The Spatial Distribution of Markups

Preliminary and Incomplete

Jonathan Becker Chris Edmond
Virgiliu Midrigan Daniel Yi Xu

April 2026

Abstract

How much does the geographic segmentation of the US economy matter for competition and market power? We answer this question using a quantitative spatial model with multi-establishment firms, oligopolistic competition, and endogenously variable markups, calibrated to match US manufacturing data across 170 Economic Areas. Our benchmark model implies an aggregate markup of 1.26 with substantial variation across locations, from 1.22 in greater New York to 1.35 in Honolulu. The welfare costs of markups are large, 5.8% on average, and considerably higher than the 3.7% implied by an equivalent model without geography and spatial frictions. The welfare costs are also very unevenly distributed, ranging from about 1% or less in large central locations like New York to as much as 20% in Honolulu.

Keywords: competition, misallocation, multi-establishment firms, trade flows, gravity.

JEL classifications: D4, E2, F1, L1, O4.

Acknowledgments: We particularly thank our discussants Mayara Felix, Jake Zhao and Sebastian Heise for their detailed comments and thank audience participants at the 2023 FRB Dallas conference on international economics, the 2023 OzMac workshop, the 2024 PHBS Shenzen workshop on macroeconomics and finance, the 2025 workshop of the Australian Macroeconomics Society, the 2026 ASSA meetings, and seminar participants at the University of Nottingham—Ningbo. Earlier drafts of this paper circulated under the title "Local Concentration, National Concentration, and the Spatial Distribution of Markups".

Affiliations: Jonathan Becker, Stony Brook University (jonathan.becker@stonybrook.edu). Virgiliu Midrigan, New York University and NBER (virgiliu.midrigan@nyu.edu). Daniel Yi Xu, Duke University and NBER (daniel.xu@duke.edu).

1 Introduction

How much does the geographic segmentation of the US economy matter for competition and market power? Despite considerable improvements in transportation and the introduction of new technologies that make it easier for firms to sell in more locations, trade in goods across the US remains quite segmented with substantial home bias in shipment patterns even in the absence of formal trade barriers (Wolf, 2000; Hillberry and Hummels, 2008). Does this pattern simply reflect the concentration of production in a few highly productive locations? Or would more trade *within* the US increase competition, reduce markups, and increase aggregate productivity? Most work on the macroeconomic consequences of markups treats the domestic economy as a single, perfectly integrated national market. Does that matter?

We answer these questions using a quantitative spatial model with oligopolistic competition and endogenously variable markups. Our model features many geographically segmented locations and heterogeneous firms that can, in general, source goods from multiple locations and sell in multiple locations. Firms are heterogeneous in their productivity and in the number and geographic location of their establishments. Taking wages in each location as given, each firm chooses an optimal, comprehensive production plan for their set of establishments that determines that firm's effective marginal cost of producing for each possible destination market. Given these firm-and-destination-specific marginal costs, oligopolistic competition as in Atkeson and Burstein (2008) determines the markups each firm charges in each of its destination markets. Equilibrium wages in each location are determined by local labor market clearing conditions.

Most work on the macroeconomic consequences of markups (e.g., Baqaee and Farhi, 2020; De Loecker, Eeckhout and Mongey, 2021; Edmond, Midrigan and Xu, 2023) assumes frictionless trade within a country, and hence implicitly assumes that the spatial structure of economic activity within a country is irrelevant for competition and market power. Similarly, most work on quantitative spatial economics (e.g., Allen and Arkolakis, 2014; Redding and Rossi-Hansberg, 2017) assumes either perfect competition or monopolistic competition with constant markups — and so again implicitly assumes that the spatial economy does not interact with the determinants of competition and market power. By contrast, our model allows for extensive geographic variation in economic conditions within and across sectors and locations and determines firm-and-location-specific markups endogenously, in general equilibrium, jointly with other macro outcomes.

Our analysis focuses on manufactured goods, for which we can bring together detailed data on firms, establishments, and trade flows across locations. This is also a setting where goods vary enormously in their tradeability, production is geographically concentrated, and firms range from single plants to large multi-establishment enterprises selling nationwide. One might guess that for a well-developed economy like the US, and especially for

manufactured goods, spatial frictions would be modest and not very important for the overall macroeconomy. Whether this intuition holds up is a quantitative question. And, as we show, spatial frictions turn out to be quantitatively important even in this setting.

We calibrate our model to match the operations of some 220,000 US manufacturing firms organized into 363 6-digit NAICS sectors, with an average of some 600 firms per sector. We take our geographical locations to be 170 Economic Areas as constructed by the Bureau of Economic Analysis. While most firms are small and have only one establishment, larger firms tend to have multiple establishments and multiple-establishment firms produce in a broad range of locations. We ensure every firm in the model reproduces a real firm's *geographic footprint* — we place each firm's establishments exactly where they appear in the data. We choose the parameters of our model governing the distribution of productivity across firms and the correlation between a firm's establishment count and its productivity to match key facts on national sales concentration and the share of employment accounted for by multi-establishment firms. We parameterize sector-specific iceberg trade costs so that the model exactly reproduces sector-specific gravity regressions using state-to-state trade flows from the Commodity Flows Survey for 3-digit NAICS manufacturing sectors.

The fact that we match these sector-specific gravity effects is important. These gravity effects determine the quantitative significance of the spatial frictions in each sector, i.e., they determine which sectors produce goods that are intrinsically less tradeable and, within a given sector, which locations are more central and which are more remote. In equilibrium, these spatial frictions play a key role in determining the spatial distribution of production and sales concentration. Intuitively, we find that production concentration is larger and more dispersed than sales concentration and that this effect is more pronounced in sectors with low spatial frictions.[1] More generally, we provide results for the full spectrum of possible trade costs, from autarky to free trade. The marginal effect of reducing trade costs on market power is very large near autarky and diminishes as locations become increasingly well integrated. Our benchmark calibration is much closer to the free-trade end of this spectrum, with more modest effects on markups, but even so the remaining spatial frictions are quantitatively important for competition and market power.

Our model generates a wide range of outcomes across US locations. Consider greater New York and San Francisco. Both are large and highly productive, but our model implies that New York has low sales concentration and low markups while San Francisco has high sales concentration and high markups. What accounts for this? Both are so productive that they end up importing relatively little. What distinguishes them are stark differences in the amount of competition amongst their respective domestic producers. In the data, we

[1]For example, in our benchmark model the local sales Herfindahl-Hirschman Index (HHI) is, on average, about 0.15, higher than the national sales HHI of 0.10 but lower than the local production HHI of 0.36 reported by Autor, Patterson and Van Reenen (2023). Reassuringly, the ordering of concentration implied by our model is also consistent with the findings of Benkard, Yurukoglu and Zhang (2026) who study concentration in finely disaggregated consumer survey data.

observe less production concentration in New York and so through the lens of our model infer more head-to-head competition and lower markups. By contrast we observe more production concentration in San Francisco and so infer less head-to-head competition and higher markups. Despite the high markups, San Francisco remains a relatively cheap place to buy manufactures because of the overall low cost of producing there.

For smaller, less productive locations which have to import almost everything, what matters is the amount of competition among their potential suppliers. And this is where geography and spatial frictions play a key role, determining what locations supply them and at what cost. Consider Honolulu and tiny Scottsbluff near the Nebraska/Wyoming border. Both import almost everything. But Honolulu has relatively high markups while Scottsbluff has relatively low markups. This is because Honolulu is extremely remote — the trade costs facing potential suppliers are large — and its imports are concentrated among a few sources, limiting the amount of head-to-head competition among importers. Scottsbluff, by contrast, imports from a broader and more evenly matched set of sources who compete head-to-head. Despite the low markups, Scottsbluff remains an expensive place because the cost of delivering goods there is relatively high. Moreover, such locations are especially vulnerable to changes in trade costs. We show that if trade costs were substantially higher, Scottsbluff would *flip* from being a low-markup, highly competitive destination for goods to being one of the least competitive places in the country.

How much does any of this matter? Overall, we find that these spatial frictions are a quantitatively significant determinant of the macroeconomic losses due to market power. Our benchmark model implies an economy-wide aggregate markup of 1.26 with considerable variation across locations, ranging from 1.22 in greater New York to 1.35 in Honolulu. If we calibrate the model to match the same national concentration but abstract from geography and spatial frictions we find a much lower aggregate markup of 1.18 as well as much less markup dispersion, and hence lower productivity losses due to misallocation. Abstracting from geography and spatial frictions leads to a quantitatively significant understatement of the macroeconomic losses associated with market power. Roughly speaking, the model without geography and spatial frictions acts as if the competitive conditions of New York (or better) prevailed everywhere in the country. While the difference between an aggregate markup of 1.26 and 1.18 may appear modest, aggregate markups are notoriously difficult to move in this class of models — achieving a comparable reduction would require at least a ten-fold increase in the number of firms in every sector.

In line with this, we find that accounting for geography and spatial frictions leads to substantially higher estimates of the welfare costs of markups. We measure these welfare costs by asking how much the representative consumer in each location would gain from eliminating markups. We find that the average welfare cost is high, about 5.8% in consumption-equivalent terms, considerably higher than the 3.7% we find in an equivalent model that abstracts from geography and spatial frictions. While our benchmark economy

is much closer to the free-trade end of the spectrum than to autarky, with relatively modest amounts of intranational trade acting to substantially restrain markups in most locations, overall these spatial frictions are nonetheless quantitatively important. Moreover the welfare costs are very unevenly distributed, ranging from about 1% in locations like greater New York to as much as 20% in locations like Honolulu. The large costs in some locations are driven not by a lack of trade in general, but by the specific geographic and production structures that leave some locations with too few competitive suppliers. Abstracting from geography and spatial frictions leads to not just a sizeable understatement of the average welfare costs of markups, but also a complete inability to characterize this large variation in the burden of market power across locations.

We also show that our model is consistent with key facts stressed in the recent literature on concentration. In particular, we show that our model generates endogenously diverging trends in national concentration and local concentration, as in Rossi-Hansberg, Sarte and Trachter (2020). For this exercise, we consider an exogenous 20% reduction in trade costs, chosen to match Coşar, Osotimehin and Popov (2024), who find a long-run decrease of 15-20% for US manufacturing from 1963 to 2017. This decrease in trade costs allows the largest, most productive firms to sell to more locations than they did previously. Our model interprets the 'diverging trends' in concentration as *convergence* towards a better-integrated national market, with national concentration increasing from low initial levels while local concentration decreases from high initial levels. In the limit as trade costs vanish, so that there is effectively free trade in goods, there is simply a single national market for each good and no distinction between national and local concentration. In short, in response to decreasing trade costs, national sales concentration is rising but markets are becoming more competitive, markups are falling, and aggregate productivity is rising. In this sense, trends in national concentration may be a poor guide to trends in aggregate market power.

Finally, we ask to what extent can worker mobility mitigate the welfare costs of markups. Our benchmark model features workers that are immobile across locations. In an extension we consider a setup where workers have heterogeneous preferences for different location-specific amenities, pinning down labor supply to each location. We find that, when we parameterize the model to match the same initial allocation of labor across locations as in our benchmark, the welfare costs of markups end up being almost identical. With worker mobility, eliminating markups leads to larger changes in consumption in the most attractive locations, but these locations also receive labor inflows so that the changes in consumption per worker are almost exactly the same as in our benchmark model.

Macroeconomics of market power. Our paper builds on the literature studying the macro consequences of market power, including Baqaee and Farhi (2020), De Loecker, Eeckhout and Unger (2020), De Loecker, Eeckhout and Mongey (2021), and Edmond, Midrigan and Xu (2023). These papers study the macro implications of changes in the

aggregate markup and changes in markup dispersion but treat the domestic economy as a single spatially integrated whole. We show that abstracting from spatial frictions like this leads to a quantitatively significant understatement of the macro effects of market power.

Spatial economics. Our paper also builds on the recent literature on quantitative spatial economics, including Allen and Arkolakis (2014), Caliendo and Parro (2015), Redding (2016), Redding and Rossi-Hansberg (2017), and Fajgelbaum, Morales, Serrato and Zidar (2019). This literature has developed rich quantitative models for studying the spatial distribution of economic activity and the effects of changes in trade policy for trade flows and welfare across space, but typically assumes either perfect competition or monopolistic competition with constant markups. We show that allowing markups to respond endogenously to the spatial economic structure has quantitatively important implications for the level and distribution of market power across locations.

Spatial misallocation. Our work is closely related to two recent papers on the spatial distribution of markups. Like us, Asturias, García-Santana and Ramos (2019) develop an Atkeson and Burstein (2008) model with many locations — which they use to assess the importance of better transportation infrastructure in India — but unlike us they abstract from multi-establishment firms and do not develop the implications of their model for trends in local concentration. Similarly, Franco (2023) studies spatial misallocation across cities in a model of monopolistic competition with Kimball (1995) demand. He emphasizes the sorting of firms across locations, which we abstract from, but does not consider the implications of multi-establishment firms for the spatial distribution of markups.

Trends in concentration. This paper contributes to the extensive literature on the causes and consequences of the rise in concentration in the US since the early 1980s, following Grullon, Larkin and Michaely (2019), Autor, Dorn, Katz, Patterson and Van Reenen (2020), Amiti and Heise (2021), Ganapati (2021), and many others. In an influential paper using National Establishment Time Series (NETS) data, Rossi-Hansberg, Sarte and Trachter (2020) argue that local concentration has been declining even while national concentration has risen. Further work using the US Census of Retail Trade by Smith and Ocampo (2024) argues that *both* national and local sales concentration have been rising since the early 1990s. Similarly, using the US Economic Census more broadly, Autor, Patterson and Van Reenen (2023) find that local sales concentration has risen but that local employment concentration has fallen. But, as argued by Benkard, Yurukoglu and Zhang (2026), measures of concentration using Census data focus on the classification of economic activity by *production*, not by *consumption* and it is the availability of good substitutes for consumers that ultimately determines how much market power producers have. Benkard et al. find decreasing local sales concentration in finely disaggregated consumer survey data. Neiman and Vavra

(2023) report a similar decrease in sales concentration which they interpret as arising due to increasingly 'niche' consumption patterns.

Multi-establishment production. This paper also contributes to the recent literature on the increasing importance of multi-establishment firms, following Jia (2008), Holmes (2011), Basker, Klimek and Van (2012), Foster, Haltiwanger, Klimek, Krizan and Ohlmacher (2016), Cao, Hyatt, Mukoyama and Sager (2022), and many others. While this literature originally focused on retail trade, this phenomenon has become increasingly important for services too, as in Hsieh and Rossi-Hansberg (2023).

2 Model

The economy consists of many heterogeneous locations. Across the economy there are many heterogeneous firms that, in general, can source goods from multiple locations and sell in multiple locations. Firms compete *oligopolistically* in their destination markets. The economy is *geographically segmented* in two ways: (i) labor is immobile across locations,[2] with location-specific wages pinned down by local labor market clearing conditions, and (ii) goods shipments are subject to iceberg trade costs.

2.1 Environment

There are J locations indexed by $j, k = 1, ..., J$. There is a continuum of sectors indexed by $s \in [0, 1]$. Within each sector there is a finite $N(s)$ firms indexed by $i = 1, ..., N(s)$ that compete oligopolistically in their destination markets. Trade in goods is subject to sector-specific iceberg trade costs $\tau_{jk}(s) \geq 1$ with $\tau_{jj}(s) = 1$. A notational convention that we maintain throughout is that location j refers to the *source* of a good and location k refers to a *destination* so that $\tau_{jk}(s)$, say, refers to the sector-specific cost of shipping from j to k.

Location-specific final good. In each destination market k there is a non-tradeable final good produced under perfectly competitive conditions. This location-specific final good is given by a CES aggregate *across sectors*

$$C_k = \left(\int_0^1 C_k(s)^{\frac{\theta-1}{\theta}} \, ds \right)^{\frac{\theta}{\theta-1}}, \qquad \theta > 1 \tag{1}$$

[2]We discuss an extension with *labor mobility* across locations in Section 7 below.

Then *within sectors*, output is given by a CES aggregate across the $N(s)$ firms in sector s

$$C_k(s) = \left(\sum_{i=1}^{N(s)} C_{ik}(s)^{\frac{\gamma-1}{\gamma}} \right)^{\frac{\gamma}{\gamma-1}}, \qquad \gamma > \theta \tag{2}$$

We assume $\gamma > \theta$ so that goods are more substitutable within sectors than across sectors.

Firms source from establishments in multiple locations. Each firm i selling in destination market k sources goods from establishments in multiple locations $j = 1, ..., J$. Specifically, we assume that each firm i supplies k with a CES aggregate *across establishments*

$$C_{ik}(s) = \left(\sum_{j=1}^{J} c_{ijk}(s)^{\frac{\eta-1}{\eta}} \right)^{\frac{\eta}{\eta-1}}, \qquad \eta \geq \gamma \tag{3}$$

Two special cases are worth noting: (i) the case $\eta = \gamma$ where goods are equally substitutable within and across firms (within a given sector), and (ii) the case $\eta = +\infty$ where a given firm will source all of its output from its least-cost establishment.

Location-specific representative consumer. Each location j is populated by L_j identical workers each endowed with E_j efficiency units of labor. Labor is *immobile* across locations. Each worker inelastically supplies their E_j units of labor to the local labor market and receives location-specific wage W_j per efficiency unit. Aggregating the budget constraints of workers in location j gives

$$P_j C_j = W_j E_j L_j + \Pi_j \tag{4}$$

where Π_j denotes aggregate profits paid out to workers in location j.

Distribution of profits. We assume that firm ownership is perfectly diversified across locations with profits paid out in proportion to labor income

$$\Pi_j = \bar{\pi} W_j E_j L_j \tag{5}$$

This implies that every location has the same labor and profit income shares

$$\left(\frac{1}{1+\bar{\pi}}, \frac{\bar{\pi}}{1+\bar{\pi}} \right) \tag{6}$$

for some constant $\bar{\pi} \geq 0$ to be determined in equilibrium.

Demand system. This nested-CES setup implies that the demand for goods sourced from location j to be sold at destination k by firm i in sector s is given by

$$c_{ijk}(s) = \left(\frac{\tau_{jk}(s)p_{ijk}(s)}{P_{ik}(s)}\right)^{-\eta} \underbrace{\left(\frac{P_{ik}(s)}{P_k(s)}\right)^{-\gamma}\left(\frac{P_k(s)}{P_k}\right)^{-\theta} C_k}_{=C_{ik}(s)} \tag{7}$$

As usual, the location-specific final good and sector-level price indexes are given by

$$P_k = \left(\int_0^1 P_k(s)^{1-\theta}\, ds\right)^{\frac{1}{1-\theta}}, \qquad P_k(s) = \left(\sum_{i=1}^{N(s)} P_{ik}(s)^{1-\gamma}\right)^{\frac{1}{1-\gamma}} \tag{8}$$

But in this setup there is now also an index for aggregating prices across establishments within each firm. This firm-level composite price is given by

$$P_{ik}(s) = \left(\sum_{j=1}^{J} (\tau_{jk}(s)p_{ijk}(s))^{1-\eta}\right)^{\frac{1}{1-\eta}} \tag{9}$$

We implicitly set $p_{ijk}(s) = +\infty$ for any firm i that does not sell in location k.

Production. Firm i in sector s is endowed with productivity $z_{ij}(s) \geq 0$ for goods produced at location j. For simplicity we assume that the output shipped by firm i from source j to destination k is linear in labor

$$y_{ijk}(s) = z_{ij}(s) l_{ijk}(s) \tag{10}$$

Resource constraints. Given the sector-specific iceberg trade costs $\tau_{jk}(s) \geq 1$, the resource constraints on the flow of output from j to k are simply

$$y_{ijk}(s) = \tau_{jk}(s) c_{ijk}(s) \tag{11}$$

Marginal cost. Taking the wage rate W_j as given, firm i can source goods from j for any destination k at marginal cost

$$\frac{W_j}{z_{ij}(s)} \tag{12}$$

Profits. Since a firm can supply destination k with goods sourced from establishments at any location j, the firm's profits from sales at k are given by

$$\Pi_{ik}(s) = \sum_{j=1}^{J} \left(p_{ijk}(s) - \frac{W_j}{z_{ij}(s)}\right) y_{ijk}(s) \tag{13}$$

A firm's total profits are then given by $\Pi_i(s) = \sum_{k=1}^{J} \Pi_{ik}(s)$. This objective is separable across destinations k and hence the firm maximizes total profits by maximizing profits in each destination k separately.

We characterize the firm's profit maximizing strategy in each destination k in two steps: (i) taking as given the firm's composite price for its destination market, $P_{ik}(s)$, we determine the least-cost way of servicing that destination with one unit of the firm's composite good, $C_{ik}(s) = 1$, then (ii) we characterize how the firm's price $P_{ik}(s)$ is determined through oligopolistic competition with the other firms servicing destination k. The first step implicitly gives us a characterization of the allocation of production across locations within a given firm. Given the first step, the second step is a nested-CES oligopoly problem familiar from Atkeson and Burstein (2008) and Edmond, Midrigan and Xu (2015).

Within-firm allocation. Taking as given $P_{ik}(s)$ and $C_{ik}(s) = 1$, for step (i) firm i chooses prices $p_{ijk}(s)$ for $j = 1, ..., J$ to minimize the total cost of servicing destination k

$$\sum_{j=1}^{J} \frac{W_j}{z_{ij}(s)} y_{ijk}(s) = \sum_{j=1}^{J} \frac{\tau_{jk}(s) W_j}{z_{ij}(s)} \left(\frac{\tau_{jk}(s) p_{ijk}(s)}{P_{ik}(s)} \right)^{-\eta} \underbrace{C_{ik}(s)}_{=1} \tag{14}$$

subject to the firm-level price index (9). The Lagrangian for this problem can be written

$$\mathcal{L} = \sum_{j=1}^{J} \frac{\tau_{jk}(s) W_j}{z_{ij}(s)} \left(\frac{\tau_{jk}(s) p_{ijk}(s)}{P_{ik}(s)} \right)^{-\eta} - \lambda_{ik}(s) \sum_{j=1}^{J} \left(\left(\frac{\tau_{jk}(s) p_{ijk}(s)}{P_{ik}(s)} \right)^{1-\eta} - 1 \right) \tag{15}$$

where $\lambda_{ik}(s) \geq 0$ denotes the multiplier on the firm's constraint. The first order conditions for interior solutions simplify to

$$\eta \frac{\tau_{jk}(s) W_j}{z_{ij}(s)} = (\eta - 1) \lambda_{ik}(s) \left(\frac{\tau_{jk}(s) p_{ijk}(s)}{P_{ik}(s)} \right) \tag{16}$$

Rearranging this we see that, at the optimum, source prices satisfy

$$p_{ijk}(s) = \mu_{ik}(s) \frac{W_j}{z_{ij}(s)}, \qquad \mu_{ik}(s) = \frac{\eta}{\eta - 1} \left(\frac{P_{ik}(s)}{\lambda_{ik}(s)} \right) \tag{17}$$

Hence the least-cost way to service destination k is to set a *destination-specific* markup $\mu_{ik}(s)$ that applies uniformly regardless of the source location j. The firm 'prices to market' in a way that reflects the demand and competitive conditions specific to market k. But by making the markup independent of j the firm *avoids distorting allocations within the firm*.

Plugging this expression for the source prices $p_{ijk}(s)$ back into the firm-level price index

(9) and eliminating the multiplier gives

$$P_{ik}(s) = \mu_{ik}(s) \cdot \mathsf{MC}_{ik}(s) \tag{18}$$

where

$$\mathsf{MC}_{ik}(s) = \left(\sum_{j=1}^{J} \left(\frac{\tau_{jk}(s) W_j}{z_{ij}(s)} \right)^{1-\eta} \right)^{\frac{1}{1-\eta}} \tag{19}$$

denotes the firm's marginal cost of servicing destination k with one unit of the composite good $C_{ik}(s)$. With this characterization of the within-firm allocation in hand, we can now turn to the strategic interactions between firms in each destination k.

Oligopolistic competition. For step (ii) we then need to characterize how the firm's price $P_{ik}(s)$ is determined through oligopolistic competition with the other firms servicing destination k. Given the within-firm allocation we can use (7), (13) and (18) to write the firm's profits from destination k

$$\begin{aligned} \Pi_{ik}(s) &= \left(P_{ik}(s) - \mathsf{MC}_{ik}(s)\right) C_{ik}(s) \\ &= \left(P_{ik}(s) - \mathsf{MC}_{ik}(s)\right) \left(\frac{P_{ik}(s)}{P_k(s)} \right)^{-\gamma} \left(\frac{P_k(s)}{P_k} \right)^{-\theta} C_k \end{aligned} \tag{20}$$

with each firm internalizing the effect of their price $P_{ik}(s)$ on the sector-level price index $P_k(s)$ in (8). Given our characterization of the within-firm allocation in the first step, this second step is a standard nested-CES oligopoly problem familiar from Atkeson and Burstein (2008) and Edmond, Midrigan and Xu (2015).

As is well known, this implies that each firm sets a markup of the form

$$\mu_{ik}(s) = \frac{\varepsilon_{ik}(s)}{\varepsilon_{ik}(s) - 1} \tag{21}$$

where the demand elasticity $\varepsilon_{ik}(s)$ facing firm i is endogenous to the firm's sales share in destination k. For our benchmark model we assume that each destination market is characterized by *Cournot competition*. With this specification, the demand elasticity works out to be a sales-weighted harmonic average of the elasticities of substitution within and across sectors

$$\varepsilon_{ik}(s) = \left(\omega_{ik}(s) \frac{1}{\theta} + (1 - \omega_{ik}(s)) \frac{1}{\gamma} \right)^{-1} \tag{22}$$

where $\omega_{ik}(s)$ denotes the market share of firm i in destination market k

$$\omega_{ik}(s) := \frac{P_{ik}(s)C_{ik}(s)}{\sum_{i=1}^{N(s)} P_{ik}(s)C_{ik}(s)} = \frac{P_{ik}(s)^{1-\gamma}}{\sum_{i=1}^{N(s)} P_{ik}(s)^{1-\gamma}} \tag{23}$$

Since the elasticity of substitution across firms is larger than across sectors, $\gamma > \theta$, the demand elasticity $\varepsilon_{ik}(s)$ facing a firm is lower for firms with larger market shares in destination k. Intuitively, firms that are small within a given market are mostly competing with other firms within the same sector and so face a relatively high demand elasticity, approaching the within-sector elasticity γ as $\omega_{ik}(s) \to 0$. At the other extreme, firms that are large within a given market are mostly competing with firms in other sectors and so face a relatively low demand elasticity, approaching the across-sector elasticity θ as $\omega_{ik}(s) \to 1$.

While intuitive, this discussion is incomplete. It simply takes market shares $\omega_{ik}(s)$ as exogenous and traces out the implications of those market shares for markups $\mu_{ik}(s)$. But in this model, markups and market shares are *jointly determined* as part of a larger fixed-point problem. To solve this problem, it turns out to be convenient to first combine (21) and (22) to write the inverse markup as a linear function of the sales share

$$\frac{1}{\mu_{ik}(s)} = 1 - \frac{1}{\varepsilon_{ik}(s)} = \frac{\gamma - 1}{\gamma} - \left(\frac{1}{\theta} - \frac{1}{\gamma}\right)\omega_{ik}(s) \tag{24}$$

From which we see that indeed a firm's markup is *strictly increasing* in its market share.

To obtain the second condition we need, we substitute prices $P_{ik}(s) = \mu_{ik}(s)\mathsf{MC}_{ik}(s)$ into (23) to get

$$\omega_{ik}(s) = \frac{(\mu_{ik}(s)\mathsf{MC}_{ik}(s))^{1-\gamma}}{\sum_{i=1}^{N(s)}(\mu_{ik}(s)\mathsf{MC}_{ik}(s))^{1-\gamma}} \tag{25}$$

Here we see that, conditional on other firms' markups, each firm's market share is *strictly decreasing* in its markup. Together, equations (24) and (25) are two equations in two unknowns that jointly determine the markups $\mu_{ik}(s)$ and market shares $\omega_{ik}(s)$ for each i, k and s. Notice that the interactions between firms within a given market enter only through the denominator in (25) and that market shares are homogenous of degree zero in the markups.

Eliminating the market shares between these we have a single fixed point condition

$$\frac{1}{\mu_{ik}(s)} = \frac{\gamma - 1}{\gamma} - \left(\frac{1}{\theta} - \frac{1}{\gamma}\right)\frac{(\mu_{ik}(s)\mathsf{MC}_{ik}(s))^{1-\gamma}}{\sum_{i=1}^{N(s)}(\mu_{ik}(s)\mathsf{MC}_{ik}(s))^{1-\gamma}} \tag{26}$$

This condition implicitly determines the distribution of markups $\mu_{ik}(s)$ within and across locations as a function of the distribution of marginal costs $\mathsf{MC}_{ik}(s)$ within and across

locations. The marginal costs $\mathsf{MC}_{ik}(s)$ are exogenous to each firm but, because they depend on the wages W_j, still need to be determined in equilibrium.

General equilibrium. The equilibrium of the model is pinned down by labor market clearing in each local labor market. The labor market in each location j clears when the total supply of efficiency units of labor $E_j L_j$ equals the total labor demand in that location

$$E_j L_j = \int_0^1 \sum_{k=1}^{J} \sum_{i=1}^{N(s)} l_{ijk}(s)\, ds \tag{27}$$

Solving the model. We solve the model by Newton methods on a high-dimensional system of nonlinear equations. We initialize the algorithm using the solution of an alternative model which is identical except that profits are distributed uniformly across locations rather than proportionately. This alternative model can be solved much more efficiently and provides a good initial guess for the Newton algorithm we use to solve the benchmark model. The full details of our computational procedure are given in the Appendix.

We next briefly outline how the spatial distribution of markups $\mu_{ik}(s)$ affects aggregate productivity within and across locations.

2.2 *Aggregation*

Underlying all our aggregation results is an endogenous bilateral *productivity network*, a collection of productivity levels $\bar{z}_{jk}(s)$ for each sector s that forms a graph on the nodes j, k with directed edges from origins j to destinations k. The results below are obtained for the benchmark case $\eta = \gamma$, where goods are equally substitutable within and across firms. For this case we can give an exact closed-form decomposition of sector-level productivity $Z_k(s)$ at each destination k in terms of this underlying bilateral productivity network $\bar{z}_{jk}(s)$.[3]

Productivity network. To derive this productivity network, first let $\bar{c}_{jk}(s)$ denote the composite formed from goods shipped from j to k within a given sector

$$\bar{c}_{jk}(s) = \left(\sum_{i=1}^{N(s)} c_{ijk}(s)^{\frac{\gamma-1}{\gamma}} \right)^{\frac{\gamma}{\gamma-1}} \tag{28}$$

To write the bilateral composite $\bar{c}_{jk}(s)$ this way we make use of the fact that with $\eta = \gamma$ the quantity c_{ijk} is proportional to $p_{ijk}^{-\gamma}$ across firms i for any source-destination pair

[3] More generally, with $\eta > \gamma$, our expressions for firm-level markups (no distortions within the firm) and sector-level markup aggregates go through, but the bilateral decomposition is not available in closed form.

j, k. Now let $\bar{l}_{jk}(s) = \sum_i l_{ijk}(s)$ denote the labor used to produce this composite and let $\bar{y}_{jk}(s) = \tau_{jk}(s)\bar{c}_{jk}(s)$ denote the amount of this composite that has to be produced at j for $\bar{c}_{jk}(s)$ to arrive at k. The productivity network is then given by the collection of $\bar{z}_{jk}(s) := \bar{y}_{jk}(s)/\bar{l}_{jk}(s)$. In keeping with this notation, let $\bar{\mu}_{jk}(s)$ denote the implied markup, satisfying $\bar{p}_{jk}(s) = \bar{\mu}_{jk}(s)W_j/\bar{z}_{jk}(s)$ where

$$\bar{p}_{jk}(s) = \left(\sum_{i=1}^{N(s)} p_{ijk}(s)^{1-\gamma}\right)^{\frac{1}{1-\gamma}} = \left(\sum_{i=1}^{N(s)} \left(\frac{\mu_{ik}(s)}{z_{ij}(s)}\right)^{1-\gamma}\right)^{\frac{1}{1-\gamma}} W_j \tag{29}$$

is the price index for the composite good. As in Edmond, Midrigan and Xu (2015, 2023), we then obtain

$$\bar{z}_{jk}(s) = \left(\sum_{i=1}^{N(s)} \left(\frac{\mu_{ik}(s)}{\bar{\mu}_{jk}(s)}\right)^{-\gamma} z_{ij}(s)^{\gamma-1}\right)^{\frac{1}{\gamma-1}} \tag{30}$$

Notice that in the special case of no dispersion in markups this reduces to a standard CES productivity index that depends only on the exogenous firm-level productivity $z_{ij}(s)$. More generally, markup dispersion reduces $\bar{z}_{jk}(s)$ below this benchmark. Notice also that $\bar{z}_{jk}(s)$ does not depend on the trade costs $\tau_{jk}(s)$. This is because, within a given sector s, the trade costs $\tau_{jk}(s)$ apply to all firms shipping from j to k equally.

With this expression for $\bar{z}_{jk}(s)$ in hand, the markup $\bar{\mu}_{jk}(s)$ on the composite good is

$$\bar{\mu}_{jk}(s) = \left(\sum_{i=1}^{N(s)} \frac{1}{\mu_{ik}(s)}\,\omega_{ijk}(s)\right)^{-1} = \frac{\sum_{i=1}^{N(s)} \mu_{ik}(s)^{1-\gamma}\, z_{ij}(s)^{\gamma-1}}{\sum_{i=1}^{N(s)} \mu_{ik}(s)^{-\gamma}\, z_{ij}(s)^{\gamma-1}} \tag{31}$$

where

$$\omega_{ijk}(s) = \left(\frac{p_{ijk}(s)}{\bar{p}_{jk}(s)}\right)^{1-\gamma} = \left(\frac{\mu_{ik}(s)}{z_{ij}(s)}\,\frac{\bar{z}_{jk}(s)}{\bar{\mu}_{jk}(s)}\right)^{1-\gamma} \tag{32}$$

denotes the sales share of firm i in shipments from j to k. In short, as in Edmond, Midrigan and Xu (2015, 2023), the markup on the composite good is a sales-weighted harmonic average of the firm-level markups $\mu_{ik}(s)$ for destination k.

Location-specific productivity and markups. With the bilateral productivity network $\bar{z}_{jk}(s)$ and associated markups $\bar{\mu}_{jk}(s)$ determined, we can aggregate further to get measures of location-specific productivity and markups. To this end, let $\bar{L}_k(s) := \sum_j \bar{l}_{jk}(s)$ denote the labor used *across all source locations* j to produce for destination k. Then let $\bar{Z}_k(s) := C_k(s)/\bar{L}_k(s)$ denote the real consumption at destination k per unit of this total labor input. Using $\eta = \gamma$, we can write $C_k(s)$ as an aggregate of the bilateral composite $\bar{c}_{jk}(s)$ and then express the productivity index $\bar{Z}_k(s)$ in terms of the bilateral productivity network $\bar{z}_{jk}(s)$

and markups $\bar{\mu}_{jk}(s)$, giving

$$\bar{Z}_k(s) = \left(\sum_{j=1}^{J} \left(\frac{\bar{\mu}_{jk}(s)}{\bar{\mu}_k(s)} \right)^{-\gamma} \left(\frac{\bar{z}_{jk}(s)}{\tau_{jk}(s)} \right)^{\gamma-1} \left(\frac{W_j}{\bar{W}_k(s)} \right)^{-\gamma} \right)^{\frac{1}{\gamma-1}} \tag{33}$$

where $\bar{\mu}_k(s)$ denotes the location-specific markup, which again can be written as sales-weighted harmonic average of the underlying $\bar{\mu}_{jk}(s)$, and where $\bar{W}_k(s) = \sum_j W_j \bar{l}_{jk}(s)/\bar{L}_k(s)$ is the weighted average of source wages. This expression for $\bar{Z}_k(s)$ is similar to that given for the productivity nodes $\bar{z}_{jk}(s)$ in equation (30) above, but differs in two ways. First, the expression for $\bar{Z}_k(s)$ also depends on trade costs $\tau_{jk}(s)$, since trade costs reduce the contributions that high-productivity sources j make to destinations k that are costly to ship to. Second, the expression for $\bar{Z}_k(s)$ also depends on the relative wage $W_j/\bar{W}_k(s)$, which is absent from the expression for $\bar{z}_{jk}(s)$ since those productivity nodes refer to firms who are all paying the same wage W_j to produce in j.[4]

Overall we see that markup dispersion reduces aggregate productivity, both because markup dispersion *across firms* within a given destination k directly reduces productivity $\bar{z}_{jk}(s)$ at each node in the productivity network, as in equation (30), and because for any given network of $\bar{z}_{jk}(s)$, markup dispersion *across locations*, reduces aggregate productivity $\bar{Z}_k(s)$ at each destination k, as in equation (33). In this sense, the spatial dispersion in markups creates *endogenous misallocation*, both across firms and across locations.

3 Quantifying the model

In this section we outline our benchmark parameterization and calibration strategy and present our model's implications for national and local sales concentration.

3.1 Benchmark parameterization

Our geographical locations are Economic Areas (EAs) constructed by the US Bureau of Economic Analysis. There are $J = 170$ EAs in our data. Each EA is built around an urban core, either larger Metropolitan Statistical Areas (MSAs) or smaller Micropolitan Statistical Areas, along with adjacent counties with strong commuting ties. Importantly, these EAs are built to reflect the fact that economic activity spans administrative/jurisdictional boundaries. For example, the Chicago EA includes both the Chicago metropolitan area along with other counties in Illinois, Indiana, and Wisconsin where workers have strong connections to Chicago. EAs are very diverse in size, ranging from the greater Los Angeles

[4]This relative wage is also missing from the equivalent expression in Edmond, Midrigan and Xu (2015). They study a two-location model with a form of *aggregate symmetry* so that $W_j = \bar{W}_k(s)$ for $j, k = 1, 2$.

area which accounts for around 14.6% of total employment, all the way down to Scottsbluff, Nebraska (near the Wyoming border) which accounts for 0.0012% of total employment:

1	Los Angeles-Riverside-Orange County	14.6%	of total employment
2	New York-North New Jersey-Long Island	7.2%	
3	Chicago-Gary-Kenosha	6.9%	
	⋮		
169	San Angelo, TX	0.0013%	
170	Scottsbluff, NE-WY	0.0012%	

Sectors. We calibrate our model to match the operations of firms in 363 NAICS 6-digit manufacturing sectors, examples of which include *breakfast cereal* (sector 311230), *ready-mix concrete* (327320), *aircraft engine & engine parts* (336412), *optical instrument & lens manufacturing* (333314), and *wood kitchen cabinet & countertop manufacturing* (337110).

Labor supply. For each $j = 1, ..., J$ we measure the number of workers L_j as manufacturing employment from the County Business Patterns (CBP) aggregated to the EA level. We likewise measure the wage bill from the CBP aggregated to the EA level and choose the efficiency units E_j for each location so that the wage bill in our model for each location $W_j E_j L_j$ matches the wage bill measured in the data for that location.

Firms and establishments. On average there are about 600 firms nationally per NAICS 6-digit sector, most of which are small. But there is considerable variation in the number of firms across sectors, ranging from a low of $N(s) = 8$ firms nationally for *custom roll forming* (sector 332114), a specialized manufacturing activity focused on the metal forming process, to $N(s) = 11,492$ for *machine shops* (332710) and $N(s) = 14,279$ for *commercial printing* (323111), which covers printing of stationery, advertising materials, etc. All together we have $N = \sum_s N(s) = 219,365$ firms. We set the number of establishments for each firm by counting the number of EA locations where firm i has at least one establishment in National Establishment Time Series (NETS) county-level data aggregated to the EA level. So, for example, if a given firm has an establishment in each of two counties that belong to the same EA we call that 'one' establishment for the purposes of our model.

Firms and establishment locations. We then populate the economy by placing each firm's establishments exactly where they appear in the data, so every model firm mirrors a real firm's geographic footprint. Let $\mathcal{J}_i(s) \subseteq \{1, ..., J\}$ denote the set of locations where firm i has establishments and let $n_i(s) = |\mathcal{J}_i(s)| = \sum_j \mathbf{1}\{j \in \mathcal{J}_i(s)\}$ denote the number of locations where firm i has establishments. We take this set $\mathcal{J}_i(s)$ exactly from the data.

Firm-level productivity fixed effects. We abstract from location-specific firm productivity effects and assume that firm-level productivity can be written

$$z_{ij}(s) = z_i(s) \cdot \mathbf{1}\{j \in \mathcal{J}_i(s)\} \tag{34}$$

In other words, a firm's productivity is equal to the firm-level productivity fixed effect $z_i(s)$ in all locations where it has an establishment and zero in all locations where it has no establishments. Since we have taken the set of locations $\mathcal{J}_i(s)$ for each firm directly from the data, the only thing left to do is to assign these firm-level productivity fixed effects.

To assign the firm-level productivity fixed effects, we simulate a large number of paired uniform ranks (u, v) using a Gumbel copula

$$\mathcal{C}(u, v) = \exp\left(-\left[(-\ln u)^{\frac{1}{1-\rho}} + (-\ln v)^{\frac{1}{1-\rho}}\right]^{1-\rho}\right) = \text{Gumbel Copula}(\rho) \tag{35}$$

We then transform the first rank into a Pareto productivity draw $z = F^{-1}(u)$ where

$$F(z) = 1 - z^{-\xi} = \text{Pareto}(\xi)$$

and transform the second rank into an empirical draw of establishment counts. For each establishment-count category, e.g., all firms with n establishments, we take the simulated pool of productivity-count pairs and condition on that count to obtain candidate productivity draws. We then randomly sample the exact number of real firms in that category and assign these productivity levels to the corresponding firms in their real geographic locations.

EXAMPLE. Suppose in the data there are 1123 firms that have 10 establishments. In the model, we can place the establishments of each of these 1123 firms into the exact EA that they are in the data. But what we do not observe is their firm-level productivity. From the Gumbel simulation, however, we have an extremely large set of (productivity, establishment-count) pairs. We drop all pairs whose establishment count is not 10, which leaves us with a still very large sample of productivity draws from the conditional distribution of firm-level productivity (conditional on an establishment count of 10). We then randomly sample 1123 productivity draws with replacement from this conditional distribution and assign them to the 1123 firms we have already placed in their real geographic locations.

With this procedure in place, we are left with two parameters to pin down, the the Pareto tail ξ and the Gumbel rank correlation parameter ρ, as discussed further below.[5]

Trade costs. Following Caliendo, Parro, Rossi-Hansberg and Sarte (2018), we parameterize the sector-specific iceberg trade costs $\tau_{jk}(s)$ by assuming a sector-specific log-linear relationship between trade costs and physical distance d_{jk}

$$\ln \tau_{jk}(s) = \delta(s) \ln d_{jk} \tag{36}$$

This gives us a further set of parameters to pin down, the trade cost coefficients $\delta(s)$.

3.2 Calibration

Calibration strategy. We assign values to several conventional parameters that are held constant throughout all our quantitative exercises. We calibrate the remaining parameters internally using the simulated method of moments. We calibrate the parameters governing the distribution of firm-level productivity and the operations of multi-establishment firms to match establishment-level data from the US Census of Manufactures. We calibrate the parameters governing spatial trade frictions by requiring that the model reproduce gravity regressions based on the Commodity Flow Survey (CFS).

Assigned parameters. For parsimony, we assume that the elasticity of substitution across goods within a firm equals the elasticity of substitution across firms within a given sector, $\eta = \gamma$, i.e., goods produced by different establishments under the umbrella of the same firm are just as substitutable for one another as are goods sold by other firms within the same sector. This leaves us with two elasticities of substitution. Following Edmond, Midrigan and Xu (2023), we set the across-sector elasticity of substitution to $\theta = 1.25$ and the within-sector elasticity to $\gamma = 10$ so that the model matches the sector-level relationship between inverse markups and sales concentration observed in US manufacturing data.[6] This relationship can be obtained by multiplying both sides of equation (24) by market

[5] For the Gumbel copula in (35), the parameter $\rho \in [0, 1]$ corresponds to the robust rank correlation coefficient known as 'Kendall's tau' commonly used to summarize dependence in heavy-tailed distributions (Nelsen, 2006). If $\rho = 0$ the copula simplifies to $\mathcal{C}(u, v) = uv$ so that the ranks are independent, If $\rho \to 1$ the copula approaches $\mathcal{C}(u, v) = \min[u, v]$ so that the ranks are perfectly dependent. For simplicity we refer to ρ as the Gumbel correlation parameter.

[6] Edmond, Midrigan and Xu (2023) calibrate their parameters to match the slope coefficient of this relationship, estimated in differences over time, jointly with their other parameters which target measures of concentration in 4-digit US Census of Manufactures data. For their preferred specifications, they obtain estimates of the across-sector elasticity of substitution θ between 1.15 and 1.35 and estimates of the within-sector elasticity γ between 7 and 13. We set $\theta = 1.25$ and $\gamma = 10$ as the rough midpoints of these ranges.

Table 1: Parameterization

Parameter		Value	Target
Assigned Values			
Elas. subs. across sectors	θ	1.25	Sector-level markups and concentration
Elas. subs. within sectors	$\gamma = \eta$	10	(Edmond, Midrigan and Xu, 2023)
Method of Moments			
Pareto tail firm productivity	ξ	10.35	National concentration
Gumbel rank correlation	ρ	0.81	Employment share multi-estab firms
Trade cost wrt distance	$\delta(s)$		Gravity coeff. 3-digit NAICS

shares $\omega_{ik}(s)$ and summing across firms and locations to get

$$\frac{1}{\mu(s)} = \frac{\gamma - 1}{\gamma} - \left(\frac{1}{\theta} - \frac{1}{\gamma}\right) \text{HHI}(s) \tag{37}$$

where $\text{HHI}(s)$ denotes the sector's Herfindahl-Hirschman index of sales concentration.[7]

Calibrated parameters. We are left with the need to calibrate the Pareto tail, the Gumbel correlation, and the trade cost coefficients for each sector

$$\xi \,,\, \rho \,,\, \delta(s)$$

We calibrate these parameters internally using the simulated method of moments. We jointly target (i) measures of national sales concentration, to pin down the Pareto tail parameter ξ, (ii) measures of the employment share of multi-establishment firms to pin down the Gumbel correlation parameter ρ, and (iii) sector-level gravity regressions to pin down the trade cost coefficients, $\delta(s)$. Importantly, we calibrate the model using measures of *national* sales concentration, not local concentration.

(i) NATIONAL SALES CONCENTRATION. We target the average top-4 national sales shares and national sales HHI. In the US Census of Manufactures, the average top-4

[7] That is, if we let ω_i denote the sales share of firm i, then the HHI $:= \sum_i \omega_i^2$. For example, if there are N firms with $\omega_i = 1/N$ then the HHI $= 1/N$.

national sales share for 6-digit NAICS sectors is 42% while the average national HHI is 0.10.

(ii) OPERATIONS OF MULTI-ESTABLISHMENT FIRMS. In the US Census of Manufactures, about 4% of firms are multi-establishment and these multi-establishment firms account for about 54% of employment.

(iii) GRAVITY REGRESSIONS. Recall that $\bar{p}_{jk}(s)\bar{y}_{jk}(s)$ denotes the value of shipments from j to k. We estimate sector-specific gravity regressions of the form

$$\ln(\bar{p}_{jk}(s)\bar{y}_{jk}(s)) = \gamma_j(s) + \gamma_k(s) + \beta(s)\ln d_{jk} + \epsilon_{jk}(s) \tag{38}$$

where $\gamma_j(s)$,$\gamma_k(s)$ denote sector-specific source and destination fixed effects. We estimate these gravity regressions using county-to-county trade flows from the Commodity Flows Survey aggregated to the EA level for each 3-digit NAICS manufacturing sector. Our estimated slope coefficients $\beta(s)$, reported in the Appendix, measure how sensitive trade flows are to geographical distance. Goods that are more easily tradeable, such as *computers & electronics* (sector 334) and *electric equipment & appliances* (335), have estimated $\beta(s)$ that are small in magnitude. Goods that are less easily tradeable, such as *wood* (321), *petroleum, asphalt and coal* (324) and *non-metallic minerals* (327) have large negative estimated $\beta(s)$.

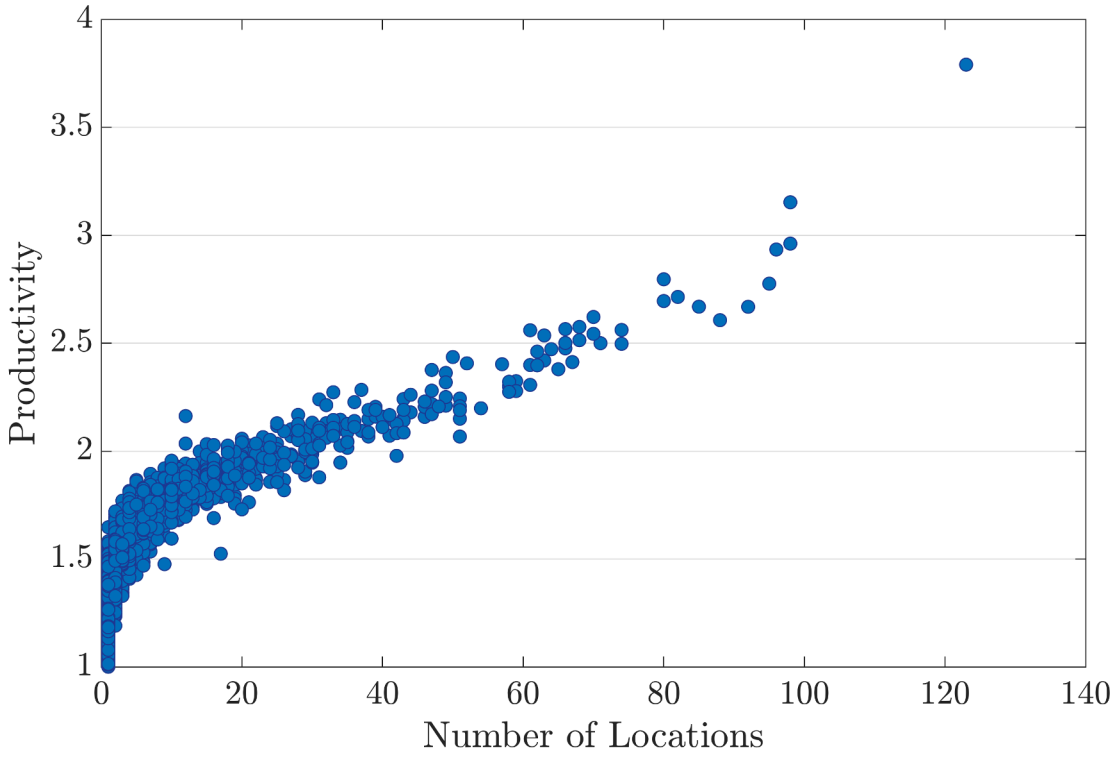
In the model we simulate data for each of our 363 6-digit sectors and, for each sector s, calculate the total value of shipments from j to k as

$$\bar{p}_{jk}(s)\bar{y}_{jk}(s) = \sum_{i=1}^{N(s)} p_{ijk}(s)y_{ijk}(s). \tag{39}$$

To be consistent with our empirical gravity regressions from the Commodity Flows Survey, we aggregate these simulated shipment flows to a 3-digit cluster of sectors and choose the parameters $\delta(s)$ in our specification (36) so that the estimated $\beta(s)$ in the model gravity regressions match their empirical counterparts from (38).

We report our internally calibrated parameters governing the productivity distribution and the operations of multi-establishment firms across locations in Table 1. Jointly with our other parameters, our model matches the data on national sales concentration with a Pareto tail $\xi = 10.35$, implying considerably thinner tails than the model of oligopolistic competition in Edmond, Midrigan and Xu (2023), which abstracts from spatial frictions. Our model matches the 54% employment share of multi-establishment firms with a Gumbel correlation of $\rho = 0.81$ between a firm's productivity fixed effect $z_i(s)$ and its number of establishments $n_i(s)$, as illustrated in Figure 1.

Figure 1: Firm-Level Productivity $z_i(s)$ and Establishment-Count $n_i(s)$



Model fit. We report key moments in the data and their model counterparts in Table 2, highlighting in red the moments targeted by our calibration procedure. The model does a good job of reproducing the average amount of national sales concentration, matching the national sales HHI exactly and slightly overshooting the national top-4 sales share. The model also reproduces the employment share of multi-establishment firms almost exactly. We report the 3-digit gravity coefficients $\beta(s)$ we estimate from the CFS and their model counterparts in Figure 2. Importantly, our model exactly reproduces the sector-level gravity effects that pin down our spatial trade frictions.

Model validation. In Table 2 we also report some key moments that were not targeted in our calibration exercise. In the data, local production is much more concentrated than national sales, the local production HHI is 0.36 compared to the national sales HHI of 0.10. Our model reproduces this fact almost exactly. That said, the model undershoots the national top-20 share and the sales share of multi-establishment firms.

Table 2: Model Fit

Moments [targeted]	Data	Model
National Concentration		
Top 4 sales share	0.42	0.44
Top 20 sales share	0.73	0.65
HHI sales	0.10	0.10
Local Concentration		
HHI production	0.36	0.37
Multi-Establishment Firms		
Fraction multi-establishment firms	0.03	0.03
Employment share of multi-establishment firms	0.54	0.53
Sales share of multi-establishment firms	0.62	0.55

4 Spatial distribution of concentration and markups

In this section we present our model's implications for the distribution of markups and sales concentration across the 170 EAs in our data. We first present our model's implications for local *sales* concentration and show that variation in local sales concentration across locations is not well explained by variation in local *production* concentration. We then present our model's implications for variation in markups across locations. Finally, we develop a number of additional measures of market competition and contestability and present a series of examples using these measures to understand the wide range of outcomes across US locations implied by our model.

4.1 Local sales concentration

Of key interest in our framework is how much competition firms face in the destination markets that they sell to. In less competitive markets, dominant firms will be able to charge high markups. Sales concentration in local markets is not something we can directly observe with the Census data. But given that our model does a good job of reproducing national sales concentration and matches local production concentration by construction, it seems natural to use the model to infer the amount of local sales concentration.

We report our benchmark model's implication for local sales concentration in Table 3.

Figure 2: Gravity Coefficients $\beta(s)$ in Data and Model

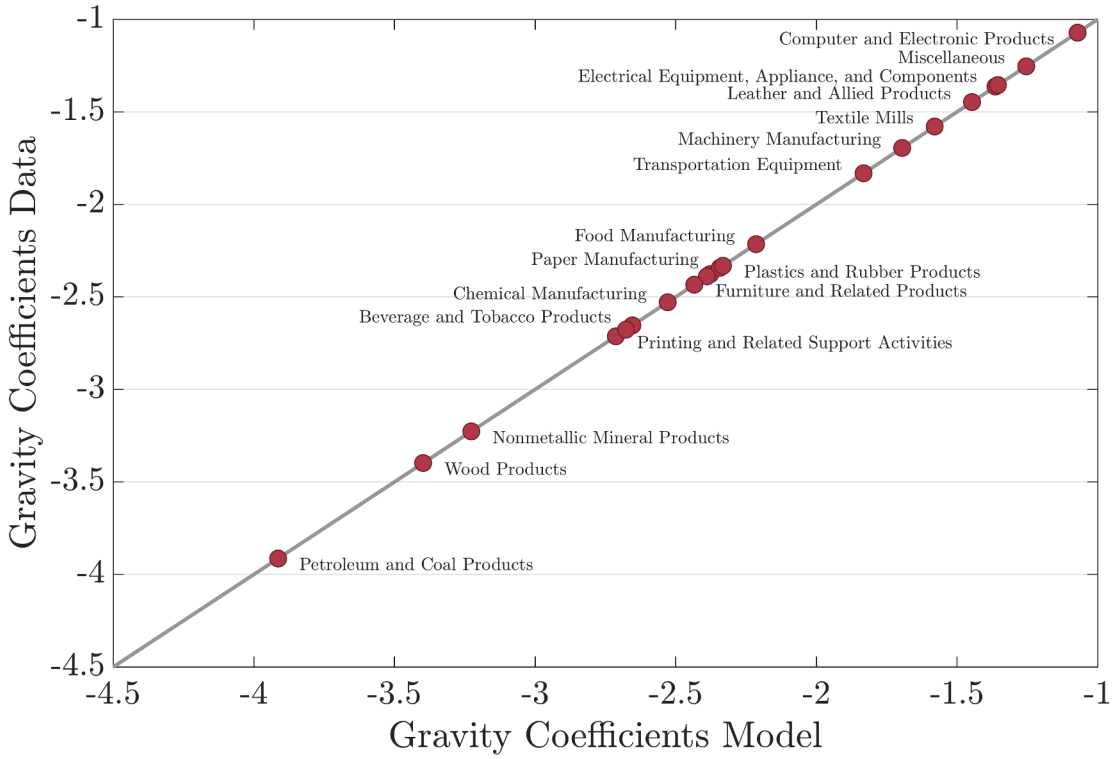


Table 3: Local Sales Concentration

Moment	Model
Top 1 sales share	0.27
Top 4 sales share	0.58
Top 20 sales share	0.85
HHI sales	0.15

Intuitively, we find that:

National Sales		**Local Sales**		**Local Production**
HHI $= 0.10$	$<$	HHI $= 0.15$	$<$	HHI $= 0.37$
least concentrated				*most concentrated*

Since the inverse of the HHI corresponds to the number of equally-sized firms, to interpret these concentration statistics more intuitively, it is as if there are on average 10 equally-sized firms nationally, about 6-7 equally-sized firms selling locally, and just under 3 equally-sized firms producing locally. The fact that local sales concentration is less than local production concentration reflects the fact that most goods are at least somewhat tradeable. While the production of goods may be quite concentrated at specific source locations, destination markets generally receive goods from a range of sources, pushing local sales concentration lower than production concentration. That said, the fact that local sales concentration is greater than national sales concentration reflects the fact that goods cannot be traded *frictionlessly*, i.e., the economy is genuinely geographically segmented. In the rest of this section we document these results more systematically.

4.2 Sales concentration vs. production concentration

Our model predicts that local sales concentration is both lower and less dispersed than local production concentration. This is because goods that are easily tradeable can be shipped from the most productive source locations to almost any destination market, increasing the amount of competition amongst producers of tradeable goods in those markets. By contrast, goods that are less easily tradeable will be shipped to a more limited set of destinations, inhibiting the amount of competition amongst producers of less-tradeable goods in those markets. Because of this, measures of local production concentration — of the kind readily computed from data on shipments — may provide a poor guide to the amount of competition firms face in the locations where people live and consume.

To see this, Figure 3 reports average local production concentration (as measured by HHIs) across the 170 EAs in our data. These range from around 0.16 to nearly 1, indicating areas where many goods are produced by a single firm. Figure 4 reports the average local sales concentration (again measured by HHIs) in our model. These range from around 0.12 to 0.17. In short, the sales concentration in our model is indeed both much lower on average and much less variable than production concentration in the data. Moreover production concentration and sales concentration are not strongly correlated across locations. For example, greater New York has both low production and sales concentration, while the greater Seattle area has only moderate production concentration but some of the highest sales concentration in our model, as discussed in more detail below.

Figure 5 reports the scatter of observed local production concentration from our data against the local sales concentration HHIs for the 170 EAs in our model. If local production concentration was a good predictor of local sales concentration, we would expect this scatter to be clustered tightly around the $45°$-line. But instead the scatter is close to a horizontal line — local production concentration is simply not very informative about local sales concentration. If anything, the slope coefficient is slightly negative, indicating

Figure 3: Local Production Concentration (HHI)

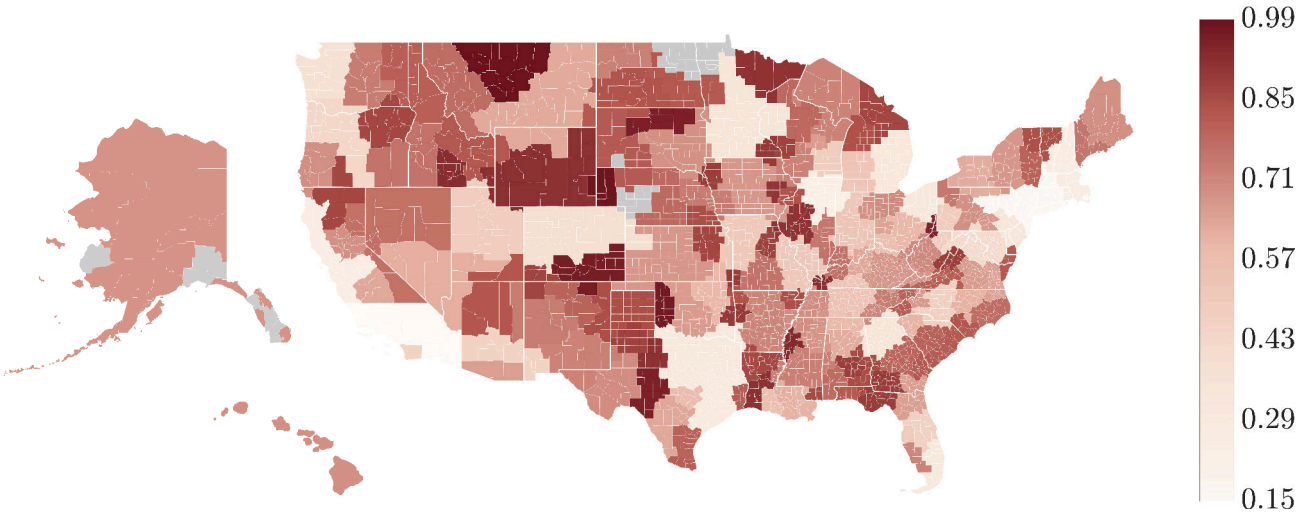
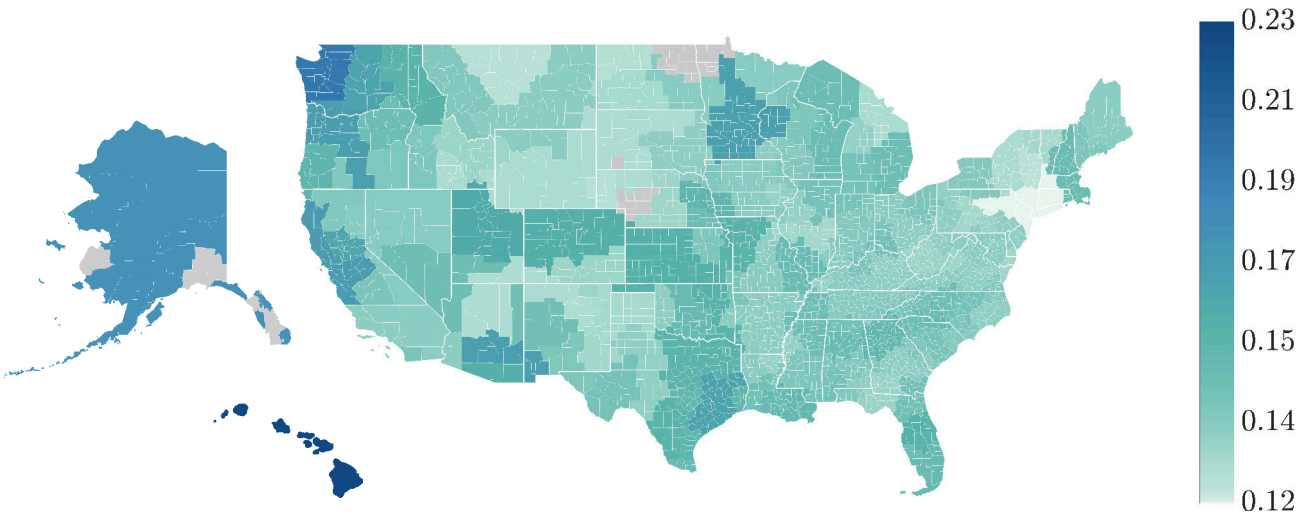


Figure 4: Local Sales Concentration (HHI)



that having higher production concentration predicts that a location will have *lower* sales concentration. In this sense, our model implies that the usual kind of local production concentration that we can readily measure with data on shipments is simply not very informative about the local sales concentration that matters for competition and measures of market power.

Spatial frictions and concentration. To further highlight the importance of tradeability, Figure 6 reports the scatter of local production HHIs in our data against the local sales HHIs in our model when we split 3-digit sectors into 'high gravity' sectors with high spatial frictions that are relatively costly to ship across locations, e.g., *petroleum & coal* products or *wood* products and 'low gravity' sectors with low spatial frictions that are much less costly to ship across locations, e.g., *computer & electronic* products. On average, high gravity sectors have higher levels of sales concentration than low gravity sectors. In high gravity sectors, trade costs can substantially limit the amount of competition — especially for smaller and more geographically remote locations. High gravity sectors also have sales concentration that is considerably more variable across locations than low gravity sectors. Indeed we see that for low gravity sectors, the local sales HHIs are tightly clustered at just above 0.1, i.e., just above the national sales HHI. For these low gravity sectors, goods are traded in something much closer to a single national market.

Figure 5: Production HHI $\neq$ Sales HHI

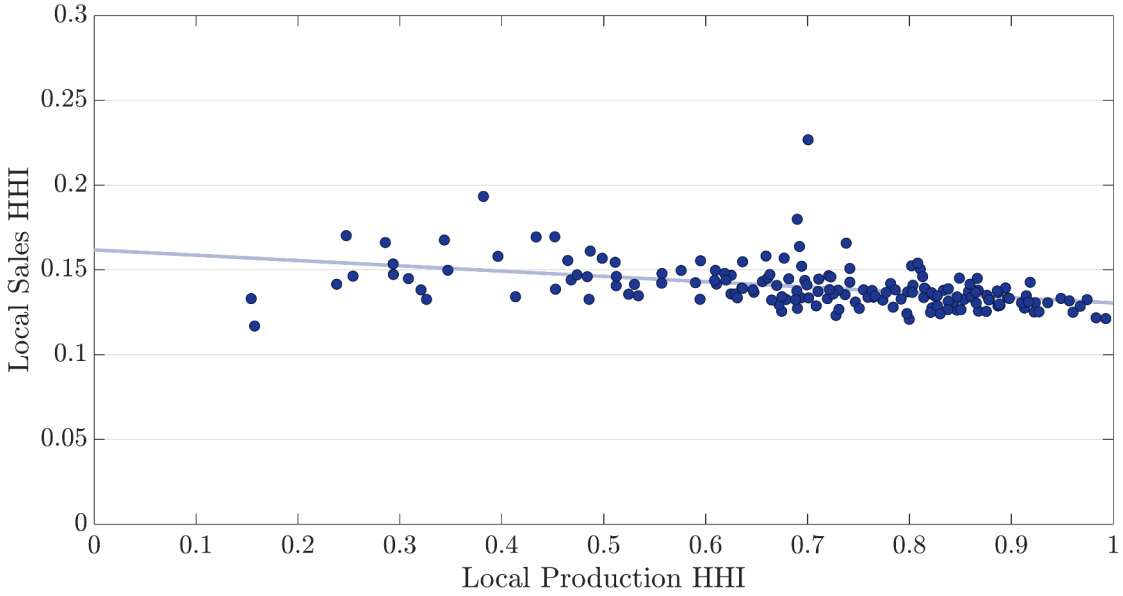


Figure 6: High Gravity Sectors Have Higher Sales HHI

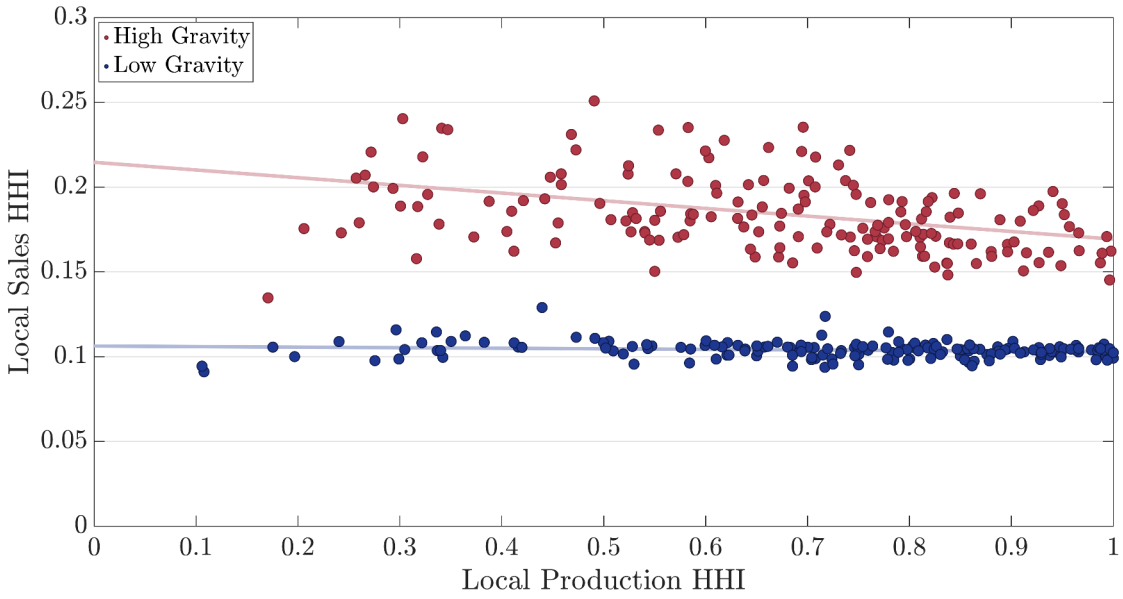
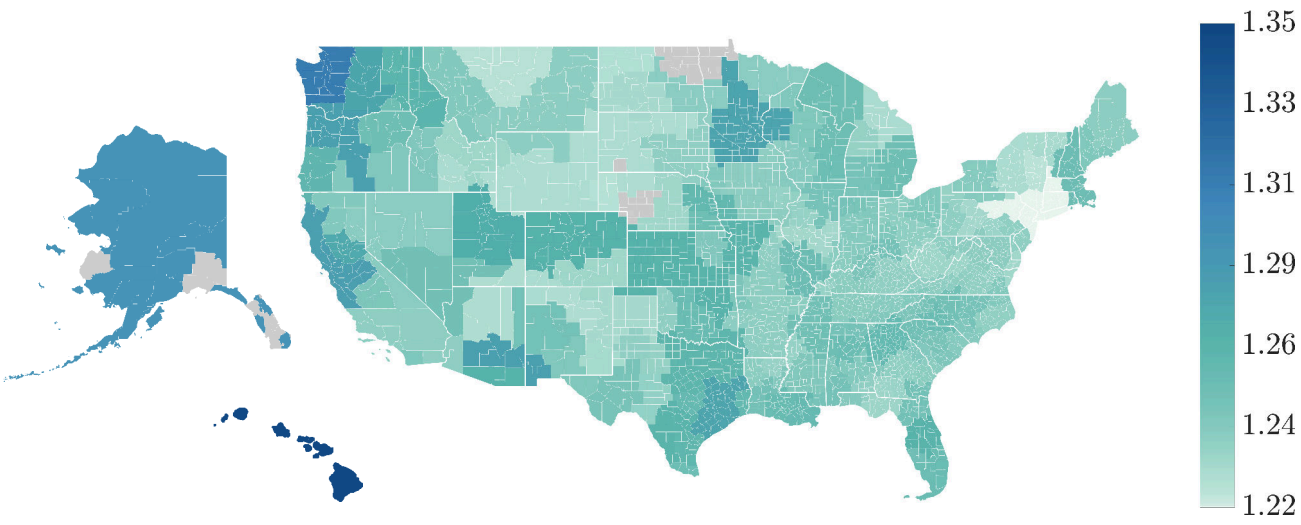


Figure 7: The Spatial Distribution of Markups



4.3 *Markups*

In our model, a location's sales-weighted harmonic average markup μ_k is in one-to-one correspondence with the location's sales-weighted arithmetic average HHI.[8] Because of this, the spatial variation in markups is qualitatively exactly the same as the spatial variation in local sales concentration documented above. For completeness, we now briefly summarize the spatial distribution of markups in our benchmark model.

The aggregate markup (unconditional sales-weighted harmonic average) in our model is 1.26. Figure 7 reports our model's implications for the full geographic variation in average markups μ_k across the 170 EAs in our data. These range from a low of 1.22 in the greater New York area, through San Antonio with an average markup of 1.26 to highs of 1.29 in Anchorage and 1.35 in Honolulu. These location-specific averages μ_k mask extensive variation in markups $\mu_{ik}(s)$ across sectors s within each location k and across firms i within each sector s. Across all 6-digit NAICS sectors in our model, the lowest percentile of markups is 1.13, the median markup is 1.23, and the top percentile of markups is 1.62.

[8] Similar to (37), this follows by aggregating (24) over firms and sectors to get the location-level relationship

$$\frac{1}{\mu_k} = \frac{\gamma - 1}{\gamma} - \left(\frac{1}{\theta} - \frac{1}{\gamma}\right) \text{HHI}_k$$

between average markups μ_k and average sales concentration. Both these variables are of course endogenous, jointly determined in general equilibrium. There is no sense in which one independently causes the other.

4.4 *Understanding variation in sales concentration and markups*

If raw production concentration is not very helpful in understanding the spatial variation in sales concentration and markups, what does explain these outcomes? Table 4 reports key statistics for a selection of the 170 EAs chosen to highlight the main economic forces in our model. The locations are ordered by size and span a wide range, from Los Angeles with manufacturing employment roughly 25 times the economy-wide average down to Scottsbluff at roughly 1/500th of the average. Across locations we observe two broad patterns. First, for large productive locations that have substantial domestic production and import relatively little, what matters most for local sales concentration is the structure of home supply — in particular, how many firms compete head-to-head, and how dominant the lowest-cost producer is. Second, for small unproductive locations that are forced to import almost everything, what matters most is the structure of import competition — how many outside sources are realistically cost-competitive and how concentrated import sourcing is across origins. Even so, within each broad category there is a perhaps surprisingly wide range of outcomes. New York and Seattle are both large and have similar import shares but have quite different markup levels. Honolulu and Scottsbluff are both small and import almost everything but end up with very different levels of local sales concentration and markups.

Key statistics. Table 4 reports for each selected location k employment L_k, the average markup, μ_k, the average local sales HHI, the average share of expenditure on imports, the average number of plants, the average fraction of sectors with domestic production, and a number of measures of market *contestability* for each destination k.

Market contestability. Let $\mathsf{c}_{ij}(s) = W_j/z_{ij}(s)$ denote a firm's unit cost at source j and let $\mathsf{C}_j(s) = \left(\sum_i \mathsf{c}_{ij}(s)^{1-\gamma}\right)^{1/(1-\gamma)}$ denote the implied CES source cost index. Likewise let $\mathsf{C}_{jk}(s) = \tau_{jk}(s)\mathsf{C}_j(s)$ denote the trade-cost-inclusive cost index for goods delivered from source j to destination k. Then let $\mathsf{B}_k(s)$ denote the costs of the *best* outside source for k

$$\mathsf{B}_k(s) := \min_{j \neq k} \mathsf{C}_{jk}(s) \tag{40}$$

and let $j_k^*(s)$ denote the source location that achieves this minimum. For each location k we then compute and report in Table 4 the following summary statistics:

- CONTEST: The average number of outside sources j where $\mathsf{C}_{jk}(s)$ is within 10% of home cost $\mathsf{C}_k(s)$.
- TOP HOME COST SHARE: The average share of costs accounted for by the lowest cost home producer.

- BEST GAP: The log delivered cost advantage of the best outside source

$$\begin{aligned}\ln \mathsf{B}_k(s) - \ln \mathsf{C}_k(s) &= \ln \mathsf{C}_{j^*,k}(s) - \ln \mathsf{C}_k(s) \\ &= \ln \tau_{j^*,k}(s) + (\ln \mathsf{C}_{j^*}(s) - \ln \mathsf{C}_k(s))\end{aligned}$$

- TRADE COST: The contribution of the trade cost term $\ln \tau_{j^*,k}(s)$ to the Best Gap.
- COMPETITIVENESS: The contribution of the log difference in source costs $\ln \mathsf{C}_{j^*}(s) - \ln \mathsf{C}_k(s)$ to the Best Gap.

With these statistics in hand we review outcomes for a number of illustrative locations. First lets consider the similarities and differences between some large, productive locations.

New York. The greater New York area is extremely large, with high employment and a large number of plants covering almost every sector. In our model, New York's local sales concentration and markups are low, with an average markup of $\mu_k = 1.22$ and a sales HHI of 0.12. The import share is relatively low, 0.36. On average the best outside source of supply has delivered costs 0.23 log points higher than the costs of domestic supply, and on average the best outside source of supply is only slightly, -0.04 log points, more competitive than home supply. Cost-realistic outside options are rare: the average number of other locations that have delivered costs within 10% of domestic costs is 0.284, i.e., there is generally less than 1 effective outside competitor location. The low import share therefore reflects the difficulty of penetrating the New York market given the combination of sizeable trade costs and low domestic costs. But despite the lack of import discipline, concentration stays low because home supply features extensive head-to-head competition — there are roughly 61 plants per sector on average, and the lowest-cost home plant accounts for 'only' about 51% of home supply in cost terms. Overall, greater New York's low local sales concentration reflects the strong competition amongst its many domestic producers, not import contestability.

But not all large productive locations have strong competitive home conditions.

Seattle. The greater Seattle area is also large and productive with home production in over 90% of sectors. Like New York, imports provide little discipline, the best outside source has delivered costs 0.38 log points above home costs and the number of cost-realistic outside locations is just 0.41. But unlike New York, in our model Seattle has high markups and sales concentration, with an average markup of $\mu_k = 1.31$ and a sales HHI of 0.19, the highest we see among the large locations. What makes Seattle different is the structure of its home supply. The lowest-cost home plant accounts for considerably more of home supply, about 66% in cost terms compared to 51% for the lowest-cost home producer in

New York, and there are only about 10 plants per sector compared to about 61 in New York. With fewer plants and a more dominant low-cost producer in each sector, there is less head-to-head competition among domestic firms, which translates directly into higher concentration and higher markups. In short, Seattle shows that being large and productive is not sufficient for low concentration, what really matters is the depth and competitiveness of home supply.

What about some smaller, less productive locations? Are they fated to have weak competitive conditions?

Scottsbluff. Scottsbluff is a tiny location on the Nebraska/Wyoming border. Home production is minimal, with fewer than one plant per sector on average and home production present in only 13% of sectors. As a result, Scottsbluff is a near-complete importer with an import share of essentially 1.00. One might expect such dependence on imports to leave Scottsbluff vulnerable to high markups, but in fact the opposite is true — the average markup is $\mu_k = 1.23$ and the sales HHI is 0.12, both among the lowest we find. This happens because Scottsbluff's imports are intensely contested. Even though Scottsbluff is in the middle of the country, it happens to be roughly equidistant from many production centers. The number of cost-realistic outside locations is nearly 48, one of the largest we see. With so many suppliers competing head-to-head, no single firm dominates. That said, while Scottsbluff has strong competitive conditions, the low markups do not mean that Scottsbluff is cheap. The CES price index works out to be $P_k = 0.92$, well above locations like New York ($P_k = 0.70$) and Minneapolis ($P_k = 0.76$) which have similar average markups, reflecting the high cost of delivering goods to a remote location with almost no domestic production.

But sometimes being small and remote really just means what you expect.

Honolulu. Honolulu is a mid-sized location that, like Scottsbluff, has limited home production and is heavily reliant on imports, with an import share of 0.62. But where Scottsbluff has low sales concentration because of its diversified and highly-contested imports, Honolulu has the opposite: in our model, Honolulu's local sales HHI is 0.23 and the average markup is $\mu_k = 1.35$, among the highest we find. Scottsbluff is remote, but not overwhelmingly so, and more importantly is comparably close to many competing sources of supply. Honolulu is *much* more remote. Honolulu's trade costs are much larger than elsewhere — the trade cost component of the best outsider gap is 0.98, roughly double that of any mainland location. Even though the best outside source is substantially more cost-competitive than home in net terms, about -0.25 log points, the trade costs overwhelm this advantage, leaving the best outsider 0.74 log points more expensive than

home overall. Because of this, cost-realistic outside options are rare (there are only about 0.92 such locations) and home supply, where it exists, is concentrated (the lowest-cost home plant accounts for 79% of home supply in cost terms). The combination of remoteness, low contestability, and incomplete home production coverage allows a few suppliers to dominate, sustaining the highest markups and concentration of any location in our model.

These four examples illustrate the key forces that operate, in various combinations, across all 170 locations in our model. Locations with low contestability and concentrated home supply tend to have higher markups, while locations with either deep home competition or diversified import sourcing tend to have lower markups — though as the contrast between New York and Seattle or Scottsbluff and Honolulu makes clear, the finer details of each location's geography and production structure still play an important role in determining their individual location-level outcomes.

But how much does all this matter in the aggregate?

Table 4: Characteristics of Select Locations

Location	L_k	μ_k	HHI	Import Share	Plants	Coverage	Contest	Top 1 Share	Best Gap	
									Trade Cost	Comp.
Los Angeles	24.9	1.24	0.13	0.22	56.79	0.98	0.223	0.505	0.277	0.043
New York	12.3	1.22	0.12	0.36	61.02	0.98	0.284	0.511	0.270	−0.044
San Francisco	7.84	1.28	0.17	0.26	24.98	0.95	0.143	0.598	0.306	0.084
Minneapolis	4.49	1.28	0.17	0.39	14.71	0.93	0.597	0.650	0.473	−0.091
Seattle	3.24	1.31	0.19	0.37	10.21	0.91	0.409	0.657	0.426	−0.049
Phoenix	2.71	1.28	0.17	0.50	8.61	0.86	1.109	0.680	0.372	−0.107
San Antonio	0.641	1.26	0.16	0.77	3.85	0.76	5.107	0.766	0.336	−0.342
Honolulu	0.298	1.35	0.23	0.62	1.74	0.55	0.917	0.786	0.983	−0.249
Anchorage	0.289	1.29	0.18	0.77	0.99	0.36	0.339	0.782	0.906	−0.240
Redding	0.027	1.24	0.14	0.97	0.52	0.27	17.43	0.856	0.367	−0.642
Minot	0.010	1.23	0.12	0.97	0.24	0.15	20.07	0.897	0.498	−0.560
Scottsbluff	0.002	1.23	0.12	1.00	0.18	0.13	47.63	0.935	0.442	−0.916

Notes: Locations are ordered by L_k (manufacturing employment relative to the economy-wide average). Markup is the sales-weighted average markup μ_k. Sales HHI is the sales-weighted average local Herfindahl index. Import Share is the sales-weighted average import penetration. Plants is the average number of production plants per sector. Home Coverage is the expenditure-weighted fraction of sectors with home production. Contest. is the number of outside origins within 10% of home delivered cost. Top 1 Home Cost Share is the cost share of the lowest-cost home plant. Trade Cost and Competitiveness are the two components of the log delivered cost advantage of the best outsider relative to home: Trade Cost is $\log \tau_{j^*k}$ and Competitiveness is $\log C_{j^*} - \log C_k$, so that the best outsider gap $\log B_k - \log C_k$ equals Trade Cost + Competitiveness. All statistics are sales-weighted averages across sectors. Location names correspond to BEA Economic Areas: Los Angeles = Los Angeles-Riverside-Orange County, CA-AZ; New York = New York-North New Jersey-Long Island, NY-NJ-CT-PA-MA-VT; San Francisco = San Francisco-Oakland-San Jose, CA; Minneapolis = Minneapolis-St. Paul, MN-WI-IA; Seattle = Seattle-Tacoma-Bremerton, WA; Phoenix = Phoenix-Mesa, AZ-NM; San Antonio = San Antonio, TX; Honolulu = Honolulu, HI; Anchorage = Anchorage, AK; Redding = Redding, CA-OR; Minot = Minot, ND; Scottsbluff = Scottsbluff, NE-WY.

5 Quantitative importance of spatial frictions

In this section we present our main quantitative results on the aggregate effects of spatial frictions and geographic variation in market power and concentration. We first compute equilibrium outcomes for two extreme cases, intranational 'autarky' where trade costs are so high as to prohibit any trade between locations, and 'free trade' where there are no trade costs at all. Our benchmark calibration lies between these extremes, closer to the free trade end. We use these results to build intuition for which locations are most disadvantaged by spatial frictions. We then show that abstracting from spatial frictions leads to a quantitatively significant *understatement* of the macroeconomic losses associated with market power. With these results in hand, we show that our model can generate *endogenously divergent* trends in national and local concentration in response to falling trade costs — a fact that has been extensively debated in the recent literature. Finally, we document our model's implications for the consumption gains from reductions in trade costs and show that the smallest, poorest, most remote locations gain substantially more than the largest, richest, most economically central locations.

5.1 From autarky to free trade

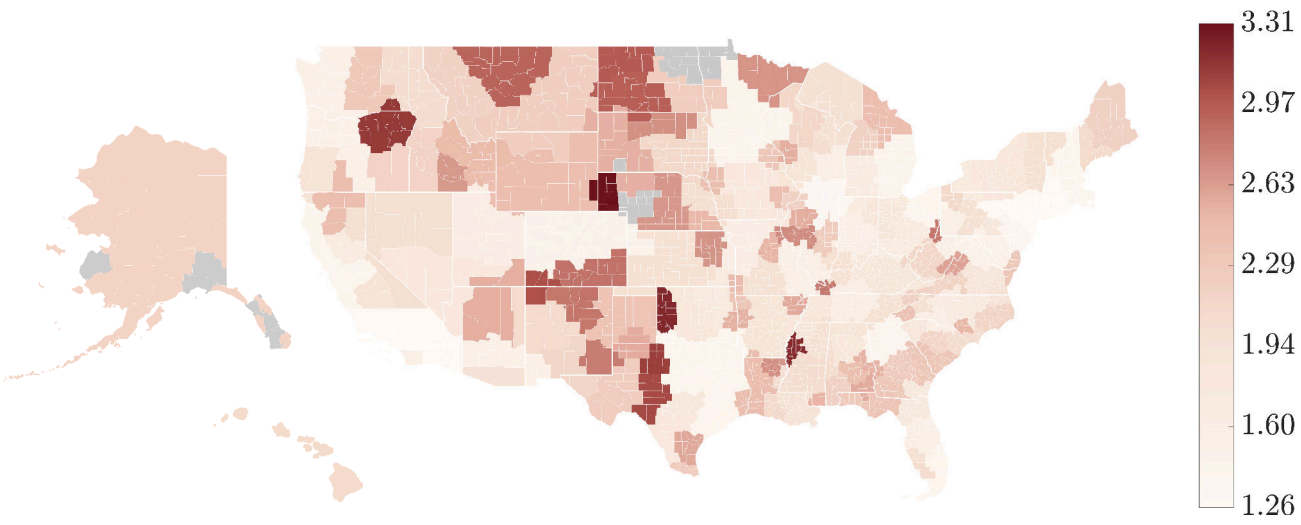
What would happen to market power and economic activity across the US if there was much less trade between locations? Which locations would be most affected by this?

Autarky. To build intuition for this, we consider the case of intranational autarky, setting $\tau_{jk}(s) = +\infty$, holding other parameters fixed. Figure 8 reports our model's implications for average markups μ_k across the 170 EAs in our model for this extreme case. Overall, markups are much larger and much more variable than in our benchmark model. This is because without trade, local sales concentration mirrors local production concentration. Smaller, less productive locations have much higher sales concentration and hence higher markups than in our benchmark model with realistic trade flows.

For example, under autarky Honolulu's average markup is 2.09 as opposed to 1.35 in our benchmark. Scottsbluff is even more affected. Remember that in our benchmark calibration both Honolulu and Scottsbluff import almost everything, but Scottsbluff benefited from being near-equidistant to many competitive sources of supply. Under autarky these hypothetical sources of supply are inoperative and Scottsbluff's average markup is 3.31, the highest of all locations under autarky, as opposed to 1.23 in our benchmark. In other words, shutting trade between locations down *flips* Scottsbluff from a surprisingly low-markup location, benefitting from ruthless competition between importers, to the highest-markup, least competitive place in the whole country.

The lack of trade matters less for larger, more productive locations. New York's average

Figure 8: Markups in Autarky



markup is 1.26 under autarky as opposed to 1.22 in our benchmark, consistent with our previous finding that New York's low markups are driven by its deep, competitive home market conditions not import competition. Locations like Seattle are a bit in between, with an average markup of 1.50 under autarky as opposed to 1.31 in our benchmark. This bigger jump is consistent with our previous finding that Seattle has more concentrated home market conditions. Shutting down trade matters more for Seattle than for New York but not nearly as much as for the really small and remote locations like Scottsbluff and Honolulu.

Free trade. Under free trade, $\tau_{jk}(s) = 1$, there is no geographic segmentation in product markets. Firms in each sector s compete in a single national market and set the same markup in every destination. Because of this, under free trade average markups are the same in every location, $\mu_k = \mu$. This common average markup level works out to be $\mu = 1.21$.[9] In our benchmark model large, productive locations have average markups around 1.22 (in New York) or 1.24 (Los Angeles) and so are already quite close to this free trade limit. Smaller, less productive locations are further away from this free trade limit, but overall our benchmark calibration is closer to the free trade case than the autarky case.

[9]Even under free trade there is of course still extensive variation in markups across firms within a given sector and across sectors, as discussed below.

5.2 *Abstracting from spatial frictions understates market power*

We next quantify the significance of geography and spatial frictions for aggregate outcomes. To do this, we use an otherwise equivalent model that abstracts from geography and spatial frictions but matches the same facts on national sales concentration as our benchmark model. The results of this exercise are given in Table 5, which reports the distribution of sector-level markups and the aggregate markup. Across all sectors, markups in the model without geography are both lower and less dispersed than in our benchmark model with geography. Overall the model without geography implies both a lower aggregate markup, down from 1.26 to 1.19, and lower productivity losses due to misallocation. Markup dispersion falls considerably, with the log p90/p50 markup ratio down from about $\ln(1.41/1.23) = 0.137$ in our benchmark model to $\ln(1.29/1.18) = 0.089$ in the model without geography — a fall of about one-third. In this sense, abstracting from spatial frictions leads to a quantitatively significant *understatement* of the macroeconomic losses associated with market power.

A macro model of market power that assumes away intranational geography and spatial frictions is equivalent to pushing the economy to the free trade limit and hence overstates the strength of competition in many places, even quite large ones. Roughly speaking, the model without geography and spatial frictions acts as if the competitive conditions of New York (or even better) prevail everywhere in the country.

Gravity. Another way to see the importance of geography and spatial frictions is to consider the differences between sectors which produce highly tradeable goods and sectors which produce intrinsically much less tradeable goods. To do this, we split our 3-digit sectors into 'low gravity' sectors, facing weak spatial frictions, and 'high gravity' sectors, facing strong spatial frictions, and then compute the ratio of the markup in each sector in our benchmark model to its counterpart markup in the model without geography. The distribution of these relative markups for the two categories, low gravity and high gravity, is shown in Figure 9. In the benchmark model with geography, sector-level markups are larger and more dispersed and this effect is concentrated in high gravity sectors. For low gravity sectors, geography hardly matters — goods are traded in something close to a single national market even in our benchmark, so removing spatial frictions makes little difference.

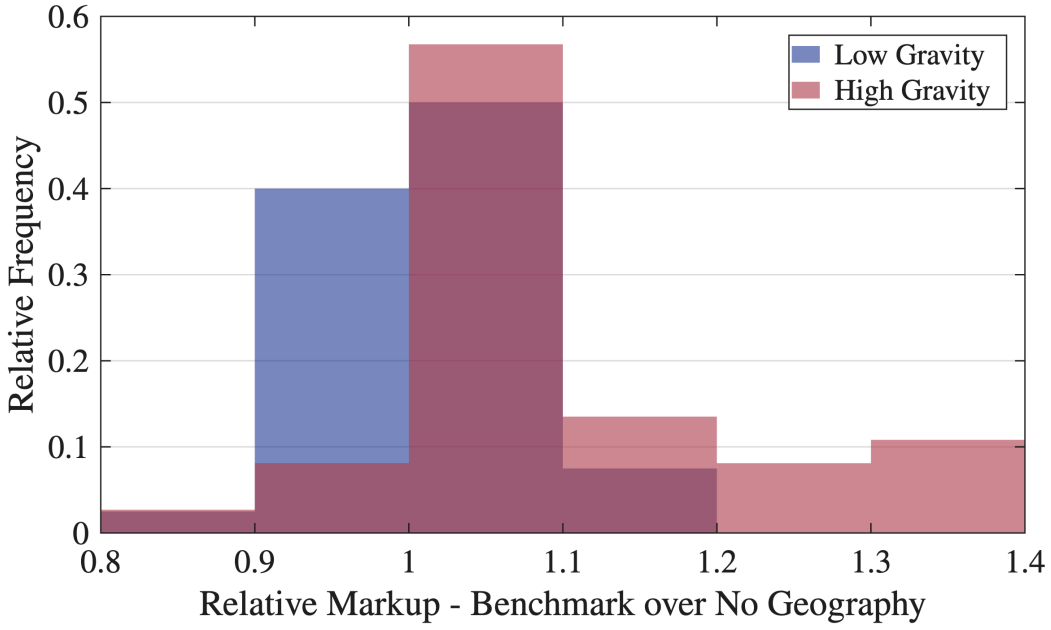
5.3 *Divergence between local and national concentration*

Recent empirical research has documented significantly different trends in national sales concentration and local sales concentration. National sales concentration, along with

Table 5: Markup Distribution, No Geography

Percentile	Benchmark Model	No Geography
p01	1.13	1.13
p10	1.15	1.14
p25	1.18	1.16
p50	1.23	1.18
p75	1.30	1.22
p90	1.41	1.29
p99	1.62	1.50
Aggregate Markup	1.26	1.19

Figure 9: Role of Gravity



local production concentration, has been on the rise since the early 1980s. But local sales concentration has been on the decline. We now show that these divergent trends in national and local sales concentration emerge naturally in our model when intranational trade costs are falling over time.

Reduction in trade costs. Over time, improvements in transportation technology and infrastructure should decrease trade costs, i.e., gravity effects should be becoming weaker. Consistent with this, using interregional trade data, Coşar, Osotimehin and Popov (2024) find a 15 to 20% decrease in manufacturing distance elasticities from 1963 to 2017. Since our benchmark model is calibrated to current data, to replicate the conditions of 1963 we increase trade cost elasticities $\delta(s)$ uniformly by 20% for all 3-digit NAICS manufacturing sectors. Table 6 reports the effects of such changes in trade costs on concentration. Moving forward from 1963 to the present, the model predicts that in response to a reduction in trade costs the average national top-4 sales share increases modestly from 0.43 to 0.44 while the average local top-4 sales share decreases from 0.61 to 0.58. In short, the model predicts that a reduction in trade costs endogenously drives national sales concentration and local sales concentration in opposite directions. Intuitively, in response to lower trade costs, the most productive firms expand by accessing more distant markets. This increases national sales concentration and increases local production concentration. But this pattern of expansion also leads to more competition in destination markets and hence lower local sales concentration.

Divergence or convergence? It is conventional in this literature to refer to these opposite movements in national and local sales concentration as a form of *divergence* — and we have used this language too. But from the perspective of our model, the opposite-signed changes in national and local sales concentration in response to a change in trade costs are in fact a kind of *convergence*, with national sales concentration increasing from low initial levels and local sales concentration decreasing from high initial levels. In the limit as intranational trade costs disappear, the distinction between national and local sales concentration also disappears. This can be seen in the last column of Table 6, which shows that in this free trade limit the national concentration moments exactly equal the local concentration moments. An observer of this process would see a stark pattern of increasing national concentration and decreasing local concentration, e.g., with the average national top-4 sales share starting at 0.46 in our benchmark economy and increasing from below to its limit of 0.57 while the average local top-4 sales share starts at 0.69 in our benchmark economy and decreases from above to the same limit. In this limit, there is effectively a single national market for each good.

5.4 Implications for markups and consumption

This exercise makes clear that a reduction in intranational trade costs can lead to an increase in national sales concentration. We now show that nonetheless this reduction in trade costs makes markets more competitive and increases consumption per worker in most locations, despite the increase in national sales concentration. The increase in national sales concentration is simply a byproduct of the most productive firms being able to sell in more locations. But what really matters for competitive conditions is how much competition these firms face in the markets where they sell their goods.

Changes in markup distribution. To understand these changes in competitive conditions, Table 7 reports the effects of changes in intranational trade costs on the sector-level markup distribution and the aggregate markup as we move along the spectrum between autarky and free trade. Since markups are determined in large part by the amount of competition in local markets, and local sales concentration is declining as trade costs decrease from 1963 to the present, it is not surprising that we find that the reduction in trade costs leads to both lower markups and lower markup dispersion — and hence lower productivity losses due to misallocation. For example, starting in 1963 with 20% higher trade costs and moving forward to our benchmark economy, we find that the aggregate markup decreases from 1.27 to 1.26 while markup dispersion, as measured by the log p90/p50 ratio, decreases from about 0.150 to about 0.137. Further decreases in trade costs in turn lead to further decreases in the aggregate markup and markup dispersion. That said, while the qualitative direction of these changes is clear, the changes are modest — e.g., with the aggregate markup decreasing by about 0.8% in response to a 20% reduction in trade costs from 1963 to our benchmark economy. But as we will now see, these modest changes in the sector-level markup distribution mask considerable heterogeneity in markup changes across locations.

Spatial heterogeneity in markup changes. To see the heterogeneity in markup changes across locations, Figure 10 reports the percentage decrease in average markups across our 170 EAs in response to a 20% reduction in intranational trade costs. There is considerable spatial variation in markup changes, ranging from around a 0.5% decrease for greater New York and around a 0.6% decrease for greater Los Angeles to around a 1.8% or 1.9% decrease in rural Utah, Colorado and Kansas. In other words, the markup changes in more remote locations are up to around 4 times as large as the markup changes in more central locations. Consistent with our previous results, the pro-competitive effects of a given reduction in trade costs are strongest for locations where competitive conditions are weakest to begin with and more modest for locations like New York where domestic producers already compete intensely (see Edmond, Midrigan and Xu, 2015, for further

discussion of these pro-competitive effects).

Table 6: Changes in Trade Costs

	Autarky	+20%	Benchmark	−20%	Free Trade
Increasing National Sales Concentration ⇑					
Top 4 share	0.32	0.43	0.44	0.45	0.49
HHI sales	0.07	0.10	0.10	0.10	0.11
Increasing Local Production Concentration ⇑					
HHI production	0.31	0.36	0.37	0.38	0.40
Decreasing Local Sales Concentration ⇓					
Top 4 share	0.77	0.61	0.58	0.56	0.49
HHI sales	0.31	0.16	0.15	0.13	0.11

Table 7: Effects of Changes in Trade Costs on Markup Distribution

Percentile	Autarky	+20%	Benchmark	−20%	Free Trade
p01	1.19	1.13	1.13	1.12	1.12
p10	1.28	1.16	1.15	1.14	1.13
p25	1.36	1.19	1.18	1.17	1.15
p50	1.48	1.24	1.23	1.21	1.19
p75	1.65	1.32	1.30	1.28	1.26
p90	1.95	1.44	1.41	1.39	1.34
p99	3.92	1.66	1.62	1.58	1.52
Aggregate Markup	1.52	1.27	1.26	1.24	1.21

Figure 10: Percentage Markup Decrease from 20% Reduction in Trade Cost

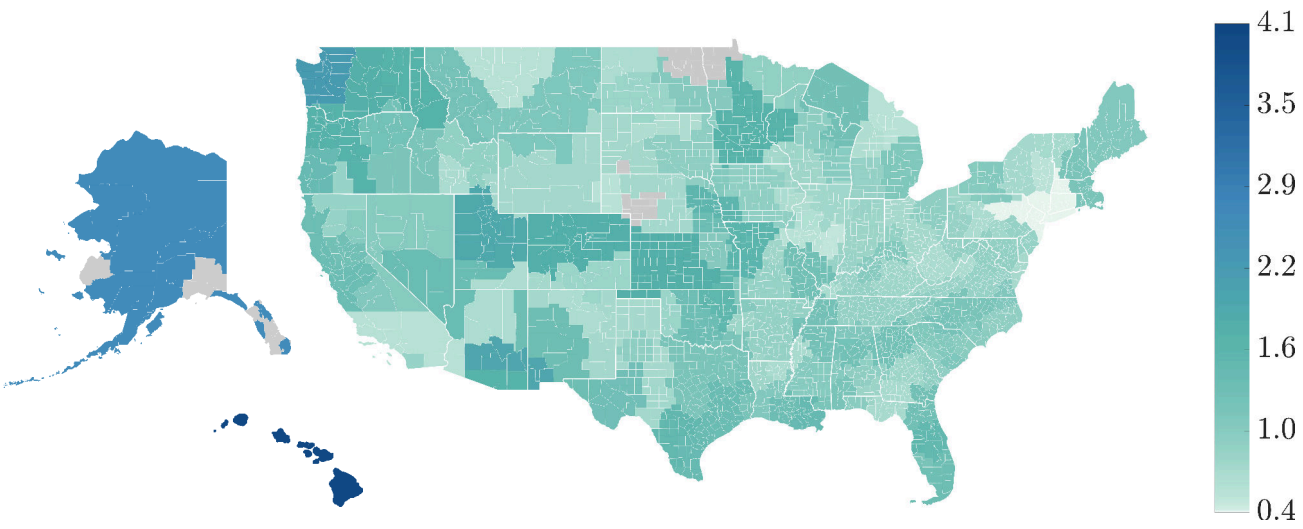
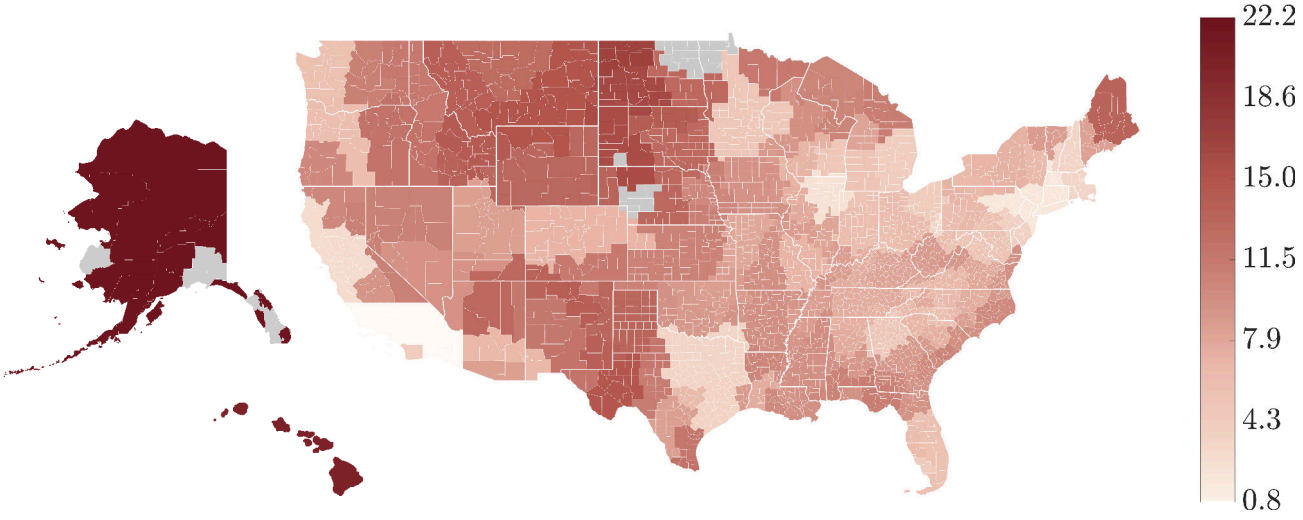


Figure 11: Percentage Consumption Gain from 20% Reduction in Trade Cost



Spatial heterogeneity in consumption gains. Figure 11 reports the percentage increase in final consumption C_k for each of our 170 EAs in response to a 20% reduction in intranational trade costs.[10] There is considerably more spatial variation in consumption gains, ranging from around a 1.6% increase in greater New York to increases of nearly 16% in remote parts of North and South Dakota. That is, there are nearly 10-times differences in consumption gains across locations as compared to the 4-times differences in markup decreases across locations. The largest consumption gains are generally in remote locations with more modest gains in coastal and large metro areas.

To be clear, these consumption gains from the reduction in intranational trade costs are driven by both the pro-competitive effects — with reduced markup dispersion reducing the productivity losses due to misallocation — and the standard Ricardian gains from trade that we would have in an otherwise equivalent model with constant markups. In the next section we isolate the markup channel by calculating the consumption-equivalent welfare gains from eliminating markups *holding trade costs fixed.*

6 Welfare costs of markups

In this section we quantify the welfare costs of markups in each location. We measure these welfare costs by asking how much the representative consumer in each location would gain in consumption from policies that eliminate markups. Since our benchmark model features inelastic factor supply, all of the gains from eliminating markups are due to eliminating markup *dispersion*, i.e., to reducing the productivity losses due to *misallocation*. We find that the average welfare costs of markups are large, about 5.8% in consumption-equivalent terms and vary considerably across locations, from 1% or less in the most competitive locations to more than 20% in the least competitive locations. We then use our aggregation results to decompose these productivity losses into three distinct dimensions of misallocation: (i) *across-firms*, across firms i within source locations j, (ii) *across-sources*, across source locations j within destination k, and (iii) *across-sectors*, across sectors s within destination k. We find that the across-firm dimension is the largest, accounting for over half of total misallocation, and is remarkably stable across locations. The across-source and across-sector dimensions account for almost all the geographic variation in misallocation.

6.1 Measuring the welfare costs of markups

We measure the welfare costs of markups by asking how much the representative consumer in each location would gain from a policy that eliminates markups.

[10] Since labor L_k for each location is fixed in our benchmark model, this is equivalent to the percentage increase in consumption per worker C_k/L_k.

Eliminating markups. A simple policy that eliminates markups is to pay each firm i in sector s a destination-specific sales subsidy $\chi_{ik}(s) \geq 1$ to induce the firm to set establishment-level prices equal to establishment-level marginal cost. From (17) we get

$$\chi_{ik}(s) \cdot p_{ijk}(s) = \mu_{ik}(s) \cdot \frac{W_j}{z_{ij}(s)} \tag{41}$$

which evidently induces marginal-cost pricing when the subsidy equals the markup, $\chi_{ik}(s) = \mu_{ik}(s)$ for all j. The subsidy is independent of establishment location j because the firm's optimal markup is independent of establishment location. To isolate the welfare costs of markup distortions we assume that these subsidies are funded by lump-sum taxes.

Costs of markups. Table 8 reports our main results for the welfare costs of markups in our benchmark model. On average, the representative consumer would gain 5.8% in consumption-equivalent terms from eliminating markups. Since factor supply is inelastic in our benchmark model, this is entirely due to the location-specific productivity gains from the reduced misallocation that results from eliminating markup dispersion so that relative prices are aligned with relative marginal costs everywhere. The median is similar, 5.6%. When we compute the costs of markups in our model without geography, as discussed in Section 5.2 above, we find that the representative consumer would gain 3.7% in consumption-equivalent terms from eliminating markups.[11] So again we see that abstracting from spatial frictions leads to a quantitatively significant understatement of the welfare costs of markups. Just as importantly, there is also substantial variation in the welfare costs of markups across locations — the welfare costs of markups are generally large and *very unevenly distributed*.

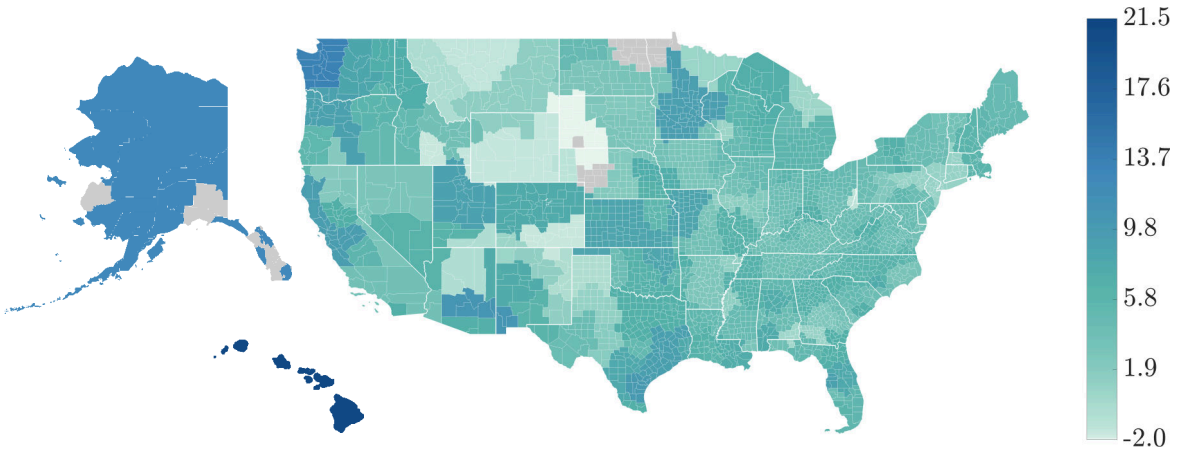
Spatial heterogeneity in the welfare costs of markups. To see this unevenness, Figure 12 reports the full geographic variation in the consumption gains from eliminating markups across our 170 EAs. Naturally, the welfare costs of markups are larger in locations where markups are larger to begin with. Controlling for initial markups, we also find that (i) locations which have lower initial levels of consumption per worker experience larger gains, and (ii) locations which have higher initial trade shares experience larger gains from eliminating markups. In this sense, policies that eliminate markups do not just have important aggregate welfare benefits, they have stark implications for the relative gains experienced by different geographic areas, with smaller, poorer, more remote locations benefitting from the elimination of markups to a considerably greater extent than larger, richer, more central locations. Overall, we find that in the smallest, poorest, most remote locations the welfare costs of markups can be as large as 20% in consumption-equivalent

[11] Likewise Edmond, Midrigan and Xu (2023) find that the aggregate productivity losses due to misallocation in their oligopolistically competitive model that abstracts from spatial frictions are about 3-4%.

Table 8: Eliminating Markups: Benchmark Model

Percentile	Consumption Gain, %
p01	0.9
p10	3.6
p25	3.9
p50	5.6
p75	6.9
p90	9.1
p99	14.5
Average	5.8

Figure 12: Percentage Consumption Gain from Eliminating Markups



terms — say 3-4 times as large as the average — but can be small, or even occasionally negative, in the largest, richest, most central locations.[12]

[12]Recall that we assume that profit income is distributed in proportion to labor income, i.e., that profit income is relatively high in high income locations. But large, highly competitive locations will have low markups. This means that when markups are eliminated, such locations experience a relatively small benefit from increased competition that can be overwhelmed by the loss of profit income resulting in a consumption loss. As can be seen from Figure 12 this is a quite rare occurrence .

6.2 Spatial misallocation

The welfare costs of markups are driven by *productivity losses due to misallocation*, i.e., markup variation that distorts relative prices away from relative marginal costs. In our model, there is misallocation along multiple dimensions, across firms, across locations, and across sectors. We now use our aggregation results from Section 2.2 above to quantify these different dimensions of misallocation.

Efficient productivity. First note that if markup variation is eliminated then the bilateral productivity network $\bar{z}_{jk}(s)$ from equation (30) above simplifies to the familiar CES index

$$\bar{z}_j^*(s) = \left(\sum_{i=1}^{N(s)} z_{ij}(s)^{\gamma-1} \right)^{\frac{1}{\gamma-1}} \tag{42}$$

This is the *efficient* productivity level at each source j in sector s. Notice that this efficient level of productivity is independent of the destination k. Without markup distortions, the allocation across firms at a given source is the same regardless of which destinations they serve. Using these efficient levels of productivity for each source j we can compute the sector-level productivity that would prevail at each destination k if markup variation is eliminated.

$$\bar{Z}_k^*(s) = \left(\sum_{j=1}^{J} \left(\frac{\bar{z}_j^*(s)}{\tau_{jk}(s)} \right)^{\gamma-1} \left(\frac{W_j}{\bar{W}_k(s)} \right)^{-\gamma} \right)^{\frac{1}{\gamma-1}} \tag{43}$$

Importantly, we perform this calculation *holding wages fixed* at their benchmark equilibrium levels W_j. As discussed in the Appendix, we hold wages fixed to isolate the *direct* productivity effects of markup dispersion from the *indirect* effects that operate through shifts in relative wages. We then aggregate across sectors s to get a measure of the efficient level of productivity for each destination k, again keeping wages fixed at their benchmark equilibrium levels

$$\bar{Z}_k^* = \left(\int_0^1 \bar{Z}_k^*(s)^{\theta-1} \left(\frac{\bar{W}_k(s)}{\bar{W}_k} \right)^{-\theta} ds \right)^{\frac{1}{\theta-1}} \tag{44}$$

Misallocation. We then measure misallocation at each destination k and each sector s by

$$\hat{Z}_k(s) := \frac{\bar{Z}_k(s)}{\bar{Z}_k^*(s)} \leq 1 \tag{45}$$

and similarly measure overall misallocation at each destination k by $\hat{Z}_k := \bar{Z}_k / \bar{Z}_k^*$.

Results. Overall we find that the productivity losses due to misallocation range from 4.4% in greater New York to 14.2% in Honolulu. Locations with extensive misallocation due to markup dispersion also tend to be locations with high markup levels — the correlation between misallocation and average markups $\bar{\mu}_k$ is 0.84, rising to 0.85 if we exclude the outliers Anchorage and Honolulu. Figure 13 reports the full geographic variation in these losses across the 170 EAs.

Three dimensions of misallocation. Our model features three distinct dimensions of misallocation: (i) *across-firms*, across firms i within source locations j, (ii) *across-sources*, across source locations j within destination k, and (iii) *across-sectors*, across sectors s within destination k. What kind of misallocation is the most important? Table 10 decomposes the productivity losses along these three dimensions for a selection of locations.

The across-firm component is generally the largest, averaging 3.0% across locations, and is quite stable geographically, ranging from 2.6% in Anchorage to 3.7% in Seattle. By contrast the across-source and across-sector components are generally smaller, 1.9% and 0.7% respectively, but much more variable across locations — they account for almost all of the cross-sectional variation in total misallocation. In a model without spatial frictions, only the across-firm dimension would be present, and average losses would be roughly 3% everywhere. The additional losses from the across-source and across-sector dimensions are due to the spatial frictions in our model and in this sense capture the *interactions* between market power and the spatial economy.

(i) ACROSS-FIRMS: Markup dispersion across firms producing at the same source j reduces bilateral productivity $\bar{z}_{jk}(s)$ below the efficient benchmark $\bar{z}^*_j(s)$. This dimension of misallocation is most important at *large production hubs* where many firms colocate and compete with heterogeneous markups. Small locations with one or fewer firms per sector have essentially no across-firm misallocation.

(ii) ACROSS-SOURCES: Markup dispersion across source locations j serving a given destination k distorts the allocation of consumption across sources. Sources with above-average markups are over-priced, reducing their effective contribution to the destination's consumption. This dimension of misallocation is most important for *remote destinations* whose suppliers have heterogeneous markups.

(iii) ACROSS-SECTORS: Markup dispersion across sectors at a given destination k distorts the sectoral composition of consumption. Sectors with above-average markups are again effectively over-priced, reducing their consumption share below the efficient level. This further amplifies misallocation, particularly for geographically isolated locations.

Table 10 summarizes these results. Roughly speaking, across-firm misallocation is around 3% everywhere. The spatial variation in total misallocation is driven by the across-source and across-sector variation, especially in the tails of the distribution — i.e., Honolulu and Anchorage. Nonetheless, because these locations are small these large misallocation losses in the tails have little effect on average misallocation along each dimension.

Across-sector amplification. Since preferences are concave, a mean-preserving spread in sector-level losses makes the representative consumer worse off. Consider two locations, both with an average loss of 7% where in one location the sector-level losses are tightly clustered around 7% while in the other location they vary from nearly 0% to 20%. The second location will have a much worse aggregate outcome. Honolulu is exactly this case: its sector-level losses range from near 0% to over 20%, and as a result its aggregate loss (14.2%) is nearly double its sector-level average (7.4%). New York, by contrast, has tightly concentrated sector-level losses (most between 0–5%), so its aggregate (4.4%) is close to its sector-level average (3.9%). Figure 14 illustrates this directly.

7 Labor mobility

In this section we consider an extension of our benchmark model to allow for workers to move across locations. We use a setup in the spirit of Lagakos and Waugh (2013) and Kline and Moretti (2014) where workers have different preferences for different location-specific amenities, characterized by idiosyncratic differences in Fréchet draws, to pin down labor supply to each location. One might reasonably guess that labor mobility acts to *mitigate* the welfare costs of markups — workers can now move to locations that provide higher wages and/or lower prices. But quantitatively we find that labor mobility does little to mitigate the welfare costs of markups once we parameterize the model with labor mobility to match the same initial allocation of labor across locations as in our benchmark calibration.

Location-specific amenities. Each location $j = 1, ..., J$ is characterized by a location-specific amenity value $A_j > 0$ that is common to all workers (e.g., the location's climate). Each worker is characterized by a vector of idiosyncratic amenity draws, one for each location

$$\boldsymbol{v} = (v_1, v_2, ..., v_J) \tag{46}$$

Specifically, we assume that each v_j is drawn IID from a standard Fréchet distribution

$$\text{Prob}[v_j \leq v] = \exp(-v^{-\sigma}), \qquad \sigma > 0 \tag{47}$$

Figure 13: Productivity Losses from Markup Misallocation

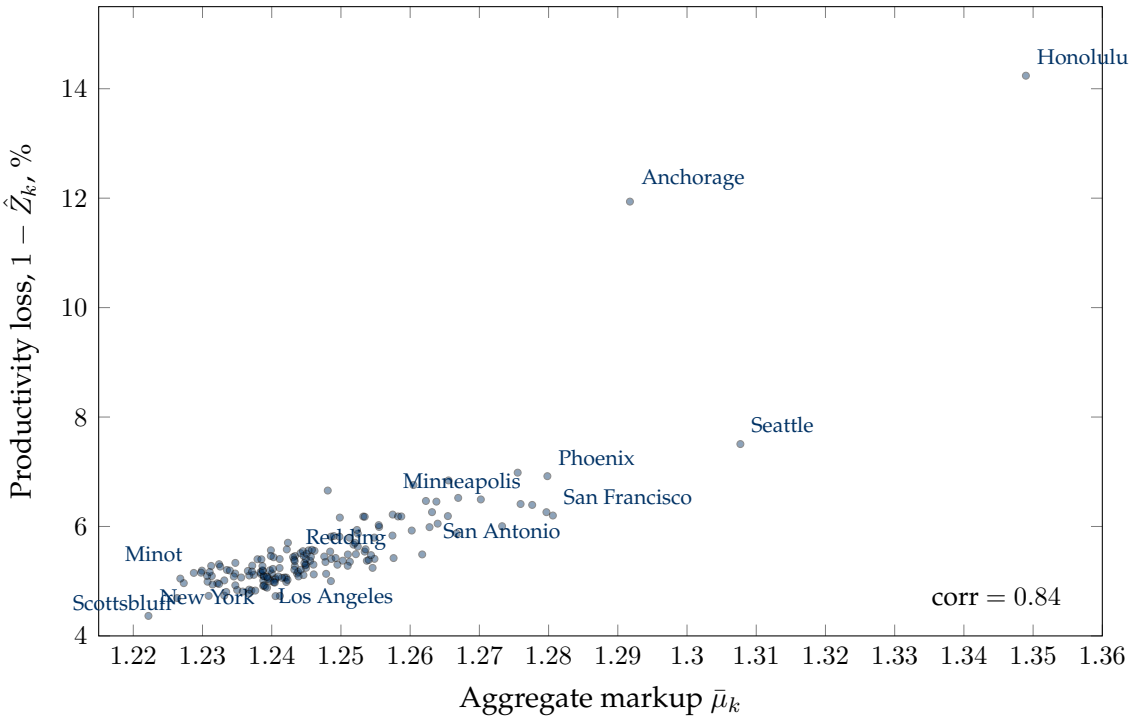


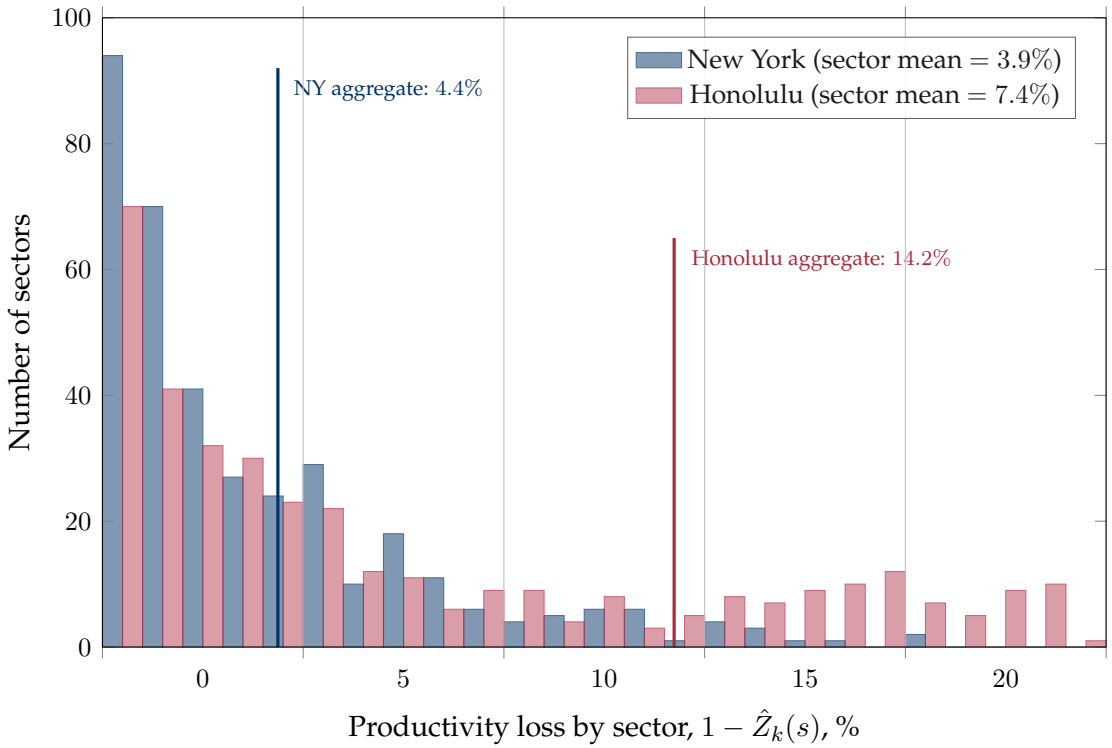
Table 9: Decomposition of Productivity Losses from Misallocation

	Across-Firms	Across-Sources	Across-Sectors	Total
p01	2.6	1.0	0.2	4.7
p10	2.7	1.6	0.3	4.9
p25	2.8	1.8	0.4	5.1
p50	2.9	1.9	0.6	5.3
p75	3.1	2.1	0.7	5.6
p90	3.3	2.3	0.9	6.2
p99	3.7	3.5	5.0	11.1
Average	3.0	1.9	0.7	5.5

Table 10: Decomposition of Productivity Losses: Selected Locations

Location	Across-Firms	Across-Sources	Across-Sectors	Total
Los Angeles	3.4	0.9	0.8	5.1
New York	3.1	0.8	0.5	4.4
San Francisco	3.8	1.7	0.8	6.2
Minneapolis	3.5	2.0	1.0	6.4
Seattle	3.7	2.4	1.6	7.5
Phoenix	3.5	2.3	1.2	6.9
San Antonio	3.2	2.3	0.8	6.3
Honolulu	3.0	4.4	6.9	14.2
Anchorage	2.6	3.7	5.8	11.9
Redding	2.9	1.8	0.7	5.4
Minot	2.6	2.0	0.6	5.2
Scottsbluff	2.7	1.7	0.6	5.0

Figure 14: Sector-Level Misallocation: New York vs. Honolulu



with tail parameter σ. As in our benchmark model, a worker that supplies labor in location j provides E_j efficiency units of labor.

Location choice. Let $c_j(\boldsymbol{v})$ denote the consumption of a worker of type $\boldsymbol{v}$ if they choose location j and let $u_j(\boldsymbol{v})$ denote their payoff from this choice. Consumption satisfies the individual worker's budget constraint

$$P_j c_j(\boldsymbol{v}) = (1+\bar{\pi}) W_j E_j \tag{48}$$

where as in the benchmark model we continue to assume that profit income is paid out in proportion to labor income for some constant $\bar{\pi} \geq 0$ to be determined in equilibrium. A worker's payoff from choosing location j is then given by

$$u_j(\boldsymbol{v}) = A_j v_j(\boldsymbol{v}) c_j(\boldsymbol{v}) = A_j v_j(\boldsymbol{v})(1+\bar{\pi}) \frac{W_j}{P_j} E_j \tag{49}$$

where $v_j(\boldsymbol{v})$ denotes the specific amenity draw for location j of a worker of type $\boldsymbol{v}$. The problem of an individual worker is to choose a location j that maximizes their payoff

$$u(\boldsymbol{v}) = \max_{j=1,\ldots,J} u_j(\boldsymbol{v}) \tag{50}$$

Labor supply. Let $\bar{L} > 0$ denote the total mass of workers. Following standard Fréchet calculations, the mass of workers supplying labor to location j is given by

$$L_j = \frac{\left(A_j E_j \frac{W_j}{P_j}\right)^{\sigma}}{\sum_j \left(A_j E_j \frac{W_j}{P_j}\right)^{\sigma}} \bar{L} \tag{51}$$

Profit income per location cancels out because of our assumption that locations receive profit income in proportion to labor income. In short, we have a labor supply curve, increasing in the real wage W_j/P_j for each location, with elasticity given by the Fréchet tail parameter σ. A higher σ reduces the dispersion in idiosyncratic amenity draws and so makes relative labor supply across locations more responsive to relative differences in real wages.

Labor market clearing. The labor market in location j clears when the total supply of efficiency units of labor $E_j L_j$ equals the total labor demand in that location

$$E_j L_j = \frac{\left(A_j E_j \frac{W_j}{P_j}\right)^{\sigma}}{\sum_j \left(A_j E_j \frac{W_j}{P_j}\right)^{\sigma}} \bar{L} = \int_0^1 \sum_{k=1}^{J} \sum_{i=1}^{N(s)} l_{ijk}(s)\, ds \tag{52}$$

Table 11: Eliminating Markups: Labor Mobility

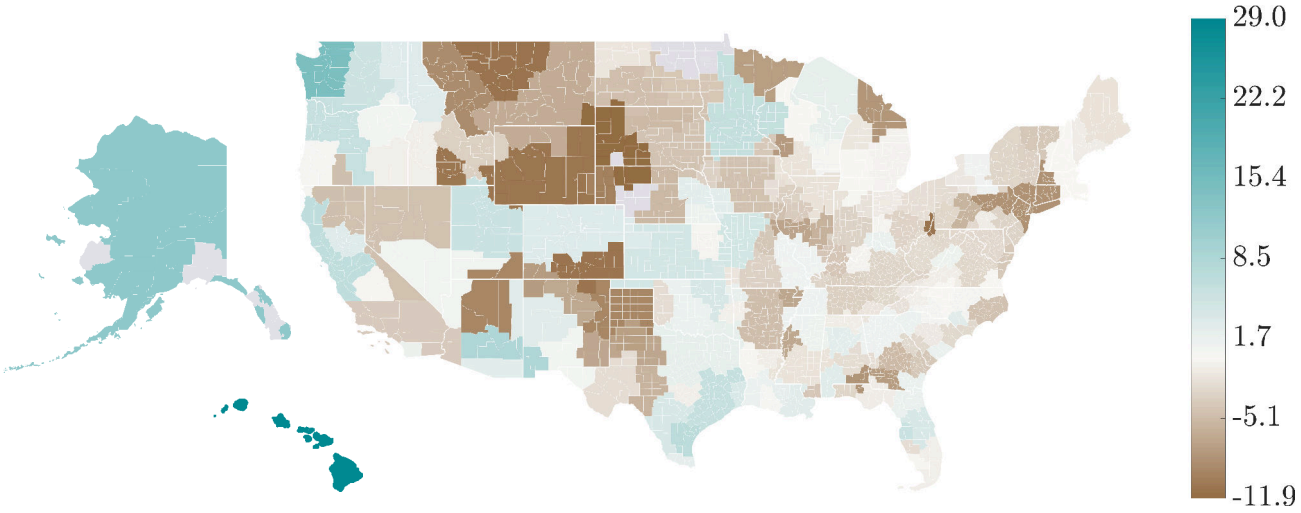
	Immobile	Mobile	
Percentile	C_j/L_j	C_j/L_j	C_j
p10	3.6	3.7	−0.4
p25	3.9	4.1	0.9
p50	5.6	5.6	5.4
p75	6.9	6.9	9.3
p90	9.1	8.9	15.6
Average	5.8	5.8	6.0

Parameterization. To solve this version of the model we fix the efficiency units of labor E_j at their benchmark values and assign a labor supply elasticity of $\sigma = 2$, in line with the range of values discussed by Fajgelbaum, Morales, Serrato and Zidar (2019). We choose the common location-specific amenity values A_j so that the model replicates manufacturing employment L_j from the County Business Patterns aggregated to the EA level. In other words, we choose A_j so that the model with labor mobility rationalizes the allocation of employment across locations in our benchmark calibration.

Quantitative results. Table 11 reports the effects of eliminating markups, across key percentiles of the distribution, for the version of the model with mobile labor and our benchmark model with immobile labor. For the model with mobile labor we report both the percentage changes in consumption per worker, C_j/L_j, and in aggregate consumption, C_j. In terms of consumption per worker, we find that labor mobility makes little difference quantitatively. For example, the average gain in consumption per worker from eliminating markups is 5.8% in both the model with mobile labor and the benchmark. Only in the upper tail of the distribution of consumption gains do we find noticeable differences, with the p90 gain in consumption per worker falling from 9.1% in the benchmark model to 8.9% in the model with labor mobility. In this sense, labor mobility mitigates the losses due to markups in those locations that are most severely impacted by markup distortions — but even in this upper tail the size of the mitigating effect is small. For most locations, labor mobility has has even smaller effects on the welfare costs of markups.

That said, the model with labor mobility implies more substantial differences in the changes in aggregate consumption and employment across locations. For example, the

Figure 15: Percentage Labor Flows from Eliminating Markups



p75 gain in consumption per worker is the same, 6.9%, in both models. But with mobile labor, that 6.9% gain can be decomposed into a 9.3% gain in aggregate consumption and an approximately 2% increase in employment. Likewise, the p90 gain in consumption per worker of 8.9% with mobile labor can be decomposed into a large 15.6% gain in aggregate consumption and a 6-7% increase in employment. The locations which gain the most from the elimination of markups see large increases in employment. Intuitively, there are large labor inflows to locations which are growing substantially. At the other end of the spectrum, the p10 gain of consumption per worker of 3.7% with mobile labor masks a 0.4% *decrease* in aggregate consumption and an approximately 4% decrease in employment. These locations still gain from the elimination of markups in consumption per worker terms, but they are shrinking both in absolute terms and relative to the rest of the economy. These implications for the reallocation of labor across locations are of course absent from our benchmark model. Figure 15 reports the full geographic variation in labor market flows across locations resulting from the elimination of markups.

8 Conclusion

We study the spatial distribution of economic activity in a quantitative model with multi-establishment firms, oligopolistic competition, and endogenously variable markups. We calibrate our model to match US Census of Manufactures firm and establishment data and intranational trade flows from the Commodity Flows Survey across 170 US Economic Areas. We show spatial frictions can have large aggregate effects, increasing both the aggregate markup and the productivity losses due to misallocation. We show that a reduction in intranational trade costs, calibrated to match long-run trends in US manufacturing, will increase national sales concentration but decrease local sales concentration. Local markets become more competitive, markups fall, and aggregate productivity rises, despite the increase in national concentration. We also show that the welfare costs of markups are large on average and very unevenly distributed. Smaller, poorer, more remote locations have costs some 20 times the costs of larger, richer, more central locations.

References

Allen, Treb and Costas Arkolakis, "Trade and the Topography of the Spatial Economy," *Quarterly Journal of Economics*, August 2014, *129* (3), 1085–1140.

Amiti, Mary and Sebastian Heise, "US Market Concentration and Import Competition," May 2021. CEPR Discussion Paper 16126.

Asturias, Jose, Manuel García-Santana, and Roberto Ramos, "Competition and the Welfare Gains from Transportation Infrastructure: Evidence from the Golden Quadrilateral of India," *Journal of the European Economic Association*, 2019, *17* (6), 1881–1940.

Atkeson, Andrew and Ariel Burstein, "Pricing-to-Market, Trade Costs, and International Relative Prices," *American Economic Review*, 2008, *98* (5), 1998–2031.

Autor, David, Christina Patterson, and John Van Reenen, "Local and National Concentration Trends in Jobs and Sales: The Role of Structural Transformation," April 2023. NBER Working Paper 31130.

_ , David Dorn, Lawrence F. Katz, Christina Patterson, and John Van Reenen, "The Fall of the Labor Share and the Rise of Superstar Firms," *Quarterly Journal of Economics*, May 2020, *135* (2), 645–709.

Baqaee, David Rezza and Emmanuel Farhi, "Productivity and Misallocation in General Equilibrium," *Quarterly Journal of Economics*, February 2020, *135* (1), 105–163.

Basker, Emek, Shawn D. Klimek, and Pham Hoang Van, "Supersize It: The Growth of Retail Chains and the Rise of the Big-Box Store," *Journal of Economics and Management Strategy*, 2012, *21* (3), 541–582.

Benkard, C. Lanier, Ali Yurukoglu, and Anthony Lee Zhang, "Concentration in Product Markets," *AEJ: Microeconomics*, Forthcoming. 2026.

Caliendo, Lorenzo and Fernando Parro, "Estimates of the Trade and Welfare Effects of NAFTA," *Review of Economic Studies*, January 2015, *82* (1), 1–44.

_ , _ , Esteban Rossi-Hansberg, and Pierre-Daniel Sarte, "The Impact of Regional and Sectoral Productivity Changes on the US Economy," *Review of Economic Studies*, 2018, *85* (4), 2042–2096.

Cao, Dan, Henry R. Hyatt, Toshihiko Mukoyama, and Erick Sager, "Firm Growth Through New Establishments," 2022. Working Paper.

Coşar, A. Kerem, Sophie Osotimehin, and Latchezar Popov, "The Long-Run Effects of Transportation Productivity on the US Economy," December 2024. NBER Working Paper 33248.

De Loecker, Jan, Jan Eeckhout, and Gabriel Unger, "The Rise of Market Power and the Macroeconomic Implications," *Quarterly Journal of Economics*, May 2020, *135* (2), 561–644.

_ , _ , and Simon Mongey, "Quantifying Market Power and Business Dynamism in the Macroeconomy," April 2021. KU Leuven Working Paper.

Edmond, Chris, Virgiliu Midrigan, and Daniel Yi Xu, "Competition, Markups, and the Gains from International Trade," *American Economic Review*, October 2015, *105* (10), 3183–3221.

_ , _ , and _ , "How Costly Are Markups?," *Journal of Political Economy*, July 2023, *131* (7), 1619–1675.

Fajgelbaum, Pablo D., Eduardo Morales, Juan Carlos Suárez Serrato, and Owen Zidar, "State Taxes and Spatial Misallocation," *Review of Economic Studies*, January 2019, *86* (1), 333–376.

Foster, Lucia, John Haltiwanger, Shawn Klimek, C.J. Krizan, and Scott Ohlmacher, "The Evolution of National Retail Chains: How We Got Here," in Emek Basker, ed., *Handbook on the Economics of Retailing and Distribution*, Elgar, 2016.

Franco, Santiago, "Output Market Power and Spatial Misallocation," November 2023. University of Chicago Working Paper.

Ganapati, Sharat, "Growing Oligopolies, Prices, Output, and Productivity," *AEJ: Microeconomics*, August 2021, *13* (3), 309–327.

Grullon, Gustavo, Yelena Larkin, and Roni Michaely, "Are US Industries Becoming More Concentrated?," *Review of Finance*, July 2019, *23* (4), 697–743.

Hillberry, Russell and David Hummels, "Trade Responses to Geographic Frictions: A Decomposition Using Micro-Data," *European Economic Review*, April 2008, *52* (3), 527–550.

Holmes, Thomas J., "The Diffusion of Wal-Mart and Economies of Density," *Econometrica*, January 2011, *79* (1), 253–302.

Hsieh, Chang-Tai and Esteban Rossi-Hansberg, "The Industrial Revolution in Services," *Journal of Political Economy: Macroeconomics*, 2023, *1* (1), 3–42.

Jia, Panle, "What Happens When Wal-Mart Comes to Town: An Empirical Analysis of the Discount Retailing Industry," *Econometrica*, November 2008, *76* (6), 1263–1316.

Kimball, Miles S., "The Quantitative Analytics of the Basic Neomonetarist Model," *Journal of Money, Credit, and Banking*, 1995, *27* (4, Part 2), 1241–1277.

Kline, Patrick and Enrico Moretti, "People, Places, and Public Policy: Some Simple Welfare Economics of Local Economic Development Programs," *Annual Review of Economics*, August 2014, *6*, 629–662.

Lagakos, David and Michael E. Waugh, "Selection, Agriculture, and Cross-Country Productivity Differences," *American Economic Review*, April 2013, *103* (2), 948–980.

Neiman, Brent and Joseph Vavra, "The Rise of Niche Consumption," *AEJ: Macroeconomics*, July 2023, *15* (3), 224–264.

Nelsen, Roger B., *An Introduction to Copulas*, 2nd ed., Springer, 2006.

Redding, Stephen J., "Goods Trade, Factor Mobility and Welfare," *Journal of International Economics*, July 2016, *101*, 148–167.

_ **and Esteban Rossi-Hansberg**, "Quantitative Spatial Economics," *Annual Review of Economics*, August 2017, *9*, 21–58.

Rossi-Hansberg, Esteban, Pierre-Daniel Sarte, and Nicholas Trachter, "Diverging Trends in National and Local Concentration," *NBER Macroeconomics Annual*, 2020, pp. 115–150.

Smith, Dominic A. and Sergio Ocampo, "The Evolution of US Retail Concentration," *AEJ: Macroeconomics*, 2024, *forthcoming.*

Wolf, Holger C., "Intranational Home Bias in Trade," *Review of Economics and Statistics*, November 2000, *82* (4), 555–563.

Appendix

This appendix documents how we construct the datasets used to discipline trade costs, firm geography, labor endowments, and the concentration targets used in the calibration.

A. Commodity Flow Survey and Gravity Estimates

We estimate sector-specific distance elasticities based on public-use microdata from the 2017 Commodity Flow Survey (CFS). The unit of observation in the raw data is an individual shipment.

Sectors and shipment values. For each shipment we observe a detailed six-digit NAICS industry code, a shipment value in current dollars, and a survey weight. We collapse the data to three-digit NAICS industries and restrict attention to manufacturing, yielding 21 three-digit manufacturing sectors.

We form origin-destination-sector shipment totals by summing shipment values using the CFS sampling weights. Because we only use the 2017 cross section, we work in current dollars and do not deflate these flows. Origin-destination pairs with no shipments do not enter the aggregated data and are therefore absent from the regressions.

Notion of location and distances. In the gravity regressions, the notion of geography is the CFS origin and destination region. For most shipments these regions coincide with individual U.S. states. In a few cases, however, a region is defined as a multi-state metropolitan area (for example, New York-Newark, Boston-Worcester-Providence, and Washington-Arlington-Alexandria). We map each shipment to an origin region and a destination region and then aggregate weighted values by region pair and three-digit sector.

We measure distance as the great-circle distance in kilometers between region centroids. For every pair of distinct regions we observe such a bilateral distance, including pairs involving the multi-state CFS regions described above, which are treated as additional regions and handled in exactly the same way. Shipments with the same origin and destination region do not enter the gravity regressions.

Gravity specification. For each three-digit manufacturing sector s, we estimate a standard gravity equation of the form

$$\log X_{jk}(s) = \gamma_j(s) + \gamma_k(s) + \beta(s)\, \log d_{jk} + \varepsilon_{jk}(s),$$

where $X_{jk}(s)$ is the shipment value from origin region j to destination region k in sector s, and d_{jk} is the bilateral distance between regions j and k. The origin and destination fixed effects $\gamma_j(s)$ and $\gamma_k(s)$ absorb all location-specific determinants of trade. The slope coefficients $\widehat{\beta}(s)$ discipline the distance-sensitivity of iceberg trade costs by sector in our quantitative model. We do not apply any additional trimming or winsorization to the CFS

flows beyond what is implicit in dropping origin–destination pairs with no shipments or nonpositive shipment values. Table .12 reports the resulting gravity coefficients.

Table .12: Sector-level distance elasticities from CFS gravity regressions

NAICS3	Coefficient	Standard Error	Description
311	-2.215^{***}	0.095	Food manufacturing
312	-2.654^{***}	0.138	Beverage and tobacco product manufacturing
313	-1.579^{***}	0.139	Textile mills
314	-2.377^{***}	0.124	Textile product mills
315	-1.363^{***}	0.118	Apparel manufacturing
316	-1.447^{***}	0.133	Leather and allied product manufacturing
321	-3.398^{***}	0.135	Wood product manufacturing
322	-2.343^{***}	0.120	Paper manufacturing
323	-2.713^{***}	0.116	Printing and related support activities
324	-3.914^{***}	0.194	Petroleum and coal products manufacturing
325	-2.529^{***}	0.109	Chemical manufacturing
326	-2.332^{***}	0.099	Plastics and rubber products manufacturing
327	-3.227^{***}	0.131	Nonmetallic mineral product manufacturing
331	-2.389^{***}	0.130	Primary metal manufacturing
332	-2.677^{***}	0.103	Fabricated metal product manufacturing
333	-1.695^{***}	0.102	Machinery manufacturing
334	-1.072^{***}	0.109	Computer and electronic product manufacturing
335	-1.355^{***}	0.109	Electrical equipment, appliance, and component manufacturing
336	-1.832^{***}	0.136	Transportation equipment manufacturing
337	-2.434^{***}	0.107	Furniture and related product manufacturing
339	-1.254^{***}	0.102	Miscellaneous manufacturing

B. Economic Areas, Distances, and Trade Costs

Our quantitative model is formulated at the level of BEA Economic Areas (EAs). We assign each U.S. county to a BEA EA using the BEA county-to-EA concordance based on the 1999 Economic Area definitions and aggregate all county-level information to the EA level. The Economic Areas covering the Gulf of Mexico, Puerto Rico, and the U.S. Virgin Islands are excluded, so that the dataset ultimately includes 170 EAs with positive manufacturing activity. We take as our set of model locations J the 170 EAs that receive positive manufacturing employment in County Business Patterns (CBP) and, by construction of the NETS sample below, at least one manufacturing establishment.

We measure the distance between any two Economic Areas in kilometers, using a symmetric matrix of bilateral distances. To normalize distances, we find the smallest distance observed between any two distinct EAs (79.6 km) and divide all distances by this minimum. We then also set within-EA distances to unity.

Sector-specific iceberg trade costs between EA j and EA k in sector s are parameterized as

$$\log \tau_{jk}(s) = \delta(s) \, \log d_{jk},$$

where d_{jk} is the normalized EA distance and $\delta(s)$ is a sector-specific distance elasticity of iceberg trade costs. We estimate a separate elasticity $\delta(s)$ for each three-digit manufacturing

sector and assign six-digit sectors in our quantitative model the value of their three-digit parent. The elasticities $\delta(s)$ are chosen so that, when we simulate the model and run analogous gravity regressions at the EA level, the implied distance coefficients in each sector match the CFS estimates $\widehat{\beta}(s)$.

C. NETS Establishments and Firm-Location Structure

To discipline the geography of firms' operations we use a 2014 snapshot of the National Establishment Time Series (NETS) dataset. We begin from a national NETS extract that covers manufacturing and contains, at the establishment level, a unique establishment identifier, a unique firm headquarter identifier, the establishment's state and county FIPS code, six-digit 2014 NAICS codes, 2014 employment, and total 2014 sales. We impose the following filters in constructing our sample of manufacturing establishments:

- We drop establishments located in U.S. territories and retain only establishments located in the 50 states and the District of Columbia.
- We require a valid firm headquarter identifier and drop all establishments without a headquarter ID.
- We restrict to manufacturing establishments, defined as six-digit NAICS codes with leading digit equal to 3.
- We drop establishments with fewer than five employees, to mitigate known issues with NETS coverage and coding for very small establishments (see Barnatchez, Crane, and Decker, 2017).
- We restrict attention to firms for which the headquarter identifier corresponds to at least one observed manufacturing establishment.
- We impose a minimum size in terms of economic activity by requiring sales of at least \$1,000 in 2014.

For each remaining manufacturing establishment we retain its identifiers, location, industry code, employment, and sales. We assign each establishment to a unique BEA Economic Area using the county-to-EA mapping described above.

We define a firm as a unique headquarter identifier. Firms are assigned to six-digit NAICS sectors based on the NAICS code of their headquarters. For each firm we then construct a binary indicator for each EA that equals one if the firm has at least one manufacturing establishment in that EA and zero otherwise. If a firm operates multiple establishments in the same EA, we treat this as a single producing location in that EA. The resulting firm-by-EA incidence matrix captures each firm's geographic footprint across the United States and is the object we feed into the model as the set of locations in which each firm operates. This matrix also underlies our computation of the fraction, employment share, and sales share of multi-establishment firms. We rely on NETS primarily for its

information on locations, industry codes, headquarter structure, and employment. We use sales only in our sample restrictions and validation exercise.

After applying our filters, the 2014 NETS manufacturing sample contains 219,365 firms operating 270,141 establishments. Of these, 213,346 firms (97.2%) operate a single manufacturing establishment, while 6,019 firms (2.8%) operate more than one. Multi-establishment firms account for 56,795 establishments, or 21.0% of all manufacturing establishments in the sample. These multi-establishment firms account for 54.3% of manufacturing employment and 62.1% of manufacturing sales, so a small minority of firms accounts for a disproportionate share of activity. Manufacturing establishments are also highly concentrated across Economic Areas: the 10th, 50th, and 90th percentiles of the EA establishment distribution are 158, 582, and 2,926 establishments, respectively. The five largest EAs together host 67,344 establishments (27.8% of the total across EAs) and the largest 25 EAs host 149,247 establishments (61.6%).

D. Labor Endowments and Wages by Economic Area

Labor endowments and wages at the EA level are constructed from County Business Patterns (CBP). We treat each dataset as a cross section and use the closest available pre-2014 CBP year with consistent EA definitions. We therefore use CBP data for 2012 and aggregate manufacturing employment and the manufacturing wage bill from counties to BEA Economic Areas using the same BEA county-to-EA mapping as above. For each EA we thus obtain total manufacturing employment and total manufacturing wage payments. From the EA-level manufacturing employment and wage bill we construct:

- the employment level L_k in each EA k, given by CBP manufacturing employment, normalized so that the cross-EA mean of L_k equals one; and
- the average wage per worker in each EA, given by the CBP manufacturing wage bill divided by CBP manufacturing employment in that EA, again normalized so that the cross-EA mean equals one.

In the model, L_k is the number of workers in EA k, which we interpret as supplying $E_k L_k$ efficiency units of labor. The location-specific efficiency shifters E_k are chosen so that, at the calibrated equilibrium, the model wage bill $W_k E_k L_k$ in each EA matches the observed CBP wage bill. Beyond dropping EAs with no manufacturing activity, we do not trim or winsorize wages or employment when constructing EA-level aggregates.

Across the 170 Economic Areas, 2012 total manufacturing employment in our data is 3,469,742 workers. Employment is highly uneven: mean EA employment is 20,410 workers, with the 10th, 50th, and 90th percentiles at 426, 4,842, and 53,163 workers, respectively. The five largest EAs together account for 1,306,358 workers (37.65% of manufacturing employment), and the largest 25 EAs account for 2,594,712 workers (74.78%). By contrast, average manufacturing wages display more moderate dispersion: the employment-weighted national average wage is \$53,480 per worker, while the EA-level 10th, 50th, and 90th percentiles of the average wage distribution are \$36,500, \$47,850, and \$60,780, respectively.