)}{\omega}\right)\sigma\left(\frac{\mathbb{C}(x)}{\omega}\right)-1}{\mu\left(\frac{\mathbb{C}(x)}{\omega}\right)}\right] - \frac{1}{\overline{\mu}(x)}\mathbb{E}_\lambda\left[\frac{\rho\left(\frac{\mathbb{C}(x)}{\omega}\right)\sigma\left(\frac{\mathbb{C}(x)}{\omega}\right)}{\mu\left(\frac{\mathbb{C}(x)}{\omega}\right)}\right]}{\chi(x)}.$$

By plugging back these terms into the first-order condition of the location problem we obtain:

$$\begin{aligned} \frac{\mathbb{C}'(x)x}{\mathbb{C}(x)} &\left[\frac{\theta\xi}{\overline{\mu}(x) - 1} + \theta(\xi+\eta)\Theta_2(x) + (\theta\xi - 1)\Lambda_2(x) - \eta\theta \right] \\ &= \theta\xi + \left[\theta(\xi+\eta)\Theta_1(x) + (\theta\xi - 1)\Lambda_1(x)\right] \frac{z'(x)x}{z(x)}, \end{aligned} \tag{D13}$$

where we explicitly wrote the expression for the elasticities $\mathcal{E}_z(x)$ and $\mathcal{E}_\mathbb{C}(x)$. Given M_e and s, equations (D12) and (D13) define a coupled system of ODE's, with two boundary conditions $z(\overline{x}) = \overline{z}$ and $z(\underline{x}) = \underline{z}$. The first boundary condition states that the most productive firms go to the most appealing locations, while the second condition implies that least productive firms locate in the least appealing cities. The solution to this ODE system is the assignment function $z(x)$ and the competition index function $\mathbb{C}(x)$.

Step 2: Existence and uniqueness of a solution to the ODE system given M_e and s. The system (D12) - (D13) is Lipschitz continuous and, therefore, has a unique solution if two initial conditions are provided (by the Picard–Lindelöf Theorem). However, this system constitutes a two-point boundary value problem, as it has one initial and one terminal condition. To address this, we transform the problem into an initial value problem and apply a shooting method.

Define $\overline{\mathbb{C}} \equiv \mathbb{C}(\overline{x})$. Given M_e and the terminal condition $z(\overline{x}) = \overline{z}$, we show that there exists $\overline{\mathbb{C}}$ such that $z(\underline{x}) = \underline{z}$. By Lipschitz continuity, $z(\underline{x})$ is a continuous function of $\overline{\mathbb{C}}$. As $\overline{\mathbb{C}} \to \infty$, all firms

are pushed toward the choke price, $\mathbb{E}_\omega[\Upsilon] \to 0$, and $z(\underline{x}) \to 0$. As $\overline{\mathbb{C}} \to 0$, $\mathbb{E}_\omega[\Upsilon]$ approaches its upper bound $\Upsilon_1 \equiv \Upsilon\left(\overline{\sigma}^{\overline{\sigma}/\varepsilon}\right)$, so $z'(x)$ remains bounded and $z(\underline{x}) \to G^{-1}\left(1 - \frac{1}{M_e \Upsilon_1}\right) > \underline{z}$ whenever $M_e \Upsilon_1 > 1$, a condition that any equilibrium satisfies by Step 3. By the intermediate value theorem, there exists at least one $\overline{\mathbb{C}}$ such that $z(\underline{x}) = \underline{z}$. Moreover, when the supports of $F_x(\cdot)$ and $G(\cdot)$ are sufficiently small, $\overline{\mathbb{C}}$ is unique. Up to first order, $z'(x)$ is constant and strictly increasing in $\overline{\mathbb{C}}$, since $\mathbb{E}_\omega[\Upsilon(\mathbb{C}/\omega)]$ is strictly decreasing in $\mathbb{C}$. Hence $z(\underline{x})$ is strictly decreasing in $\overline{\mathbb{C}}$. Given the unique $\overline{\mathbb{C}}$, the Picard–Lindelöf Theorem delivers a unique solution to the ODE system.

Step 3: Existence and uniqueness of M_e, s and $\overline{U}$. On the one hand, the aggregate labor adding-up constraint uniquely pins down $\overline{U}$. Local population cannot be solely characterized by the appeal x, as workers directly value local amenities. In particular, using the equilibrium expression for market size and wages, population in location $c = (a, b)$ is given by:

$$L\left(x(a,b), b\right) = \frac{\varrho^{\frac{1}{\gamma}}}{\eta(1+s)} M\left(x(a,b)\right)^{\frac{1+\gamma\phi}{\gamma(1+\phi)}} \chi\left(x(a,b)\right)^{\frac{1-\gamma}{\gamma(1+\phi)}} x(a,b)^{-\frac{\theta\xi}{1+\theta(1-\eta\beta)}} b^{\frac{\theta}{1+\theta(1-\eta\beta)}}.$$

The ex-ante workers mean utility, $\overline{U}$, is given then by the economy-wide labor adding-up constraint

$$\int_a \int_b L\left(x(a,b), b\right) f(a,b) db\, da = 1,$$

which can also be written as

$$\int_x \int_b L\left(x, b\right) f_{b|x}(b) f_x(x) db\, dx = 1,$$

where $f_{b|x}(\cdot)$ is the PDF of b given x. On the other hand, the free-entry condition (26) can be written as

$$c_e M_e = \int_x M(x)\, \mathbb{E}_\lambda \left[\frac{1}{\sigma\left(\frac{\mathbb{C}(x)}{\omega}\right)}\right] f_x(x) dx. \tag{D14}$$

For existence, note first that $\Upsilon(\varphi(\psi(\cdot)))$ is bounded above by Υ_1 under Assumption 1. Integrating (D12) over the support of x implies that any solution requires $M_e \geq M_e^{min} \equiv 1/\Upsilon_1$. The LHS of (D14) is increasing in M_e. As $M_e \to \infty$, (D12) implies that $\mathbb{C}(x) \to \infty$ for all x. Otherwise, $z'(x) \to 0$ for all x, violating the initial condition $z(\underline{x}) = \underline{z}$. Since profits in all locations decrease with local competition, the RHS of (D14) tends to zero as $M_e \to \infty$. At $M_e = M_e^{min}$, competition is at its weakest and the RHS is bounded away from zero. Provided the entry cost is not too large, the LHS at M_e^{min} lies below the RHS. By the intermediate value theorem, there exists at least one $M_e \in [M_e^{min}, \infty)$ satisfying (D14), for a given s.

To prove the uniqueness of M_e, we assume that the supports of $F(\cdot,\cdot)$ and $G(\cdot)$ are small enough. This implies that the support of $F_x(\cdot)$ is also small enough. This assumption makes possible using a first-order approximation of the ODE system (D12) - (D13). Therefore, to a first order,

$$M_e z'(x) \approx \frac{f_x(\underline{x})}{g(\underline{z})} \frac{1}{\mathbb{E}_\omega \left[\Upsilon \left(\frac{\underline{\mathbb{C}}}{\omega} \right) \mid \underline{z} \right]} = \kappa_1 \frac{1}{\Upsilon_{\mathbb{C}}(\underline{\mathbb{C}})},$$

where $\kappa_1 \equiv f_x(\underline{x})/g(\underline{z})$ depends only on parameters, and the function $\Upsilon_{\mathbb{C}}(\cdot)$ is strictly decreasing. Integrating this expression gives

$$M_e = \frac{\kappa_1(\overline{x} - \underline{x})}{\overline{z} - \underline{z}} \frac{1}{\Upsilon_{\mathbb{C}}(\underline{\mathbb{C}})}. \tag{D15}$$

Similarly, the labor adding-up constraint implies that $L(x, b) \approx 1$ up to first order. Combining the equilibrium expressions for market size, wages and rents then yields $M(\underline{x})^{1+\zeta} \approx \eta(1+s)(\underline{a}/\varrho)^{\frac{1}{\gamma}} \chi(\underline{\mathbb{C}})^{-\zeta}$, with $\zeta \equiv \frac{1-\gamma}{\gamma(1+\phi)}$. Using $\mathbb{E}_\lambda[1/\sigma] = 1 - 1/\overline{\mu}$, the free-entry condition (D14) becomes, up to first order,

$$c_e M_e = \kappa_2(\overline{x} - \underline{x}) \left[1 - \frac{1}{\overline{\mu}(\underline{\mathbb{C}})} \right] \chi(\underline{\mathbb{C}})^{-\frac{\zeta}{1+\zeta}}, \tag{D16}$$

where $\kappa_2 \equiv f_x(\underline{x}) \left[\eta(1+s)(\underline{a}/\varrho)^{\frac{1}{\gamma}} \right]^{\frac{1}{1+\zeta}}$ depends only on parameters and s. By equating (D15) and (D16), one obtains

$$\left[1 - \frac{1}{\overline{\mu}(\underline{\mathbb{C}})} \right] \chi(\underline{\mathbb{C}})^{-\frac{\zeta}{1+\zeta}} \Upsilon_{\mathbb{C}}(\underline{\mathbb{C}}) = \frac{c_e \kappa_1}{\kappa_2(\overline{z} - \underline{z})}. \tag{D17}$$

The LHS of (D17) is strictly decreasing in $\underline{\mathbb{C}}$. Under MSLD, $\overline{\mu}(\cdot)$ is strictly decreasing, so the first factor decreases with $\underline{\mathbb{C}}$. When $\gamma \geq \beta$, as in the estimated parameters, $\chi(\cdot)$ increases with $\underline{\mathbb{C}}$, so the second factor also decreases. The last factor is strictly decreasing. Hence, (D17) uniquely determines $\underline{\mathbb{C}}$, and (D15) uniquely determines M_e, given s.

Finally, we show existence and uniqueness of s. The mutual fund rebates the profits of land developers as a flat wage subsidy. Using $R(x)^{1+\phi} = M(x)\chi(x)$ and the labor income $WL = M/(\eta(1+s))$ in every city, the mutual fund budget constraint is

$$\frac{s}{\eta(1+s)} \int_x M(x) f_x(x)\, dx = \frac{1}{1+\phi} \int_x M(x)\chi(x) f_x(x)\, dx.$$

Up to first order, $\chi(x) \approx \chi(\underline{\mathbb{C}})$ and the constraint reduces to $\frac{s}{1+s} = \frac{\eta\chi(\underline{\mathbb{C}})}{1+\phi}$. The LHS is strictly increasing in s from zero to one. Holding $\overline{\mu}(\underline{\mathbb{C}})$ at its first-order value, the RHS is strictly decreasing in s and converges to $[\alpha + \eta(\gamma - \beta)/(\gamma\overline{\mu})]/(1+\phi) < 1$, a condition that holds at the estimated parameters. Hence, there exists a unique s. Given s, the first-order system determines $(\underline{\mathbb{C}}, M_e)$, and the adding-up constraint then determines $\overline{U}$.

D.4 Extension of Corollary 1

The proof in the quantitative framework is analogous to the baseline case. We consider a wide support for $G(\cdot)$ and a small support for $F(\cdot)$, and index cities by their competition index $\mathbb{C}$. Optimal location choices imply

$$\frac{1}{\overline{\mu}\left(\mathbb{C}\right)-1}=\frac{\eta}{\xi}+\frac{\xi+\eta}{\xi}\mathcal{E}_{\overline{\delta}}(\mathbb{C})+\mathcal{E}_{x}(\mathbb{C})-\frac{\theta\xi-1}{\theta\xi}\mathcal{E}_{\chi}(\mathbb{C}).$$

Replacing the expressions for $\mathcal{E}_{\overline{\delta}}(\mathbb{C})$ and $\mathcal{E}_{\chi}(\mathbb{C})$ from the extension of Proposition 2 yields

$$\frac{\theta\xi}{\overline{\mu}\left(\mathbb{C}\right)-1}=\theta\xi\mathcal{E}_{x}(\mathbb{C})+\left[\theta(\xi+\eta)\Theta_{1}\left(\mathbb{C}\right)+(\theta\xi-1)\Lambda_{1}\left(\mathbb{C}\right)\right]\mathcal{E}_{z}(\mathbb{C})-\theta(\xi+\eta)\Theta_{2}\left(\mathbb{C}\right)-(\theta\xi-1)\Lambda_{2}\left(\mathbb{C}\right)+\eta\theta.$$

In the limiting economy, $\mathcal{E}_{x}(\mathbb{C})=0$. Hence, the first-order condition above defines a nondegenerate assignment $z(\mathbb{C})$:

$$\frac{\theta\xi}{\overline{\mu}\left(\mathbb{C}\right)-1}=\left[\theta(\xi+\eta)\Theta_{1}\left(\mathbb{C}\right)+(\theta\xi-1)\Lambda_{1}\left(\mathbb{C}\right)\right]\mathcal{E}_{z}(\mathbb{C})-\theta(\xi+\eta)\Theta_{2}\left(\mathbb{C}\right)-(\theta\xi-1)\Lambda_{2}\left(\mathbb{C}\right)+\eta\theta.$$

Supermodularity of expected profits also holds in this limiting economy. Therefore, $z'(\mathbb{C})>0$ (positive assortative matching). Moreover, combining the first-order condition with the market size elasticity yields

$$\mathcal{E}_{M}(\mathbb{C})=\frac{1}{\theta(\xi+\eta)}\left[\eta\theta\,\frac{\overline{\mu}(\mathbb{C})}{\overline{\mu}(\mathbb{C})-1}-\left(\theta\xi-1\right)\mathcal{E}_{\chi}(\mathbb{C})\right],$$

which is positive whenever $\left(\theta\xi-1\right)\mathcal{E}_{\chi}(\mathbb{C})<\eta\theta\,\overline{\mu}(\mathbb{C})/(\overline{\mu}(\mathbb{C})-1)$. When $\varepsilon=0$, $\mathcal{E}_{\chi}(\mathbb{C})=0$ and $\overline{\mu}/(\overline{\mu}-1)=\overline{\sigma}$, so the condition holds trivially. In this case the market size elasticity reduces to $\mathcal{E}_{M}(\mathbb{C})=\eta\overline{\sigma}/(\xi+\eta)$. This mirrors the expression in the proof of Corollary 1, with ξ in place of α.

D.5 Equilibrium Conditions Quasi-Optimal Policy

This section derives the equilibrium conditions of the quasi-optimal policy in the extended model. The policy implements the firm subsidy (31) of Proposition 4, with firms indexed by their final productivity ω and locations indexed by their appeal x. Contrary to the baseline model, the subsidy is financed with a flat tax on labor income, constant across locations. As in the decentralized equilibrium, workers also receive the profits of land developers as a flat wage subsidy. Hence, workers receive transfers proportional to their wages and their income is $W(x)(1+t)$, where t denotes the wage subsidy net of the tax that finances the policy. We only present the conditions that change relative to the decentralized equilibrium of the extended model.

Markup removal. First, using $\iota(x)=\mathbb{D}(x)/\mathbb{P}(x)=\overline{\delta}(x)$, net profits after transfers are the extended-model analogue of (C26):

$$\widehat{\pi}(\omega,x)=\left[\Upsilon\left(\frac{y(\omega,x)}{Y(x)}\right)-\frac{\mathbb{C}(x)}{\omega}\frac{y(\omega,x)}{Y(x)}\right]\overline{\delta}(x)M(x). \tag{D18}$$

The first-order condition with respect to the relative quantity is

$$\Upsilon'\left(\frac{y(\omega,x)}{Y(x)}\right)=\frac{\mathbb{C}(x)}{\omega}. \tag{D19}$$

Since relative prices satisfy $\psi(\cdot)=\Upsilon'(\cdot)$ along the demand curve, (D19) implies that $p(\omega,x)/\mathbb{D}(x)=\mathbb{C}(x)/\omega$. Every firm prices at marginal cost, so $\mu(\omega,x)=1$ for all firms and $\overline{\mu}(x)=1$ in all locations.

Location ODE. Firms choose locations based on expected net profits, $\widehat{\Pi}(z,x)=\mathbb{E}_{\omega}[\widehat{\pi}(\omega,x)]$. The first-order condition of the location problem is

$$\mathcal{E}_{\overline{\delta}}(x)+\mathcal{E}_{M}(x)=\frac{\mathcal{E}_{\mathbb{C}}(x)}{\overline{\delta}(x)-1},$$

where the right-hand side follows from applying the envelope theorem to (D18), as in Appendix C.3. Markup removal makes $\chi(x)$ constant across locations:

$$\widehat{\chi}\equiv\frac{\alpha}{\eta}+\frac{1-\gamma}{\gamma}\frac{1}{\eta(1+t)}+\frac{(1-\beta)\gamma-\beta(1-\gamma)}{\gamma}.$$

Hence, $\mathcal{E}_{\chi}(x)=0$ and the terms $\Lambda_1(x)$ and $\Lambda_2(x)$ vanish. Market size retains the form (D10), with t in place of s and $\widehat{\chi}$ in place of $\chi(\cdot)$, which implies

$$\mathcal{E}_{M}(x)=\frac{\eta}{\xi}\mathcal{E}_{\mathbb{C}}(x)+\frac{\eta}{\xi}\mathcal{E}_{\overline{\delta}}(x)+1.$$

Moreover, since firms price at marginal cost, the pass-through of the competition index to relative prices is one. Then, $\mathcal{E}_{\overline{\delta}}(x)=\Theta_1(x)\mathcal{E}_{z}(x)-\widehat{\Theta}_2(x)\mathcal{E}_{\mathbb{C}}(x)$, with $\Theta_1(x)$ as in the extension of Proposition 2 and

$$\widehat{\Theta}_2(x)=\mathbb{E}_{\lambda}\left[\sigma\left(\frac{\mathbb{C}(x)}{\omega}\right)\left(\frac{1}{\overline{\delta}(x)}-\frac{1}{\mu\left(\frac{\mathbb{C}(x)}{\omega}\right)}\right)\right],$$

where $\mu(\cdot)$ denotes the markup function implied by the demand elasticity and not the equilibrium markup, which equals one. Combining these expressions yields the location ODE under the policy:

$$\frac{\mathbb{C}'(x)x}{\mathbb{C}(x)}\left[\frac{\theta\xi}{\overline{\delta}(x)-1}+\theta(\xi+\eta)\widehat{\Theta}_2(x)-\eta\theta\right]=\theta\xi+\theta(\xi+\eta)\Theta_1(x)\frac{z'(x)x}{z(x)}. \tag{D20}$$

The assignment ODE (D12) is unchanged, with all Kimball objects evaluated at the relative prices

$\mathbb{C}(x)/\omega$. Relative to (D13), the term $\overline{\delta}(x)$ replaces $\overline{\mu}(x)$ because firm profits are now given by the consumer surplus they create.

Free entry. Net profits are given by (D18) evaluated at (D19). Using $\mathbb{E}_\lambda\left[\delta\left(\mathbb{C}(x)/\omega\right)-1\right] = \overline{\delta}(x)-1$ and following the same steps as in (D14), the free-entry condition becomes

$$c_e M_e = \int_x M(x)\left[\overline{\delta}(x)-1\right] f_x(x)\,dx, \tag{D21}$$

which is the extended-model analogue of the entry condition under the optimal policy in Appendix C.3.

Government budget constraint. Since firms price at marginal cost, operating profits are zero and net profits equal the transfers received. Then, (D21) implies that the aggregate cost of the firm subsidy is $c_e M_e$. Using $R(x)^{1+\phi} = M(x)\widehat{\chi}$ and labor income $WL = M/(\eta(1+t))$, the national portfolio budget constraint is

$$\frac{t}{\eta(1+t)}\int_x M(x) f_x(x)\,dx = \frac{1}{1+\phi}\int_x M(x)\,\widehat{\chi}\,f_x(x)\,dx - c_e M_e. \tag{D22}$$

The left-hand side is the net transfer to workers. The right-hand side collects the profits of land developers net of the subsidy paid to firms. Given M_e and t, the ODE system (D12) and (D20) determines $z(x)$ and $\mathbb{C}(x)$ with the same boundary conditions as in the decentralized equilibrium. Conditions (D21) and (D22) determine M_e and t, and the labor adding-up constraint determines $\overline{U}$.

D.6 Equilibrium Conditions EZ Policy

This section derives the equilibrium conditions of the EZ program in the extended model. Local producers receive a wage-bill subsidy $S(x)$, a smooth function of appeal that starts at 20% in the targeted locations and declines continuously to zero in untargeted locations. As in the quasi-optimal policy, the subsidy is financed with a flat tax on labor income, and workers receive the profits of land developers as a flat wage subsidy. Worker income is $W(x)(1+t)$, where t denotes the wage subsidy net of the tax that finances the policy. We only present the conditions that change relative to the decentralized equilibrium of the extended model.

Production costs. Local producers pay $(1-S(x))W(x)$ per unit of labor. The marginal cost becomes $\nu[(1-S(x))W(x)]^\beta R(x)^{1-\beta}/\omega$ and the competition index is

$$\mathbb{C}(x) = \nu\frac{[(1-S(x))W(x)]^\beta R(x)^{1-\beta}}{\mathbb{D}(x)}.$$

The pricing problem is otherwise unchanged, so markups and all Kimball objects retain their form, evaluated at the relative prices $\mathbb{C}(x)/\omega$. By the labor first-order condition, the gross wage bill of local producers is $\beta M(x)/[(1-S(x))\overline{\mu}(x)]$, while their structures bill is unaffected. Wages retain

the form (D9) because traded-good producers are not subsidized.

Market clearing. Local labor-market clearing implies traded-good output

$$Q^T(x) = \frac{M(x)}{\gamma}\left[\frac{1}{\eta(1+t)} - \frac{\beta}{(1-S(x))\overline{\mu}(x)}\right],$$

and housing-land market clearing yields $R(x)^{1+\phi} = M(x)\chi^S(x)$, with

$$\chi^S(x) \equiv \frac{\alpha}{\eta} + \frac{1-\gamma}{\gamma}\frac{1}{\eta(1+t)} + \frac{(1-\beta)\gamma(1-S(x)) - \beta(1-\gamma)}{\gamma(1-S(x))\overline{\mu}(x)}.$$

Since the ideal price index picks up the factor $(1-S(x))^{\beta}$, market size becomes

$$M(x) = \left[\frac{\eta(1+t)^{1+\theta}}{\overline{U}^{\theta}\nu^{\eta\theta}\varrho^{\frac{1+\theta(1-\eta\beta)}{\gamma}}} \frac{\mathbb{C}(x)^{\eta\theta}\overline{\delta}(x)^{\eta\theta}}{(1-S(x))^{\eta\beta\theta}\chi^S(x)^{\theta\xi-1}}\right]^{\frac{1}{\theta\xi}} x.$$

Location ODE. The location first-order condition retains the form (D12), as the subsidy affects profits only through the competition index. The subsidy adds a direct channel to the elasticities of market size and $\chi^S(\cdot)$:

$$\begin{aligned}
\mathcal{E}_M(x) &= \frac{\eta}{\xi}\mathcal{E}_{\mathbb{C}}(x) + \frac{\eta}{\xi}\mathcal{E}_{\overline{\delta}}(x) + 1 - \frac{\theta\xi-1}{\theta\xi}\mathcal{E}_{\chi}(x) + \frac{\eta\beta}{\xi}\frac{S'(x)x}{1-S(x)},\\
\mathcal{E}_{\chi}(x) &= -\Lambda_1^S(x)\,\mathcal{E}_z(x) + \Lambda_2^S(x)\,\mathcal{E}_{\mathbb{C}}(x) - \frac{(1-\gamma)\beta}{\gamma(1-S(x))^2\overline{\mu}(x)\chi^S(x)}S'(x)x,
\end{aligned}$$

where $\Lambda_1^S(x)$ and $\Lambda_2^S(x)$ are given by $\Lambda_1(x)$ and $\Lambda_2(x)$ with $\chi^S(x)$ in place of $\chi(x)$ and the coefficient

$$\frac{(1-\gamma)\beta - (1-\beta)\gamma(1-S(x))}{\gamma(1-S(x))}$$

in place of $[(1-\gamma)\beta - (1-\beta)\gamma]/\gamma$. Collecting terms, the location ODE becomes

$$\begin{aligned}
&\frac{\mathbb{C}'(x)x}{\mathbb{C}(x)}\left[\frac{\theta\xi}{\overline{\mu}(x)-1} + \theta(\xi+\eta)\Theta_2(x) + (\theta\xi-1)\Lambda_2^S(x) - \eta\theta\right]\\
&\qquad = \theta\xi + \left[\theta(\xi+\eta)\Theta_1(x) + (\theta\xi-1)\Lambda_1^S(x)\right]\frac{z'(x)x}{z(x)} + \Lambda_3(x)\,S'(x)x, \qquad \text{(D23)}
\end{aligned}$$

with

$$\Lambda_3(x) \equiv \frac{\eta\beta\theta}{1-S(x)} + \frac{(\theta\xi-1)(1-\gamma)\beta}{\gamma(1-S(x))^2\overline{\mu}(x)\chi^S(x)}.$$

Since $S'(x) \leq 0$, the new term lowers the growth of competition around the targeted locations.

Free entry. Firm profits are still a fraction $1/\sigma(\mathbb{C}(x)/\omega)$ of sales, so the free-entry condition retains the form (D14).

Government budget constraint. The national portfolio collects the profits of land developers and pays the net transfer to workers plus the wage-bill subsidy, which costs $S(x)$ times the gross wage bill of local producers in each location. Using $R(x)^{1+\phi} = M(x)\chi^S(x)$ and labor income $WL = M/(\eta(1+t))$, the budget constraint is

$$\frac{t}{\eta(1+t)}\int_x M(x)f_x(x)\,dx + \beta\int_x \frac{S(x)}{(1-S(x))\overline{\mu}(x)}M(x)f_x(x)\,dx = \frac{1}{1+\phi}\int_x M(x)\,\chi^S(x)f_x(x)\,dx. \tag{D24}$$

Given M_e and t, the ODE system (D12) and (D23) determines $z(x)$ and $\mathbb{C}(x)$ with the same boundary conditions as in the decentralized equilibrium. The free-entry condition (D14) and the budget constraint (D24) determine M_e and t, and the labor adding-up constraint determines $\overline{U}$.

D.7 Decomposition of Local TFP Changes

This section derives the decomposition of the change in local TFP under a policy into four channels: allocative efficiency, aggregate entry, firm relocation, and firm productivity. The derivation applies to any policy that leaves the Kimball aggregator unchanged, including the quasi-optimal policy and the EZ program.

Local TFP as a function of three city-level objects. Let $q(\omega,x) \equiv y(\omega,x)/Y(x)$ denote the relative quantity of a firm with productivity ω in a city of appeal x. Along the demand curve, $q(\omega,x) = \varphi(\psi_r(\mathbb{C}(x)/\omega))$, where the pricing rule r maps relative marginal costs into relative prices: $\psi_r = \psi$ under markup pricing, as in the decentralized equilibrium and under the EZ program, whose wage subsidy enters only through $\mathbb{C}(x)$ (Appendix D.6); and $\psi_r(u) = u$ under the marginal-cost pricing induced by the quasi-optimal policy, by (D19). Let $N(x) \equiv M_e\, g(z(x))\, z'(x)/f_x(x)$ denote the mass of local producers in a city of appeal x. For a pricing rule r, a mass of local producers N, and a sorting productivity z, the Kimball normalization (D12) and local TFP in (39) read

$$N\int_\omega \Upsilon\left(\varphi\left(\psi_r\left(\frac{\mathbb{C}}{\omega}\right)\right)\right) g(\omega|z)\,d\omega = 1, \qquad \mathcal{Z}(r,N,z) = \left[N\int_\omega \frac{\varphi\left(\psi_r\left(\frac{\mathbb{C}(r,N,z)}{\omega}\right)\right)}{\omega}\, g(\omega|z)\,d\omega\right]^{-1}, \tag{D25}$$

where $\mathbb{C}(r,N,z)$ denotes the solution of the normalization. This solution is unique because $\psi_r(\cdot)$ is increasing, $\varphi(\cdot)$ is decreasing, and $\Upsilon(\cdot)$ is increasing, so that the left-hand side is strictly decreasing

in $\mathbb{C}$. In any equilibrium e, $\mathcal{Z}^e(x) = \mathcal{Z}(r^e, N^e(x), z^e(x))$, since $\mathbb{C}^e(x)$ solves the normalization at $N^e(x)$ and $z^e(x)$. The competition index is therefore not an independent determinant of local TFP: a policy changes $\mathcal{Z}(x)$ only through the pricing rule, the mass of local producers, and the productivity of the firms that sort into x. Finally, the mass of local producers is the product of aggregate entry and the density of entrants across cities,

$$N(x) = M_e\, s(x), \qquad s(x) \equiv \frac{g(z(x))\, z'(x)}{f_x(x)}, \qquad \int_x s(x) f_x(x)\, dx = 1, \tag{D26}$$

so that $s(x) f_x(x)$ is the share of entrants that locate in cities of appeal x.

Channels. Consider a policy that moves the economy from a baseline equilibrium, denoted by superscript 0, to a policy equilibrium, denoted by superscript 1. By (D25) and (D26), the change in local TFP is the joint effect of the changes in the four inputs $\theta(x) \equiv (r, M_e, s(x), z(x))$, which define the channels: (i) *allocative efficiency*, the change in the pricing rule at fixed $N(x)$ and $z(x)$, which corresponds to the within-city misallocation measure of Edmond, Midrigan, and Xu (2023); (ii) *aggregate entry*, the change in M_e at fixed $s(x)$; (iii) *firm relocation*, the change in $s(x)$ at fixed M_e; and (iv) *firm productivity*, the change in $z(x)$, which shifts the local productivity distribution $g(\omega|z(x))$. Under the EZ program $r^1 = r^0$, so channel (i) is identically zero.

Shapley decomposition. Because the inputs interact through $\mathbb{C}(r, N, z)$, the contribution of each channel depends on the values at which the other inputs are held. The Shapley value (Shapley, 1953) removes this dependence by averaging over all orders in which the inputs can be switched from baseline to policy values. Let $K = \{r, M_e, s, z\}$ and, for $S \subseteq K$, let $\theta^S(x)$ denote the vector of inputs with policy values for the inputs in S and baseline values for the others. Define $v_x(S) \equiv \mathcal{Z}(\theta^S(x))$ using (D25), so that $v_x(\emptyset) = \mathcal{Z}^0(x)$ and $v_x(K) = \mathcal{Z}^1(x)$. The contribution of channel $k \in K$ is

$$\begin{aligned} \phi_k(x) &= \sum_{S \subseteq K \setminus \{k\}} \frac{|S|!\ (|K| - |S| - 1)!}{|K|!} \left[v_x\left(S \cup \{k\}\right) - v_x(S) \right] \\ &= \frac{1}{|\Pi_K|} \sum_{\pi \in \Pi_K} \left[v_x\left(P_k^\pi \cup \{k\}\right) - v_x\left(P_k^\pi\right) \right], \end{aligned} \tag{D27}$$

where Π_K is the set of orderings of K and P_k^π is the set of inputs that precede k in ordering π. The second expression shows that $\phi_k(x)$ is the average over switching orders of the sequential contribution of k, and it implies that the contributions add up exactly to the change in local TFP, $\sum_{k \in K} \phi_k(x) = \mathcal{Z}^1(x) - \mathcal{Z}^0(x)$. We report the contributions in percent of baseline TFP, $100\, \phi_k(x) / \mathcal{Z}^0(x)$, which add up to the percent change in local TFP.

Nesting of aggregate entry and firm relocation. Since M_e and $s(x)$ enter (D25) only through their product $N(x)$, we treat them as a block. Let $\Pi_K^N \subset \Pi_K$ denote the orderings in which M_e and s are adjacent, and compute (D27) averaging over Π_K^N instead of Π_K.[42] This is equivalent

[42] This is the Owen value of the game with a priori unions $\{r\}$, $\{M_e, s\}$, and $\{z\}$ (Owen, 1977).

to computing the Shapley values over the three inputs $\{r, N, z\}$ and splitting the contribution of $N(x)$ by switching M_e and $s(x)$ in both orders within each step. Consequently, the contributions of allocative efficiency and firm productivity coincide with their three-input Shapley values, and the contributions of aggregate entry and firm relocation add up to the contribution of the mass of local producers.

Computation. For each city on the grid of x used to solve the model, $v_x(S)$ is evaluated at the 2^4 combinations of baseline and policy inputs, each requiring only the solution of the normalization in (D25) for $\mathbb{C}$, so no additional equilibrium needs to be computed. The mixed combinations of M_e and $s(x)$ correspond to $N(x) = M_e^1 s^0(x)$ and $N(x) = M_e^0 s^1(x)$.

E Markup Estimation

This section expands on the details for markup estimation of Section 4.4.2.

E.1 Derivations for Estimating Equation

Let p_{jc} denote the price charged by establishment j in city c. Let $\mathbb{P}_c$ and $\mathbb{D}_c$ denote the ideal and competition price indices in city c, respectively. By the Lerner formula, the markup of establishment j in c can be written as a function of its relative price:

$$\mu_{jc} = \mu\left(\frac{p_{jc}}{\mathbb{D}_c}\right).$$

I use the demand system to express the relative price as a function of the establishment's sales share. Define the sales share of establishment j in city c as

$$s^y_{jc} \equiv \frac{p_{jc}y_{jc}}{\sum_{j'\in c} p_{j'c}y_{j'c}}.$$

Using that total expenditure on local goods in c is given by $\mathbb{P}_c Y_c L_c$, this share can be written as

$$\begin{aligned} s^y_{jc} &= \frac{p_{jc}y_{jc}L_c}{\mathbb{P}_c Y_c L_c}, \\ &= \frac{p_{jc}y_{jc}}{\mathbb{P}_c Y_c}. \end{aligned}$$

Using the residual demand curve (7), this expression becomes

$$\begin{aligned} s^y_{jc} &= \frac{p_{jc}}{\mathbb{D}_c}\frac{y_{jc}}{Y_c}\frac{\mathbb{D}_c}{\mathbb{P}_c}, \\ &= \frac{p_{jc}}{\mathbb{D}_c}\varphi\left(\frac{p_{jc}}{\mathbb{D}_c}\right)\frac{\mathbb{D}_c}{\mathbb{P}_c}. \end{aligned}$$

To express the relative price $p_{jc}/\mathbb{D}_c$ as a function of the sales share s^y_{jc} and the ratio $\mathbb{P}_c/\mathbb{D}_c$, it is sufficient to show that the function $x\varphi(x)$ is strictly monotone. Indeed,

$$\begin{aligned} \frac{d\left(x\varphi(x)\right)}{dx} &= \varphi(x) + x\varphi'(x), \\ &= -\varphi(x)\left(\sigma(x) - 1\right), \end{aligned}$$

where I use (14) for the definition of $\sigma(\cdot)$. Since firms choose prices such that $\sigma(x) > 1$, it follows that $x\varphi(x)$ is strictly decreasing and therefore invertible. Let $\zeta(\cdot)$ denote its inverse.

It then follows that the relative price can be written as a function of the sales share and the price

indices, and therefore markup can be expressed as

$$\mu_{jc} = \mu\left(\zeta\left(s_{jc}^{y}\frac{\mathbb{P}_c}{\mathbb{D}_c}\right)\right).$$

Considering a third-degree polynomial in $s_{jc}^{y}(\mathbb{P}_c/\mathbb{D}_c)$, I approximate this markup function as

$$\log\mu\left(\zeta\left(s_{jc}^{y}\frac{\mathbb{P}_c}{\mathbb{D}_c}\right)\right) = \varsigma_1 s_{jc}\frac{\mathbb{P}_c}{\mathbb{D}_c} + \varsigma_2\left(s_{jc}^{y}\frac{\mathbb{P}_c}{\mathbb{D}_c}\right)^2 + \varsigma_3\left(s_{jc}^{y}\frac{\mathbb{P}_c}{\mathbb{D}_c}\right)^3 + \nu_{jc}^{*},$$

where ν_{jc}^{*} is an approximation error that goes to zero once one considers higher polynomial terms.[43] Crucially, Assumption 3 implies that the approximation in does not have a constant term. Intuitively, because of the choke price, firms with zero sales share face an infinite elasticity of demand and, therefore, have markups equal to one. Replacing back the markup approximation into the estimating equation gives:

$$\log\alpha_{jc}^{l} = \log\beta + \underbrace{\varsigma_{1,c}s_{jc}^{y} + \varsigma_{2,c}(s_{jc}^{y})^2 + \varsigma_{3,c}(s_{jc}^{y})^3}_{\equiv\log\mu_{jc}} + (\nu_{jc} - \nu_{jc}^{*}),$$

where $\varsigma_{1,c} \equiv -\varsigma_1(\mathbb{P}_c/\mathbb{D}_c)$, $\varsigma_{2,c} \equiv -\varsigma_2(\mathbb{P}_c/\mathbb{D}_c)^2$, and $\varsigma_{3,c} \equiv -\varsigma_3(\mathbb{P}_c/\mathbb{D}_c)^3$.

E.2 Consistency of Sieve Estimator

This section outlines the necessary conditions for the consistency of the sieve estimator used in (43) for the markup estimation. Theorem 4 in Newey (1997) provides conditions for root-n consistency of a polynomial (power series) sieve estimator. It is straightforward to extend these conditions to the current setting.

First, note that Assumption 3 implies that firms charge finite prices. Therefore, the price index $\mathbb{D}_c$ is bounded. Moreover, this assumption also implies that active firms have a strictly positive market share, $s_{jc}^{Y} > 0$. Thus, the object $s_{jc}^{y}(\mathbb{P}_c/\mathbb{D}_c)$ is bounded away from zero and Assumption 8 in Newey (1997) is satisfied.

Second, note that the derivative of the markup function is given by

$$\frac{d\mu(\zeta(a))}{dx} = \frac{\mu'(\zeta(a))}{\varphi(\zeta(a)) + \zeta(a)\varphi'(\zeta(a))},$$

where we used the fact that $\zeta(\cdot)$ is the inverse of the function $x\varphi(a)$. Note that the order of differentiability of the functions $\mu(\cdot)$ and $\varphi(\cdot)$ is given by the order of differentiability of the Kimball aggregator $\Upsilon(\cdot)$. In particular, if $\Upsilon(\cdot)$ is differentiable of order d, the functions $\mu(\cdot)$ and $\varphi(\cdot)$ are differentiable of order $d-1$. Hence, Assumption 9 in Newey (1997) is satisfied with $s = d-1$.

[43] The error term is given by $\nu_{jc}^{*} = \log\mu\left(\zeta\left(s_{jc}^{y}\frac{\mathbb{P}_c}{\mathbb{D}_c}\right)\right) - \varsigma_1 s_{jc}\frac{\mathbb{P}_c}{\mathbb{D}_c} - \varsigma_2\left(s_{jc}^{y}\frac{\mathbb{P}_c}{\mathbb{D}_c}\right)^2 - \varsigma_3\left(s_{jc}^{y}\frac{\mathbb{P}_c}{\mathbb{D}_c}\right)^3$.

Lastly, we are required to impose the condition that $\text{Var}(\mu|s^y, \mathbb{P}, \mathbb{D})$ is bounded. Under these conditions, Theorem 4 applies and the error term $\upsilon = O(K^{-d+1})$ where K is degree of the polynomial approximation. Higher smoothness of $\Upsilon(\cdot)$ guarantees a quicker convergence rate.

E.3 Multi-Sector Estimation

In this section, I extend the baseline estimation framework to allow for multiple sectors.

The multi-sector estimation procedure assumes that consumers have Cobb-Douglas preferences over sector-specific bundles of local varieties. Formally, the bundle of local varieties $Y(a)$ is a Cobb-Douglas aggregator of sector-specific bundles:

$$Y(a) = \prod_{n=1}^{N} Y_n(a)^{\theta_n}, \qquad \sum_{n=1}^{N} \theta_n = 1,$$

where n denotes the sector. Each sector-specific bundle $Y_n(a)$ is, in turn, implicitly defined by a Kimball aggregator:

$$\int_{\omega} \Upsilon_n \left(\frac{y_n(\omega, c)}{Y_n(a)} \right) dG_{n,c}(\omega) = 1,$$

where $y_n(\omega, c)$ is consumption in city c of a variety produced by a firm with productivity ω in sector n, $G_{n,c}(\omega)$ is the productivity distribution of sector n in city c, and the Kimball aggregator $\Upsilon_n(\cdot)$ is sector specific. Under this formulation, all derivations from Section 4 extend, with the key difference that the markup function, sales shares, and price indices are now sector specific:

$$\mu_{jnc} = \mu_n \left(\zeta_n \left(s_{jnc} \frac{\mathbb{P}_{nc}}{\mathbb{D}_{nc}} \right) \right). \tag{E1}$$

Moreover, I allow production functions to differ across sectors:

$$y_{jnc} = \omega_{jnc} e^{\nu_{jnc}} l_{jnc}^{\beta_n} s_{jnc}^{1-\beta_n}. \tag{E2}$$

Equations (E1) and (E2) imply the multi-sector estimating equation

$$\log \alpha_{jnc}^{l} = \log \beta_n + \varsigma_{1,nc} s_{jnc}^{y} + \varsigma_{2,nc} (s_{jnc}^{y})^2 + \varsigma_{3,nc} (s_{jnc}^{y})^3 + \upsilon_{jnc}, \tag{E3}$$

where $\varsigma_{1,nc}$, $\varsigma_{2,nc}$, and $\varsigma_{3,nc}$ are city-sector-specific coefficients that absorb the variation in $\mathbb{P}_{nc}/\mathbb{D}_{nc}$.

E.4 Labor Market Power

This section extends the estimation framework to allow local producers to exercise market power in local labor markets, mirroring the product-market derivation in Appendix E.1.

Labor Supply and Markdowns. Workers in city c view jobs at different local producers as imperfect substitutes. Total labor supplied to local producers, L_c^N, is implicitly defined by a Kimball aggregator, $\int_{j\in c} \Upsilon^l\left(l_{jc}/L_c^N\right) dj = 1$, where $\Upsilon^l(\cdot)$ is strictly increasing and convex with $\Upsilon^l(0) = 0$. Workers allocate hours across establishments to maximize labor income subject to this constraint, which yields the labor supply to establishment j:

$$\frac{l_{jc}}{L_c^N} = \varphi^l\left(\frac{w_{jc}}{\mathbb{D}_c^l}\right), \tag{E4}$$

where $\varphi^l(\cdot) \equiv \left(\Upsilon^{l\prime}\right)^{-1}(\cdot)$ is strictly increasing and $\mathbb{D}_c^l$ is a wage competition index implicitly defined by $\int_{j\in c} \Upsilon^l\left(\varphi^l\left(w_{jc}/\mathbb{D}_c^l\right)\right) dj = 1$. The ideal wage index $\mathbb{W}_c$ is defined through the local wage bill, $\mathbb{W}_c L_c^N \equiv \int_{j\in c} w_{jc} l_{jc}\, dj$. The wage elasticity of labor supply to an establishment, $\sigma^l(x) \equiv x\varphi^{l\prime}(x)/\varphi^l(x)$ evaluated at $x = w_{jc}/\mathbb{D}_c^l$, is finite and varies with the relative wage whenever $\Upsilon^l(\cdot)$ is not isoelastic. Establishments take (E4) as given when choosing employment, so the first-order condition for labor equates the wage to a markdown over the marginal revenue product of labor, $\beta p_{jc} y_{jc}/(\mu_{jc} l_{jc})$. The ratio estimator (40) therefore becomes

$$\alpha_{jc}^l = \frac{\beta}{\mu_{jc}\,\varrho_{jc}} e^{-\nu_{jc}}, \qquad \varrho_{jc} \equiv 1 + \frac{1}{\sigma^l\left(w_{jc}/\mathbb{D}_c^l\right)} \equiv \varrho\left(\frac{w_{jc}}{\mathbb{D}_c^l}\right), \tag{E5}$$

where $\varrho_{jc} \geq 1$ is the markdown, the ratio of the marginal revenue product of labor to the wage, and $\alpha_{jc}^l = (w_{jc} l_{jc})/(p_{jc} y_{jc})$ is the establishment's own payroll over sales.

Control Function and Estimating Equation. Let $s_{jc}^l \equiv w_{jc} l_{jc} / \sum_{j'\in c} w_{j'c} l_{j'c}$ denote the wage-bill share of establishment j among local producers in city c. By (E4) and the definition of $\mathbb{W}_c$, $s_{jc}^l = x\varphi^l(x)\,\mathbb{D}_c^l/\mathbb{W}_c$ with $x = w_{jc}/\mathbb{D}_c^l$. Since $d\left(x\varphi^l(x)\right)/dx = \varphi^l(x)\left(1 + \sigma^l(x)\right) > 0$, the function $x\varphi^l(x)$ is invertible, with inverse $\zeta^l(\cdot)$, and the markdown can be written as a function of the wage-bill share and the wage indices, $\varrho_{jc} = \varrho\left(\zeta^l\left(s_{jc}^l \mathbb{W}_c/\mathbb{D}_c^l\right)\right)$, the labor-market analogue of (42). Approximating the log markdown with a third-degree polynomial in $s_{jc}^l(\mathbb{W}_c/\mathbb{D}_c^l)$, as in Appendix E.1 for the log markup, and taking logs of (E5) yields

$$\log \alpha_{jc}^l = \log\beta + \underbrace{\sum_{k=1}^{3} \varsigma_{k,c}\left(s_{jc}^y\right)^k}_{\equiv -\log\mu_{jc}} + \underbrace{\sum_{k=1}^{3} \varsigma_{k,c}^l\left(s_{jc}^l\right)^k}_{\equiv -\log\varrho_{jc}} + \upsilon_{jc}, \tag{E6}$$

where $\varsigma_{k,c}^l \equiv -\varsigma_k^l\left(\mathbb{W}_c/\mathbb{D}_c^l\right)^k$ are city-specific coefficients that absorb the variation in the wage indices, and υ_{jc} collects the measurement and approximation errors. I estimate (E6) by OLS, with markups given by (44) and markdowns by $\hat{\varrho}_{jc} = \exp(-\sum_k \hat{\varsigma}_{k,c}^l (s_{jc}^l)^k)$. They are separately identified because, by the first-order condition for labor, s_{jc}^l/s_{jc}^y is inversely proportional to $\mu_{jc}\varrho_{jc}$, so sales and wage-bill shares are not collinear within a city. In the multi-sector version of Appendix E.3, sales shares are city-sector specific while wage-bill shares remain city specific, since workers supply labor to local producers in all sectors. These estimates underlie Figure F3 and Table F2.

E.5 General Production Function

This section derives the markup estimating equation under a general production function. Because data on production inputs are limited outside Manufacturing, I derive the estimating equation using the more detailed data available for this sector. Formally, consider a Hicks-neutral production function in labor, materials, energy, and capital:

$$y_{jc} = \omega_{jc} e^{\nu_{jc}} \mathcal{F}\left(l_{jc}, m_{jc}, e_{jc}, k_{jc}\right), \tag{E7}$$

where $\mathcal{F}(\cdot)$ is a continuously differentiable function, m_{jc} denotes materials, e_{jc} denotes energy and k_{jc} denotes capital. For Manufacturing, capital includes commercial structures, consistent with the baseline markup estimation. Under this specification, the elasticities of output with respect to labor, materials and energy are given by

$$\begin{aligned}
\frac{\partial \log y_{jc}}{\partial \log l_{jc}} &= \frac{\partial \mathcal{F}\left(l_{jc}, m_{jc}, e_{jc}, k_{jc}\right)}{\partial l_{jc}} \frac{l_{jc}}{\mathcal{F}\left(l_{jc}, m_{jc}, e_{jc}, k_{jc}\right)} \\
\frac{\partial \log y_{jc}}{\partial \log m_{jc}} &= \frac{\partial \mathcal{F}\left(l_{jc}, m_{jc}, e_{jc}, k_{jc}\right)}{\partial m_{jc}} \frac{m_{jc}}{\mathcal{F}\left(l_{jc}, m_{jc}, e_{jc}, k_{jc}\right)}, \\
\frac{\partial \log y_{jc}}{\partial \log e_{jc}} &= \frac{\partial \mathcal{F}\left(l_{jc}, m_{jc}, e_{jc}, k_{jc}\right)}{\partial e_{jc}} \frac{e_{jc}}{\mathcal{F}\left(l_{jc}, m_{jc}, e_{jc}, k_{jc}\right)}.
\end{aligned}$$

In all three cases, the Hicks-neutral specification implies that the output elasticity with respect to a variable input is a function only of the inputs of production:

$$\frac{\partial \log y_{jc}}{\partial \log l_{jc}} \equiv \beta_l\left(l_{jc}, m_{jc}, e_{jc}, k_{jc}\right), \quad \frac{\partial \log y_{jc}}{\partial \log m_{jc}} \equiv \beta_m\left(l_{jc}, m_{jc}, e_{jc}, k_{jc}\right), \quad \frac{\partial \log y_{jc}}{\partial \log e_{jc}} \equiv \beta_e\left(l_{jc}, m_{jc}, e_{jc}, k_{jc}\right).$$

Therefore, under the production function (E7), the markup estimating equation (43) becomes

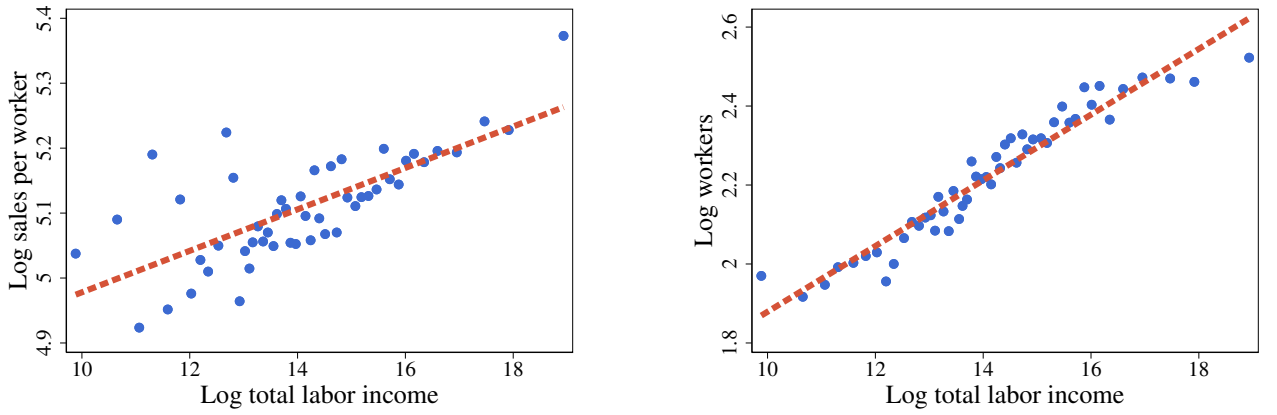
$$\log \alpha_{jc}^{x} = \beta_x\left(l_{jc}, m_{jc}, e_{jc}, k_{jc}\right) + \varsigma_{1,c} s_{jc}^{y} + \varsigma_{2,c}\left(s_{jc}^{y}\right)^2 + \varsigma_{3,c}\left(s_{jc}^{y}\right)^3 + \upsilon_{jc}, \tag{E8}$$

where x denotes the input of production, $x \in \{l, m, e\}$. As with the markup function, I approximate $\beta_x\left(l_{jc}, m_{jc}, e_{jc}, k_{jc}\right)$ semi-parametrically using a third-order polynomial in its arguments, following Gandhi, Navarro, and Rivers (2020).

F Empirical Appendix

F.1 Additional Figures

Figure F1: Revenue per worker and average employment across commuting zones: firms operating in a single commuting zone

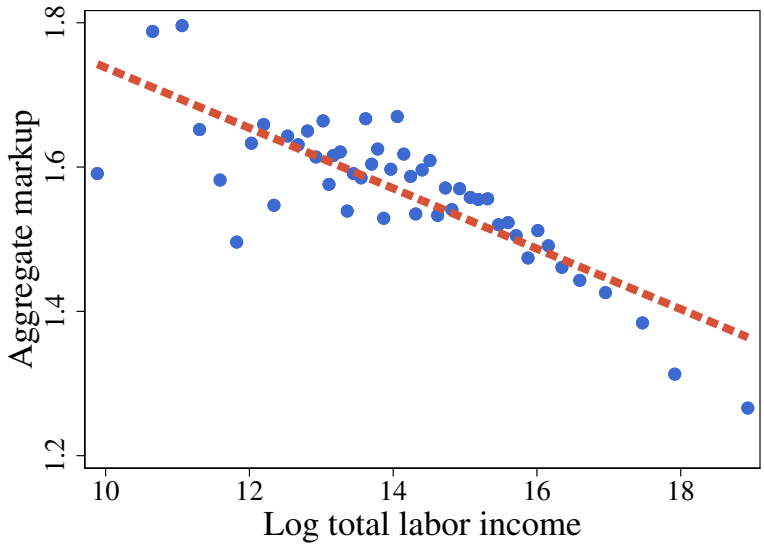


(a) Sales per worker and total labor income

(b) Average employment and total labor income

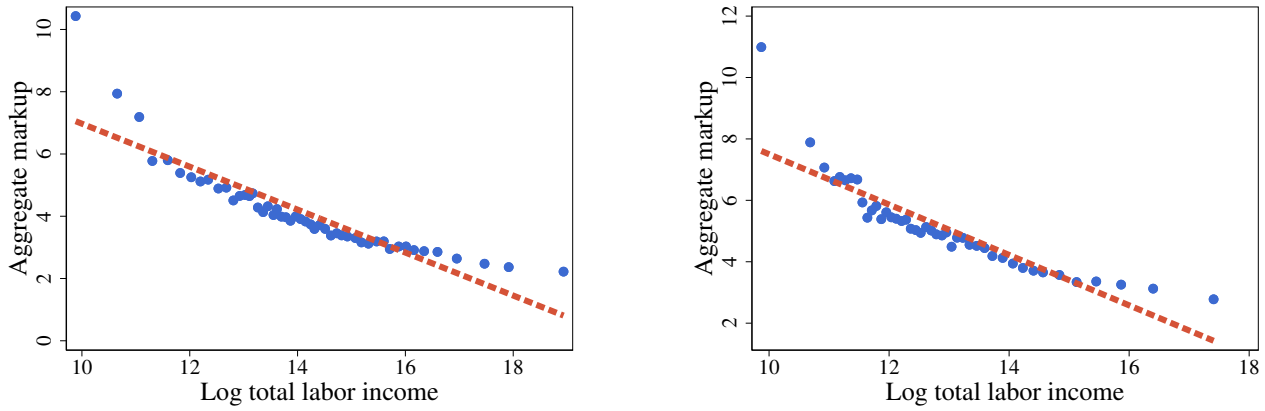
Notes: Figures F1(a) and F1(b) reproduce Figures 1(a) and 1(b) restricting the sample to establishments belonging to firms that operate in a single commuting zone. Both bin-scatters consider 50 equally sized commuting zone bins according to their total labor income, and both averages are computed among establishments in local industries. The dashed line is a linear fit across bins.

Figure F2: Aggregate markup and market size, lagged-share instrument



Notes: Figure F2 reproduces the commuting zone panel of Figure 2 with establishment markups estimated from (43) using lagged sales shares as instruments for contemporaneous sales shares. Commuting zone size is defined as total labor income. The bin-scatter considers 50 equally sized commuting zone bins according to their total labor income. Commuting zone aggregate markup is the sales-weighted harmonic mean in (45). The dashed line is a linear fit across bins.

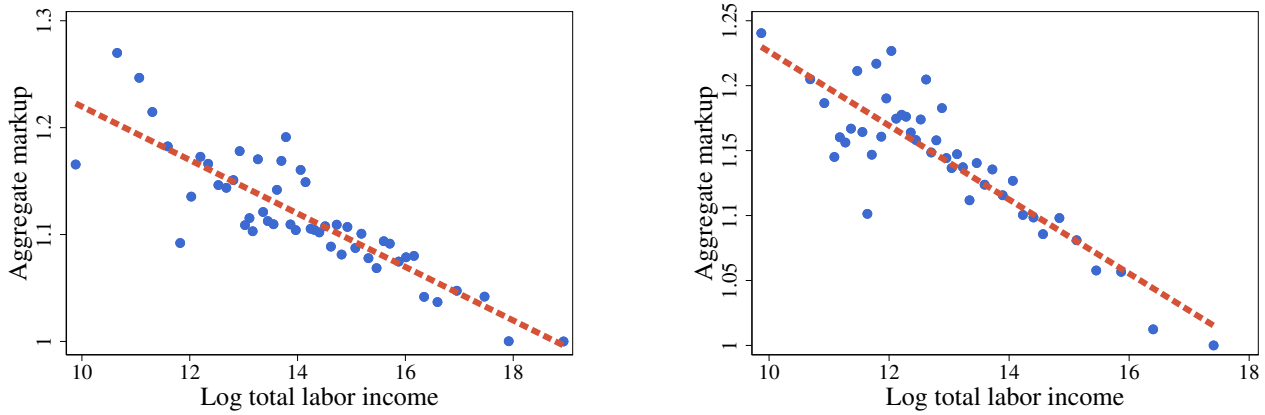
Figure F3: Aggregate markup and market size, controlling for labor market power



(a) Commuting zones (b) Counties

Notes: Figures F3(a) and F3(b) reproduce Figure 2 with establishment markups estimated from (43) controlling for labor market power through a flexible polynomial in the establishment's wage-bill share. City size is defined as total labor income. Both bin-scatters consider 50 equally sized bins according to total labor income; in the county panel, disclosure rules pooled the smallest bins. City aggregate markup is the sales-weighted harmonic mean in (45). The dashed line is a linear fit across bins.

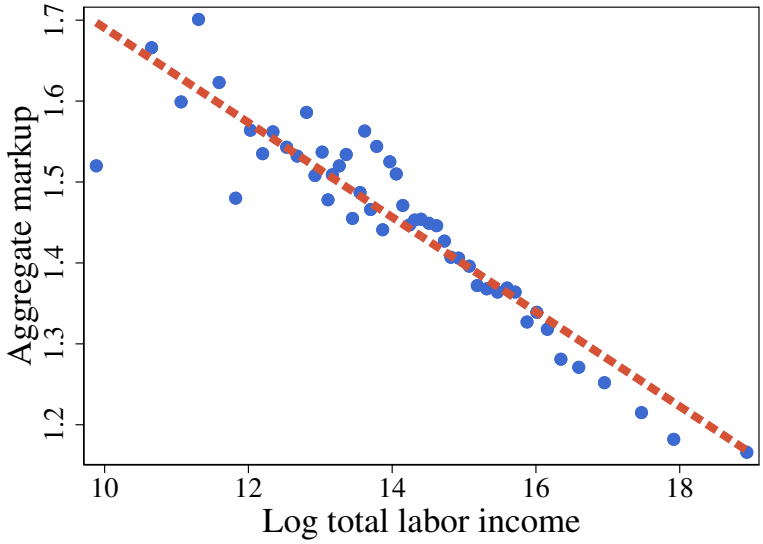
Figure F4: Aggregate markup (inverse labor share) and market size



(a) Commuting zones (b) Counties

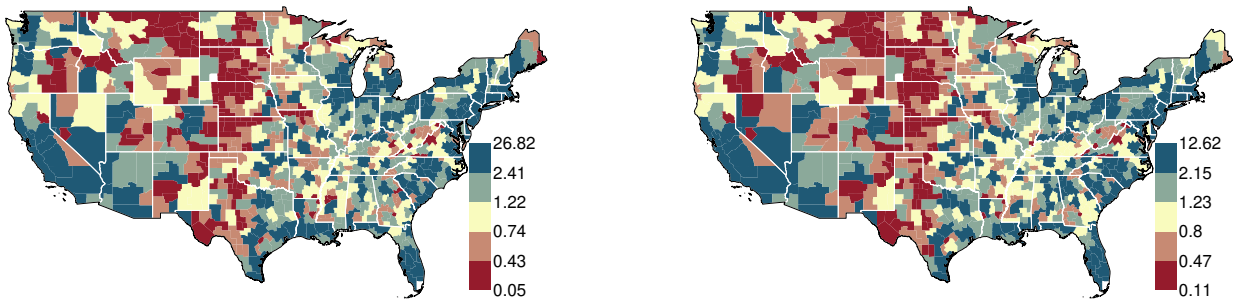
Notes: Figures F4(a) and F4(b) reproduce Figure 2 with establishment markups defined as the inverse of the labor share, $\mu_{jc} = 1/\omega_{jc}$, with $\omega_{jc} \equiv \ell_{jc} w_{jc}/(p_{jc} y_{jc})$. Because markup levels are not identified under this measure, each panel is normalized by the aggregate markup of the largest-market bin. City size is defined as total labor income. Both bin-scatters consider 50 equally sized bins according to total labor income; in the county panel, disclosure rules pooled the smallest bins. City aggregate markup is the sales-weighted harmonic mean in (45). The dashed line is a linear fit across bins.

Figure F5: Aggregate markup and market size: firms operating in a single commuting zone



Notes: Figure F5 reproduces the commuting zone panel of Figure 2 restricting the sample to establishments belonging to firms that operate in a single commuting zone. Commuting zone size is defined as total labor income. The bin-scatter considers 50 equally sized commuting zone bins according to their total labor income. Commuting zone aggregate markup is the sales-weighted harmonic mean in (45). The dashed line is a linear fit across bins.

Figure F6: Local Productivity and Local Amenities

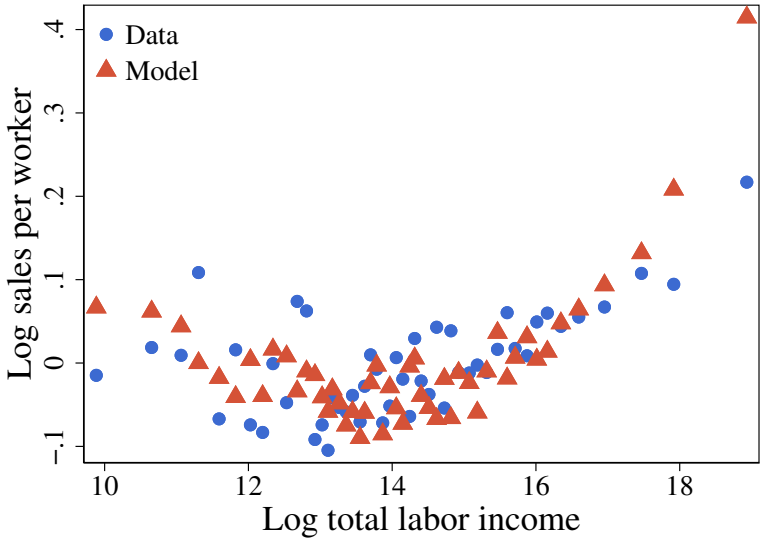


(a) Traded Good Productivity

(b) Local Amenities

Notes: Figure F6(a) shows the model implied local traded good productivity and Figure F6(b) shows the model implied local amenities. Commuting zones omitted in the analysis are not colored.

Figure F7: Sales per Worker: Model and Data



Notes: Figure F7 plots data and model average sales per worker across the 50 commuting zone bins of Figure 1(a).

F.2 Additional Tables

Table F1: Summary Statistics: Markups

	Mean	p10	p90
	(1)	(2)	(3)
Baseline - OLS	1.055	1.001	1.085
Lagged-share IV	1.056	1.001	1.094

Notes: Table F1 displays summary statistics for the establishment-level markups according to (44).

Table F2: Markups Controlling for Labor Market Power

	Dep. var.: Log aggregate markup			
	2002	2007	2012	2017
	(1)	(2)	(3)	(4)
	Panel A: Commuting Zones			
Log total labor income	-0.1579***	-0.1469***	-0.1530***	-0.1532***
	(0.0044)	(0.0039)	(0.0046)	(0.0045)
Observations	700	700	700	700
R-squared	0.79	0.773	0.764	0.766
	Panel B: Counties			
Log total labor income	-0.1603***	-0.1520***	-0.1657***	-0.1561***
	(0.0047)	(0.0048)	(0.0051)	(0.0047)
Observations	3200	3200	3200	3200
R-squared	0.41	0.386	0.426	0.396

Notes: Table F2 displays the average elasticity of city aggregate markup and city size. Establishment markup is estimated from (43) controlling for labor market power through a flexible polynomial in the establishment's wage-bill share. City aggregate markup is the sales-weighted harmonic mean in (45), and city size is defined as total labor income. Panel A shows the mean elasticity defining cities as commuting zones. Panel B shows the mean elasticity defining cities as counties. Robust standard errors in parenthesis. *10% level, **5% level, ***1% level.

Table F3: Markups as Inverse Labor Share

	Dep. var.: Log aggregate markup			
	2002	2007	2012	2017
	(1)	(2)	(3)	(4)
	Panel A: Commuting Zones			
Log total labor income	-0.0230***	-0.0257***	-0.0284***	-0.0195***
	(0.0027)	(0.0032)	(0.0035)	(0.0032)
Observations	700	700	700	700
R-squared	0.167	0.163	0.17	0.101
	Panel B: Counties			
Log total labor income	-0.0301***	-0.0313***	-0.0312***	-0.0076**
	(0.0030)	(0.0029)	(0.0028)	(0.0030)
Observations	3200	3200	3200	3200
R-squared	0.076	0.081	0.067	0.005

Notes: Table F3 displays the average elasticity of city aggregate markup and city size. Establishment markup is defined as the inverse of the labor share, $\mu_{jc} = 1/\omega_{jc}$, with $\omega_{jc} \equiv \ell_{jc} w_{jc}/(p_{jc} y_{jc})$. City aggregate markup is the sales-weighted harmonic mean in (45), and city size is defined as total labor income. Panel A shows the mean elasticity defining cities as commuting zones. Panel B shows the mean elasticity defining cities as counties. Robust standard errors in parenthesis. *10% level, **5% level, ***1% level.

F.3 Markups Across Cities for additional Sectors

Table F4 reports the estimated elasticity of county-sector aggregate markups with respect to county size across two-digit NAICS sectors. Establishment-level markups are estimated using equation (E3). The results reveal substantial heterogeneity across sectors. Construction, Manufacturing, Wholesale, Real Estate, Information, and Healthcare display positive elasticities, while the remaining sectors display negative elasticities. Strikingly, the elasticity is statistically different from zero in every sector, suggesting substantial variation in the degree of local competition across counties within sectors.

The results also shed light on which sectors drive the negative aggregate relationship between markups and city size documented in Figure 2. The bottom row of each panel reports the average employment share of each sector in total local-industry employment across counties. Healthcare, Retail, and Accommodation and Food Services account for the largest employment shares. Retail and Accommodation and Food Services display negative elasticities, suggesting that they are important drivers of the aggregate relationship in Figure 2, while the small positive elasticity in Healthcare partially offsets these effects.

Table F5 repeats the analysis using commuting zones instead of counties to define cities. The pattern is similar. Construction, Wholesale, Real Estate, and Healthcare display positive and statistically significant elasticities, while the elasticities for Manufacturing and Information are not statistically different from zero. Importantly, the three sectors with the largest shares of local

employment display the same signs as in the county-level analysis.

Table F4: County Aggregate markup and County Size Regressions by Sector

			Panel A		
	Construction	Manufacturing	Wholesale	Retail	Transportation
	(1)	(2)	(3)	(4)	(5)
Log total labor income	0.0417***	0.0152***	0.0972***	-0.0243***	-0.0481***
	(0.0025)	(0.0052)	(0.0047)	(0.0020)	(0.0035)
Observations	3100	1800	2400	3100	2800
R-squared	0.089	0.004	0.118	0.056	0.07
Avg. Local Emp. Share	0.0798	0.013	0.0206	0.241	0.0259
			Panel B		
	Information	Finance	Real Estate	PST Services	AWR Services
	(1)	(2)	(3)	(4)	(5)
Log total labor income	0.0090***	-0.0038**	0.0385***	-0.0377***	-0.0583***
	(0.0028)	(0.0018)	(0.0036)	(0.0023)	(0.0030)
Observations	2700	3100	2800	3000	2900
R-squared	0.004	0.001	0.039	0.086	0.105
Avg. Local Emp. Share	0.0163	0.0464	0.0156	0.0273	0.055
			Panel C		
	Education Services	Healthcare	Arts, Entertainment and Recreation	Accommodation and Food Services	Other Services
	(1)	(2)	(3)	(4)	(5)
Log total labor income	-0.0378***	0.0207***	-0.0104***	-0.0218***	-0.0255***
	(0.0050)	(0.0014)	(0.0026)	(0.0017)	(0.0026)
Observations	1400	3100	2100	3100	3100
R-squared	0.034	0.058	0.007	0.068	0.032
Avg. Local Emp. Share	0.00378	0.274	0.0112	0.143	0.0499

Notes: Table F4 displays the average elasticity of county aggregate markup and county size. County aggregate markup is defined as (45) and county size is defined as total labor income. Dependent variable in all columns and panels is log county aggregate markup. County aggregate markup is computed using establishments in local industries within the specific sector. Sectors are defined as 2-digits NAICS sectors. PST Services: Professional, Scientific, and Technical Services. AWR Services: Administrative and Support and Waste Management and Remediation Services. Average local employment share is the average employment share of the sector across counties among the total employment of local industries. Robust standard errors in parenthesis. *10% level, **5% level, ***1% level.

Table F5: Commuting Zone Aggregate Markup and Commuting Zone Size Regressions by Sector

			Panel A		
	Construction	Manufacturing	Wholesale	Retail	Transportation
	(1)	(2)	(3)	(4)	(5)
Log total labor income	0.033***	0.001	0.086***	-0.020***	-0.038***
	(0.004)	(0.007)	(0.006)	(0.003)	(0.006)
Observations	750	600	700	750	700
R-squared	0.126	0	0.236	0.086	0.093
Avg. Local Emp. Share	0.0709	0.0104	0.0186	0.237	0.0217
			Panel B		
	Information	Finance	Real Estate	PST Services	AWR Services
	(1)	(2)	(3)	(4)	(5)
Log total labor income	-0.003	-0.005**	0.024***	-0.032***	-0.050***
	(0.004)	(0.003)	(0.005)	(0.004)	(0.005)
Observations	750	750	700	750	700
R-squared	0.001	0.007	0.048	0.141	0.171
Avg. Local Emp. Share	0.0177	0.0409	0.0167	0.0267	0.0527
			Panel C		
	Education Services	Healthcare	Arts, Entertainment and Recreation	Accommodation and Food Services	Other Services
	(1)	(2)	(3)	(4)	(5)
Log total labor income	-0.034***	0.013***	-0.015***	-0.014***	-0.027***
	(0.006)	(0.002)	(0.003)	(0.003)	(0.004)
Observations	550	750	650	750	750
R-squared	0.05	0.078	0.028	0.056	0.08
Avg. Local Emp. Share	0.00374	0.282	0.0107	0.146	0.0396

Notes: Table F5 displays the average elasticity of commuting zone aggregate markup and commuting zone size. Commuting zone aggregate markup is defined as (45) and commuting zone size is defined as total labor income. Dependent variable in all columns and panels is log commuting zone aggregate markup. Commuting zone aggregate markup is computed using establishments in local industries within the specific sector. Sectors are defined as 2-digits NAICS sectors. PST Services: Professional, Scientific, and Technical Services. AWR Services: Administrative and Support and Waste Management and Remediation Services. Average local employment share is the average employment share of the sector across commuting zones among the total employment of local industries. Robust standard errors in parenthesis. *10% level, **5% level, ***1% level.

F.4 Robustness Using the Census of Manufactures

This section presents the robustness exercises for the empirical analysis section. To perform these exercises, I use Manufacturing as it is the only sector with detailed data on different inputs of production other than labor.

F.4.1 General Production Function

I estimate markups for Manufacturing using the equation (E8) derived in Appendix E.5. In particular, I use materials and energy as flexible inputs. As highlighted by Yeh, Macaluso, and Hershbein (2022), when there is labor market power, the markdown firms charge in the labor market appears in the first-order condition for labor. Arguably, the materials and energy markets are such that firms do not have any market power in those input markets. Moreover, the data in Manufacturing also allow me to relax the assumption of constant output elasticities. Then, I follow Foster, Grim, and Haltiwanger (2016) and construct measures of labor, materials, energy, and capital usage at the establishment level. Equipped with these measures, I estimate markups using (E8), where a third-order polynomial in labor, materials, energy, and capital approximates the elasticities for materials and energy.

Estimating markups for Manufacturing using (E8) then serves for two robustness checks: 1) using an input for which producers do not have input market power, and 2) considering a general production function for which output elasticities are not constant and does not necessarily exhibit constant returns to scale.

On the one hand, Table F6 displays the results of regression between the baseline manufacturing markups estimated using (43) and the estimates using (E8). The baseline markups are estimated pooling all establishments in local industries while the estimates using (E8) only include establishments in local manufacturing industries. Moreover, I consider two definitions of a city: a commuting zone and a Commuting Zone. Columns indicate the baseline markups, whereas rows indicate the alternative estimates.

The results show that baseline markups highly correlate with the alternative estimates. Strikingly, the regression coefficients are close to one, suggesting that the baseline and the alternative markups move almost one-for-one. Nonetheless, the constant terms are positive and statistically significant than zero in all columns, suggesting that the baseline markups exhibit slightly higher levels than the alternative estimates.

Table F6: Regressions baseline markups and extensions for Manufacturing

	Baseline County	Baseline CZ	Baseline County	Baseline CZ
	(1)	(2)	(3)	(4)
Flex. PF, Energy (County)	1.036***			
	(0.004)			
Flex. PF, Energy (CZ)		1.022***		
		(0.005)		
Flex. PF, Materials (County)			1.075***	
			(0.005)	
Flex. PF, Materials (CZ)				1.061***
				(0.005)
Constant	0.341***	0.281***	0.304***	0.243***
	(0.009)	(0.007)	(0.009)	(0.007)
Observations	27500	27500	27500	27500
R-squared	0.745	0.782	0.738	0.778

Notes: Table F6 displays coefficients of a regression between the baseline markups and the alternative markup estimates for Manufacturing. The baseline markups are the ones estimated by equation (43) and pool establishments in all local industries. Alternative markup estimates are estimated using (E8) and consider only establishments in local manufacturing industries. Columns indicate the baseline markups and rows indicate the alternative estimates. Different columns and rows consider two definitions of a city: commuting zone and Commuting Zone (CZ). Robust standard errors in parenthesis. *10% level, **5% level, ***1% level.

F.4.2 Cost-Shares Approach

As an alternative procedure, I compute markups for local producers in Manufacturing by estimating the labor output elasticity from the firm's cost minimization problem. As discussed in De Loecker and Syverson (2021), under the assumption that producers operate under a constant return to scale technology, the output labor output elasticity for establishment j in city c and industry n is given by:

$$\beta_{jc} = \frac{W_{jc}\ell_{jc}}{W_{jc}\ell_{jc} + W^{m}_{n(j)}m_{jc} + W^{e}_{n(j)}e_{jc} + W^{k,e}_{n(j)}k^{e}_{jc} + W^{k,s}_{n(j)}k^{s}_{jc}}, \tag{F1}$$

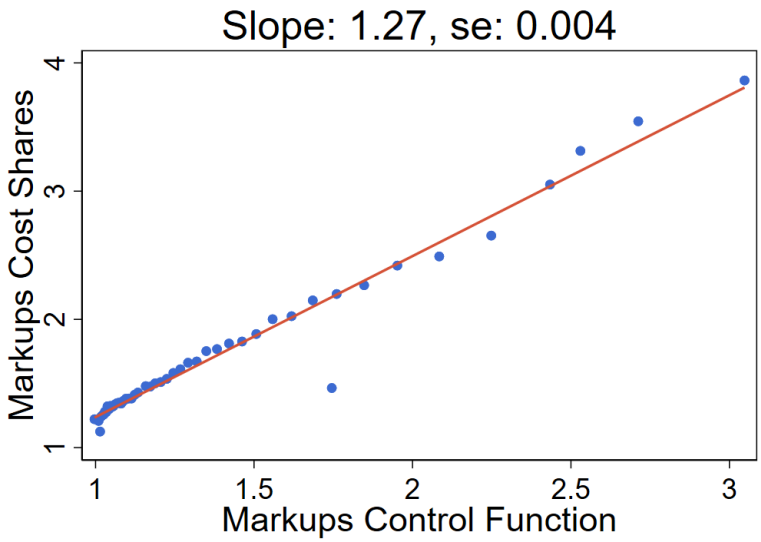
where similarly to Foster, Haltiwanger, and Tuttle (2022), I consider five inputs in production: labor, ℓ_{jc}, materials, m_{jc}, energy, e_{jc}, capital equipment, k^{e}_{jc}, and capital structures (buildings), k^{s}_{jc}. Wages, w_{jc}, are directly observed in the data. To construct industry-level prices for materials and energy I follow the methodology of Foster, Haltiwanger, and Tuttle (2022) described in Appendix B. For the capital cost, I follow Foster, Grim, and Haltiwanger (2016) and construct rental prices for equipment, $W^{k,e}_{n(j)}$, and structures, $W^{k,s}_{n(j)}$ as:

$$W^{k,e}_{n(j)} = \frac{EQKY_{n(j)}}{EQPK_{n(j)} \times EQKC_{n(j)}}, \qquad W^{k,s}_{n(j)} = \frac{STKY_{n(j)}}{STPK_{n(j)} \times STKC_{n(j)}},$$

where $EQKY_{n(j)}$ and $STKY_{n(j)}$ denote capital income for equipment and structures, $EQPK_{n(j)}$ and $STRPK_{n(j)}$ denote productive capital stock, and $EQKC_{n(j)}$, $STKC_{n(j)}$ denote capital composition, for equipment and structures respectively.

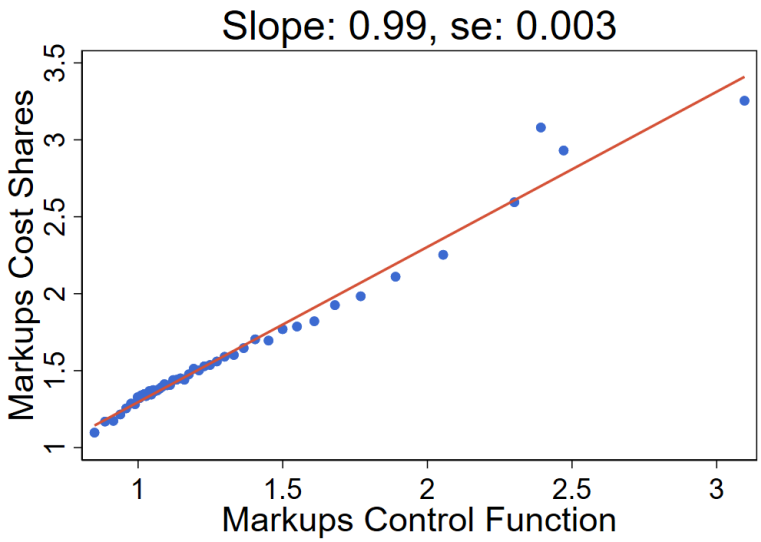
To reduce the role of measurement error, I estimate the labor output elasticity by averaging the expression given in (F1) within each 2-digit and 4-digit NAICS sector. This also follows from the fact that materials, energy, and capital costs are estimated at the sector level. I allow for these elasticities to vary over time by constructing them for every Census year. Then, after estimating these elasticities for every Census year, I compute markups by plugging them into (40). Figures F8 and F9 display the correlation between markups estimated using the control function approach in (43) and markups via cost-shares with the labor output elasticity, β, estimated within 2-digit and 4-digit NAICS industries, respectively. Both figures represent a bin-scatter with 50 equally-sized bins according to the establishment-level control function markup. For each bin, the figures display the average control function markups and the average cost-shares markup. Markups estimated via cost-shares with output elasticity at the 2-digits NAICS industry exhibit slightly higher levels but highly correlate with the control function markup. In contrast, markups estimated via cost-shares with output elasticity at the 4-digits NAICS industry exhibit levels that are close to the ones of the control function markups. Moreover, there is a correlation of one between both markup estimates.

Figure F8: Control Function and Cost-Shares based Markups (β at 2-digits NAICS industry)



Notes: Figure F8 shows a bin-scatter of the control function markups obtained from (43) and the cost-shares markups obtained from (F1) by averaging within 2-digit NAICS industries. The bin-scatter considers 50 equally sized establishment bins according to their control function markup.

Figure F9: Control Function and Cost-Shares based Markups (β at 4-digits NAICS industry)



Notes: Figure F9 shows a bin-scatter of the control function markups obtained from (43) and the cost-shares markups obtained from (F1) by averaging within 4-digit NAICS industries. The bin-scatter considers 50 equally sized establishment bins according to their control function markup.

G Quantitative Appendix

This section provides additional details and derivations for the quantitative analysis in Section 5.

G.1 GMM Estimation

This section provides the details of the GMM estimation in Section 5. First, the Klenow and Willis (2016) Kimball specification implies the following relationship between establishment j markup in commuting zone c, μ_{jc}, and the establishments sales share, s^y_{jc}:

$$\frac{1}{\mu_{jc}} + \log\left(1 - \frac{1}{\mu_{jc}}\right) = \frac{\overline{\sigma}-1}{\overline{\sigma}} - \log\overline{\sigma} + \frac{\varepsilon}{\overline{\sigma}}\log\frac{\overline{\sigma}}{\overline{\sigma}-1} - \frac{\varepsilon}{\overline{\sigma}}\log\frac{\mathbb{P}_c}{\mathbb{D}_c} + \frac{\varepsilon}{\overline{\sigma}}\log s^y_{jc}, \tag{G1}$$

Using the estimated markups, $\hat{\mu}_{jc}$, I estimate the following equation via OLS:

$$\frac{1}{\hat{\mu}_{jc}} + \log\left(1 - \frac{1}{\hat{\mu}_{jc}}\right) = \varpi + \varpi_c + \frac{\varepsilon}{\overline{\sigma}}\log s^y_{jc} + \iota_{jc}, \tag{G2}$$

where ϖ is a constant absorbing the constant terms in (G1), ϖ_c is a commuting zone fixed-effect absorbing the term $(\epsilon/\overline{\sigma})\log(\mathbb{P}_c/\mathbb{D}_c)$, and ι_{jc} is an approximation error coming from the estimation of the markups. The regression coefficient of $\log s^y_{jc}$ is an estimate of the ratio $\varepsilon/\overline{\sigma}$.

Moreover, the Klenow and Willis (2016) also implies the following system of equations in the relative quantities y_{jc}/Y_c and the ratio of price indices $\mathbb{P}_c/\mathbb{D}_c$[44]

$$\begin{aligned}
\frac{y_{jc}}{Y_c} &= \left[-\overline{\sigma}\Omega\left(-\left(s^y_{jc}\frac{\mathbb{P}_c}{\mathbb{D}_c}\frac{\overline{\sigma}}{\overline{\sigma}-1}\right)^{\frac{\epsilon}{\overline{\sigma}}}\frac{\exp\left(-\frac{1}{\overline{\sigma}}\right)}{\overline{\sigma}}\right)\right]^{\frac{\overline{\sigma}}{\epsilon}}, \\
\frac{\mathbb{P}_c}{\mathbb{D}_c} &= \sum_{i\in c}\frac{\overline{\sigma}-1}{\overline{\sigma}}\exp\left(\frac{1-\frac{y_{jc}}{Y_c}^{\frac{\varepsilon}{\overline{\sigma}}}}{\varepsilon}\right)\frac{y_{jc}}{Y_c}.
\end{aligned}$$

With the $\varepsilon/\overline{\sigma}$ estimate, the data on sales share, and a given value of $\overline{\sigma}$, the above system gives relative quantities $\frac{y_{jc}}{Y_c}(\overline{\sigma})$. With the implied relative quantities, I compute the implied markups:

$$\breve{\mu}_{jc}(\overline{\sigma}) = \frac{1}{1 - \frac{1}{\overline{\sigma}}\frac{y_{jc}}{Y_c}(\overline{\sigma})^{\frac{\varepsilon}{\overline{\sigma}}}}$$

Then, I estimate $\overline{\sigma}$ such that we minimize the distance between the predicted and the estimated markups:

[44] One can also work directly with relative prices instead with relative quantities. However, using relative quantities simplifies the system of equations.

$$\widehat{\overline{\sigma}} = \underset{\overline{\sigma}}{\operatorname{argmin}} \left\| \breve{\mu}_{jc}(\overline{\sigma}) - \hat{\mu}_{jc} \right\|$$

G.2 Estimation of City Fundamentals

Within each iteration of the SMM estimation, I invert the model to recover local productivities and amenities. Given a set of parameters, I solve the decentralized equilibrium over a grid for the appeal index x. I then use the zero-profit condition for traded-good producers in (D9) to recover traded-good productivity

$$a(c) = \varrho \, W(c)^{\gamma} \, R\left(x(c)\right)^{1-\gamma},$$

where $W(c)$ is data on the commuting zone average wages and $R(x)$ is the model-implied housing-structure rent, evaluated at the city's appeal $x(c)$ defined in (37). Similarly, I recover local amenities from the labor supply condition (D4). Substituting the ideal price index (D3) into this condition yields

$$b(c) = \nu^{\eta} \, \overline{U} \, \frac{L(c)^{\frac{1}{\theta}} \, R\left(x(c)\right)^{\alpha+\eta(1-\beta)}}{(1+s) \, W(c)^{1-\eta\beta} \left[\mathbb{C}\left(x(c)\right) \overline{\delta}\left(x(c)\right)\right]^{\eta}},$$

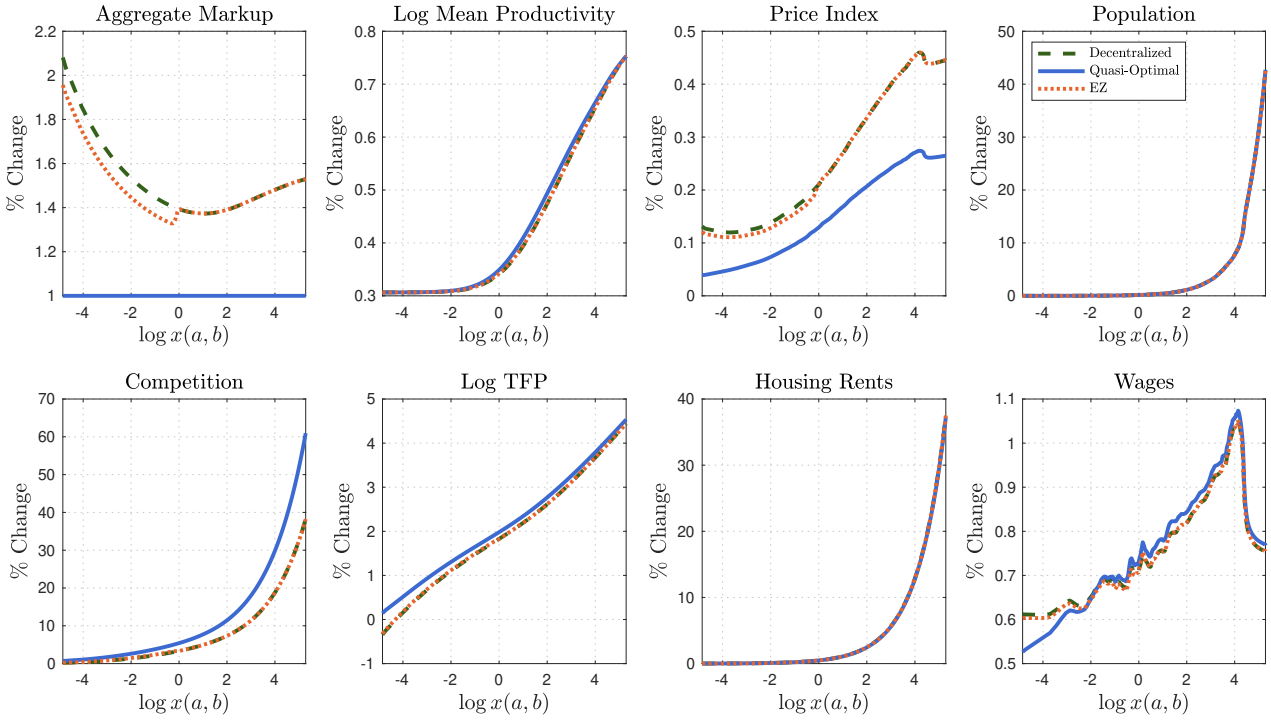
where $L(c)$ is data on commuting zones population, expressed as a share of the total workforce.

Because the appeal index depends on the recovered fundamentals, the expressions above define a fixed-point problem. Given (a, b), I construct the grid for x and solve the decentralized equilibrium. Given the resulting equilibrium objects, I then update (a, b). I iterate this mapping until convergence.

The inversion recovers city fundamentals up to two scale normalizations. I normalize local productivity by its geometric mean because the units in which wages are measured provide one degree of freedom. Similarly, I normalize local amenities by their geometric mean because the units in which $\overline{U}$ is measured provide a second degree of freedom. Under these normalizations, the inversion matches population shares exactly and wages up to a common scale factor.

G.3 Additional Figures

Figure G1: Equilibrium objects in levels under the Quasi-Optimal Policy and EZ Program



Notes: Figure G1 shows equilibrium outcomes in levels under the decentralized equilibrium, the quasi-optimal policy, and the EZ program. The x-axis represents the city's estimated appeal $x(a,b)$. Mean firm productivity and local TFP are reported in logs.