Output Market Power and Spatial Misallocation*

Santiago Franco
The University of Chicago

November 27, 2023

Updated regularly. Click here for the latest version.

Abstract

Most product industries are local. In the U.S., firms selling goods and services to local consumers account for half of total sales and more than sixty percent of jobs. Competition in these industries occurs in local product markets: cities. I propose a theory of such competition in which firms have output market power. Spatial differences in local competition arise endogenously due to the spatial sorting of heterogeneous firms. The ability to charge higher markups induces more productive firms to overvalue locating in larger cities, leading to a misallocation of firms across space. The optimal policy eliminates markups and incentivizes productive firms to relocate to smaller cities, providing a rationale for commonly used place-based policies. I use U.S. Census establishment-level data to estimate markups and to estimate the parameters of the model. I document a significant heterogeneity in markups for local industries across U.S. cities. Cities in the top decile of the city-size distribution have a fifty percent lower markup than cities in the bottom decile. I use the estimated model to quantify the general equilibrium effects of place-based policies. Policies that remove markups and relocate firms to smaller cities yield sizable aggregate welfare gains.

[*] Email: sfranco@uchicago.edu. Website: www.santiago-franco.com. I am extremely grateful to my advisors Esteban Rossi-Hansberg, Ufuk Akcigit, Chang-Tai Hsieh, and Erik Hurst. I also thank Adrien Bilal, Jonathan Dingel, and Simon Mongey for insightful discussions since early stages of the project. This paper has also benefited from discussions with Rodrigo Adao, Milena Almagro, Fernando Alvarez, Olivia Bordeu, Arshia Hashemi, Agustin Gutierrez, Greg Kaplan, Benny Kleinman, Pete Klenow, Santiago Lacouture, Marco Loseto, Sreyas Mahadevan, Hannes Malmberg, Kiminori Matsuyama, Pauline Mourot, Jose M. Quintero, Pascual Restrepo, Marcos Sora, and Chad Syverson. Any views expressed are those of the authors 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 total sales and generate 60% of the nation's total jobs.[1] Firms in these industries operate in local output markets: cities. For example, restaurants or retail stores in New York do not compete with similar establishments in Chicago or Seattle. As such, competition in these industries varies at the local level. The location choice of producers is one key determinant of the strength of local competition: some cities attract productive firms that set low prices, while other cities hardly attract any producers, leading to slack competition. When firms in these industries exert output market power, the degree of local competition becomes a crucial determinant 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 three contributions. First, I develop a model in which spatial differences in local competition arise endogenously due the location choices of heterogeneous firms charging variable markups. The equilibrium allocation is inefficient as more productive firms over-concentrate in bigger cities. Second, I use the model and U.S. Census micro-data to estimate markups for establishments in local industries. I document significant heterogeneity in markups across U.S. cities, with cities in the top decile of city-size distribution having a markup 50% lower than cities in the bottom decile. Third, I estimate the model and use it to quantify the welfare effects of the optimal policy. A utilitarian social planner chooses a policy that eliminates markups by inducing firms to produce at marginal cost and incentivizes productive firms to relocate from larger to smaller cities, providing a rationale for commonly used place-based policies.

In the model, production takes place in locations that I call cities. These locations are populated by freely mobile workers with idiosyncratic location tastes who consume local goods, housing, and a traded good. Cities differ ex-ante along two dimensions. First, they differ in local amenities, which affect workers' utility of residing in a specific city. Second, they differ in local productivity, which determines efficiency in producing the traded good. This good is homogeneous, produced by perfectly competitive producers in each city, and freely traded.

At the core of the model lies the location choice of heterogeneous local goods producers. Potential entrants pay an entry cost to learn their productivity and then chooses a city to operate. Once located in a city, they produce a differentiated variety using labor and commercial structures and compete monopolistically with other local producers.[2] Workers have Kimball preferences (Kimball, 1995) over the local varieties of their city, allowing each firm's price elasticity to vary with its position on its residual demand curve, resulting in heterogeneous markups. More productive firms charge lower prices and exhibit higher markups.[3]

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

[2] Traded good producers also have a technology that uses labor and commercial structures.

[3] Alternative frameworks with heterogeneous markups are models of oligopolistic competition, similar to Atkeson and Burstein (2008). More productive firms are larger and have lower demand elasticity in such models, resulting in higher markups. Kimball preferences within monopolistic competition capture this characteristic parsimoniously, and the main forces of my model can seamlessly be extended to oligopolistic competitive environments.

Local producers consider two endogenous city characteristics when choosing where to locate: city size and local competition. The size of a city is determined by the total income of workers, which, in turn, influences potential sales. Local competition is influenced by other producers' prices and the cost of production inputs: wages and land rents. The level of local competition affects the markup a firm can charge in two ways. First, consumers become more price-sensitive when other local producers charge low prices, leading to lower markups. Second, because firms experience incomplete costs-to-price pass-through, higher input costs also reduce markups. In equilibrium, larger locations are more competitive. Therefore, firm location decisions are driven by a trade-off between sales gains and markup reductions. However, more productive firms value relatively more producing in the largest cities. Due to their ability to charge higher markups, they benefit more from the increased sales opportunities in larger cities. This results in positive assortative matching: more productive firms choose to locate in larger cities where local competition is more intense.

Two opposite forces determine differences in market power across cities. First, a 'competition force' pushes markups down in bigger cities. As larger markets are endogenously more competitive, firms in those markets must charge lower markups. Second, bigger markets attract more productive firms that charge higher markups. This 'selection force' pushes markups up in bigger cities. The relative strength of these two forces determines whether bigger cities have lower or higher markups. When firms' productivity dispersion is low, the competition force dominates: big cities attract firms of similar productivity who compete in prices by charging lower markups. In contrast, when productivity dispersion is high, big cities attract producers with distinct productivities, allowing the most productive among them to face milder competition and charge higher markups.

The spatial equilibrium allocation is inefficient due to two externalities arising from firms' entry into a city. Introducing a new variety raises consumer surplus, but firms can only partially appropriate this gain, resulting in insufficient entry—a 'variety gains' externality. Simultaneously, firms impose a 'business-stealing' externality on incumbents by reducing the consumption of existing varieties, leading to excessive entry as firms do not internalize their negative impact on others' profits. The dominance of these externalities determines whether entry is excessive, insufficient, or efficient.[4] The interplay between variety gains and business-stealing externalities causes productive firms to over-concentrate in larger cities. While the gains from additional varieties are higher in smaller cities, the business-stealing effect is more pronounced in larger cities. This imbalance results in excessive entry in larger cities and insufficient entry in smaller ones, leading to spatial misallocation. Relocating productive firms from larger to smaller markets enhances aggregate welfare.

The extent of spatial misallocation is mitigated when larger cities exhibit smaller markups. In equilibrium, there is less competition in smaller places, and profits are higher in bigger places. These two factors make the variety gains externality larger in smaller cities and the business stealing effect larger in bigger cities. Consequently, spatial misallocation persists regardless of markup variations across locations. However, if markups are lower in larger cities, the impact of the 'business stealing' effect diminishes, preventing some productive firms from favoring the largest

[4] These externalities exist in models of free entry and differentiated varieties and are not specific to my framework. However, Dixit and Stiglitz (1977) show that in standard models of monopolistic competition with CES preferences, these externalities remain constant and precisely offset each other, resulting in efficient entry.

locations. In essence, lower markups in larger cities curtail firms' profits, prompting marginally productive firms to opt for smaller locations, thereby reducing spatial misallocation.

In the second part of the paper, I examine the positive implications of the model. To do so, I estimate markups for U.S. local producers. Since physical output and total input costs are not directly observed, I employ an empirical strategy to estimate markups by combining the demand system with the firm's production function similar to De Loecker (2011). I begin with the firm's first-order condition for labor, linking the labor cost share of sales to the markup and the labor output elasticity, following the approach in De Loecker and Warzynski (2012). The demand system implies that firm's markup is a function of its local sales share. I use this relationship to construct a non-parametric control function. By substituting this function into the firm's first-order condition, I jointly estimate the labor output elasticity and the markup. The within-city variation in the sales market share and the labor cost share of sales serves to identify both objects.[5]

I implement the proposed empirical strategy using U.S. Census micro-data.[6] I use data from the Longitudinal Business Database (LBD) to construct establishments' labor cost-shares of sales and data from the Economic Censuses (EC) to construct establishments' local sales shares. For my baseline 2017 sample, I estimate markups for five million local establishments. To corroborate my empirical strategy, I estimate markups for Manufacturing using the ratio estimator of De Loecker and Warzynski (2012), estimating the labor output elasticity using cost-shares as in De Loecker, Eeckhout, and Unger (2020) and Edmond, Midrigan, and Xu (2023). Both markup estimates exhibit similar levels and show a correlation close to one. Subsequently, for constructing the city-level aggregate markup, I adopt the aggregation method implied by my model, which is a sales-weighted harmonic mean of the establishment-level markups.

I find significant heterogeneity in markups for local industries across U.S. cities. Cities in the top decile of the city-size distribution have a 50% lower markup than cities in the bottom decile. This finding implies that the competition force outweighs the selection force in U.S. local industries. Furthermore, it suggests that the competitiveness of larger cities contributes to the reduction of spatial misallocation. This empirical regularity holds consistently across various years, and remains robust, whether defining a city as a county or as a Commuting Zone.

I also uncover heterogeneity in the spatial distribution of markups across sectors. Expanding the baseline empirical strategy, I incorporate variations in labor output elasticity and consumer demand across sixteen 2-digit NAICS sectors. Notably, bigger cities exhibit lower markups in nine out of the sixteen sectors. For local Retail, cities in the top decile of the city-size distribution have a 40% lower markup than cities in the bottom decile. In contrast, for local Manufacturing, cities in the top decile have a 60% higher markup than those in the bottom decile. These empirical findings suggest significant variation in the determinants of the competition and selection forces across economic sectors. Moreover, they imply that the degree of firm spatial misallocation may be more pronounced in sectors where producers in larger cities command higher markups.

[5] If firms have labor market power, the markdown appears in the firm's first-order condition. To address this issue, I control for a flexible polynomial in the firm's wage bill share. In models of labor market power like Berger, Herkenhoff, and Mongey (2022) and Trottner (2023), the markdown is a function of the firm's wage bill share.

[6] To classify establishments as local, I rely on the categorization by Delgado, Porter, and Stern (2015).

In the final section of the paper, I estimate the remaining parameters of the model to study its normative implications. Firstly, the model generates estimating equations that link demand parameters, firm markups, and sales shares. I use the estimated markups to identify the Kimball demand parameters from these equations. Secondly, I employ a Simulated Method of Moments (SMM) approach to estimate the local producer's productivity distribution and the aggregate entry cost. Leveraging the model's structure, I infer the productivity distribution parameters from the average employment per establishment across cities and determine the entry cost from the economy-wide aggregate markup. Location productivity and amenities are estimated by precisely matching population and average wages for each city. The estimated model replicates the negative relationship between markups and city size, which constitutes a non-targeted moment.

I utilize the estimated model to conduct a counterfactual exercise assessing the impact of the optimal policy. I show that the optimal policy involves a non-linear, location-specific subsidy contingent on the firm's total production, financed by a non-distortionary flat labor tax. This subsidy effectively eliminates markups by incentivizing firms to produce at marginal cost. After transfers, net profits align with the consumer surplus each firm generates, correcting variety gains and business-stealing externalities and leading to an efficient firm location. Marginally productive producers in larger cities find it optimal to relocate to smaller cities, where they can earn higher profits by generating a greater consumer surplus.

The optimal policy lowers the prices of local varieties in all cities. Because the policy eliminates markups, all local producers reduce their prices. Nevertheless, since markups were initially higher in smaller cities, these locations experience more significant price reductions than their larger counterparts. For example, prices for local producers in Franklin, FL, drop by 30%, while prices for local producers in Richmond, VA, drop by 10%. Moreover, as smaller cities also witness an influx of new producers, the local variety price index experiences further reductions. The price index in cities in the bottom quartile of the city-size distribution drops by 65%, whereas it only decreases by 30% for cities in the top quartile.

The policy also has notable implications for the spatial distribution of local producers. On one hand, the policy is effective in reallocating productive establishments from big to smaller cities. Mid-sized cities experience the largest productivity improvements, with a 15% increase in the productivity of local producers. Intuitively, marginally productive producers in the biggest markets like Los Angeles move to smaller suburban areas like Santa Clara. On the other hand, once one accounts for net entry, smaller cities undergo an even more substantial productivity surge, with Total Factor Productivity (TFP) increasing by 30% in locations like Wilcox County, GA. This, however, comes at the cost of slight reductions in TFP in larger cities as productive firms relocate. For instance, places like Manhattan experience a 5% TFP reduction.

The aggregate welfare gains of the policy are more modest than the reduction in prices and the gains in productivity. Due to the reduction in markups, producers increase the usage of commercial structures, leading to an increase in housing rents across all cities. Furthermore, the policy is costly, with workers contributing 15% of their income to finance the producer's subsidies. These two effects counterbalance the gains from lower prices and a better allocation of producers across space. All in all, the policy enhances aggregate worker welfare by 2.35%.

Related literature. This paper contributes to five strands of the literature. The first strand it relates to is the one studying competition in local output markets. Studies by Hsieh and Rossi-Hansberg (2023), Oberfield et al. (2023) , and Kleinman (2023) explore the competition of multi-establishment firms across regions, where consumers have CES preferences, resulting in constant markups for all firms. I enhance this literature by introducing a framework in which firms have endogenous variable markups. This innovation enables a study of richer local pricing dynamics, introducing an additional force influencing firm location. However, I abstract from the combinatorial problem of a firm opening branches in different locations. Additionally, 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 study by providing direct empirical measures of output market power across local markets.

The second strand is the work on firm sorting. Gaubert (2018) and Bilal (2023) investigate firm sorting through agglomeration forces and labor market frictions. I diverge from these studies by focusing on how firms sort through competitive price pressures. The policy implications of my framework align with those in Bilal (2023), who finds policies relocating firms to smaller locations beneficial. Additionally, Nocke (2006), Combes et al. (2012), and Matsuyama and Ushchev (2022) also consider settings in which firms sort through competitive price pressures, but local population is exogenous. In contrast, local population is endogenously determined in my model, enabling me to investigate additional general equilibrium implications of firm location decisions.

The third strand is the recent body of work examining the aggregate implications of markups. Edmond, Midrigan, and Xu (2023) quantify the aggregate welfare cost of markups using a model encompassing monopolistic competition with Kimball preferences and oligopolistic competition with nested-CES. I adopt their monopolistic competition market structure and extend it into a spatial framework with many output markets. However, I depart from their work by quantifying the welfare costs of markups through a new channel: the inefficient location of firms. Other studies, such as Peters (2020), De Loecker, Eeckhout, and Mongey (2022), Aghion et al. (2023), and Akcigit and Ates (2023), have analyzed the implications of markups for business dynamism. In contrast, my focus is on the effect of markups on regional aggregates, such as productivity and prices.

The fourth strand of the literature to which this paper relates is the one on misallocation. Similar to Edmond, Midrigan, and Xu (2023), in my model, more productive firms charge higher markups, creating misallocation in the form of Restuccia and Rogerson (2008) and Hsieh and Klenow (2009). However, rather than concentrating on misallocation across firms, my emphasis is on misallocation across cities. Specifically, I demonstrate how heterogeneous markups lead firms to locate inefficiently across cities, and how aggregate welfare increases by relocating firms across places.

The final strand of the literature to which I contribute is the one studying markups across space. Hottman (2021) and Anderson, Rebelo, and Wong (2018) focus on markups across cities within the Retail sector. On one hand, Hottman (2021) builds an oligopolistic competition model and estimates it using scanner data, finding that markups are lower in more populous cities, a result I also observe. On the other hand, Anderson, Rebelo, and Wong (2018) use gross margins as a proxy for markups and demonstrate that wealthier cities have higher markup. I depart from these studies in two ways. First, my markup estimation extends beyond the Retail sector to encompass all local industries in the U.S. Second, I develop a general equilibrium model that utilizes the estimated markups as inputs to conduct counterfactual analysis.

The rest of the paper is organized as follows. Section 2 lays out the theoretical framework. Section 3 explores the efficiency properties of the model, emphasizing the spatial misallocation of firms. 4 presents the empirical analysis investigating markups across cities and model predictions. Section 5 details the quantitative analysis in which I estimate the model and quantify the welfare gains of place-based policies. Finally, Section 6 concludes.

2 Model

This section develops a theory of spatial differentials in local competition, where production takes place in locations I call cities. The theory abstracts from dynamics, describing a long-run steady state of the economy.

2.1 Environment

Geography. There is a continuum of cities indexed by $c \in [0, 1]$, that differ in their local productivity, $a(c) \in [\underline{a}, \overline{a}]$, and their local amenities $b(c) \in [\underline{b}, \overline{b}]$. These characteristics are distributed with a cumulative distribution function $F(c) \equiv F(a(c), b(c))$, with density $f(c) \equiv f(a(c), b(c))$.

Workers Preferences. The whole economy is populated by $\overline{L}$ freely mobile identical workers, indexed by i. Each worker has one unit of labor, which is supplied inelastically. Worker i observes a collection of idiosyncratic location-specific preference shocks, $\varsigma_i(c)$, and decides her location of work and residence. When locating in c, worker i derives utility from consuming a bundle of local varieties $Y(c)$, housing, $H(c)$, and a freely traded good, $Q(c)$, according to:

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

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

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

where $y(z, c)$ is the per-capita consumption of a local variety produced by a firm with productivity z, $G_c(\cdot)$ is local producers productivity distribution in c, and $\Upsilon(\cdot)$ is a strictly increasing and concave function satisfying $\Upsilon(0) = 0$.[7] CES preferences are special case of (2) when $\Upsilon(x) = x^{\frac{\sigma-1}{\sigma}}$.

[7] This notation previews that in equilibrium, two firms of the same productivity and located in the same city

Kimball preferences have three advantages. First, they can generate cross-sectional variation in markups, the central object of this paper. Second, they are homothetic. Therefore, they allow us to focus on markup differences across cities due to price competition pressures and not from potential income effects.[8] Third, despite their flexibility, they remain tractable enough to characterize the model's equilibrium uniquely. In models with similar preferences like nested CES (as in Atkeson and Burstein (2008)), problems of multiple equilibria often arise. This becomes more challenging when producers have an entry decision per market, as is the case in this study.

The idiosyncratic preferences draw $\varsigma_i(c)$ is assumed to be independent, identically distributed across individuals and cities, and following a Frechet distribution with shape parameter θ.[9]

Local Varieties. A mass M_e of potential entrants pays an entry cost c_e to learn their productivity z. This productivity has common distribution with cumulative distribution function $G(\cdot)$, density function $g_(\cdot)$, and connected support $[\underline{z}, \overline{z}]$.[10] After learning their productivity, firms choose a city to produce and sell. This location decision determines the set of locally available varieties. Within a city, local producers compete in a monopolistically competitive fashion and produce according to a Cobb-Douglas production function

$$y(z, c) = z\ l(z, c)^{\beta} s(z, c)^{1-\beta}, \tag{3}$$

where $l(z, c)$ is labor, and $s(z, c)$ is commercial structures (buildings). Firms pay a common wage of $W(c)$ and commercial structures rent $R(c)$ in city c.

Traded Good. A perfectly competitive representative firm produces the homogeneous traded good in every location. This good is freely traded and used as the *numeraire*. Similarly to the local varieties technology, the traded good is produced by combining labor and commercial structure according to a Cobb-Douglas production function given by

$$Q^T(c) = a(c)\ \left(L^T(c)\right)^{\gamma} \left(S^T(c)\right)^{1-\gamma}, \tag{4}$$

where $Q^T(c)$ denotes the total production of traded good in city c, $L^T(c)$ is the traded good total employment, and $S^T(c)$ is the traded good total commercial structures demand. We use different notation for the traded good quantities produced in c, $Q^T(c)$, and the traded good workers demand, $Q(c)$. Because of trade across cities, these two quantities are different. Traded good producers compete in the same local inputs market with the local varieties producers, paying a wage $W(c)$ and a price for commercial structures $R(c)$.

make the same pricing and production decisions. Therefore, consumers consume the same amount of each variety produced by each of these firms.

[8] Although this is an important channel to study, it is left for future research. See Melitz and Ottaviano (2008) and Combes et al. (2012) for settings in which firms have endogenous variable markups due to non-homothetic linear preferences.

[9] See Bilal (2023) Appendix G.4. for a discussion of how to extend discrete choice results to a framework with a continuum of locations.

[10] The overall productivity distribution $G(\cdot)$ does not necessarily coincide with the productivity distribution in a given city, $G_c(\cdot)$. The latter is an endogenous object determined by the location choices of the local producers.

Land Developers. In every city, competitive land developers use the traded good to produce housing and commercial structures according to the isoelastic production function:

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

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. I assume that land developers use their profits to consume the final good only. However, for the counterfactual exercises, I consider an alternative formulation in which land developers' profits are aggregated into a national portfolio and rebated back to workers as a flat labor subsidy.[11]

2.2 Worker's Consumption and Location Decisions

We start by characterizing the workers' optimal consumption and location decisions. Workers solve this problem in two steps: first, conditional on locating in c, they solve for the optimal consumption quantities, which determines local utility. Then, they choose where to locate, conditional on local utility and the realization of their preference shocks.

When choosing how much of the local varieties, housing, and traded good to consume, workers face the budget constraint

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

where $\mathbb{P}(c)$ is the price of the bundle of local varieties, $R(c)$ is the housing price, and where we used the fact that the traded good is used as the numeraire.[12] The homotheticity of the 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 that result from this maximization are given by

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

Appendix A.1 shows that the per-capita consumption of an individual variety $y(z,c)$ is, in turn

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

[11] This can be interpreted as workers having a share in the national portfolio which is increasing in the level of income. See Redding and Rossi-Hansberg (2017) Section 2.7.3 for a general discussion of rebate schemes in quantitative spatial models.

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

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

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

The expression in (7) is the residual demand curve faced by local variety producers. For an individual firm, changes in other firms' prices are summarized by the price index $\mathbb{D}(c)$. In other words, firms in every location compete against the price index $\mathbb{D}(c)$ when choosing their optimal price. Therefore, $\mathbb{D}(c)$ captures the degree of local competition, and I call it the *competition price index*. In contrast, the *ideal price index*, $\mathbb{P}(c)$, which is the price of the bundle of local varieties, is given by

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

Two price indices then characterize the Kimball demand system. The competition price index $\mathbb{D}(c)$ mediates the relative consumption of different varieties, whereas the ideal price index $\mathbb{P}(c)$ determines the consumption of the overall bundle $Y(c)$ relative to other goods. In the particular case of CES, these price indices are proportional.

Workers consumption decisions (6) imply that the indirect utility of worker i in city c is given by

$$U_i(c) = u(c)\varsigma_i(c), \qquad \text{with} \qquad u(c) \equiv b(c)\frac{W(c)}{\mathbb{P}(c)^{\eta}R(c)^{\alpha}}, \tag{10}$$

where $u(c)$ is the mean utility of workers residing in c. After observing the elements of $u(c)$ and the collection of idiosyncratic location preference shocks, $\varsigma_i(c)$, workers choose the location c that maximizes $U_i(c)$. The Frechet assumption implies that the share of workers residing in i is equal to

$$\frac{L(c)}{\overline{L}} = \left(\frac{u(c)}{\overline{U}}\right)^{\theta}, \qquad \text{with} \qquad \overline{U} = \left[\int_c u(c)dF(c)\right]^{\frac{1}{\theta}}. \tag{11}$$

The expression (11) is the supply of workers in city c. When θ is larger, the idiosyncratic preference shocks, $\varsigma_i(c)$, are less dispersed, and therefore cities become closer substitutes. In equilibrium, (11) implies that a higher θ makes workers in c more sensitive to changes in the local utility level, $u(c)$. Note that, all else equal, cities with higher wages and amenities are more desirable for workers. Similarly, cities with lower housing rents and lower local varieties prices attract a higher share of workers.

2.3 Local Varieties Production and Location Decisions

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

2.3.1 Pricing Decision

Local varieties producers in c set their optimal price given their own productivity, z, and location aggregates, $Y(c)$, $\mathbb{D}(c)$, $W(c)$, and $R(c)$.

The production function in (3) implies that the marginal cost for firm z of producing one unit of output in location c is equal to $\nu(W(c)^{\beta}R(c)^{1-\beta})/z$, where ν is a constant depending solely on β.[13] Therefore, firm z located in c chooses $p(z,c)$ to maximize:

$$\Pi(z,c) = \max_{p(z,c)} \left[p(z,c)y(z,c) - \frac{\nu W(c)^{\beta}R(c)^{1-\beta}}{z} y(z,c) \right] L(c) \qquad \text{s.t} \qquad (7). \tag{12}$$

The profit function in (12) scales with the number of workers in c, because larger cities represent a larger customer base. The first-order condition of this problem is given by

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

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

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

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

$$\mathbb{C}(c) \equiv \frac{\nu W(c)^{\beta}R(c)^{1-\beta}}{\mathbb{D}(c)} \tag{15}$$

This index accounts for competition in the local input market and ins th local output market. When there is more intense competition in the input market, wages or commercial structures rents are high; therefore, $\mathbb{C}(c)$ increases. Similarly, when other local producers in location c set lower prices, profits for potential entrants decrease. This is captured by a lower price index $\mathbb{D}(c)$, which ultimately translates into a higher $\mathbb{C}(c)$.

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

[13] Formally, $\nu \equiv \frac{1}{\beta^{\beta}(1-\beta)^{1-\beta}}$.

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

Similar to the frameworks in Atkeson and Burstein (2008) and Amiti, Itskhoki, and Konings (2019), firms "price-to-market" by choosing an optimal price relative to their competitor's prices summarized by the price index $\mathbb{D}(c)$.

The optimal markup is given by the Lerner formula that combines (14) and (16)

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

Equation (17) reveals the forces that determine firms markups. On the one hand, the function $\psi(\cdot)$ is always strictly increasing. Therefore, conditional on firm productivity, an increase in local competition forces firms to charge higher relative prices. On the other hand, conditional on local competition, more productive firms charge lower relative prices. However, whether differences in relative prices due to competition or firm productivity lead to differences in markups depends on the properties of the price-elasticity function $\sigma(\cdot)$. As I discuss later, if $\sigma(\cdot)$ is increasing, then tougher competition reduces the markups for all firms in a city, and more productive firms charge higher markups within a city. In the particular case of CES, $\sigma(\cdot)$ is constant. Therefore, all firms charge the same markup in all locations regardless of the local competition or their productivity.

Similarly, the firm's optimal relative quantity is given by (7) and (16):

$$\frac{y(z,c)}{Y(c)} = \varphi\left(\psi\left(\frac{\mathbb{C}(c)}{z}\right)\right). \tag{18}$$

Finally, Appendix A.2 shows that optimal labor and commercial structures demands are given by:

$$\begin{aligned} l(z,c) &= \beta\frac{\psi\left(\frac{\mathbb{C}(c)}{z}\right)\varphi\left(\psi\left(\frac{\mathbb{C}(c)}{z}\right)\right)\mathbb{D}(c)Y(c)}{\mu\left(\frac{\mathbb{C}(c)}{z}\right)W(c)}, \\ s(z,c) &= (1-\beta)\frac{\psi\left(\frac{\mathbb{C}(c)}{z}\right)\varphi\left(\psi\left(\frac{\mathbb{C}(c)}{z}\right)\right)\mathbb{D}(c)Y(c)}{\mu\left(\frac{\mathbb{C}(c)}{z}\right)R(c)} \end{aligned} \tag{19}$$

Total labor employed by local producers, $L^N(c)$, and total structures demanded by local producers, $S^N(c)$, are given by:

$$L^N(c) = \int_z l(z,c)dG_c(z), \qquad \text{and} \qquad S^N(c) = \int_z s(z,c)dG_c(z) \tag{20}$$

2.3.2 Location Decision

Now we turn to analyze the location decision of the local varieties producers. A producer z contemplates potential profits in every city and chooses the city that delivers the highest profits.

Let $M(c)$ denote the total expenditure of local varieties in location c: $M(c) \equiv \mathbb{P}(c)Y(c)L(c)$. We refer to $M(c)$ as the *size* of the city c as it measures local producers' potential revenue in a particular location.

We can use (16) to write into the firms profits gives firms' z overall potential profits:

$$c^*(z) = \max_c \; \underbrace{\log M(c)}_{\text{Market size}} + \underbrace{\log \frac{\psi\left(\frac{\mathbb{C}(c)}{z}\right)\varphi\left(\psi\left(\frac{\mathbb{C}(c)}{z}\right)\right)}{\mathbb{P}(c)/\mathbb{D}(c)}}_{\text{Market Share}} - \underbrace{\log \frac{\mu\left(\frac{\mathbb{C}(c)}{z}\right)}{\mu\left(\frac{\mathbb{C}(c)}{z}\right)-1}}_{\substack{\text{Fraction of sales}\\ \text{going to inputs}}} \tag{21}$$

Equation (21) reveals three forces that shape the location decision of local varieties producers. The first, market size, reflects the total expenditure of local varieties in the city c. All producers prefer bigger cities because potential revenue is higher.

The second term, market share, indicates how much of the total revenue in a given city each firm can appropriate. The market share of firm z depends on the firm's relative price, $\psi(\mathbb{C}(c)/z)$, and on the demand's price indices, $\mathbb{D}(c)$ and $\mathbb{P}(c)$.[14] Producers charging a lower relative price have a higher market share, which allows them to capture a larger fraction of the total expenditure on local varieties and, therefore, have higher sales. The ratio $\mathbb{P}(c)/\mathbb{D}(c)$ captures the sales of the other local producers in c.[15] Producers prefer locations in which they can have a higher market share.

Finally, the last term (21) encodes how much of the firms' sales are going to their profits and how much goes to pay the inputs of production. In the particular CES case, this last term is constant, which implies that firms' profits are always proportional to firms' sales. Nevertheless, profits are no longer proportional to sales once one departs from CES. Because each firm chooses its optimal markup, how much revenue goes to pay the production inputs varies across producers. In particular, high markup producers turn a larger fraction of the sales into revenue. All else equal, firms value locations in which they can charge higher markups.

2.4 Traded Good Producers

Now, we turn to characterize the traded good producers problem. Recall that these producers are immobile and only make production decisions.[16] The production function (4) implies that the marginal cost of the perfectly competitive producers is $\varrho(W(c)^\gamma R(c)^{1-\gamma})/a(c)$, where ϱ is a

[14] As first recognized by Matsuyama and Ushchev (2017), the distinct characteristic of the HDIA preferences is that the firm's market share is given by the two price indices of the demand system.

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

[16] The retrained mobility of the traded good producers is immaterial for the model predictions. As these producers make zero profits in every location, they would be indifferent between locating among any cities.

constant depending on γ.[17] Therefore, the zero-profit condition in every city implies:

$$a(c) = \varrho W(c)^{\gamma} R(c)^{1-\gamma}. \tag{22}$$

In equilibrium, the cost of production in more productive cities is higher. This translates into higher wages and housing/commercial structures rents in such locations. Note that (22) has implications for the local good producers' production and location decisions. Because traded good and local producers compete in the same labor and housing markets, local producers in more productive cities face tougher competition in the input market, embedded in higher input costs. As seen from (15), tougher competition in the input market leads to a higher competition overall.

Finally, the production function (4) implies that total labor labor demand and total commercial structures demand from traed good producers in city c are given by:

$$L^T(c) = \frac{\gamma Q^T(c)}{W(c)}, \qquad S^T(c) = \frac{(1-\gamma)Q^T(c)}{R(c)}. \tag{23}$$

Because traded good producers earn zero profits in equilibrium, each production input receives a fraction of the total sales given by their output elasticicities.

2.5 Equilibrium Definition

Having laid how workers and firm make their optimal consumption, production, and location decisions, characterize the conditions for the decentralized equilibrium.

First, the land developer's production function leads to an equilibrium buildings supply equal to $R(c)^{\phi}$ and traded good demand of $\phi R(c)^{1+\phi}/(1+\phi)$. Then, local housing and local labor markets clear in every city:

$$R(c)^{\phi} = L(c)H(c) + S^N(c) + S^T(c), \qquad L(c) = L^N(c) + L^T(c), \tag{24}$$

where $H(c)$ is given by (6), $L(c)$ is given by (11), $L^N(c)$ and $S^N(c)$ are given by (19), and $L^T(c)$ and $S^T(c)$ are given by (23). Moreover, the traded good market must also clear. Recall that the traded good is consumed by workers, used to build housing and structures, and used to pay the local producers' entry costs. Hence, the traded good market clearing condition is:

$$\int_c Q^T(c)dF(c) = \int_c \left(L(c)Q(c) + \frac{\phi}{1+\phi} R(c)^{1+\phi} \right) dF(c) + c_e M_e \tag{25}$$

Because of free entry, local producers' expected profits must equal the entry cost. Furthermore, the labor market clears in the aggregate:

[17] $\varrho \equiv \frac{1}{\gamma^{\gamma}(1-\gamma)^{1-\gamma}}$.

$$\int_c \Pi(z, c^*(z))dG(z) = c_e \qquad \int_c L(c)dF(c) = \overline{L} \tag{26}$$

A decentralized equilibrium is comprised of a mass of entering local producers M_e, an ex-ante utility $\overline{U}$, an optimal price function, $\psi(\cdot)$, a location choice function, $c^*(z)$, a local productivity distribution, $G_c(\cdot)$, a wage function, $W(c)$, housing prices, $R(c)$, and population distribution, $L(c)$, such that workers maximize utility given prices, (6), local producers maximize profits given location aggregates, (16), local producers choose optimally where to locate (21) and the local productivity distribution $G_c(\cdot)$ is consistent with these location choices, local labor and housing markets clear, (24), the traded good market clears, (25), local producers make zero profits on average and aggregate labor market clears, (26). I characterize the existence, uniqueness and properties of the equilibrium in the next section.

2.6 Equilibrium Characterization

In equilibrium, cities are characterized by a combined index. Recall that, initially, a city is described by a pair of productivity of the traded good and amenities, a, b. Appendix B.1 shows that:

$$c(a, b) = a^{\frac{1+\theta(1-\eta\beta)}{\gamma}} b^{\theta} \tag{27}$$

is a local sufficient statistic for the model's outcomes. In particular, the city-level objects that determine the firms' location decisions in (21) depend solely on $c(a, b)$. Intuitively, local amenities govern local population (see (11)), and local productivity determines local wages (see (22)). Thus, city size depends on a combined index of these two characteristics. The same logic applies to local competition. Therefore, firms make production and location decisions based on c. Therefore, thorugh the rest of the paper, c denotes the combined index $c(a, b)$ rather than the particular "name" of a city. I call this combined index c the *appeal* of a city.

Formally, denoting cities by their appeal allows us to characterize location choice problem (21). The following expression anticipates that equilibrium conditions involve only continuously differentiable fixed point functionals:

$$\underbrace{\frac{\partial}{\partial c}\left(\log M(c) - \log \frac{\mathbb{P}(c)}{\mathbb{D}(c)}\right)}_{\substack{\Delta \text{ in sales from locating} \\ \text{in more appealing cities}}} = \underbrace{\frac{\mathbb{C}'(c)}{\mathbb{C}(c)} \frac{1}{\mu\left(\frac{\mathbb{C}(c)}{z}\right) - 1}}_{\substack{\Delta \text{ in fraction of sales going} \\ \text{to inputs of production}}} \tag{28}$$

The left-hand-side (LHS) of (28) encodes the benefits of locating in more appealing cities (high c cities). First, through $M(c)$, it captures how the size of cities changes when they become more attractive for firms. Second, $\mathbb{P}(c)/\mathbb{D}(c)$ captures how market share changes as c increases.[18]

[18] Appendix B.1 shows that, to a first-order, when comparing two locations, the firm price changes in each location do not affect its market share. That is, because firms are already choosing their optimal price, the envelope

The sum of these two terms represents the percentual increase in sales from locating in high c locations. Importantly, the LHS of (28) does not depend on the firm's productivity. Therefore, the net increase in sales from locating in more appealing cities is the same for all producers.

The right-hand-side (RHS) of (28) represents the costs of locating in more competitive cities and embeds the primary sorting mechanism. This term reveals that the cost of locating in cities with high competition depends on the local producer's markup. If all firms charge the same markup, the RHS of (28) is not a function of the firm's productivity z. In this scenario, which is the case when worker's have CES preferences over local varieties, the cost of locating in more competitive cities is the same for all producers and the model does not generate any sorting predictions.[19]

Once we allow the price-elasticity $\sigma(\cdot)$ to vary with the firms price, the model delivers stark sorting predictions. If $\sigma(\cdot)$ is an increasing function, within a city, more productive firms charge higher markups. This is often referred to as "Marshall's Second Law of Demand" (MSLD). Intuitively, consumers become less price elastic when facing lower prices. This allows more productive firms, who charge lower relative prices, to have higher markups. I call this complementarity between the firm's productivity and the markup, *pricing complementarities*. Because of the pricing complementarities, the RHS of (28) indicates that the cost of locating in more competitive cities is lower for high-productivity firms. When firms locate in more competitive environments, they turn a lower fraction of their sales into profits because they charge lower markups. However, the fraction of sales firms forgo locating in more competitive environments is smaller for a high z firms as she can charge a higher markup relative to a low z firm.

Of course, the competition of a city, $\mathbb{C}(c)$, is an endogenous object determined in equilibrium. It ultimately depends on the strength of local price and input competition. Whether more appealing cities are more competitive or not is determined in general equilibrium. As formalized in Proposition 1 below, under MSLD, more productive firms self-select in more appealing cities that are bigger and where competition is endogenously tougher. The main driver of this spatial sorting is the fact that more productive firms gain relatively more from increased sales a bigger city allows.

To characterize the problem (21), I restrict the attention to parametrizations $\Upsilon(\cdot)$ that satisfy MSLD. There is empirical evidence supporting this feature of the consumer preferences (see, for instance, De Loecker and Goldberg (2014) and Amiti, Itskhoki, and Konings (2019)) and its converse, although theoretically possible, seems counter-intuitive.[20] Furthermore, to facilitate the exposition, I assume the particular functional form of Klenow and Willis (2016) which satisfies MSLD. Appendix B provides conditions for general Kimball aggregators satisfying MSLD for which 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:*

theorem implies that $\partial \log \psi\left(\frac{\mathbb{C}(c)}{z}\right) \varphi\left(\psi\left(\frac{\mathbb{C}(c)}{z}\right)\right) / \partial c = 0$.

[19] Formally, the first-order condition (28) is not a function of the firms' productivity, z, and therefore it is not possible to characterize the location choice problem.

[20] Melitz (2018) offers a detailed discussion of the implications of violations of the MSLD.

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

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

I define an *assignment pair* as a pair of functions $c \to (z(c), M(c))$, where $z(c)$ is the assignment function of local producers to cities, which is the inverse of $c^*(z)$.[21] The function $M(c)$ is the equilibrium market size that supports this location choice.

Proposition 1 (Sorting)**.** *Suppose that Assumption 1 holds, and that $\xi(\overline{\sigma} - 1) > 1$, where ξ is given by:*

$$\xi \equiv 1 + \frac{\theta(\alpha + \eta\beta) + \gamma(1 + \theta(1 - \eta\beta))}{\gamma(1 + \phi)} \tag{30}$$

Then, there exists a threshold $\underline{\varepsilon}$ such that for all $\varepsilon \in (0, \underline{\varepsilon}]$ there exists a unique solution to (21). In this solution, the functions $z(c)$, $M(c)$, and $\mathbb{C}(c)$ are strictly increasing.

Proof. See Appendix B.1. □

Proposition 1 demonstrates the existence and uniqueness of the assignment between city appeal c and local producer's productivity z. It features positive assortative matching. More productive producers go to more appealing cities, $z'(c) > 0$. Furthermore, more appealing cities are bigger, $M'(c) > 0$, and have tougher competition, $\mathbb{C}'(c) > 0$. Intuitively, more appealing cities attract more workers and pay higher wages. This triggers the incentives for local producers to enter such markets. In turn, local competition increases. Because of pricing complementarities, only the most productive firms can sustain the high level of local competition. Low-productivity firms opt-out and locate in smaller cities where competition is slack.

The parameter restrictions in Proposition 1 ensure the existence of a unique assignment. The term ξ captures different congestion forces in the framework. Namely, as cities grow, the cost of labor and buildings (housing and commercial structures) also grow. Moreover, cities are not perfect substitutes for workers because of the idiosyncratic location tastes. Therefore, these forces prevent all workers and firms from locating in the most appealing city. When ε is not too large, workers' valuation from an additional variety is given by $1/(\overline{\sigma} - 1)$, which is the central agglomeration force in this framework. Therefore, the condition $\xi(\overline{\sigma} - 1) > 1$ captures the standard condition for the uniqueness of equilibrium in general equilibrium spatial models: congestion forces need to

[21] As Proposition 1 shows, $z(c)$ is strictly increasing and, therefore, its inverse is well defined.

be greater than agglomeration forces.[22] Proposition 2 formalizes this discussion by showing that under the assumptions of Proposition 1, there exists of a unique equilibrium.

Proposition 2 (Existence and uniqueness)**.** *Suppose the assumptions of Proposition 1 hold, and that the supports of $G(\cdot)$ and $F(\cdot)$ are not too large. There exists a unique decentralized equilibrium. This equilibrium exhibits positive assortative matching.*

Proof. See Appendix B.2. □

The results in Proposition 2 make it possible to shed further light on how differences in local competition depend on the local producers sorting using a particular limiting equilibrium. Suppose that the ex-ante differences across cities become arbitrarily small: all cities have arbitrarily small differences in the productivity of the traded good and local amenities. In this limiting, all cities have virtually the same appeal (27). In that case, only differences in the pool of local producers determine any ex-post differences across cities. Corollary 1 below shows that spatial differentials in local competition arise even without any ex-ante heterogeneity between locations.

Corollary 1 (Equilibrium with ex-ante identical locations)**.** *Suppose the assumptions of Proposition 2 hold. Then, there exists a limiting equilibrium in which more appealing cities are larger and more competitive as the exogenous differences in appeal, c, go to zero.*

Proof. See Appendix B.3. □

The results in Corollary 1 highlight that pricing complementarities are the mechanism that generates spatial differences in local competition. In the limiting economy, cities are ex-ante identical, but ex-post differences arise endogenously. These differences arise due to the spatial sorting of the local producers. On the one hand, because of the pricing complementarities, more productive firms can adapt better to environments with intense competition and hence locate in bigger cities in which competition is endogenously tougher. On the other hand, the least productive firms find it optimal to locate in smaller cities where competition is slack. Note that as locations do not differ ex-ante along any dimension, the size of a city and the level of local competition are driven purely by the location choice of the local producers.

This result should not be surprising as there are no technological complementarities between the local producer's productivity and the city's local productivity or amenities in the baseline economy. Local producers only value the exogenous city characteristics through their effect on city size. When there are no differences in the city's ex-ante characteristics, firms still sort across space due to the endogenous differences in city size and local competition. Of course, city size and local competition are determined in equilibrium with the location choice of local producers. The message Corollary 1 emphasizes is that the limiting economy illustrates a case in which the joint determination of these three objects leads to non-degenerate equilibrium. In this equilibrium, bigger cities are more competitive, and more productive firms are located in such places.

[22] See Redding and Rossi-Hansberg (2017) Section 3.5 for a detailed discussion on congestion and agglomeration forces in a canonical spatial quantitative model.

Formally, Corollary 1 selects a particular equilibrium of the limiting economy. If ex-ante differences across cities are precisely zero, there are infinite equilibria. However, in Corollary 1, we contemplate an economy in which cities differ ex-ante, but those differences become arbitrarily small. In the limit, there are two possible equilibrium outcomes: the mixing economy in which all cities are identical or the separating economy in which cities differ ex-post. What Corollary 1 shows is that, in the limit, the separating economy is the one that emerges due to the pricing complementarities.

2.7 Markups across Cities

Having characterized the location decision of the local varieties producers, we turn to study how these decisions shape the distribution of markups across cities. Under the assumptions of Proposition 1, local producers in c charge a markup equal to $\mu\left(\mathbb{C}(c)/z(c)\right)$. The term $z(c)$ is the assignment function from Proposition 1 and reflects the productivity level of the local producers.

Let $\mathcal{M}(c)$ denote the city-level markup, implicitly defined by the city-level labor share in local goods:

$$\frac{W(c)L^N(c)}{M(c)} = \frac{\beta}{\mathcal{M}(c)}.$$

Combining this definition with the firm-level labor share (19), one can show that the city-level markup is a sales-weighted harmonic average of the firm-level markups.[23] Nevertheless, under Proposition 1 all firms in c charge the same markup and therefore the aggregate markup $\mathcal{M}(c)$ is:

$$\mathcal{M}(c) = \mu\left(\frac{\mathbb{C}(c)}{z(c)}\right) \tag{31}$$

Equation (31) reveals two opposite forces that determine the city-level markup. On the one hand, bigger cities have tougher competition: $\mathbb{C}(c)$ is strictly increasing. Therefore, a *competition force* in bigger cities pushes the city-level markup down. On the other hand, bigger cities attract more productive firms: $z(c)$ is strictly increasing. Hence, a *selection force* in bigger cities pushes the level of markups up. The relative strength of these two forces determines if bigger cities have higher or lower markups in equilibrium.

To facilitate the characterization of the level of markups across cities, I adopt here a particular functional form of the economy-wide productivity distribution of local producers $G(\cdot)$.

Assumption 2 (Firm Productivity Distribution)**.** *The common productivity distribution of local producers is a truncated Pareto with support* $[z_L, z_H]$ *and shape parameter* δ*, namely*

$$G(z) = \frac{1-\left(\frac{z_L}{z}\right)^{\delta}}{1-\left(\frac{z_L}{z_H}\right)^{\delta}}, \qquad \delta > 0. \tag{32}$$

[23] Edmond, Midrigan, and Xu (2023) obtain an equivalent expression when defining the sector-level markup.

Corollary 2 (Markups and City Size)**.** *Suppose that Assumptions 1 and 2 hold. There exists a threshold $\underline{\delta}$ such that*

1. *If $\delta > \underline{\delta}$, $\mathcal{M}(c)$ is strictly decreasing.*
2. *If $\delta < \underline{\delta}$, $\mathcal{M}(c)$ is strictly increasing.*

Proof. See Appendix B.4. □

Proposition 2 establishes that the productivity dispersion of local producers determines the cross-sectional markup variation across cities. The productivity distribution shape parameter, δ, is a measure of how dispersed is the productivity of local producers. When the productivity of local producers is not that dispersed, bigger cities have lower markups: the competition force dominates. If, on the other hand, local producers differ too much in their productivity, then bigger cities have higher markups: the selection force dominates. Appendix B.4 shows how the results in Corollary 2 extend when considering a general economy-wide productivity distribution for local producers, $G(\cdot)$.

The results in Proposition 2 resemble recent findings in the literature of endogenous variable markups. Recall that, under Proposition 1, bigger cities are more competitive. However, depending on local producers' productivity dispersion, bigger cities can have higher or lower markups. Therefore, the level of markups in a given city should not be taken as *prima-facie* evidence of reduced competition.[24]

Finally, we can also characterize how the location choice of local producers affects the Total Factor Productivity (TFP) of a city. Formally, we consider an aggregation exercise in which the total labor and structures determine the amount of the bundle of local varieties produced in a given city. Formally, let $\mathcal{Z}(c)$ be implicitly defined by the aggregate production function:

$$Y(c) = \mathcal{Z}(c)\left(L^N(c)\right)^{\beta}\left(S^N(c)\right)^{1-\beta},$$

where $L^{NT}(c)$ and $S^{NT}(c)$ are given by (20). In the decentralized equilibrium, we have that

$$\mathcal{Z}(c) = z(c)\frac{\Upsilon\left(\varphi\left(\psi\left(\frac{\mathbb{C}(c)}{z(c)}\right)\right)\right)}{\varphi\left(\psi\left(\frac{\mathbb{C}(c)}{z(c)}\right)\right)}, \tag{33}$$

where $\varphi\left(\psi\left(\mathbb{C}(c)/z(c)\right)\right)$ is the optimal relative quantity of local producers in city c given by (18). Equation (33) reveals that local TFP has two components. On the one hand, it depends on the productivity of local producers. Cities that attract more productive firms have higher TFP. On the other hand, local TFP also depends on the production distribution within a city. Because consumers have a taste for variety, cities that attract more producers will exhibit higher TFP. The relative quantities of local producers capture this feature. In the counterfactual exercises, I

[24] Baqaee, Farhi, and Sangani (2023) and Matsuyama and Ushchev (2022) also find conditions under which larger markets could have higher markups.

explore how the number of local producers in each city magnifies differences in TFP coming from the productivity differences of local producers.

3 Efficiency

In a single location model with variable markups, there are two different margins of inefficiency. As pointed out by Baqaee, Farhi, and Sangani (2023) and Edmond, Midrigan, and Xu (2023), variable markups can lead to inefficient overall entry and misallocation of factors of productions across firms.[25] The entry inefficiency arises because profits (private return) from the marginal entrant differ from the consumer surplus their entry generates (social return). Moreover, misallocation of factors of production arises when more productive firms charge higher markups.[26] Relative to the social optimum, more productive firms are too small, and aggregate welfare could increase by reallocating production from low to high-productivity firms.

With geography, local good producers make another decision: choose where to locate. With this additional layer, the overall entry margin may be inefficient, and the city-specific entry rate could be inefficient. The equilibrium allocation in the decentralized equilibrium is inefficient because of two opposite externalities that arise with local entry.

Firms create a positive externality when entering a particular city. Because consumer values variety in local goods, when a firm enters a city, it raises consumer surplus by creating a new good. I call this externality, *variety gains externality.* Nevertheless, firms can only partially appropriate the gain in consumer surplus into their profits. This non-appropriability reduces firms' incentives to enter a particular city, leading to insufficient entry. From a social planner's perspective, we would like to have more firms in particular locations.

Firms also create a negative externality when entering a city. Because local varieties are imperfect substitutes, when a producer enters a city, it reduces the consumption of the existing varieties. Thus, firms impose a negative externality on incumbents by reducing their profits. This is a *business stealing* externality. There is excessive entry because firms do not internalize their effect on other producers. From a social planner perspective, we would like to have fewer varieties in a particular location and increase the consumption of the existing ones.

It is worth to highlight that the variety gains and the business stealing externality are not specific to my framework and are present in standard models of firm entry.[27] Nevertheless, in the commonly used models with CES preferences, these two externalities are always constant and offset each other (Matsuyama and Ushchev (2021)).

In equilibrium, whether there is too much or too little entry in a particular city depends on the strength of the variety of gains and business stealing externalities. In the spatial equilibrium described in the previous section, the variety gains externality is higher in small cities, and the business stealing externality is higher in bigger cities. Consequently, there is too much entry in

[25] Moreover, with overhead costs, there is a third margin of inefficiency: the selection cutoff in productivity.

[26] Which is the case when Marshall's second law holds.

[27] This was early pointed out by Mankiw and Whinston (1986)

bigger cities and too little in small cities. The spatial sorting of firms through pricing complementarities leads to more productive firms over-concentrating in larger markets relative to the social optimum. This inefficiency ignites a “top-down” effect on other cities, generating misallocation throughout the economy.

To better understand the spatial nature of the variety gains and the business stealing externalities, consider two locations $c_1 < c_2$. Because location c_2 is more appealing, it is bigger and more competitive, $M(c_1) < M(c_2)$ and $C(c_1) < C(c_2)$. A lower competition level in the small city reflects that it cannot attract too many local producers, and the ones that decide to operate there are of low productivity, $z(c_1) < z(c_2)$. Therefore, consumers in the small location benefit more from an additional variety than consumers in the big city. On the other hand, bigger cities can attract the more productive firms because potential profits are higher relative to small cities.[28] Therefore, the incumbents' profit loss from a marginal entrant is higher in big cities than in smaller cities. As a result, the variety gains externality dominates in smaller markets, while the business stealing externality dominates in bigger ones.

3.1 Social Planner's Problem

I formalize the previous arguments by characterizing the planner's problem. An utilitarian planner maximizes the population-weighted sum of workers' utility in every city. The planner chooses the location of local producers and population subject to workers' idiosyncratic location tastes. The planner also chooses the allocation of labor into traded good production and local varieties production in every city. Moreover, she is subject to the housing supply technology in every location. I relegate the formal definition of the planning problem to Appendix C and characterize the solution in the main text. I use SP superscripts for the solutions in the planner problem and DE superscripts for the decentralized equilibrium. The decentralized equilibrium is inefficient when the decentralized allocation does not coincide with the planning one.

Proposition 3 (Efficient Allocation)**.** *The decentralized equilibrium is inefficient. Moreover, suppose that the supports of $G(z)$ and $F(c)$ are not too large as in Proposition 2. Then, for all c:*

$$z^{SP}(c) > z^{DE}(c) \qquad \textit{for all} \qquad c \in (c_L, c_H). \tag{34}$$

Proof. See Appendix C.1. □

Proposition 3 establishes that the decentralized equilibrium is inefficient. In the baseline framework, markups generate inefficiencies through three channels. First, they distort the relative consumption between local varieties, housing, and the traded good. Second, they distort the location decisions of firms. Third, they distort the aggregate entry margin of local producers. Importantly, because local producers in each city have the same productivity, there is no misallocation within a city in the sense of Hsieh and Klenow (2009).

[28] Note that if profits were higher in smaller cities, then high productivity firms would have a profitable deviation by locating in smaller cities, which contradicts the results of Proposition 1.

Equation (34) shows that firms are misallocated across cities. Local producers in the decentralized equilibrium are not productive enough relative to the social planners' solution. More productive firms are too concentrated in bigger cities in the decentralized equilibrium. On the other hand, for any city c, the social planner selects more productive firms $z^{DE}(c) < z^{SP}(c)$. Therefore, there is a misallocation of firms across cities: aggregate welfare can increase by reallocating productive producers from big to small cities.

Whether big cities have higher or smaller markups affects firm misallocation. As Appendix C.2 formalizes, the slope of markups across cities exacerbates the business stealing externality. To understand this mechanism, it is helpful to consider two economies: one in which markups in big cities are low and another in which markups are high. In the first economy, more productive firms face the trade-off between locating in larger markets when they sell more but make lower margins. This trade-off reduces the incentives for setting production in big cities, and marginal producers find it optimal to reallocate to smaller locations. On the other hand, in the second economy, where markups are high in bigger cities, firms no longer face this trade-off: they sell more and have larger margins in such locations. Of course, low-productive producers still self-select into smaller cities because of the pricing complementarities. However, the more productive producers who can handle the high competitive pressure in bigger cities over-concentrate even more in such locations than in the first economy scenario.

As the results in the empirical section highlight, the slope of the relationship between markups and city size is then informative of the degree of misallocation in the economy. Even though firms always over-concentrate in big cities, a negative slope suggests that misallocation across space is less severe than in a situation with a positive slope.

3.2 First-best Implementation

How can efficiency in the decentralized equilibrium be restored? Proposition 4 shows that the first-best allocation can be attained by implementing a location-specific subsidy per total production. This subsidy corrects the three margins of inefficiency previously discussed: markups, misallocation of firms across cities, and overall entry. Finally, the subsidy is financed by a flat labor tax.[29]

Formally, consider $T(y, c)$ that is city-specific and depends the quantities sold, y:

$$T(y(z,c),c) = \left[\underbrace{\Upsilon\left(\frac{y(z,c)}{Y(c)}\right)}_{\text{worker's utility}} - \underbrace{\Upsilon'\left(\frac{y(z,c)}{Y(c)}\right)\frac{y(z,c)}{Y(c)}}_{\text{original revenue curve}} \right] \frac{\mathbb{D}(c)}{\mathbb{P}(c)} M(c) \tag{35}$$

The policy in (35) affects firm revenue in two ways. First, it takes away the sales from the firm's original revenue curve: the term corresponding to $\Upsilon'\left(y/Y(c)\right)\left(y/Y(c)\right)$. Second, it returns revenues proportionately to $\Upsilon\left(y/Y(c)\right)$, which measures the relative "utility" each firm generates. Under this policy, the net profits for firm z in city c are given by:

[29] That is, workers earnings in every city are equal to $(1-\tau)W(c)$, where τ is the tax.

$$\widehat{\Pi}(z,c) = \Pi(z,c) + T(y(z,c),c) = \left[\Upsilon\left(\frac{y(z,c)}{Y(c)}\right) - \frac{\mathbb{C}(c)}{z}\frac{y(z,c)}{Y(c)}\right]\frac{\mathbb{D}(c)}{\mathbb{P}(c)}M(c) \tag{36}$$

Equation (36) reveals that the transfer eliminates any incentives for firms to charge a markup. When sales come from the original revenue curve, producers are incentivized to shrink production to maximize sales. However, when sales come proportional to $\Upsilon\left(y/Y(c)\right)$, firms have the incentives to maximize the units produced. In turn, firms produce at marginal cost and exert no market power. As Proposition 4 clarifies, when firms profits are given by (36), firm misallocation across cities is also eliminated.

Proposition 4 (Optimal Policy)**.** *Under the location-specific subsidy* (35)*, the decentralized equilibrium allocation coincides with the planner solution.*

Proof. See Appendix C.3. □

Proposition 4 shows that the subsidy (36) corrects the three margins of inefficiency in the decentralized equilibrium. First, it eliminates markups, which corrects the inefficient relative consumption between the bundle of local varieties, housing, and the traded good. Second, it gives the right incentives for firms to locate efficiently. Lastly, it also corrects the overall entry into the economy. As Appendix C.3 shows, in equilibrium, firms profits (36) can be written as:

$$\Pi(z,c) = \left[\underbrace{\delta\left(\frac{y(z,c)}{Y(c)}\right)}_{\text{consumer surplus}} - 1\right]\underbrace{\Upsilon\left(\frac{y(z,c)}{Y(c)}\right)}_{\text{market share}} M(c), \tag{37}$$

where $\delta\left(\frac{y(z,c)}{Y(c)}\right)$ is the ratio of the consumer surplus to firm sales.[30] In equilibrium, firms capture a share of the total revenue in a market proportionally to workers' utility. Moreover, firms' profits exactly coincide with the consumer surplus they generate. Then, because firms are now correctly compensated for their effect on workers' utility, the location and the overall entry margin are corrected. Finally, it is worth to highlight that (35) generalizes the insights of Edmond, Midrigan, and Xu (2023). In their setting, the optimal policy for a single market is similar to (36).

This section outlined a spatial general equilibrium model in which spatial markup differences arise because of the location choice of heterogeneous local producers. The framework highlights that differences in markups across are explained by differences in local competition and the productivity of local producers. Moreover, In the next section, I study the framework's predictions using data from local producers in the United States.

[30] Formally, $\delta\left(\frac{y(z,c)}{Y(c)}\right) = \Upsilon\left(\frac{y(z,c)}{Y(c)}\right) / \left(\frac{y(z,c)}{Y(c)}\Upsilon'\left(\frac{y(z,c)}{Y(c)}\right)\right)$. See Figure 1 in Baqaee, Farhi, and Sangani (2023) for a visual representation of this object.

4 Empirical Analysis

In this section, I empirically investigate the theory predictions. When taking the model to the data, we need to take a stand of a definition of a city and on the map between locations in the model (continuum) and cities in the data (discrete). In turn, I define a city in the data as a county and I consider a county as being a collection of related locations in the model. Formally, a county is an interval $[c, c + dc]$ of cities in the model.

To conduct the empirical investigation, I first describe the U.S. establishment-level data used for all exercises. Second, I introduce the classification of traded and local (non-traded) sectors I use through the empirical exercises. Then, I perform model validation exercises. Finally, I outline and implement the empirical strategy to estimate markups and study the variation across U.S. cities.

4.1 Data

The primary dataset used in this project is the micro-data from the U.S. Census Longitudinal Business Database (LBD). This data source uses administrative employment records of every non-farm private establishment in the U.S. economy. The establishment-level variables I used are employment, wage bill, geographic location (county), industry (6-digit NAICS), and the establishment identifier.

I supplement the LBD data with sales data at the establishment level from the Economic Censuses every five years from 2002 to 2017. Specifically, I use the micro-data from the Census of Construction Industries, Manufacturing, Retail Trade, Census of Services, Wholesale Trade, Finance, Insurance and Real Estate, and the Census of Transportation, Communications and Utilities. I use the establishment identifier to link the establishment in the Economic Censuses to the establishments in the LBD. The final sample is the establishments in the LBD with matched sales data from the Economic Censuses.[31] I use 2017 as the baseline year, leaving 2002, 2007, and 2012 for robustness exercises.

I use the Census of Manufactures to perform additional markup estimation exercises. The Census of Manufactures has detailed data on establishment materials, capital (equipment and structures), and energy expenditures. Unfortunately, such detailed data is not available in the other Economic Censuses. I construct real capital, materials, and labor measures using standard procedures used in the productivity estimation literature (see Foster, Grim, and Haltiwanger (2016)).

In the baseline exercises, I associate a city with a county. Focusing on continental U.S., I include 3080 counties in my estimation. For some the robustness exercises, I define cities as Commuting Zones.[32]

[31] This is virtually the same sample used in Hsieh and Rossi-Hansberg (2023).

[32] To map counties to commuting zones, I use the crosswalk provided by Autor and Dorn (2013).

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 (baseline sample). The traded-local industries classification is based on Delgado, Porter, and Stern (2015).

4.2 Local Industries

I use the definition of Delgado, Porter, and Stern (2015) to classify establishments in the LBD as traded or local producers. Broadly, this definition classifies 6-digit NAICS industries into "Traded" or "Local" based on employment specialization, geographic concentration, and distance to final consumers. Local producers belong to industries in most of the geographic areas and sell to local consumers. On the other hand, traded industries sell to other regions and are sometimes geographically concentrated. Formally, the authors group 310 6-digit NAICS industries as Local and 778 6-digit NAICS industries as Traded.[33] Using the industry codes from the LBD, I classify an establishment as a local producer if it belongs to any of the 310 local industries.[34]

Table 1 displays summary statistics for the baseline sample. The sample includes 85% of all the LBD establishments in 2017. Establishments operating in local industries are to be smaller in number of workers, sales, and wage bill compared to their counterparts in traded industries. Nevertheless, the number of local establishments is almost three times that of traded ones. Hence, in the aggregate, local industries represent more than half of the U.S. economic activity by employing 67% of the labor force and by accounting for 52% of the total sales and 59% of total labor income.

4.3 Model Validation

Before analyzing the empirical patterns of markups across U.S. cities, I provide empirical support for the model's sorting prediction. Proposition 1 indicates that local producers are more productive in bigger cities. Because firm productivity is not observed in the data, I consider how two proxies of firm productivity relate to city size.

Labor productivity and city size. The first proxy for firm productivity is labor productivity.

[33] For the full list of 310 local NAICS industries see Cluster Mapping Project.

[34] This classification is used in Berger, Herkenhoff, and Mongey (2022).

Ideally, one would like to study output (physical quantities) per worker. However, I cannot separate prices and quantities as I only observe sales at the establishment level. Therefore, I define labor productivity at the establishment level as sales per worker.

I construct measures of labor productivity and size for every county. Using the establishment's location and local-traded classification, I compute sales per worker for establishments in local industries across all counties. Then, for every county, I compute the average sales per worker across all establishments in local industries in a given county. This is the measure of labor productivity at the county level. On the other hand, guided by the theoretical framework, I define the "size" of a county as the total income of workers residing in that county.[35]

Figure 1(a) displays a bin-scatter of counties' log labor productivity and log county size. As Proposition 1 establishes, county size and county labor productivity have a positive and significant relationship. An increase of 1% in a county's size is associated with a 0.03% increase in county labor productivity. Moreover, counties in the top decile of the county-size distribution have a labor productivity 13% higher than counties at the bottom.

Establishment size and county size. The second firm productivity proxy we consider is establishment size. Appendix A.2 shows that the model predicts that producers in bigger cities employ more workers than producers in smaller cities. To empirically investigate this prediction, I first measure establishment size in each county by computing the average establishment employment for local industries. Then, I split counties into 100 equally-sized bins according to their total labor income and compute the average establishment size across counties in each bin. Generally, we compute the average establishment size for counties across the percentiles of the county-size distribution.

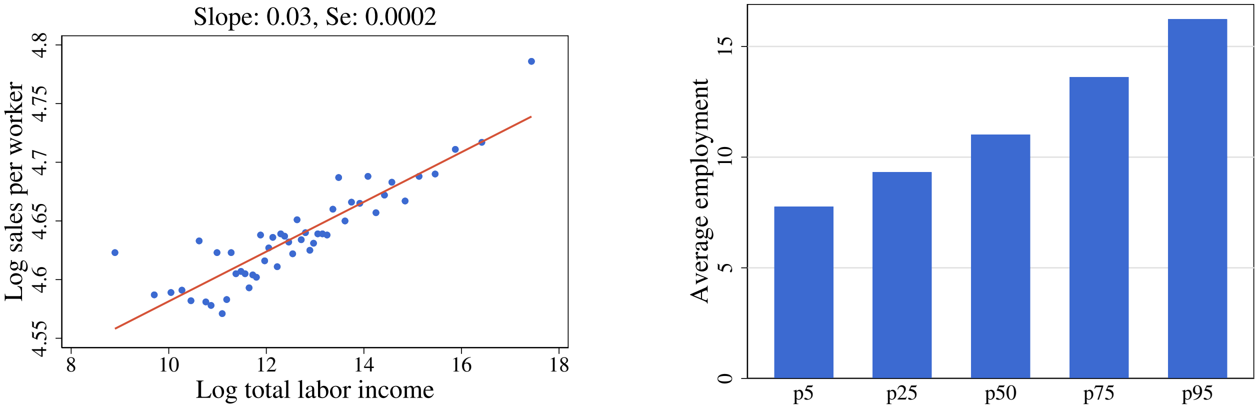
Figure 1(b) shows the average employment for local establishments in different counties across the county-size distribution. In particular, it displays the average employment for local establishments in counties in the 5th, 25th, 50th, 75th, and 95th percentiles of the city-size distribution. A typical local establishment in the smaller counties (5th percentile) has 7.76 workers; in the largest counties (95th percentile), it employs 16.23 workers. This demonstrates that average employment in local industries doubles across the county-size distribution.

4.4 Markup Estimation

This section outlines the empirical strategy to estimate markups for establishments in local industries. Using the insights of De Loecker (2011), I develop an alternative method for markup estimation that combines consumer preferences with the firm optimal input decisions. Before proceeding with the description of the method, I discuss the limitations of applying some of the existing methods in my LBD-Economic Census sample.

[35] As my model indicates, I consider all workers regardless of whether they are employed in local or traded industries.

Figure 1: Local industries revenue per worker and average employment across counties



(a) Sales per worker and total labor income

(b) Average estab. employment across counties.

Notes: Figure 1(a) shows a bin-scatter of county log average sales per worker and log total labor income for local industries. The bin-scatter considers 50 equally sized county bins according to their total labor income. Average sales per-worker is computed among establishments in local industries. Figure 1(b) displays average employment for establishments in local industries across counties in different percentiles of the county-size distribution. Different bars indicate percentiles of the county-size distribution: 5th, 25th, 50th, 75th, and 95th percentiles. The height of each bar represents the average employment of establishments in local industries across counties in each percentile. County size is defined as total labor income.

4.4.1 Existing Methods

I build on the production approach to estimate markups. Originally developed by Hall (1988) and recently extended by De Loecker and Warzynski (2012), this approach produces markup estimates using data on sales, variable input expenditures, paired with estimates of output elasticities. In contrast, an alternative procedure often refereed as the demand approach, uses data on prices and quantities to estimate the marginal cost of production. With estimates on own and cross-price elasticities across goods, markups can be recovered from the firms pricing first-order conditions after specifying the market structure under which firms compete.[36] I do not observe prices or product characteristics in the LBD or the EC's data, and therefore I cannot implement the demand approach.

I index establishments by j and counties by c in the data. Under certain regularity conditions of the firms' cost-minimization problem, markup for establishment j in county c can be expressed as the ratio of the output elasticity of a flexible input and the cost-shares of sales of that input.[37] The conditions for the ratio estimator hold in the setting outlined in Section 2 and equation (19) implies that:

$$\mu_{jc} = \frac{\beta}{\alpha^{l}_{jc}}, \tag{38}$$

[36] See Ackerberg et al. (2007) and Berry, Gaynor, and Scott Morton (2019) for excellent overviews.

[37] A flexible input is one that: 1) can be adjusted freely every period, and 2) establishments take as given the price of the input. The latter condition rules out the possibility of monopsony power that inputs the market.

where $\alpha^l_{jc} = (W_c l_{jc})/(p_{jc} y_{jc})$ is the labor expenditure share of total sales, which is observed in the data. Then, using (38), one can form an estimate of the markup by obtaining an estimate of the labor output elasticity, $\hat{\mu}_{jc} = \hat{\beta}/\alpha^l_{jc}$. This estimator is commonly called the "ratio estimator".[38] Under this approach, the main econometric challenge is to estimate production elasticities.

The first procedure to estimate output elasticities is the production function approach. Under this alternative, researchers estimate a production function by regressing output on inputs. The estimation is usually done by implementing a control function approach as in Olley and Pakes (1996), Levinsohn and Petrin (2003), Ackerberg, Caves, and Frazer (2015), and Gandhi, Navarro, and Rivers (2020), or by estimating dynamic panel models as in Arellano and Bover (1995) and Blundell and Bond (1998, 2000). Each of these approaches has costs and benefits. However, one common requirement is data on physical quantities as the output measure.[39] Unobserved output price differences confound the identification of the production function parameters when sales are used as an output measure.[40] Data on physical quantities for establishments throughout different economic sectors in the U.S. does not exist.[41] In particular, I observe sales as the output measure in the LBD-EC sample. Therefore, we cannot implement the production function approach to form the ratio estimator.

The second alternative to estimate production elasticities is the cost-share approach.[42] Relying on cost minimization conditions, an input's output elasticity equals the input's cost share of total costs times the scale elasticity.[43] In contrast to the production function alternative, this approach does not require data on physical quantities. However, it requires data on the establishment's total costs. As highlighted by De Loecker and Syverson (2021), data on total costs is rare, with capital costs being the most difficult to observe. Indeed, data on the U.S. establishment's total costs does not exist except for publicly traded companies and manufacturing establishments. Thus, I cannot implement the cost-share approach to the LBD-EC sample to estimate markups through the ratio estimator.

In sum, data limitations prevent the implementation of the ratio estimator. As an alternative, in the same spirit of De Loecker (2011), I use the demand structure from my model to overcome the identification challenge. De Loecker (2011) uses a CES demand structure to control for unobserved prices in a production function estimation context. Similarly, I use the demand structure from the theory section to construct a markup control function.[44] The following section describes in detail this alternative procedure.

[38] See De Loecker and Warzynski (2012) for a detailed derivation of this estimator.

[39] De Loecker and Syverson (2021) offer an exhaustive review of the control functions and dynamic panel models.

[40] Bond et al. (2021) discusses pitfalls of using the ratio estimator without data on physical quantities.

[41] Exemptions are Manufacturing sub-samples in Foster, Haltiwanger, and Syverson (2008) and Atalay (2014).

[42] 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.

[43] The scale elasticity is the degree of returns to scale of the production technology.

[44] The main difference between my framework and De Loecker (2011) is that my demand structure allows for variable markups.

4.4.2 Alternative Procedure

To avoid estimates dependency on functional forms specifications, I consider a general parametrization of the Kimball aggregator $\Upsilon(\cdot)$ with the only assumption of a choke price.[45]

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

$$\left(\Upsilon'\right)^{-1}\left(\overline{p}\right) = 0. \tag{39}$$

Recall from (7) that workers relative demand is given by the inverse of the derivative of the Kimball aggregator.[46] Hence, condition (39) implies the intuitive idea that a finite price exists at which workers demand zero quantities.

On the other hand, we can re-organize (38) and take logs to obtain:

$$\log \alpha^{l}_{jc} = \beta - \log \mu_{jc} \tag{40}$$

Note that the LHS of (40) is observed in the data. Therefore, one could potentially estimate markups as the residual of a regression of log labor cost share of revenue and a constant. Nevertheless, this procedure has two potential threats. First, any measurement error on the labor cost share of revenue is absorbed in the error term confounding markup estimates. Second, as Appendix D.3 illustrates, if one considers a more general production function in which output elasticities are not constant and vary with inputs, the potential correlation between markups and input usage invalidates the identification of markups as residuals from (40). I use the demand system in Assumption 3 to construct a markup control function to avoid these issues.

Let p_{jc} be the price establishment j charges in county c. Moreover, let D_c and P_c be the competition and ideal price indices in county c, respectively. Using the Lerner formula, we can μ_{jc} as a function on the relative price:

$$\begin{aligned}\mu_{jc} &= \mu\left(\frac{p_{jc}}{D_c}\right),\\ &= \frac{1}{1-\frac{1}{\sigma\left(\frac{p_{jc}}{D_c}\right)}}\end{aligned} \tag{41}$$

where $\sigma(\cdot)$ is given by (14). Because there is no price information in the U.S. micro-data, the object p_{jc}/D_c is unobserved. However, we can use the demand system to express relative prices as a function of sales market shares. Let s_{jc} be th sales share of establishment j in county c.[47]

[45] The Klenow and Willis (2016) introduced in Assumption 1 has a choke price. The CES functional form for $\Upsilon(\cdot)$ has no choke price.

[46] The concavity of $\Upsilon(\cdot)$ guarantees the existence of this inverse function.

[47] Formally, $s_{jc} \equiv p_{jc}y_{jc}/\left(\sum_{j'\in c} p_{j'c}y_{j'c}\right)$.

Appendix D.1 shows that we can express the sales share as:

$$s_{jc} = \frac{D_c}{P_c} \Upsilon' \left(\frac{p_{jc}}{D_c} \right) \frac{p_{jc}}{D_c} \tag{42}$$

Appendix D.1 further shows that the function $\Upsilon'(x)x$ is strictly increasing, and therefore we can use (42) to solve for p_{jc}/D_c as a function of the sales market share and the price index ratio P_c/D_c:

$$\frac{p_{jc}}{D_c} = \zeta \left(s_{jc} \frac{P_c}{D_c} \right), \tag{43}$$

where $\xi(x)$ is the inverse of the function $\Upsilon'(x)x$. By combining (41) and (43) we can write μ_{jc} as function of $s_{jc} \times (P_c/D_c)$:

$$\mu_{jc} = \mu \left(\zeta \left(s_{jc} \frac{P_c}{D_c} \right) \right) \tag{44}$$

The exact functional form of the markup function $\mu \circ \zeta$ depends on the parametrization of Υ. However, to maintain the estimation parsimoniously, I use a semi-parametric approximation for the markup function and use a sieve series estimator as analyzed in Chen (2007) and used in the production function estimation context by Gandhi, Navarro, and Rivers (2020). Formally, I approximate the log markup function by a third-order degree polynomial in $s_{jc} \times (P_c/D_c)$:

$$\log \mu \left(\zeta \left(s_{jc} \frac{P_c}{D_c} \right) \right) = \varsigma_1 s_{jc} \frac{P_c}{D_c} + \varsigma_2 \left(s_{jc} \frac{P_c}{D_c} \right)^2 + \varsigma_3 \left(s_{jc} \frac{P_c}{D_c} \right)^3 + \upsilon_{jc}, \tag{45}$$

where υ_{jc} is an approximation error that goes to zero once one considers higher polynomial terms. Crucially, the approximation in (45) does not have a constant term. As shown in Appendix D.1, Assumption 3 implies that when producers have a zero sales share, they charge a markup equal to one. Intuitively, because of the choke price, firms with zero sales share face an infinite elasticity of demand and, therefore, have no markups. Combining (40) and (45) yields the equation from which markups are identified:

$$\log \alpha^l_{jc} = \beta - \underbrace{\varsigma_{1,c} s_{jc} - \varsigma_{2,c} s^2_{jc} - \varsigma_{3,c} s^3_{jc}}_{\equiv \log \mu_{jc}} - \upsilon_{jc}, \tag{46}$$

where $\varsigma_{1,c} \equiv \varsigma_1 (P_c/D_c)$, $\varsigma_{2,c} \equiv \varsigma_2 (P_c/D_c)^2$, and $\varsigma_{3,c} \equiv \varsigma_3 (P_c/D_c)^3$. Because I do not observe the county price indices, I treat them as county fixed-effects, and hence (46) takes the form of a heterogeneous slopes model.

Equation (46) reveals the variation that identifies markups. Conditional on the establishment technology, β, markups are identified using within-county variation in the sales shares and the labor cost share of sales. Within a county, establishments with high sales shares and low labor expenditure share have higher markups. The choke price allows us to separate the markup approximation function's constant (zero) from the labor output elasticity β. Thus, the levels of markup are correctly identified.

Appendix D.2 shows how to extend the estimation when considering multiple sectors and controlling for potential labor market power.[48] In particular, when estimating markups by sector, the estimating equation takes the form of:

$$\log \alpha^l_{jnc} = \beta_n - \varsigma_{1,nc} s_{jnc} - \varsigma_{2,nc} s^2_{jnc} - \varsigma_{3,nc} s^3_{jnc} - \upsilon_{jnc}, \tag{47}$$

where n index sector. In contrast to (46), the sector estimation uses the within-county-sector variation in sales share and labor cost share of sales to recover markups.

4.4.3 Markups for Establishments in Local Industries

This section presents the results for the markup estimation outlined in Section 4.4.2. Table 2 presents summary statistics for establishments in different local industries. The first row displays the results from the estimation for all local industries in (46). The remaining rows present markup estimates for the sector estimation in (47). For the sector estimation, I define a sector as a 2-digit NAICS industry.

There is considerable heterogeneity in markups across establishments in local industries. The median establishment in local industries has a markup of 1.43, which is the lines of the findings of De Loecker, Eeckhout, and Unger (2020) for publicly traded companies. However, I find significant cross-sectional heterogeneity, with establishments in the top decile of the markup distribution charging a markup seven times higher than establishments in the bottom decile.

There is also significant markup heterogeneity across local industries. On the one hand, the median and the 10th markup percentile are similar across different local sectors. Nonetheless, sectors like Manufacturing, Wholesale, and Information exhibit a mean markup significantly higher than the other sectors. These sectors also exhibit a larger p90 - p10 gap than the others.

4.4.4 Local Industries Markups across Cities

I now turn to the main empirical analysis of this section: markups across cities. I construct the sales-weighted harmonic mean of establishment markups to compute the county-level markup. Formally, following the implications of the theory in Section 2, the county aggregate markup $\mathcal{M}_c$ is defined as:

$$\mathcal{M}_c = \left(\sum_{j \in c} s_{jc} \frac{1}{\mu_{jc}} \right)^{-1}, \tag{48}$$

[48] In the baseline estimations, I control for potential labor market power by adding a flexible polynomial in the establishments' wage bill share to (46).

Table 2: Summary Statistics: Markups for Establishments in Local Industries

	Mean (1)	Median (2)	p10 (3)	p90 (4)
All Local	2.57	1.43	1.07	7.78
Local Construction	1.79	1.13	1.01	3.25
Local Manufacturing	2.56	1.43	1.07	7.77
Local Wholesale	3.28	1.79	1.05	7.76
Local Retail	2.4	1.35	1.04	7.78
Local Transportation and Warehousing	1.89	1.16	1.02	3.82
Local Information	2.7	1.41	1.05	7.75
Local Finance and Insurance	1.67	1.17	1.03	2.75
Local Real Estate	1.29	1.1	1.02	1.85
Local Profesional, Scientific and Technical Services	1.24	1.08	1.01	1.69
Local Administrative	1.41	1.1	1.01	2.22
Local Education Services	1.11	1.06	1.01	1.3
Local Healthcare	1.54	1.14	1.02	2.48
Local Arts, Entertainment, and Recreation	1.36	1.13	1.02	2.07
Local Accommodation and Food Services	1.32	1.17	1.03	1.87
Local Other Services	1.18	1.08	1.01	1.52

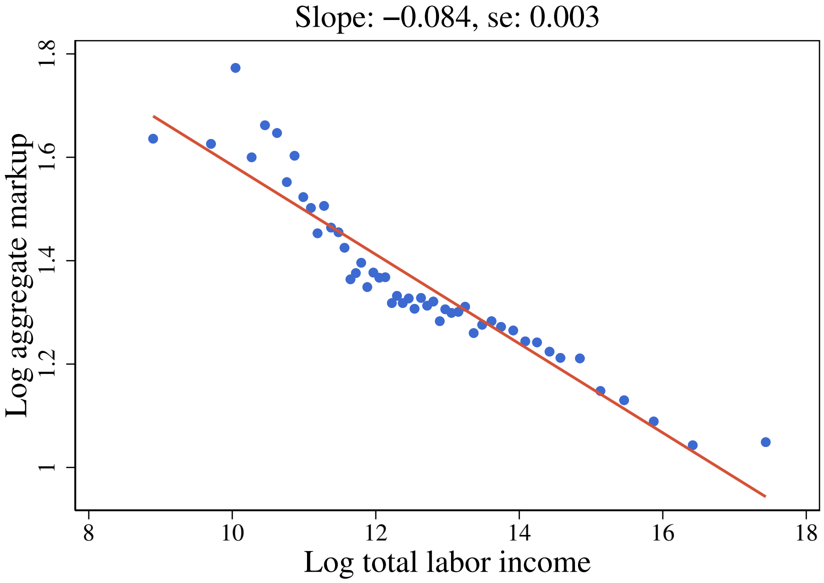
Notes: Table 2 displays summary statistics for the estimated markups using (46). The first row considers all local establishments. The following rows display statistics for establishments in local industries for 2-digits NAICS sectors. Columns p10 and p90 denote the 10th and 90th percentile of the markup distribution, respectively. The traded-local industries classification is based on Delgado, Porter, and Stern (2015).

where the sum is taking over the local establishments in county c.

Figure 2 shows the relationship between county aggregate markup and county size. The figure displays a clear empirical pattern: bigger counties have a markup significantly lower than their smaller counterparts. Counties like Manhattan or Cook County (Chicago) have a markup 50% lower than small counties like Highland, VA, or Armstrong, TX. Moreover, an increase of 1% in a county's size is associated with a decrease in the county markup of 0.084%. Table 3 shows that empirical findings are not particular to 2017 or defining a city as a county. The negative pattern between county markup and county size is also present when considering other years and defining a city in the data as a Commuting Zone.

Figure 2 sheds light on the mechanisms that shape the distribution of markups across cities. On the one hand, the empirical regularities shown in Section 4.3 show that bigger locations attract more productive producers. However, Figure 2 shows that markups in such locations are significantly lower than in small locations. Through the lens of the theory, the competition force dominates the selection force. Even though bigger cities attract more productive local producers, competition in those locations is high enough to restrain the market power of local producers. Furthermore, guided by the results in Proposition 2, this finding suggests that the dispersion of local producers' productivity is lower relative to the local characteristics of cities.

Figure 2: County aggregate markup and county size



Notes: Figure 2 shows a bin-scatter of county log aggregate markup and log total labor income. County size is defined as total labor income. The bin-scatter considers 50 equally sized county bins according to their total labor income. County aggregate markup is a sales-weighted harmonic mean of the local establishment's (48).

The negative relationship between county markup and county size also sheds light on the spatial misallocation of local producers. Proposition 3 highlights that if markups were higher in bigger cities, the misallocation of local producers would be exacerbated. Nonetheless, Figure 2 sends a reassuring message that markups in larger cities are lower than in smaller cities. The results suggest that the competition force governing local producers' location decisions prevents establishments from over-locating in bigger cities. Competition in bigger cities is intense enough to prevent local producers from charging higher markups and induces marginal producers to locate in smaller cities.

Although the primary goal of the current section is to analyze markups for all local industries, I turn now to a sector-specific analysis of markups across cities. Although I focus on the Retail and Manufacturing, Appendix E.1 shows results for the other sectors.

Figure 3 shows markups across cities for local Retail and Manufacturing. The patterns for Retail resemble the ones from all local industries. Local Retail producers in big cities charge a markup 60% lower than local Retail producers in the smallest cities. This finding is unsurprising as local Retail accounts for one-quarter of local local industries' employment. Hence, it is reasonable to think that local Retail producers are one of the drivers of the dynamics displayed in Figure 2.

Markups for local Manufacturing producers are higher in larger cities. Contrary to the results for all local industries, local Manufacturers in the bigger counties charge a markup two times higher than producers in the smallest cities. Furthermore, the local Manufacturing markup distribution across cities unveils two additional findings. First, the forces that govern competition and selection of local producers seem to vary across sectors. Second, the markup estimation procedure outlined in Section 4.4.2 does not mechanically deliver lower markups in big cities.

Table 3: Average City Markup elasticity with respect to City Size

	Dep. var.: Log aggregate markup			
	2002	2007	2012	2017
	(1)	(2)	(3)	(4)
	Panel A: Counties			
Log total labor income	-0.0987***	-0.0897***	-0.0931***	-0.0834***
	(0.0032)	(0.0034)	(0.0033)	(0.0031)
Observations	3100	3100	3100	3100
R-squared	0.368	0.328	0.363	0.323
	Panel B: Commuting Zones			
Log total labor income	-0.0671***	-0.0662***	-0.0651***	-0.0609***
	(0.0029)	(0.0026)	(0.0030)	(0.0029)
Observations	750	750	750	750
R-squared	0.598	0.61	0.526	0.551

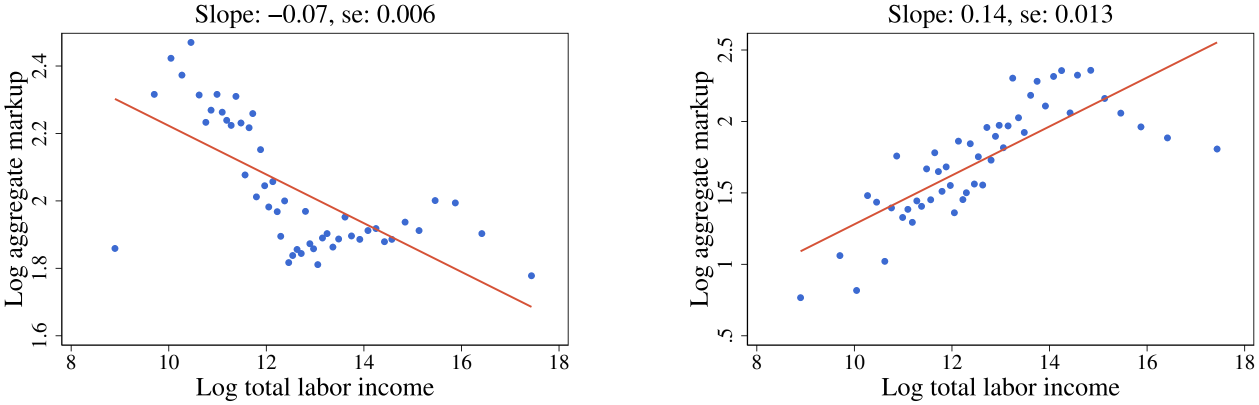
Notes: Table 3 displays the average elasticity of county aggregate markup and city size. City aggregate markups is defined as (48) and city size is defined as total labor income. Panel A shows the mean elasticity defining cities as counties. Panel B shows the mean elasticity defining cities as Commuting Zones. Robust standard errors in parenthesis. *10% level, **5% level, ***1% level.

The different empirical patterns for local Retail and local Manufacturing inform the spatial misallocation across sectors. Because local manufacturing markups in larger markets are higher, the spatial misallocation in Manufacturing may be more considerable than the misallocation in Retail. The selection force in local Manufacturing is significantly stronger than in local Retail. Intuitively, the productivity differences across local Manufacturing plants are much more significant than those across local Retail producers.

Robustness. I use the Manufacturing sector to perform robustness exercises that provide additional empirical support for the findings of this section. In contrast to other censuses, the Census of Manufactures has detailed data on production inputs. In particular, I observe materials, energy, and capital expenditure measures. The detailed data allows me to investigate the variation in markups across cities with two variations of the baseline estimation. Appendix E.2 shows the results of these alternative exercises.

First, one potential threat to the markup estimating equation (46) is that labor might not be fully flexible. I tackle this concern by estimating this equation using materials and energy as flexible inputs. Second, I relax the Cobb-Douglas assumption by considering a general production function. Under this approach, output elasticities are no longer constant and can be a function of the production inputs. In particular, I approximate the output elasticity by a flexible polynomial in labor, materials, energy, and capital as in Gandhi, Navarro, and Rivers (2020). Under these two alternative procedures, the estimated markups highly correlate with the baseline estimates. I also obtain the same cross-sectional city variation as in Figure 3(b).

Figure 3: County aggregate markup and county size for Local Retail and Local Manufacturing



(a) Retail

(b) Manufacturing

Notes: Figure 3(a) shows a bin-scatter of county log aggregate markup for Local Retail and log total labor income. Figure 3(b) shows a bin-scatter of county log aggregate markup for Local Manufacturing and log total labor income. County aggregate markup for Retail is a sales-weighted harmonic mean of the Retail local establishment's (48). County aggregate markup for Manufacturing is a sales-weighted harmonic mean of the Manufacturing local establishment's (48). In both figures, county size is defined as total labor income. Both bin-scatters consider 50 equally sized county bins according to their total labor income.

This section outlined the empirical approach to estimating markups using U.S. micro-data and showed the resulting markup estimates. There is significant cross-section heterogeneity in markups across local producers. Moreover, there is also a significant heterogeneity in the markups across cities, with bigger cities having lower markups than smaller ones. The results support the idea that local competition and local producers' productivity vary tremendously across space and, in turn, offer empirical support for the economic forces proposed in the theoretical framework. We now turn to the quantitative investigation that measures the welfare effects of place-based policies.

5 Quantitative Analysis

In this section, I estimate the model and use it to quantify the general equilibrium effects of place-based policies. Similarly to the empirical analysis section, I define a city in the model as a county. Focusing on the continental U.S., the quantitative exercises consider 3,080 counties. Even though the different forces highlighted in the model may act differently across sectors, as shown in 3(a), I estimate the parameters for establishments in all local industries. Similarly, the counterfactual exercises abstract from sector heterogeneity across local industries.

5.1 Model Estimation

The model has 12 parameters, which I divide into three groups. Three parameters in the first group are externally calibrated using standard values from the literature. The second group comprises five parameters estimated using Generalized Method of Moments (GMM). The model delivers estimating equations for each of the parameters. A Simulated Method of Moments (SMM) routine estimates four parameters in the third group. I target the establishment's average employment

Table 4: First Group: External Calibration

Parameter	Description	Source	Value
α	Housing expenditure share	Davis and Ortalo-Magne (2011)	0.24
ϕ	Housing supply elasticity	Saiz (2010)	1.75
θ	Dispersion location preferences	Fajgelbaum et al. (2018)	1.73

Notes: Table 4 displays parameter values for the first estimation block.

across counties displayed in Figure 1(b) and the economy-wide aggregate markup.

City exogenous characteristics, traded good productivity and amenities, are recovered non parametrically by exactly matching employment and average wages per county.[49]

Externally Calibrated Parameters (3 parameters). This group has three parameters: the housing expenditure share α, buildings supply elasticity ϕ, and the idiosyncratic location preference tastes dispersion, θ.

Table 4 summarizes the values for the three parameters. The housing expenditure share takes the value reported by Davis and Ortalo-Magne (2011), $\alpha = 0.24$. The housing supply elasticity is set to $\phi = 1.75$, the unweighted median elasticity of Saiz (2010). Finally, I set the dispersion of the location idiosyncratic preference tastes to $\theta = 1.73$, which is the baseline value estimated by Fajgelbaum et al. (2018) for the U.S.

GMM Estimated Parameters (5 parameters). There are five parameters in this group: the local goods expenditure share η, the Kimball demand parameters $\overline{\sigma}$, ε, and the local and traded good producer output elasticities, β and γ. Table 5 summarizes the results.

The local goods expenditure share, η, is estimated using the local goods sales share reported in Table 1. Given the housing expenditure share value, $\alpha = 0.24$, an aggregate sales share of 52% implies a value of $\eta = 0.39$.

I estimate the Kimball demand parameters in two steps. I provide a summary of the variation that allows me to estimate these parameters. Appendix D.4 provides the estimation details. First, the Kimball preferences imply that markups and sales shares are related through a log-linear equation. From this equation, and using the markup estimates and data on sales shares, I recover an estimate of the ratio $\varepsilon/\overline{\sigma}$. Intuitively, this ratio is estimated using within-city variation on sales shares and markups.[50]

Equipped with an estimate of the ratio $\varepsilon/\overline{\sigma}$, I develop an iterative GMM procedure that estimates $\overline{\sigma}$. Using a similar logic to the one used in equation (42), the Kimball demand implies a system of equations for relative quantities and price indices as functions of sales market shares. Given an initial guess for $\overline{\sigma}$, I solve this system to obtain relative quantities. Then, I use the relative

[49] For the counterfactuals, I use the non-parametric estimates to fit a joint log normal distribution for a and b.

[50] Edmond, Midrigan, and Xu (2023) use a similar strategy to estimate $\varepsilon/\overline{\sigma}$. The key difference between their estimation and mine is that I use within-city variation in sales shares and markups. Because Edmond, Midrigan, and Xu (2023) do not have a spatial component in their setting, they use the pooled variation in sales shares and markups for all firms in the economy.

Table 5: Second Group: GMM

Parameter	Description	Moment	Estimate
η	Local goods expenditure share	Local establishments sales share	0.39
ε	Demand super-elasticity	Markups and sales share	1.38
$\overline{\sigma}$	Demand elasticity	Markups and implied relative quantities	2.26
β	Labor output elasticity (local)	Local establishments labor FOC	0.22
γ	Labor output elasticity (traded)	Traded establishments labor FOC	0.29

Notes: Table 5 summarizes the GMM estimation results for the second block of parameters.

quantities to compute an implied markup, using (16) and (41). I estimate $\overline{\sigma}$ by minimizing the distance between the implied markups and the markups estimates from (46).

Lastly, we turn to the estimation of the labor output elasticities. On the one hand, for the establishments in local industries, I estimate β from the markup estimation equation (46). On the other hand, I estimate γ for the establishments in traded industries from (23). This equation states that γ equals the labor cost share of sales.

SMM Estimated Parameters (four parameters). The parameters in the last group are the ones governing the local producers' productivity distribution in (32), z_L, z_H, and δ, in addition to the entry cost c_e. Table 6 displays the estimated parameters and the goodness of fit.

I estimate the four parameters via SMM. I target the establishment's average employment reported in Table 1(b). However, to avoid taking a stand on the units in which labor is measured in the model, I target the average establishment employment for counties in the 25th, 50th, 75th, and 95th percentiles relative to the average establishment employment in the 5th percentile. This yields four moments. Additionally, I target the economy-wide aggregate markup for local industries. Following Yeh, Macaluso, and Hershbein (2022), I define the economy-wide markup as a population-weighted average of the county-level markups. I compute this object using the markup estimates from Section 4. In total, I am over-identified by having five moments and four parameters.

Formally, let $\Theta = (z_L, z_H, \delta, c_e)$ be the vector of parameters to estimate. I implement the SMM by minimizing the squared percent distance between the model-simulated moments, $\Psi^m(\Theta)$, and their empirical counterparts, Ψ^d:

$$\min_{\Theta} \sum_{i=1}^{5} \left(\frac{\Psi_i^m(\Theta) - \Psi_i^d}{0.5\left(\Psi_i^m(\Theta) + \Psi_i^d\right)} \right)^2 .$$

I employ the TikTak algorithm for global optimization of Arnoud, Guvenen, and Kleineberg (2019) to search over the parameter space.[51] In every iteration of the optimization routine, I invert the model to recover non-parametric estimates of a_c and b_c using (11) and (22). Appendix D.5 discusses the inversion procedure.

Even though all parameters are jointly identified, it is possible to shed light on which moments

[51] I use 2000 starting points and a simplex search method for local optimization.

Table 6: Third Group: SMM

Moments		Model	Data
Ratio avg. estab. employment p25/p5		1.62	1.31
Ratio avg. estab. employment p50/p5		1.67	1.50
Ratio avg. estab. employment p75/p5		1.80	1.87
Ratio avg. estab. employment p95/p5		2.21	2.23
Aggregate markup		2.19	3.28
Description	Parameter	Estimate	
Min. productivity	z_L	10.61	
Max. productivity	z_H	59.44	
Shape parameter	δ	4.850	
Entry cost	c_e	0.041	

Notes: Table 6 presents SMM estimation results and goodness of fit for the third block of parameters.

help to identify each parameter. The counties' average employment relative to the smaller counties identifies the local producers' productivity distribution parameters: z_L, z_H, and δ. Conversely, the aggregate markup identifies the entry cost.

Table 6 shows the goodness of fit of the SMM estimation. The model is flexible enough to match the average employment of counties in the 25th, 50th, 75th, and 95th percentiles of county size distribution relative to counties in the 5th percentile. Nevertheless, the model falls short when matching the aggregate markup. Solving the decentralized equilibrium involves challenging fixed point algorithms with systems of highly non-linear equations within them. Improving the match between the model aggregate markup and the one estimated in the data is still a work in progress.

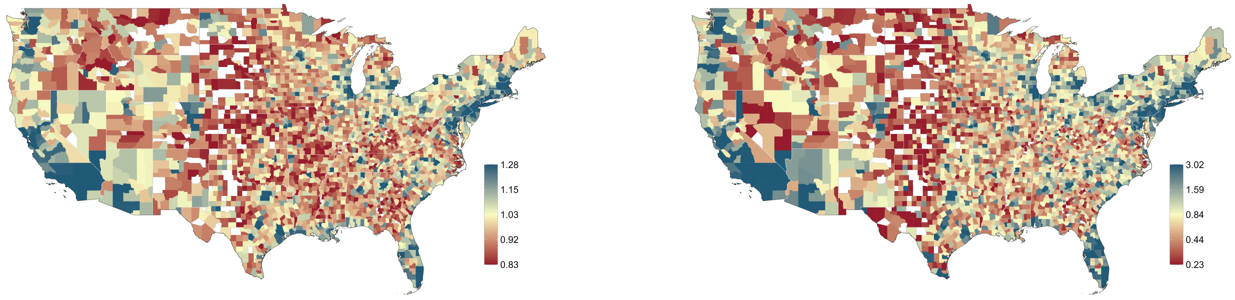
5.2 Model to the Data: the Decentralized Equilibrium

In this section, I solve for the decentralized equilibrium using the estimated parameters and show that the model can quantitatively account for spatial markup differences.

First, Figure 4 displays the results from the model inversion. Local productivity of the traded good is displayed on the left panel, while local amenities are displayed on the right. Overall, both county characteristics are highly correlated. Because the inversion exactly matches wages and population, the model rationalizes high-wage counties as having high traded good productivity. These counties are typically in the upper east coast, southern Florida, the Midwest, and southern California. On the other hand, counties with large populations are rationalized to have higher local amenities. In contrast, a significant fraction of the southern counties have high local estimated amenities.

Second, Figure 5 displays the model implied markups. It is worth mentioning that even though the economy-wide aggregate markup is one of the targeted moments, the markup's cross-sectional variation is not constrained by the estimation. Therefore, this figure serves as an over-identifying exercise. Figure 5(a) is the model equivalent of Figure 2. On the one hand, the model can qualitatively replicate the negative relationship between county aggregate markup and county size. However, on the other hand, the model estimated elasticity of aggregate markup with respect

Figure 4: Local Productivity and Local Amenities



(a) Local Productivity

(b) Local Amenities

Notes: Figure 4(a) shows the model implied local traded good productivity and Figure 4(b) shows the model implied local amenities. Counties omitted in the analysis are not colored.

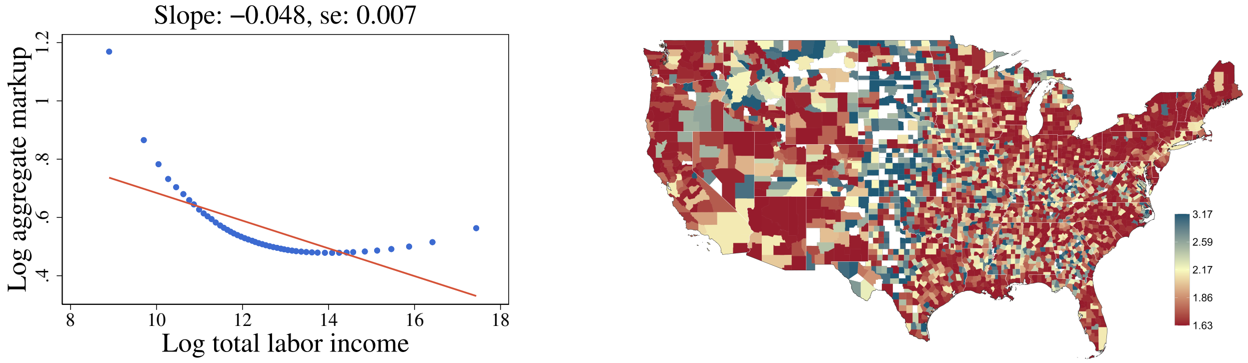
to county size is -0.048, which is lower than the one estimated with the data. Indeed, there is a level effect that the model fails to capture, and therefore, it predicts markups somewhat low. Nonetheless, the model can capture the relative markup difference between the smallest and biggest counties illustrated in Figure 2: counties in the top decile of total labor income distribution exhibit a markup 50% lower than counties in the bottom decile.

Figure 5(b) shows the model implied markups for all counties in the U.S. Small counties typically located in the south and central parts of the U.S. have markups from 2.17 to 3.17. On the contrary, big counties like Manhattan, Chicago, or Los Angeles display markups that are almost twice as small as those in small locations.

Figure 6 shows that the model can qualitatively account for different empirical regularities across cities. The blue solid line across panels illustrates different economic outcomes in decentralized equilibrium (Laissez-faire case). First, in line with the findings of Combes et al. (2012), the model predicts that bigger cities are more productive. The productivity advantage of bigger cities comes through two channels: it attracts more productive firms and displays higher local TFP. Equation (33) shows that local TFP accounts for the productivity of local producers, but also it increases with the number of firms in a location. Indeed, TFP in bigger counties doubles TFP in smaller counties. Second, the bottom-left panel displays the local varieties price index, $\mathbb{P}(c)$. Consistent with the findings of Handbury and Weinstein (2014), bigger cities have a lower price index. The model also accounts for the fact that bigger counties have higher housing rents.

Figure 6 also shows the cross-sectional differences in local competition across counties. The top-right panel documents a significant heterogeneity in the competition index across counties in the estimated model. As illustrated by the central panels at the bottom of the figure, the tougher competition in these locations is partially explained by intense competition in the inputs markets. Furthermore, firm prices in such locations are also lower than in smaller counties, magnifying the differences in the local competition index.

Figure 5: Markups in the Decentralized Equilibrium



(a) County agg. markup and county size

(b) Markups across counties

Notes: Figure 5(a) shows the model equivalent of Figure 2. Figure 5(b) shows markups across counties in the decentralized equilibrium. Counties omitted in the analysis are not colored.

5.3 Place-based Policy Counterfactual

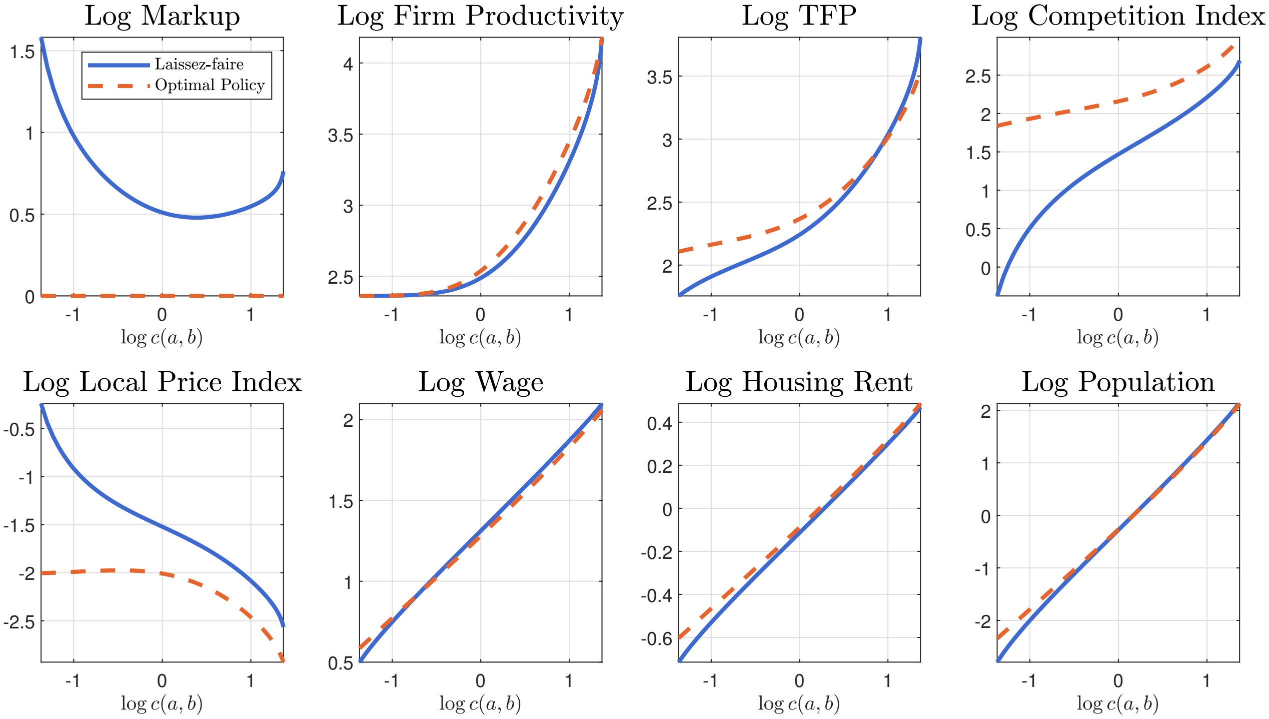
This final section studies a policy counterfactual. Formally, I investigate the aggregate effects of implementing the optimal policy of Proposition 4.

The location-specific subsidy (35) implements the optimal policy and achieves the first-best allocation. Recall that this transfer corrects three margins of inefficiency in the decentralized equilibrium: removes output price distortions, corrects the inefficient location of local producers, and generates an efficient aggregate entry rate. A non-distortionary tax on workers finances this subsidy. Moreover, I use the equivalent formulation of the baseline model in which profits from local developers are rebated back to workers in a non-distortionary flat earnings subsidy.

Figure 6 displays the cross-sectional patterns of the equilibrium under the laissez-faire and the optimal policy. Consistent with the results from Proposition 4, the optimal policy removes markups in all locations. Furthermore, the policy makes marginally productive producers relocate to smaller locations. More than 90% of the counties experience a productivity boost due to this policy. Nonetheless, this comes at the expense of productivity losses in larger locations. By removing markups, the policy also makes the price index fall everywhere. Nevertheless, the price index in smaller counties experience a more prominent decrease because of two channels. First, these locations initially had higher markups and, therefore, experience more considerable reductions in prices. Second, as these locations experience an influx of producers, the increase in local varieties also causes the price index to fall relatively more than in larger locations.

As a result of the policy, smaller cities expand. The bottom half of Figure 6 shows that smaller cities experience an increase in wages, housing rents, and population. The spatial reallocation of firms increases labor demand in smaller cities, which creates an upper pressure on wages and causes a relocation of workers to such locations. In larger counties, this spatial reconfiguration slightly decreases wages but has milder effects on population. Interestingly, housing rents increase in all locations. The reason is that reducing markups induces firms to increase production and augment their input demand. As firms demand more commercial structures, housing rents rise.

Figure 6: Model's solution in the Decentralized Equilibrium and the Optimal Policy

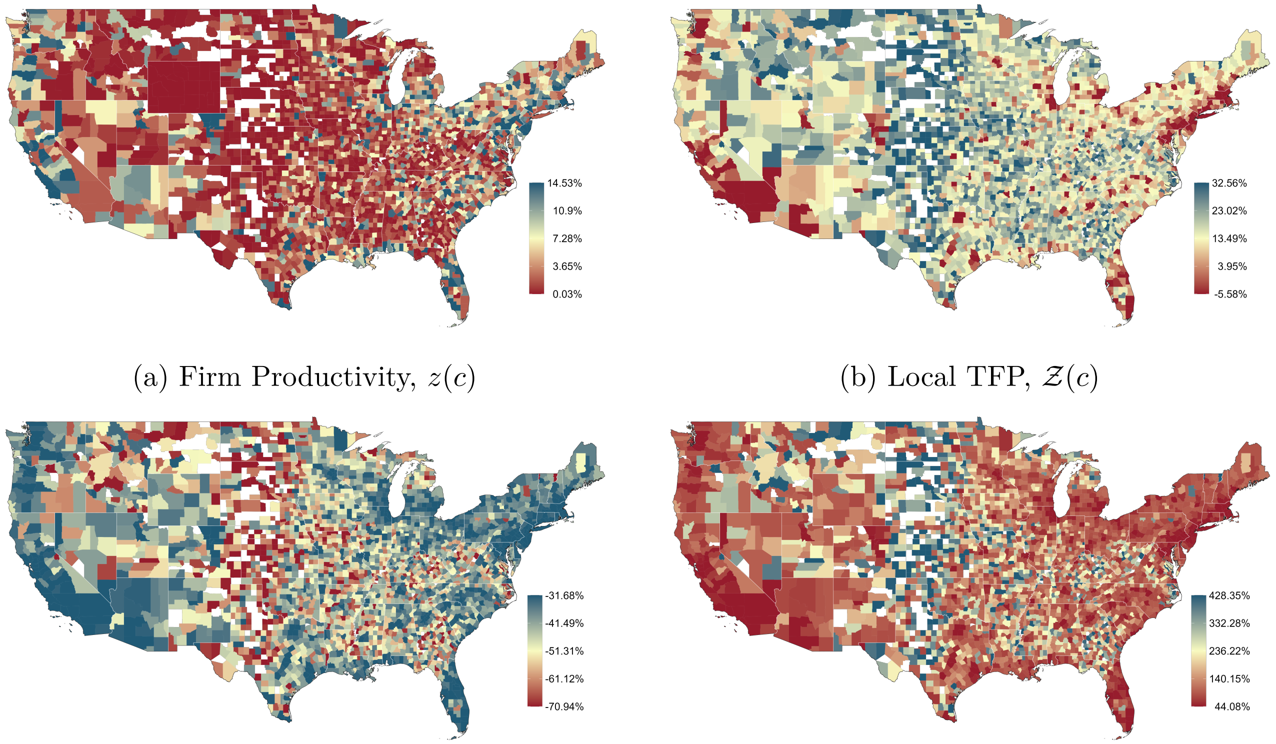


Notes: Figure 6 shows model solution in the decentralized equilibrium and under the optimal policy for all U.S. local industries. The x-axis represents the city's estimated appeal $c(a,b)$.

To highlight the spatial effects of the optimal policy, Figure 7 maps changes at the local level for different equilibrium outcomes. First, Figure 7(a) shows the change in productivity of local producers. There are two crucial messages this graph conveys. First, the local productivity level in the biggest and smallest counties remains unchanged. Local producers in Manhattan or rural small counties do not change their location decisions due to the policy. The reason is that positive assortative matching still holds in the optimal policy. Therefore, the very best producers are still located in the bigger cities, and the very least productive in the smallest. Second, counties that experience the most significant productivity increases are those located next to the most extensive locations. Indeed, the productivity of local producers in adjacent counties to large urban areas like Miami, Chicago, and Los Angeles increases by almost 15%. This reflects the "top-down" effect of the optimal policy: marginal producers relocate from big cities to marginally smaller ones. This reallocation effect also occurs once we move down across smaller counties.

Figure 7(b) shows the change in local TFP. In addition to accounting for the productivity level of local producers, total TFP also accounts for the total number of firms. Interestingly, as the policy also changes the number of producers in each city, changes in local productivity are magnified through the number of local producers, which leads to more significant changes in total TFP. On the one hand, smaller counties, primarily located in the inner part of the country, experience a significant increase in TFP. For instance, Floyd, TX, reaches a productivity increase of 30%. On the other hand, highly populated counties like Los Angeles suffer a mild decrease of around 5%. Changes in the number of producers primarily drive changes in TFP for the biggest and smallest counties. Interestingly, Figure 7(b) shows the pattern that rural counties seem to be the ones

Figure 7: Changes in firm productivity, local TFP, competition index and ideal price index.



(a) Firm Productivity, $z(c)$

(b) Local TFP, $\mathcal{Z}(c)$

(c) Ideal Price Index, $\mathbb{P}(c)$

(d) Competition Index, $\mathbb{C}(c)$.

Notes: Figures show the percentual changes in local equilibrium objects in the optimal policy relative to the Laissez-faire scenario. Figure 7(a) shows the percentual change in local firm productivity. Figure 7(b) displays the percentual change in local TFP. Figure 7(d) illustrates the percentual change in the competition index, and Figure 7(c) describes the percentual change in the ideal price index.

that benefit the most from the policy. Along these lines, the policy speaks to the discussion of Urban vs Rural development, suggesting that the policies that aim to boost commercial activity in under-developed rural areas may be beneficial.

Figure 7(c) displays the spatial heterogeneity in local prices. Counties in the coastal parts of the country and around the Great Lakes region experience a reduction in the local prices of around 30%. Strikingly, southern and central counties witness a reduction in the price index close to 70%. There are two reasons behind this significant disparity in price reduction. First, the small counties are experiencing a more considerable reduction in markups. In such locations, markups go from 3.1 to 1. Second, these counties are the ones experiencing a large influx of new local producers, which further decreases the local price index.

The last panel, Figure 7(d), displays the change in the competition index $\mathbb{C}(c)$. Recall that the index $\mathbb{C}(c)$ has two components. On the one hand, it reflects the prices of local producers. As seen from Figure 7(c), local prices fall everywhere because firms no longer charge a markup over marginal cost. Therefore, local competition increases everywhere as local producers charge lower prices. On the other hand, the competition index also captures the price of the local inputs: labor and structures. Because the policy incentivizes firms to increase production, all firms demand more of these inputs, causing wages and land rents to increase. The rise in local input prices further increases competition in all counties. Nonetheless, similarly to the pattern in Figure 7(c), small counties experience a larger increase in competition as they experience higher price reductions.

The results from the counterfactual exercise illustrate the benefits of a policy that reallocates producers from big to small cities. Qualitatively, these results are similar to the ones in Bilal (2023). Nevertheless, the rationale for such types of policy in Bilal (2023) is different from the ones considered in this study. While the rationale in Bilal (2023) comes from labor market frictions, I offer theoretical and empirical support for these policies based on the premise that output market power causes local producers to locate inefficiently. Both papers contrast with the findings Gaubert (2018), who finds that incentivizing producers to locate in smaller locations is detrimental because of agglomeration externalities. A potentially interesting future research avenue would be explicitly combining to explicitly combine all these mechanisms and asses their relative importance.

Finally, the optimal policy yields an aggregate welfare gain of 2.36%. This magnitude is within the range of results of both Bilal (2023) and Gaubert (2018). However, the gain in welfare is lower than in Edmond, Midrigan, and Xu (2023). Of course, the framework they consider differs significantly from the one in this paper, with the spatial component being the key difference.

6 Conclusion

This paper has developed a new theory of endogenous competition across cities. The theory sheds light on the mechanisms that govern the ability of local producers to exert output market power. Differences in markups across space arise due to differentials in local competition and the productivity of local firms. Pricing complementarities are the central driver of the location choice of heterogeneous producers. As a result, more productive firms over-value locating in bigger cities, and spatial misallocation arises. This view emphasizes that relocating production from bigger to smaller cities increases aggregate welfare.

The paper also provides empirical evidence on markup heterogeneity across the U.S. The structure of my model allows me to estimate markups for all the establishments that operate in local markets. Producers in larger cities have significantly lower markups than producers in smaller cities. This empirical regularity is informative of the degree of firm spatial misallocation.

Finally, I use the model to quantify the welfare gains of place-based policies. Policies that eliminate markups yield sizable welfare gains by eliminating price distortion and relocating firms from big to small cities. The view that output market power creates firm misallocation across cities helps to reconcile the intuition of place-based policies.

The methodology proposed in this paper can readily be used to study the determinants of local competition across different sectors. I empirically illustrate that different sectors have different patterns for markups across cities. These patterns suggest that the magnitude of the economic forces that determine firm location and local competition might differ across sectors. This has implications for the degree of spatial misallocation across sectors. The degree of firm misallocation across cities would be worse in sectors where markups are higher in bigger cities. Studying general equilibrium counterfactuals for different sectors is left for future research. A quantitative assessment of the general equilibrium effects of place-based policies for different sectors could be used to inform industrial policy.

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. 2018). *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. (Visited on 10/25/2023).

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.

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.

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.

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.

Chen, Xiaohong (2007). “Chapter 76 Large Sample Sieve Estimation of Semi-Nonparametric Models”. In: ed. by James J. Heckman and Edward E. Leamer. Vol. 6. Handbook of Econometrics. Elsevier, pp. 5549–5632.

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.

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

Edmond, Chris, Virgiliu Midrigan, and Daniel Yi Xu (2023). “How Costly Are Markups?” In: *Journal of Political Economy* 131.7, pp. 1619–1675.

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.

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.

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.

Handbury, Jessie and David E. Weinstein (Sept. 2014). “Goods Prices and Availability in Cities”. In: *The Review of Economic Studies* 82.1, pp. 258–296.

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 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 (2023). “Wage Inequality and the Spatial Expansion of Firms”. In.

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

Levinsohn, James and Amil Petrin (Apr. 2003). “Estimating Production Functions Using Inputs to Control for Unobservables”. In: *Review of Economic Studies* 70, pp. 317–341.

Mankiw, N. Gregory and Michael D. Whinston (1986). “Free Entry and Social Inefficiency”. In: *The RAND Journal of Economics* 17.1, pp. 48–58. (Visited on 10/21/2023).

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.

Melitz, Marc J. and Gianmarco I. P. Ottaviano (Jan. 2008). “Market Size, Trade, and Productivity”. In: *The Review of Economic Studies* 75.1, pp. 295–316.

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.

Olley, G. Steven and Ariel Pakes (1996). “The Dynamics of Productivity in the Telecommunications Industry”. In: *Econometrica* 64.6, pp. 1263–1297.

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 Nicholas Trachter (2020). “Diverging Trends in National and Local Concentration”. In.

Saiz, Albert (Aug. 2010). “The Geographic Determinants of Housing Supply*”. In: *The Quarterly Journal of Economics* 125.3, pp. 1253–1296.

Trottner, Fabian (2023). “Unbundling Market Power”. In.

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

A Derivations

A.1 Local Goods Demand

For any given $Y(c)$ consumer minimize total expenditure on local goods, subject to the utility constraint (2):

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

where $\lambda^W(c)$ is a Lagrange multiplier. The first order condition with respect to the consumption of a single variety $y(z,c)$ is:

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

Defining the competition price index as:

$$\mathbb{D}(c) \equiv \frac{\lambda(c)}{Y(c)},$$

we can write the inverse demand for a single variety $y(z,c)$ as:

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

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

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

With these expressions, the competition price index $\mathbb{D}(c)$ is given by:

$$\int_z \Upsilon\left(\varphi\left(\frac{p(z,c)}{\mathbb{D}(c)}\right)\right)dG_c(z) = 1$$

Finally, we can express the ratio of $\mathbb{D}(c)$ and $\mathbb{P}(c)$ as a function or relative prices as

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

A.2 Input Demands

Input demands follow from the firms cost minimization problem. Taking input prices as given, firms minimize total input expenditure subject to a certain level of production. The Lagrangian associated to this problem is:

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

where $\lambda(z,c)$ is a Lagrange multiplier that equals the marginal cost of production. Taking first-order conditions and recognizing that $\mu(z,c) \equiv p(z,c)/\lambda(z,c)$ we obtain:

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

Substituting in the optimal relative price, $\psi(\mathbb{C}(c)/z)$, and the optimal relative quantities $\varphi(\psi(\mathbb{C}(c)/z))$ gives (19).

A.3 Klenow and Willis (2016) Derivations

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

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

Moreover, the optimal relative price function is

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

where λ is a constant, $\Omega(\cdot)$ is the main branch of the Lambert-W function.[52] The optimal relative quantity is

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

Note that, as in any model of monopolistic competition, local producers will never producer in the inelastic area of the demand curve. Formally, producers set $p(c,z)/\mathbb{D}(c)$ such that

$$\sigma\left(\frac{p(z,c)}{\mathbb{D}(c)}\right) \geq 1 \iff \frac{p(z,c)}{\mathbb{D}(c)} \geq \exp\left(\frac{1+\varepsilon\log\frac{\overline{\sigma}-1}{\overline{\sigma}} - \overline{\sigma}}{\varepsilon}\right)$$

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

Similarly, the Klenow and Willis (2016) has a choke price given by the requirement that optimal relativa quantities are positive. From the expression for $y(z,c)/Y(c)$, the condition on the relative price is

$$\frac{p(z,c)}{\mathbb{D}(c)} \leq \exp\left(\frac{1+\varepsilon \log \frac{\overline{\sigma}-1}{\overline{\sigma}}}{\varepsilon}\right)$$

B Proofs

B.1 Proof of Proposition 1

The proof of Proposition 1 is structured in three steps. First, we show that there is positive assortative matching conditional on the location's competition index $\mathbb{C}(c)$. Second, we characterized general equilibrium objects when firms sort based on competition. Third, we construct a location's single index which is a sufficient statistic for the firm's location decisions. Then we 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. Note that the optimal relative price and hence the optimal markup depend directly on the location c only through the aggregate object $\mathbb{C}(c)$. Hence, following the insights of Bilal (2023), we index locations by their competition index $\mathbb{C}$ rather than by their appeal c.[53] In doing so, we momentarily consider the inverse function $c(\mathbb{C})$. The profit function (3) takes the form:

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

where substituted in the optimal relative prices and quantities from the profit maximization problem. Under this alternative formulation, firms sort based on the competition index $\mathbb{C}$ rather than the appeal of a city c. To prove that there is stricly increasing assignment between $\mathbb{C}$ and z, we use the methods of standard assignment problems (i.e. Galichon (2016)). In particular, note that because of the envelope theorem,

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

Moreover, replacing the expression from the first-order condition (13) implies:

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

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

Thus,

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

We see that the profit function is log-sumpermodular if and only if the elasticity of demand is increasing the relative price. Therefore, under Marshall's second law, which holds in the Klenow and Willis (2016) specification, there is a strictly assignment function between z and $\mathbb{C}$, $z(\mathbb{C})$.

Step 2: general equilibrium objects. We now derive expressions for the general equilibrium objects under the PAM between z and $\mathbb{C}$ result. First, using the tools from Costinot and Vogel (2010), the definition of the competition price index (8) implies:

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

where $f_{\mathbb{C}}(\mathbb{C})$ is the equilibrium density of $\mathbb{C}$. With this expression, the ratio $P(\mathbb{C})/D(\mathbb{C})$ in (A1) is given by:

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

To ease notation, I define $\delta(\mathbb{C}/z(\mathbb{C}))$ as the inverse of the ratio above:

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

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

$$\begin{aligned} P(\mathbb{C}) &= D(\mathbb{C})\delta\left(\frac{\mathbb{C}}{z(\mathbb{C})}\right), \\ &= \nu\frac{W(\mathbb{C})^{\beta}R(\mathbb{C})^{(1-\beta)}}{\mathbb{C}}\delta\left(\frac{\mathbb{C}}{z(\mathbb{C})}\right) \end{aligned} \tag{B3}$$

On the other hand, labor labor supply (11) implies that market size, $M(\mathbb{C})$, is expressed as:

$$M(\mathbb{C}) = \frac{\eta\overline{L}}{\overline{U}}\frac{b(\mathbb{C})^{\theta}W(\mathbb{C})^{1+\theta}}{P(\mathbb{C})^{\eta\theta}R(\mathbb{C})^{\alpha\theta}} \tag{B4}$$

The housing land market clearing condition is:

$$R(\mathbb{C})^{\phi} = \frac{\alpha}{\eta}\frac{M(\mathbb{C})}{R(\mathbb{C})} + (1-\gamma)\frac{Q^T(\mathbb{C})}{R(\mathbb{C})} + (1-\beta)\frac{M(\mathbb{C})}{\mu\left(\frac{\mathbb{C}}{z(\mathbb{C})}\right)R(\mathbb{C})}, \tag{B5}$$

where the terms on the right-hand side correspond to worker's, traded good producers, and local producers total housing consumption, respectively. Using the labor input demands, local labor market clearing implies:

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

Using the definition of $M(\mathbb{C})$ and solving for the traded-good production gives:

$$Q^T(\mathbb{C}) = \frac{M(\mathbb{C})}{\gamma}\left[\frac{1}{\eta} - \frac{\beta}{\mu\left(\frac{\mathbb{C}}{z(\mathbb{C})}\right)}\right] \tag{B6}$$

Therefore, replacing the above expression for $Q^T(\mathbb{C})$ into (B5) gives:

$$R(\mathbb{C})^{\phi} = \frac{\alpha}{\eta}\frac{M(\mathbb{C})}{R(\mathbb{C})} + \frac{(1-\gamma)}{\gamma}\frac{M(\mathbb{C})}{R(\mathbb{C})}\left[\frac{1}{\eta} - \frac{\beta}{\mu\left(\frac{\mathbb{C}}{z(\mathbb{C})}\right)}\right] + \frac{1-\beta}{\mu\left(\frac{\mathbb{C}}{z(\mathbb{C})}\right)}\frac{M(\mathbb{C})}{R(\mathbb{C})}$$

Solving for $R(\mathbb{C})$ gives the equilibrium land rents:

$$R(\mathbb{C}) = M(\mathbb{C})^{\frac{1}{1+\phi}}\chi\left(\frac{\mathbb{C}}{z(\mathbb{C})}\right)^{\frac{1}{1+\phi}}, \tag{B7}$$

with

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

The zero profit condition of the traded good producers imply that equilibrium wages are given by:

$$W(\mathbb{C}) = \left(\frac{a(\mathbb{C})}{R(\mathbb{C})^{(1-\gamma)}\varrho}\right)^{\frac{1}{\gamma}}. \tag{B8}$$

Substituting (B7), (B3), and (B8) into the market size expression (B4) and solving for $M(\mathbb{C})$ gives:

$$M(\mathbb{C}) = \left[\frac{\eta\overline{L}}{\overline{U}\nu^{\eta\theta}}\frac{\mathbb{C}}{\delta\left(\frac{\mathbb{C}}{z(\mathbb{C})}\right)\chi\left(\frac{\mathbb{C}}{z(\mathbb{C})}\right)^{\xi-1}}a(\mathbb{C})^{\frac{1+\theta(1-\eta\beta)}{\gamma}}b(\mathbb{C})^{\theta}\right]^{\frac{1}{\xi}}, \tag{B9}$$

where $\xi \equiv \frac{\gamma(1+\phi+\theta(\alpha+\eta(1-\beta)))+(1-\gamma)(1+\theta(1-\eta\beta))}{\gamma(1+\phi)}$.[54]

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:

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

$$\sigma\left(\psi\left(\frac{\mathbb{C}}{z(\mathbb{C})}\right)\right) - 1 = \mathcal{E}_{\delta}(\mathbb{C}) + \mathcal{E}_{M}(\mathbb{C}), \tag{B10}$$

where the first equation comes from (B2) (definition of competition price index) and the second from the first-order condition of the firm's location problem (B1). 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 (B2) and (B9). The the term on the left-hand side and the first term on the right-hand side depends on primitives of the model, $z(\mathbb{C})$ and $\mathbb{C}$. On the contrary, the last term on the right-hand side, depends on primitives of the model, $z(\mathbb{C})$, $\mathbb{C}$ 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 good producers make their sorting decisions based on a uni-dimensional index $c(a, b)$ rather than considering the two dimensions of location heterogeneity separately:

$$c = c(a, b) \equiv a^{\frac{1+\theta(1-\eta\beta)}{\gamma}} b^{\theta}$$

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 $c(\mathbb{C})$ is increasing in general equilibrium. We can use (B9) to write the first-order condition (B10) as:

$$\sigma\left(\psi\left(\frac{\mathbb{C}}{z(\mathbb{C})}\right)\right) - 1 = \mathcal{E}_{\delta}(\mathbb{C}) + \frac{1}{\xi}\left[1 + \mathcal{E}_{c}(\mathbb{C}) - \mathcal{E}_{\delta}(\mathbb{C}) - (\xi - 1)\mathcal{E}_{\chi}(\mathbb{C})\right],$$
$$= \frac{\xi - 1}{\xi}\left[\mathcal{E}_{\delta}(\mathbb{C}) - \mathcal{E}_{\chi}(\mathbb{C})\right] + \frac{1}{\xi}\left[1 + \mathcal{E}_{c}(\mathbb{C})\right]$$

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

$$\overline{\sigma} - 1 = \frac{1}{\xi}\left[1 + \mathcal{E}_{c}(\mathbb{C})\right],$$

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

which implies that $\mathcal{E}_c(\mathbb{C}) > 0$ if $\xi\left(\overline{\sigma}-1\right) > 1$. Therefore, when $\xi\left(\overline{\sigma}-1\right) > 1$, there exists a region of the space parameter where ε is small and PAM is obtained in general equilibrium: both $z(\mathbb{C})$ and $c(\mathbb{C})$ are strictly increasing, and therefore more productive firms locate in more appealing cities.

B.2 Proof of Proposition 2

The proof 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, s and $\overline{U}$. Third, show existence and uniqueness of general equilibrium objects.

Step 1: ODE system for local goods producers. Impose Assumption 1, and the assumptions of Proposition 1. Then, PAM between firm and location productivity is obtained. The definition of the competition price index (8) implies:

$$z'(c) = \frac{M_e f(c)}{g(z(c))} \frac{1}{\Upsilon\left(\varphi\left(\psi\left(\frac{\mathbb{C}(c)}{z(c)}\right)\right)\right)} \tag{B11}$$

On the one hand, the FOC of the location problem (21) is:

$$\mathcal{E}_{\mathbb{C}}(c) = \left[\mu\left(\psi\left(\frac{\mathbb{C}(c)}{z(c)}\right)\right) - 1\right]\left[\mathcal{E}_\delta(c) + \mathcal{E}_M(c)\right] \tag{B12}$$

On the other hand, (B9) gives:

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

We start deriving expressions for each of the terms in the expressions above. First, equation (A1) implies:

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

with

$$\Theta\left(\frac{\mathbb{C}(c)}{z(c)}\right) = \rho\left(\frac{\mathbb{C}(c)}{z(c)}\right)\left[\sigma\left(\psi\left(\frac{\mathbb{C}(c)}{z(c)}\right)\right) - 1\right]\left[\mu\left(\frac{\mathbb{C}(c)}{z(c)}\right)/\delta\left(\frac{\mathbb{C}(c)}{z(c)}\right) - 1\right]. \tag{B13}$$

Second, the definition of $\chi(\cdot)$ implies:

$$\mathcal{E}_\chi(c) = \Lambda\left(\frac{\mathbb{C}(c)}{z(c)}\right)\mathcal{E}_{\mathbb{C}}(c) - \Lambda\left(\frac{\mathbb{C}(c)}{z(c)}\right)\mathcal{E}_z(c),$$

with

$$\Lambda\left(\frac{\mathbb{C}(c)}{z(c)}\right) \equiv \frac{\left(1-\rho\left(\frac{\mathbb{C}(c)}{z(c)}\right)\right)}{\left(\alpha+\frac{1-\gamma}{\gamma}\right)\mu\left(\frac{\mathbb{C}(c)}{z(c)}\right)+\eta\left((1-\beta)-\frac{\beta(1-\gamma)}{\gamma}\right)} \tag{B14}$$

Combining these expressions gives

$$\mathcal{E}_M(c)=\frac{\eta\theta}{\xi}\mathcal{E}_{\mathbb{C}}(c)+\frac{\eta\theta}{\xi}\Theta\left(\frac{\mathbb{C}(c)}{z(c)}\right)\left[\mathcal{E}_z(c)-\mathcal{E}_{\mathbb{C}}(c)\right]-\frac{\xi-1}{\xi}\Lambda\left(\frac{\mathbb{C}(c)}{z(c)}\right)\left[\mathcal{E}_{\mathbb{C}}(c)-\mathcal{E}_z(c)\right]+1,$$

Moreover, the definition of the competition index further implies that $\mathcal{E}_{\mathbb{D}}(c)=\beta\mathcal{E}_W(c)+(1-\beta)-\mathcal{E}_R(c)-\mathcal{E}_{\mathbb{C}}(c)$. Finally, the elasticity of wages $\mathcal{E}_W(c)$ is given by (22). Combining these conditions into the location FOC, gives:

$$\begin{aligned}\frac{C'(c)c}{\mathbb{C}(c)}&\left[\alpha\left(\xi\left(\frac{\mathbb{C}(c)}{z(c)}\right)-1\right)+(\alpha+\eta)\Theta\left(\frac{\mathbb{C}(c)}{z(c)}\right)+\frac{\xi-1}{\xi}\Lambda\left(\frac{\mathbb{C}(c)}{z(c)}\right)-\eta\theta\right]\\&=\xi+\left[(\alpha+\eta)\Theta\left(\frac{\mathbb{C}(c)}{z(c)}\right)+\frac{\xi-1}{\xi}\Lambda\left(\frac{\mathbb{C}(c)}{z(c)}\right)\right]\frac{z'(c)a}{z(c)},\end{aligned} \tag{B15}$$

where we explicitly wrote the expression for the elasticities $\mathcal{E}_z(c)$ and $\mathcal{E}_{\mathbb{C}}(c)$. Equations (B11) - (B15) define a coupled system of ODE's, with two boundary conditions $z(\overline{c})=\overline{z}$ and $z(\underline{c})=\underline{c}$. The first boundary condition states that the most productive firms go to the most appealing cities locations, while the second condition implies that least productive firms locate in tge least appealing cities. The solution to this ODE system is the assignemnt function $z(c)$, the competition index function $\mathbb{C}(c)$, and the market size function $M(c)$ that determine the equilibrium in the local sector, given M_e and $\overline{U}$.

Step 2: Existence of a solution to the ODE system given M_e and $\overline{U}$. Inspection of the the system (B11) - (B15) indicate that the systems satisfies standard regularity conditions for a unique solution given the general equilibrium objects M_e and $\overline{U}$. In particular, the system is Lipschitz continuous. In particular, there exists $\underline{C}(M_e,\overline{U})$ such that $z(\underline{C})=\underline{z}$.[55]

Step 3: Existence of M_e and $\overline{U}$ and uniqueness On the one hand, aggregate labor market clearing condition uniquely pins down $\overline{U}$:

$$\int_c L(c)f(c)dc=\overline{L} \qquad \longrightarrow \qquad \overline{U}=\left[\int_{\underline{c}}^{\overline{c}}\left(\frac{b(c)W(c)}{\mathbb{P}(c)^{\eta}R(c)^{\alpha}}\right)f(c)c\right]^{\frac{1}{\theta}}$$

On the other hand, M_e is pin down by the traded-good market clearing condition (25):

[55] See Appendix B.5 in Bilal (2023) for a detailed discussion of the regularity conditions that guarantee a solution to these type of systems.

$$c_e M_e = \int_c \left[Q^T(c) - L(c)Q(c) - \frac{\phi}{1+\phi} R(c)^{1+\phi} \right] f(c) dc$$

Note that te LHS of the above expression is an increasing function of M_e. Now suppose that the supports of F_c and G_z are small enough. This assumption makes posible using a first-order approximation of the the equilibrium expressions for $Q^T(c)$, $L(c)Q(c)$ and $R(c)$ in (B6), (6), and (B5). These approximations imply that the RHS a decreasing function of M_e. Therefore, there exists a unique M_e such that the traded good market clearing condition is satisfied.

B.3 Proof of Corollary 1

The limiting case economy considers a scenario with a wide support for $G(\cdot)$, but a small support of $F(\cdot)$. In that case, we index locations by their competition index, $\mathbb{C}$, rather than by their appeal c. Using the The first-order condition (B10) implies:

$$\xi \left[\sigma \left(\psi \left(\frac{\mathbb{C}}{z(\mathbb{C})} \right) \right) - 1 \right] = 1 + (\xi - 1) \left[\mathcal{E}_\delta(\mathbb{C}) - \mathcal{E}_\chi(\mathbb{C}) \right] + \mathcal{E}_c(\mathbb{C}).$$

On the other hand, the definition of $\delta(\cdot)$ implies that:

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

where $\Theta(\cdot)$ is defined as in (B13). Similarly,

$$\mathcal{E}_\chi(\mathbb{C}) = \Lambda \left(\frac{\mathbb{C}}{z(c)} \right) - \Lambda \left(\frac{\mathbb{C}}{z(\mathbb{C})} \right) \mathcal{E}_z(\mathbb{C}),$$

with $\Lambda(\cdot)$ defined as in (B14). Combining these expressions yields

$$\begin{aligned} \xi \left[\sigma \left(\psi \left(\frac{\mathbb{C}}{z(\mathbb{C})} \right) \right) - 1 \right] = 1 + (\xi - 1) \left[\Theta \left(\frac{\mathbb{C}}{z(\mathbb{C})} \right) + \Lambda \left(\frac{\mathbb{C}}{z(c)} \right) \right] \mathcal{E}_z(\mathbb{C}) \\ - (\xi - 1) \left[\Theta \left(\frac{\mathbb{C}}{z(\mathbb{C})} \right) + \Lambda \left(\frac{\mathbb{C}}{z(c)} \right) \right] + \mathcal{E}_c(\mathbb{C}). \end{aligned}$$

In the limiting economy, we shrink the support of $F(\cdot)$. This implies that $\mathcal{E}_c(\mathbb{C}) = 0$. Hence, this first-order condition of the local producers defines a non-degenerate assignment $z(\mathbb{C})$:

$$\xi \left[\sigma \left(\psi \left(\frac{\mathbb{C}}{z(\mathbb{C})} \right) \right) - 1 \right] = 1 + (\xi - 1) \left[\Theta \left(\frac{\mathbb{C}}{z(\mathbb{C})} \right) + \Lambda \left(\frac{\mathbb{C}}{z(c)} \right) \right] \mathcal{E}_z(\mathbb{C}) - (\xi - 1) \left[\Theta \left(\frac{\mathbb{C}}{z(\mathbb{C})} \right) + \Lambda \left(\frac{\mathbb{C}}{z(c)} \right) \right]$$

The super-modularity properties of the profit function $\Pi(z, \mathbb{C})$ derived in Appendix B.1 still hold in this limiting economy. Hence, there is positive assortative matching, $z'(\mathbb{C}) > 0$, and more competitive cities are bigger, $M'(\mathbb{C}) > 0$.

B.4 Proof of Corollary 2

Using the elasticity notation, (31) implies

$$\mathcal{E}_{\mathcal{M}}(c) = \mathcal{E}_{\mu}\left(\frac{\mathbb{C}(c)}{z(c)}\right)\left[\mathcal{E}_C(c) - \mathcal{E}_z(c)\right]$$

Recall that $\mu(\cdot)$ is a decreasing function, and hence, $\mathcal{E}_{\mu}\left(\frac{\mathbb{C}(c)}{z(c)}\right) < 0$. Therefore, we get that

$$\mathcal{E}_{\mathcal{M}}(c) < (>)\, 0 \iff \mathcal{E}_C(c) > (<)\, \mathcal{E}_z(c)$$

Equation (B15) shows that:

$$\begin{aligned}\mathcal{E}_C(c)&\left[\alpha\left(\sigma\left(\frac{\mathbb{C}(c)}{z(c)}\right) - 1\right) + (\alpha+\eta)\Theta\left(\frac{\mathbb{C}(c)}{z(c)}\right) + \frac{\xi - 1}{\xi}\Lambda\left(\frac{\mathbb{C}(c)}{z(c)}\right) - \eta\theta\right] \\ &= \xi + \left[(\alpha+\eta)\Theta\left(\frac{\mathbb{C}(c)}{z(c)}\right) + \frac{\xi-1}{\xi}\Lambda\left(\frac{\mathbb{C}(c)}{z(c)}\right)\right]\mathcal{E}_z(c),\end{aligned}$$

Then, solving for $\mathcal{E}_C(c)$ in the above equation yields that:

$$\mathcal{E}_z(c) > (<)\, \mathcal{E}_C(c) > \iff \mathcal{E}_z(c) > (<) \frac{\xi}{\alpha\left(\sigma\left(\frac{\mathbb{C}(c)}{z(c)}\right) - 1\right) - \eta\theta}$$

Furthermore, equation (B11) implies the following expression for the productivity elasticity with respect city's appeal:

$$\mathcal{E}_z(c) = \frac{f(c)c}{M_e g(z(c))z(c)} \frac{1}{\Upsilon\left(\varphi\left(\psi\left(\frac{\mathbb{C}(c)}{z(c)}\right)\right)\right)}$$

Hence,

$$\begin{aligned}\mathcal{E}_z(c) > (<)\, \mathcal{E}_C(c) > &\iff \frac{f(c)c}{M_e g(z(c))z(c)} \frac{1}{\Upsilon\left(\varphi\left(\psi\left(\frac{\mathbb{C}(c)}{z(c)}\right)\right)\right)} > (<) \frac{\xi}{\xi\left(\sigma\left(\frac{\mathbb{C}(c)}{z(c)}\right) - 1\right) - \eta\theta}, \\ &\iff \frac{1}{g(z(c))z(c)} > (<) \frac{\xi\Upsilon\left(\varphi\left(\psi\left(\frac{\mathbb{C}(c)}{z(c)}\right)\right)\right)}{\xi\left(\sigma\left(\frac{\mathbb{C}(c)}{z(c)}\right) - 1\right) - \eta\theta} \frac{M_e}{f(c)c}\end{aligned}$$

Moreover, under Assumption 2, the density-weighted productivity $g(z)z$ has the following closed form

$$g(z)z = \delta \frac{\left(\frac{z_L}{z}\right)^{\delta}}{1 - \left(\frac{z_L}{z_H}\right)^{\delta}}$$

With these expressions, we start by characterizing the conditions for the first inequality. Under the conditions of Proposition 1, we have that $\xi\left(\sigma\left(\frac{\mathbb{C}(c)}{z(c)}\right) - 1\right) - \eta\theta > 1 - \eta\theta$. Also, the parametric specification for $\Upsilon(\cdot)$ implies that $\Upsilon\left(\varphi\left(\psi\left(\frac{\mathbb{C}(c)}{z(c)}\right)\right)\right) < \overline{\Upsilon}$, where $\overline{\Upsilon}$ is a constant that depends on ε and $\overline{\sigma}$. Recall that both traded good productivities and local amenities are defined over a bounded space. Therefore, there exists $\underline{c}$ such that $c > \underline{c}$ for all c. Finally, let $\underline{f}$ be the lower bound of the city density, which is exogenous. Thus, we obtain that $\mathcal{E}_z(c) > \mathcal{E}_C(c)$ if

$$\delta < \underbrace{\frac{(1-\eta\theta)\,\underline{f}\underline{c}}{\xi\overline{\Upsilon}}}_{\equiv\overline{\delta}} \frac{1}{M_e}$$

With a sufficiently low entry cost, the term $1/M_e$ is large. Therefore, we conclude that if $\delta < \overline{\delta}$ and the entry cost is not too large, $\mathcal{E}_z(c) > \mathcal{E}_C(c)$ and hence city aggregate markup is increasing in city appeal c.

C Efficiency

C.1 Social Planner's Problem

An utilitarian planner aims to maximize the sum of the utility levels of all the individuals in the economy. For the consumption side, the planner chooses traded good, $Q(c)$, local good, $Y(c)$, and housing consumption, $H(c)$, for every location. For the production side, the planner chooses the number of workers in the traded and local good sectors, $L^T(c)$ and $L^{NT}(c)$, which determine total population, $L(c) = L(c)^T + L^{NT}(c)$. Moreover, she also chooses materials, energy, and capital total usage in both sectors, $M^T(c)$, $E^T(c)$, $K^T(c)$, $M^{NT}(c)$, $E^{NT}(c)$, and $K^{NT}(c)$. For the location of local goods producers, I anticipate that the planner chooses PAM. Hence, she chooses the matching function $z(c)$, the slope of this function, and the mass of entrants M_e. To simplify the derivations, I follow an alternative formulation in which instead of choosing directly the slope of the assignment function, the planner chooses $\zeta(c)$, where $\zeta(c) \equiv \frac{z'(c)g(z(c))}{f(c)}$.[56] Finally, the planner also chooses the exit cutoff z^* such that $z(\overline{a}) = \overline{z}$ and $z(\underline{a}) = z^*$. The planner therefore maximizes the objective function:

$$\Omega = \int_{\underline{a}}^{\overline{a}} \left(L^T(c) + L^{NT}(c)\right) \left(\frac{Y(c)}{\eta}\right)^{\eta} \left(\frac{H(c)}{\alpha}\right)^{\alpha} \left(\frac{Q(c)}{1-\eta-\alpha}\right)^{1-\eta-\alpha} f(c)da,$$

[56] Note that with $z(c)$ and $\zeta(c)$, one can recover the true slope of the assignment function $z'(c) = \zeta(c)f(c)/g(z(c))$.

subject to the constraints:

$$
\begin{gathered}
\int_{\underline{a}}^{\overline{a}} \left(L^T(c) + L^{NT}(c)\right) f(c)da = \overline{L}, \\
\int_{\underline{a}}^{\overline{a}} a(L^T(c))^{\gamma}(M^T(c))^{\gamma_m}(E^T(c))^{\gamma_e}(K^T(c))^{\gamma_k} f(c)da = c_e M_e + \int_{\underline{a}}^{\overline{a}} \left\{Q(c)\left(L^T(c) + L^{NT}(c)\right)\right\} f(c)da \\
+ \int_{\underline{a}}^{\overline{a}} \left\{\Omega^e\left(E^T(c) + E^{NT}(c)\right) + M^T(c) + M^{NT}(c) + \Omega^k\left(K^T(c) + K^{NT}(c)\right)\right\} f(c)da \\
H(c)\left(L^T(c) + L^{NT}(c)\right) = 1 \;\forall\, a \\
\Upsilon\left(\frac{z(c)(L^{NT}(c))^{\beta}(M^{NT}(c))^{\beta_m}(E^{NT}(c))^{\beta_e}(K^{NT}(c))^{\beta_k}}{M_e\zeta(c)\left(L^T(c) + L^{NT}(c)\right)Y(c)}\right) M_e\zeta(c) = 1 \;\forall\, i \in a, \;\forall\, a \\
\int_{\underline{a}}^{a} \zeta(x) f(x)dx = G(z(c)) \;\forall\, a \\
\left(\frac{Y(c)}{\eta}\right)^{\eta}\left(\frac{H(c)}{\alpha}\right)^{\alpha}\left(\frac{Q(c)}{1-\eta-\alpha}\right)^{1-\eta-\alpha} = \overline{U} \;\forall\, i \in a, \;\forall\, a
\end{gathered}
$$

The first constraint corresponds to the aggregate labor market clearing. The second constraint is the aggregate resource constraint. The third constraint states that the land markets clear in every location. The fourth constraint corresponds to the local resource constraint, coming from the local goods Kimball preferences. The fifth constraint is an adding up condition coming from the definition of $\zeta(c)$, and the last constraint corresponds to the free mobility condition. The planner's Lagrangian is given by:

$$
\begin{aligned}
\mathcal{L}^P = & \int_{\underline{a}}^{\overline{a}} \left(L^T(c) + L^{NT}(c)\right) \left(\frac{Y(c)}{\eta}\right)^{\eta} \left(\frac{H(c)}{\alpha}\right)^{\alpha} \left(\frac{Q(c)}{1-\eta-\alpha}\right)^{1-\eta-\alpha} f(c)da, \\
& + \varkappa_1 \left[\overline{L} - \int_{\underline{a}}^{\overline{a}} \left(L^T(c) + L^{NT}(c)\right) f(c)da\right] \\
& + \varkappa_2 \Bigg[\int_{\underline{a}}^{\overline{a}} a(L^T(c))^{\gamma}(M^T(c))^{\gamma_m}(E^T(c))^{\gamma_e}(K^T(c))^{\gamma_k} f(c)da - c_e M_e - f M_e(1 - G(z^*)) \\
& \quad - \int_{\underline{a}}^{\overline{a}} \left\{Q(c)\left(L^T(c) + L^{NT}(c)\right) + M^T(c) + M^{NT}(c) + \Omega^e\left(E^T(c) + E^{NT}(c)\right) + \Omega^k\left(K^T(c) + K^{NT}(c)\right)\right\} f(c)d \\
& + \int_{\underline{a}}^{\overline{a}} \varsigma_1(c)\left[1 - H(c)\left(L^T(c) + L^{NT}(c)\right)\right] f(c)da \\
& + \int_{\underline{a}}^{\overline{a}} \varsigma_2(c)\left[\Upsilon\left(\frac{z(c)(L^{NT}(c))^{\beta}(M^{NT}(c))^{\beta_m}(E^{NT}(c))^{\beta_e}(K^{NT}(c))^{\beta_k}}{M_e\zeta(c)\left(L^T(c) + L^{NT}(c)\right)Y(c)}\right) M_e\zeta(c) - 1\right] f(c)da \\
& - \int_{\underline{a}}^{\overline{a}} \varsigma_3(c)G(z(c))f(c)da + \int_{\underline{a}}^{\overline{a}} (\overline{\varsigma}_3 - \vartheta(c))\zeta(c)f(c)da, \\
& + \int_{\underline{a}}^{\overline{a}} \varsigma_4(c)\left[\overline{U} - \left(\frac{Y(c)}{\eta}\right)^{\eta}\left(\frac{H(c)}{\alpha}\right)^{\alpha}\left(\frac{Q(c)}{1-\eta-\alpha}\right)^{1-\eta-\alpha}\right] f(c)da
\end{aligned}
$$

where $\varkappa_1$, $\varkappa_2$, $\varsigma_1(c)$, $\varsigma_2(c)$, $\varsigma_3(c)$, and $\varsigma_4(c)$ are lagrange multipliers.[57] I integrated by parts the constraint on $\zeta(c)$ with multiplier $\varsigma_3(c)$, and defined $\overline{\varsigma}_3 \equiv \int_{\underline{a}}^{\overline{a}} \varsigma_3(c)f(c)da$ and $\vartheta(c) \equiv \int_{\underline{a}}^{a} \varsigma_3(x)f(x)dx$.

Consumption and housing. The first-order conditions with respect to $Y(c)$, $H(c)$, and $Q(c)$ are respectively:

$$\left[1 - \frac{\varsigma_4(c)}{L(c)}\right]\eta\frac{U(c)}{Y(c)} = \varsigma_2(c)\frac{\Upsilon'(q(c))q(c)}{Y(c)}\zeta(c), \tag{C1}$$

$$\left[1 - \frac{\varsigma_4(c)}{L(c)}\right]\alpha\frac{U(c)}{H(c)} = \varsigma_1(c), \tag{C2}$$

$$\left[1 - \frac{\varsigma_4(c)}{L(c)}\right](1-\eta-\alpha)\frac{U(c)}{Q(c)} = \varkappa_2, \tag{C3}$$

where $U(c) \equiv \left(\frac{Y(c)}{\eta}\right)^{\eta}\left(\frac{H(c)}{\alpha}\right)^{\alpha}\left(\frac{Q(c)}{1-\eta-\alpha}\right)^{1-\eta-\alpha}$, and $q(c) \equiv \frac{z(c)(L^{NT}(c))^{\beta}(M^{NT}(c))^{\beta_m}(E^{NT}(c))^{\beta_e}(K^{NT}(c))^{\beta_k}}{\zeta(c)L(c)Y(c)}$, with $L(c) \equiv L^T(c) + L^{NT}(c)$. Normalizing $\varkappa_2$ to be one, these first-order conditions imply that:

[57] I can shut down the free mobility constraint by setting $\varsigma_4(c) = 0$ for all a.

$$\begin{aligned}\mathbb{P}(c)Y(c) &= \eta\overline{U}\mathbb{P}(c)^{\eta}\varsigma_1(c)^{\alpha},\\ \varsigma_1(c)H(c) &= \alpha\overline{U}\mathbb{P}(c)^{\eta}\varsigma_1(c)^{\alpha},\\ Q(c) &= (1-\eta-\alpha)\overline{U}\mathbb{P}(c)^{\eta}\varsigma_1(c)^{\alpha},\end{aligned}$$

with:

$$\begin{aligned}\mathbb{P}(c) &\equiv \frac{\mathbb{D}(c)}{\delta(q(c))},\\ \mathbb{D}(c) &\equiv \varsigma_2(c)Y(c)L(c)\end{aligned}$$

Moreover, the ratio $\varsigma_4(c)/L(c)$ is given by:

$$\frac{\varsigma_4(c)}{L(c)} = 1 - \mathbb{P}(c)^{\eta}\varsigma_1(c)^{\alpha}$$

Labor. Define the planner's shadow wage $W^*(c)$ as:

$$W^*(c) = \varkappa_1 - \overline{U}\frac{\varsigma_4(c)}{L(c)} \tag{C4}$$

Then, the first-order conditions with respect to $L^T(c)$, and $L^{NT}(c)$ are respectively:

$$\gamma\frac{Q^T(c)}{L^T(c)} = W^*(c), \tag{C5}$$

$$\beta\frac{\mathbb{P}(c)Y(c)L(c)}{L^{NT}(c)} = W^*(c),, \tag{C6}$$

where $Q^T(c) \equiv a(L^T(c))^{\gamma}(M^T(c))^{\gamma_m}(E^T(c))^{\gamma_e}(K^T(c))^{\gamma_k}$, and where we used (C1) - (C3) for the definition of the planner's shadow wage, $W^*(c)$.

Materials, energy and capital. The first-order conditions with respect $M^T(c)$, $M^{NT}(c)$, $E^T(c)$, $E^{NT}(c)$, $K^T(c)$, and $L^{NT}(c)$ are respectively:

$$
\begin{aligned}
\gamma_m \frac{Q^T(c)}{M^T(c)} &= 1, \\
\beta_m \frac{\mathbb{P}(c)Y(c)L(c)}{M^{NT}(c)} &= 1, \\
\gamma_e \frac{Q^T(c)}{E^T(c)} &= \Omega^e, \\
\beta_e \frac{\mathbb{P}(c)Y(c)L(c)}{E^{NT}(c)} &= 1\Omega^e, \\
\gamma_K \frac{Q^T(c)}{K^T(c)} &= \Omega^k, \\
\beta_k \frac{\mathbb{P}(c)Y(c)L(c)}{K^{NT}(c)} &= 1\Omega^k
\end{aligned}
$$

Combining (C5) with the rest of the traded good inputs equilibrium conditions gives the planner's counterpart of (22):

$$W^*(c) = \left(\frac{a}{\varrho}\right)^{\frac{1}{\gamma}}$$

Similarly, combining (C4) with the rest of the non-traded good inputs equilibrium conditions gives the optimal condition for $q^*(c)$:

$$\Upsilon'(q^*(c)) = \frac{C^*(c)}{z(c)}, \tag{C7}$$

where $C^*(c)$ is the planner's competition index given by:

$$C^*(c) \equiv \nu \frac{(W^*(c))^{\beta}}{D^*(c)}.$$

Allocation of non-traded varieties producers. The first-order conditions with respect $z(c)$ and $\zeta(c)$ are respectively:

$$
\begin{aligned}
\varsigma_2(c) \frac{\Upsilon'(q(c))q(c)\zeta(c)}{z(c)} &= \varsigma_3(c) g(z(c)), \\
\varsigma_2(c) \left[\Upsilon(q(c)) - \Upsilon'(q(c))q(c)\right] &= \vartheta(c) - \overline{\varsigma}_3
\end{aligned}
$$

Define $J(c) \equiv \vartheta(c) - \overline{\varsigma}_3$. Therefore, we have that $J'(c) = \varsigma_3(c) f(c)$ and we can re-write the first-order condition with respect to $z(c)$ as:

$$\varsigma_2(c)\frac{\Upsilon'(q(c))q(c)\zeta(c)}{z(c)} = \frac{J'(c)}{J(c)}(\vartheta(c) - \bar{\varsigma}_3)\frac{g(z(c))}{f(c)}$$

We can combine the above equation with the first-order condition with respect to $\zeta(c)$ to get:

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

where we used the definition of $\zeta(c)$ and $\delta(q(c))$. Furthermore, re-write the first-order condition with respect to $\zeta(c)$ as:

$$J(c) = D^*(c)L(c)Y(c)\Upsilon(q(c))\frac{\delta(q(c)) - 1}{\delta(q(c))}$$

Then,

$$\mathcal{E}_J(c) = \mathcal{E}_{D^*}(c) + \mathcal{E}_L(c) + \mathcal{E}_Y(c) + \mathcal{E}_q(c)\left[\frac{1}{\delta(q(c))} + \frac{\mathcal{E}_\delta(q(c))}{\delta(q(c)) - 1}\right]$$

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

$$\mathcal{E}_\delta(q(c)) = \frac{1}{\delta(q(c))} + \frac{1}{\sigma(q(c))} - 1, \tag{C9}$$

$$\mathcal{E}_{q^*}(c) = \sigma(q(c))\left(\mathcal{E}_z(c) - \mathcal{E}_{C^*}(c)\right) \tag{C10}$$

where we used the definition of $\delta(q(c))$ for the first expression and (C7) for the second. Define $\Theta^*(c)$ as:

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

With this notation:

$$\begin{aligned}
\mathcal{E}_{\delta^*}(c) &= \mathcal{E}_\delta(q(c))\mathcal{E}_{q^*}(c), \\
&= \mathcal{E}_\delta(q(c))\sigma(q(c))\left(\mathcal{E}_z(c) - \mathcal{E}_{C^*}(c)\right), \\
&= (\sigma(q(c)) - 1)\left(\frac{\mu(q(c))}{\delta(q(c))} - 1\right)\left(\mathcal{E}_z(c) - \mathcal{E}_{C^*}(c)\right), \\
&= \Theta^*(c)\left(\mathcal{E}_z(c) - \mathcal{E}_{C^*}(c)\right),
\end{aligned}$$

Combining these expressions gives:

$$\mathcal{E}_J(c) = \mathcal{E}_{D^*}(c) + \mathcal{E}_L(c) + \mathcal{E}_Y(c) + \frac{\mathcal{E}_z(c)}{\delta(q(c)) - 1} - \frac{\mathcal{E}_{C^*}(c)}{\delta(q(c)) - 1} \tag{C11}$$

Equating (C8) and (C11) implies:

$$\mathcal{E}_{C^*}(c) = (\delta(q(c)) - 1)\left[\mathcal{E}_{\delta^*}(c) + \mathcal{E}_{M^*}(c)\right], \tag{C12}$$

where $M^*(c) \equiv P^*(c)Y(c)L(c)$ is the planner's market size. This expression closely resembles the one for the decentralized equilibrium (B12). To find an expression for the planner's market size $M^*(c)$, first note that land market clearing implies:

$$L(c) = \frac{\varsigma_1(c)}{\alpha \overline{U}\mathbb{P}(c)^{\eta}\varsigma_1(c)^{\alpha}} \tag{C13}$$

Replacing this expression into the optimal consumption decisions, we get that:

$$M^*(c) = \frac{\eta}{\alpha}\varsigma_1(c)$$

On the other hand, the definition of $W^*(c)$ implies that:

$$\varsigma_1(c) = \left[\frac{W^*(c) + \overline{U} - \varkappa_1}{\overline{U}\mathbb{P}(c)^{\eta}}\right]^{\frac{1}{\alpha}}$$

Denoting $\check{W}^*(c) \equiv W^*(c) + \overline{U} - \varkappa_1$, we obtain the following expression for the planner's market size:

$$\begin{aligned}
\mathcal{E}_{M^*}(c) &= \frac{1}{\alpha}\left[\mathcal{E}_{\check{W}^*}(c) - \eta\mathcal{E}_{\mathbb{P}}(c)\right], \\
&= \frac{1}{\alpha}\left[\frac{W^*(c)}{\check{W}^*(c)}\frac{1}{\gamma} - \frac{\eta\beta}{\gamma} + \eta\mathcal{E}_{C^*}(c) + \eta\mathcal{E}_{\delta}(c)\right]
\end{aligned}$$

Replacing the resulting expression into (C12) and solving for $\mathcal{E}_{C^*}(c)$ gives:

$$\mathcal{E}_{C^*}(c)\left(\alpha - \eta(\delta(q(c)) - 1)\right) = (\delta(q(c)) - 1)\left[\frac{1 - \eta\beta}{\gamma} + \Xi(q(c), \overline{U}, \varkappa_1)\right],$$

with

$$\Xi(q(c), \overline{U}, \varkappa_1) = (\alpha+\eta)\mathcal{E}_{\delta^*}(c) - \frac{1}{\gamma}\frac{\overline{U}-\varkappa_1}{W^*(c)+\overline{U}-\varkappa_1}$$

Closing the social planner's problem. Note that after solving the planner's ODE's we ge $C^*(c)$ and $z(c)$. This pins down $q^*(c)$ by (C7), $\mathbb{D}(c)$ and $\mathbb{P}(c)$.

Inputs first-order conditions imply that:

$$\mathbb{P}(c)Y(c)L(c) + Q^T(c) = W(c)L(c) + M(c) + \Omega^e E(c) + \Omega^e E(c)$$

Therefore, the aggregate resource constraint becomes:

$$\int_{\underline{a}}^{\overline{a}} \left[\mathbb{P}(c)Y(c)L(c) + Q(c)L(c)\right] f(c)da = \int_{\underline{a}}^{\overline{a}} W(c)L(c)f(c)da$$

This condition states that, in the aggregate, traded and non-traded goods total revenue is equal to the aggregate wage-bill. Combining (C13) with the expressions for $Y(c)$ and $Q(c)$ we get that $\mathbb{P}(c)Y(c)L(c) = (\eta/\alpha)\varsigma_1(c)$ and $Q(c)L(c) = ((1-\eta-\alpha)/\alpha)\varsigma_1(c)$. Therefore:

$$\int_{\underline{a}}^{\overline{a}} \frac{1-\alpha}{\alpha}\varsigma_1(c)f(c)da = \int_{\underline{a}}^{\overline{a}} W(c)L(c)f(c)da$$

Which using the equilibrium conditions can be written as:

$$(1-\alpha)\int_{\underline{a}}^{\overline{a}} \left[\frac{W^*(c)+\overline{U}-\varkappa_1}{\mathbb{P}(c)^\eta}\right]^{\frac{1}{\alpha}} f(c)da = \int_{\underline{a}}^{\overline{a}} W(c)\left[\frac{\left(W^*(c)+\overline{U}-\varkappa_1\right)^{1-\alpha}}{\mathbb{P}(c)^\eta}\right]^{\frac{1}{\alpha}} f(c)da \tag{C14}$$

Finally, use (C13) and the aggregate labor market clearing to write:

$$\int_{\underline{a}}^{\overline{a}} \left[\frac{\left(W^*(c)+\overline{U}-\varkappa_1\right)^{1-\alpha}}{\mathbb{P}(c)^\eta}\right]^{\frac{1}{\alpha}} f(c)da = \overline{L} \tag{C15}$$

Equations (C14) and (C15) pin down $\overline{U}$ and $\varkappa_1$.

Decentralized ODE. The FOC of the location problem (21) is:

$$\mathcal{E}_{\mathbb{C}}(c) = \left[\mu\left(q(c)\right)-1\right]\left(\mathcal{E}_\delta(c)+\mathcal{E}_M(c)\right)$$

With $\mathcal{E}_\delta(c) = \mathcal{E}_\delta(q(c))\mathcal{E}_q(c)$, where

$$\mathcal{E}_\delta(q(c)) = \frac{1}{\delta(q(c))} + \frac{1}{\sigma(q(c))} - 1, \tag{C16}$$

$$\mathcal{E}_q(c) = \rho(q(c))\sigma(q(c))\left(\mathcal{E}_z(c) - \mathcal{E}_{\mathbb{C}}(c)\right) \tag{C17}$$

Importantly, $\mathcal{E}_\delta(q(c))$ comes from the definition of the object $\delta(\cdot)$ and $\mathcal{E}_q(c)$ comes from the first-order condition of the firm. Therefore, defining:

$$\Theta(c) \equiv \rho(q(c))(\sigma(q(c)) - 1)\left(\frac{\mu(q(c))}{\delta(q(c))} - 1\right),$$

we get that:

$$\begin{aligned}
\mathcal{E}_\delta(c) &= \mathcal{E}_\delta(q(c))\mathcal{E}_q(c),\\
&= \mathcal{E}_\delta(q(c))\rho(q(c))\sigma(q(c))\left(\mathcal{E}_z(c) - \mathcal{E}_{\mathbb{C}}(c)\right),\\
&= \rho(q(c))(\sigma(q(c)) - 1)\left(\frac{\mu(q(c))}{\delta(q(c))} - 1\right)\left(\mathcal{E}_z(c) - \mathcal{E}_{\mathbb{C}}(c)\right),\\
&= \Theta(c)\left(\mathcal{E}_z(c) - \mathcal{E}_{\mathbb{C}}(c)\right),
\end{aligned}$$

Market size is given by:

$$\begin{aligned}
\mathcal{E}_M(c) &= \frac{1}{\alpha}\mathcal{E}_W(c) - \frac{\eta}{\alpha}\mathcal{E}_P(c),\\
&= \frac{1}{\alpha}\mathcal{E}_W(c) - \frac{\eta}{\alpha}\left(\mathcal{E}_D(c) - \mathcal{E}_\delta(c)\right),\\
&= \frac{1}{\alpha}\mathcal{E}_W(c) - \frac{\eta}{\alpha}\left(\beta\mathcal{E}_W(c) - \mathcal{E}_{\mathbb{C}}(c) - \mathcal{E}_\delta(c)\right),\\
&= \frac{1-\eta\beta}{\alpha}\mathcal{E}_W(c) + \frac{\eta}{\alpha}\mathcal{E}_{\mathbb{C}}(c) + \frac{\eta}{\alpha}\mathcal{E}_\delta(c)
\end{aligned}$$

Plugging the expressions for $\mathcal{E}_\delta(c)$ and $\mathcal{E}_M(c)$ into the first-order condition gives:

$$\mathcal{E}_{\mathbb{C}}(c) = \frac{\mu\left(q(c)\right) - 1}{\alpha + (\mu(q(c)) - 1)((\alpha+\eta)\Theta(c) - \eta)}\left[\frac{1-\eta\beta}{\gamma} + (\alpha+\eta)\Theta(c)\mathcal{E}_z(c)\right] \tag{C18}$$

C.2 Misallocation from Increasing Markups

This section shows how misallocation of firms across space is exacerbated by increasing markups on city size.

C.3 First-best Implementation

In this section. I show how the location-specific subsidy (35) implements the first-best allocation in three steps. First, I show that the subsidy removes markups. Second, I show that the policy induces firms to locate optimally. Third, I show that subsidy also corrects the aggregate entry margin.

Step 1: Markup removal. First, (12) and (35) imply that the net profits after transfers are given by (36). The first-order condition of this problem gives the optimal relative quantities $y(z,c)/Y(c)$:

$$\Upsilon'\left(\frac{y(z,c)}{Y(c)}\right) = \frac{\mathbb{C}(c)}{z}. \tag{C19}$$

Compared to (13), we see that the above expression does not have any markup. Moreover, the expression above equals the social planner first-order condition from the previous section.

Step 2: Optimal Firm Location. Regarding firm's location decisions, we first show that there is still PAM under the subsidy (35). Using the envelope theorem, we have from (36) that:

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

where we used (C19) and the definition of $\delta(\cdot)$. Then, we get that:

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

where the second term on the RHS is positive because of the first-order condition (C19). Therefore, $\widehat{\Pi}(z,c)$ is log-supermodular if and only if $\delta'(\cdot) > 0$ which is true for any $\Upsilon(\cdot)$ satisfying MSLD. This establish PAM between z and $\mathbb{C}(c)$. I now provide conditions under $\mathbb{C}(c)$ is increasing in general equilibrium. The first-order condition of (36) with respect to c is:

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

The term $\mathcal{E}_M(c)$ is the same as before, except that, because firms no longer charge any markup, the function $\chi(\cdot)$ becomes a constant and therefore $\mathcal{E}_\chi(c) = 0$. Hence,

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

Moreover, (C19) implies that:

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

with

$$\Theta^*\left(\frac{\mathbb{C}(c)}{z(c)}\right) = \left[\sigma\left(\psi^*\left(\frac{\mathbb{C}(c)}{z(c)}\right)\right) - 1\right]\left[\mu\left(\frac{\mathbb{C}(c)}{z(c)}\right)/\delta\left(\frac{\mathbb{C}(c)}{z(c)}\right) - 1\right],$$

where $\psi^*\left(\frac{\mathbb{C}(c)}{z(c)}\right)$ is the optimal pricing decision implied by (C19). Putting all together, we get that

$$\mathcal{E}_{\mathbb{C}}(c)\left[\frac{\xi}{\delta\left(\frac{y(z,c)}{Y(c)}\right) - 1} + (\xi + \eta\theta)\Theta^*\left(\frac{\mathbb{C}(c)}{z(c)}\right) - \eta\theta\right] = \xi + (\xi + \eta\theta)\Theta^*\left(\frac{\mathbb{C}(c)}{z(c)}\right)\mathcal{E}_z(c). \tag{C20}$$

Equation (C20) reveals two insights. First, when $\Upsilon(\cdot)$ takes the Klenow and Willis (2016) specification and $\varepsilon \to 0$, the expression collapses to

$$\mathcal{E}_{\mathbb{C}}(c)\left[\xi(\overline{\sigma} - 1) - \eta\theta\right] = \xi,$$

which further implies that, under the policy, $\mathbb{C}'(c) > 0$ if and only if $\xi(\overline{\sigma} - 1) > \eta\theta$. Note that the conditions of Proposition 1 imply this result. The second insight is that (C20) coincides with the social planner location decision. Thus, firms sort optimally under the considered policy.

Step 3: Optimal Entry. It is straightforward to show that the policy generates an efficient economy-wide aggregate entry rate. Indeed, (37) implies that the free-entry condition under the policy is

$$\int_c \left[\delta\left(\frac{y(z(c),c)}{Y(c)}\right) - 1\right]\Upsilon\left(\frac{y(z(c),c)}{Y(c)}\right)M(c)f(c)dc = c_e M_e,$$

which coincides with the social planner's entry condition.

D Additional Derivations for Estimation

This section further explores the framework for markup estimation of Section 4.4.2.

D.1 Derivations for Markup Estimation

In this section, I derive additional results for the baseline markup estimation. Furthermore, it shows how to derive (E5) under the Klenow and Willis (2016) functional form.

D.2 Multi-Sector Estimation and Labor Market Power

In this section, I show how to extend the baseline estimation framework when having multiple sectors. Moreover, it shows how can we control for potential labor market power.

The multi-sector markup estimation procedure considers a framework in which consumers have Cobb-Douglas preferences over different bundles of local varieties within a sector. Formally, the bundle of local varieties $Y(c)$ is a Cobb-Douglas aggregator of sector-specific bundles:

$$Y(c) = \prod_{n=1}^{N} Y_n(c),$$

where n denotes the sector. Furthermore, each of the sector bundles $Y_n(c)$ is implicitly defined by a Kimball aggregator:

$$\int_z \Upsilon_n \left(\frac{y_n(z,c)}{Y_n(c)} \right) dG_{n,c}(z) = 1,$$

where $y_n(z,c)$ is the consumption in city c of a variety produced by a firm with productivity z in sector n, $G_{n,c}(z)$ is the local productivity distribution of sector n in city c, and the Kimball aggregator $\Upsilon_n(\cdot)$ now is sector-specific. Under this alternative formulation, all derivations from Section 4 extend, with the caveat that the markup function (44) is sector specific

$$\mu_{jnc} = \mu_n \left(\zeta_n \left(s_{jnc} \frac{P_{nc}}{D_{nc}} \right) \right), \tag{E1}$$

with markups, sales shares and price indices being sector specific as well. Moreover, under the multi-sector formulation, we allow different sectors to have different production functions

$$y_n(z,c) = z l_n(z,c)^{\beta_n} s_n(z,c)^{1-\beta_n}. \tag{E2}$$

Equations (E1) and (E2) imply the multi-sector estimating equation (47).

D.3 General Production Function

This section derives the markup estimating equation with a general production. Because data in sectors other than Manufacturing is limited, I derive the estimation equation using the data available for Manufacturing. Formally, consider a Hicks-neutral production function in labor, materials, energy and capital:

$$y(z,c) = z\mathcal{F}\left(l(z,c), m(z,c), e(z,c), k(z,c)\right), \tag{E3}$$

where $\mathcal{F}(\cdot)$ is a continuously differentiable function, $m(z,c)$ denotes materials, $e(z,c)$ denotes energy and $k(z,c)$ denotes capital. Under this specification, the elasticities of output with respect to labor, materials and energy are given by

$$\frac{\partial \log y(z,c)}{\partial \log l(z,c)} = \frac{\partial \mathcal{F}\left(l(z,c), m(z,c), e(z,c), k(z,c)\right)}{\partial l(z,c)} \frac{l(z,c)}{\mathcal{F}\left(l(z,c), m(z,c), e(z,c), k(z,c)\right)}$$
$$\frac{\partial \log y(z,c)}{\partial \log m(z,c)} = \frac{\partial \mathcal{F}\left(l(z,c), m(z,c), e(z,c), k(z,c)\right)}{\partial m(z,c)} \frac{m(z,c)}{\mathcal{F}\left(l(z,c), m(z,c), e(z,c), k(z,c)\right)},$$
$$\frac{\partial \log y(z,c)}{\partial \log e(z,c)} = \frac{\partial \mathcal{F}\left(l(z,c), m(z,c), e(z,c), k(z,c)\right)}{\partial e(z,c)} \frac{e(z,c)}{\mathcal{F}\left(l(z,c), m(z,c), e(z,c), k(z,c)\right)}.$$

Note that in all cases, the Hicks-neutral assumption on the production function implies that the output elasticity with respect to any variable input is just a function of the inputs of production:

$$\frac{\partial \log y(z,c)}{\partial \log l(z,c)} \equiv \kappa_l\left(l(z,c), m(z,c), e(z,c), k(z,c)\right),$$
$$\frac{\partial \log y(z,c)}{\partial \log m(z,c)} \equiv \kappa_m\left(l(z,c), m(z,c), e(z,c), k(z,c)\right)$$
$$\frac{\partial \log y(z,c)}{\partial \log e(z,c)} \equiv \kappa_e\left(l(z,c), m(z,c), e(z,c), k(z,c)\right).$$

Therefore, under the general production function (E3), the markup estimating equation (46) takes the form of:

$$\log \alpha^l_{jc} = \kappa_x\left(l(z,c), m(z,c), e(z,c), k(z,c)\right) - \varsigma_{1,c} s_{jc} - \varsigma_{2,c} s^2_{jc} - \varsigma_{3,c} s^3_{jc} - \upsilon_{jc}, \tag{E4}$$

where x denotes the input of production, $x \in \{l, m, e\}$. The function $\kappa_x\left(l(z,c), m(z,c), e(z,c), k(z,c)\right)$ can be semi-parametric approximated as the markup function. Formally, I approximate this function by a third-order polynomial in its arguments as in Gandhi, Navarro, and Rivers (2020).

D.4 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 county c, μ_{jc}, and the establishments sales share, s_{jc}:

$$\frac{1}{\mu_{jc}} + \log\left(1 - \frac{1}{\mu_{jc}}\right) = \frac{\overline{\sigma} - 1}{\overline{\sigma}} - \log \overline{\sigma} + \frac{\varepsilon}{\overline{\sigma}} \log \frac{\overline{\sigma}}{\overline{\sigma} - 1} - \frac{\varepsilon}{\overline{\sigma}} \log \frac{P_c}{D_c} + \frac{\varepsilon}{\overline{\sigma}} \log s_{jc}, \tag{E5}$$

Using the estimated markups, $\hat{\mu}_{jc}$, I estimate the following equation via OLS:

$$\frac{1}{\hat{\mu}_{jc}} + \log\left(1 - \frac{1}{\hat{\mu}_{jc}}\right) = \varpi + \varpi_c + \frac{\varepsilon}{\overline{\sigma}} \log s_{jc} + \iota_{jc}, \tag{E6}$$

where ϖ is a constant absorbing the constant terms in (E5), ϖ_c is a county fixed-effect absorbing the term $(\epsilon/\overline{\sigma}) \log(P_c/D_c)$, and ι_{jc} is an approximation error coming from the estimation of the markups. The regression coefficient of $\log s_{jc}$ 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 P_c/D_c

$$\begin{aligned}
\frac{y_{jc}}{Y_c} &= \left[-\overline{\sigma}\Omega\left(-\left(s_{jc}\frac{P_c}{D_c}\frac{\overline{\sigma}}{\overline{\sigma}-1}\right)^{\frac{\varepsilon}{\overline{\sigma}}}\frac{\exp\left(-\frac{1}{\overline{\sigma}}\right)}{\overline{\sigma}}\right)\right]^{\frac{\overline{\sigma}}{\epsilon}}, \\
\frac{P_c}{D_c} &= \sum_{i\in c}\frac{\overline{\sigma}-1}{\overline{\sigma}}\exp\left(\frac{1-\frac{y_{jc}}{Y_c}^{\frac{\varepsilon}{\overline{\sigma}}}}{\varepsilon}\right)\frac{y_{jc}}{Y_c}.
\end{aligned}$$

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:

$$\check{\mu}_{jc}(\overline{\sigma}) = \frac{1}{1 - \frac{1}{\overline{\sigma}}\frac{y_{jc}}{Y_c}(\overline{\sigma})^{\frac{\varepsilon}{\overline{\sigma}}}}$$

Then, I estimate $\overline{\sigma}$ such that we minimize the distance between the predicted and the estimated markups:

$$\hat{\overline{\sigma}} = \underset{\overline{\sigma}}{\operatorname{argmin}} \|\check{\mu}_{jc}(\overline{\sigma}) - \hat{\mu}_{jc}\|$$

D.5 Estimation of City Fundamentals

Within the loop of the SMM estimation, I perform the model inversion to obtain local productivities and amenities. Given parameters, I solve the decentralized equilibrium and use (22) to recover traded good productivity

$$a(c) = \varrho W(c)^{\gamma} R(c)^{1-\gamma},$$

where $W(c)$ is data on the county average wages and $R(c)$ is the model's implied housing rent. Similarly, I obtain local productivities from the labor supply condition (11). Replacing the equilibrium objects of the model in such expression gives

$$b(c) = \frac{L(c)^{\frac{\lambda_1}{\theta}}}{W(c)^{\lambda}_{2}} \nu^{\eta} \left(\frac{\overline{U}}{\overline{L}}\right)^{\frac{1}{\theta}} \frac{\chi \left(\frac{\mathbb{C}(c)}{z(c)}\right)^{\frac{\alpha+\eta\beta}{1+\phi}}}{\left(\mathbb{C}(c)\delta\left(\frac{\mathbb{C}(c)}{z(c)}\right)\right)^{\eta}},$$

where $L(c)$ and $W(c)$ are data on counties population and average wages, and λ_1 and λ_2 are constants given by

$$\lambda_1 = 1 + \frac{\theta(\alpha+\eta\beta)}{1+\phi}, \qquad \text{and} \qquad \lambda_2 = \frac{(1+\phi)(1-\eta\beta)-(\alpha+\eta\beta)}{1+\phi}.$$

E Additional Empirical Results

E.1 Markups Across Cities for additional Sectors

Table G1 displays the estimated mean elasticity of county aggregate markup and county size for different 2-digit NAICS sectors. The establishment-level markup is estimated using equation (47). The results show heterogeneity across sectors in the estimated elasticity of county-sector aggregate markup and county size. Manufacturing, Wholesale, Transportation, Information and Education Services display a positive elasticity, while the other sectors display a negative elasticity. Strikingly, the elasticity for all sectors is statistically significant than zero, suggesting that there is indeed large variation in the degree of local competition across counties in all sectors.

The results also shed light on the sectors that are driving the aggregate negative relationship of markup and city size displayed in Figure 2. The bottom row in each panel show the average employment share of each sector among the total employment of local industries across counties. The sectors with higher local employment share are Retail, Healthcare, and Accommodation and Food Services. The three sectors display a negative elasticity, suggesting that they are the main drivers behind the results in Figure 2

Table G1: 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	-0.0427***	0.1496***	0.0817***	-0.0706***	0.0448***
	(0.0077)	(0.0131)	(0.0114)	(0.0059)	(0.0089)
Observations	3100	1900	2400	3100	2800
R-squared	0.011	0.045	0.014	0.064	0.006
Avg. Local Emp. Share	0.0796	0.013	0.0203	0.241	0.0262

Panel B

	Information	Finance	Real Estate	PST Services	AWR Services
	(1)	(2)	(3)	(4)	(5)
Log total labor income	0.1218***	-0.0428***	-0.0151*	-0.0565***	-0.0157*
	(0.0099)	(0.0083)	(0.0086)	(0.0081)	(0.0089)
Observations	2800	3100	2800	3000	2900
R-squared	0.035	0.009	0.001	0.018	0.001
Avg. Local Emp. Share	0.0162	0.0464	0.0157	0.0274	0.0553

Panel C

	Education Services	Healthcare	Arts, Entertainment and Recreation	Accommodation and Food Services	Other Services
	(1)	(2)	(3)	(4)	(5)
Log total labor income	0.1384***	-0.0639***	0.1567***	-0.0718***	-0.1115***
	(0.0162)	(0.0097)	(0.0123)	(0.0073)	(0.0076)
Observations	1400	3100	2100	3100	3100
R-squared	0.041	0.016	0.054	0.044	0.076
Avg. Local Emp. Share	0.00379	0.273	0.0112	0.143	0.05

Notes: Table G1 displays the average elasticity of county aggregate markup and county size. County aggregate markup is defined as (48) 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. ARW 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.

E.2 Robustness Exercises

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.

I estimate markups for Manufacturing using the equation (E4) derived in Appendix D.3. 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 (E4), where a third-order polynomial in labor, materials, energy, and capital approximates the elasticities for materials and energy.

Estimating markups for Manufacturing using (E4) 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 G2 displays the results of regression between the baseline manufacturing markups estimated using (46) and the estimates using (E4). The baseline markups are estimated pooling all establishments in local industries while the estimates using (E4) only include establishments in local manufacturing industries. Moreover, I consider two definitions of a city: a county 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.

On the other hand, Table (G3) shows the elasticity of county aggregate markup to county size using the alternative estimates. This table replicates the results in Figure 3(b). The elasticity of aggregate markup to county size remains almost unchanged when considering the alternative markup estimates. Indeed, the elasticity in Figure 3(b) is 0.14 while the elasticities reported in Table (G3) are 0.141 for materials and 0.135 for energy. The results are reassuring in two ways. First, the flexible polynomial that controls for potential market power in the baseline estimation indeed corrects for any potential monopsony power in the labor market. Second, the Cobb-Douglas technology assumption in the baseline estimation seems not restrictive as considering a more flexible production function yields similar county markup and size elasticities.

Table G2: Regressions baseline markups and extensions for Manufacturing

	Baseline County (1)	Baseline CZ (2)	Baseline County (3)	Baseline CZ (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*** (0.009)	0.281*** (0.007)	0.304*** (0.009)	0.243*** (0.007)
Observations	27500	27500	27500	27500
R-squared	0.745	0.782	0.738	0.778

Notes: Table G2 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 (46) and pool establishments in all local industries. Alternative markup estimates are estimated using (E4) 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: county and Commuting Zone (CZ). Robust standard errors in parenthesis. *10% level, **5% level, ***1% level.

Table G3: County Aggregate Markup and County Size for Alternative Manufacturing Markups Estimates

	Log agg. Markup Materials (1)	Log agg. Markup Energy (2)
Log labor income	0.141*** (0.006)	0.135*** (0.006)
Observations	1800	1800
R-squared	0.183	0.17

Notes: Table G3 displays the average elasticity of county aggregate markup and county size, using the alternative estimates for Manufacturing. Alternative markup estimates are estimated using (E4) and consider only establishments in local manufacturing industries. The dependent variable in Column (1) is the log county aggregate markup using materials as the flexible input in (E4). The dependent variable in Column (2) is the log county aggregate markup using energy as the flexible input in (E4). County size is defined as total labor income. Robust standard errors in parenthesis. *10% level, **5% level, ***1% level.