# The Spatial Distribution of Markups

Jonathan Becker Chris Edmond
Virgiliu Midrigan Daniel Yi Xu

September 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. We find that spatial frictions have large effects on competition. The reduction in effective competition they create is equivalent to removing a randomly chosen 90% of firms from each sector of an otherwise identical economy without geography. The welfare costs of markups are correspondingly large, 5.7% on average, considerably higher than the 3.7% implied by that economy without geography. The welfare costs are also very unevenly distributed, ranging from about 1% or less in large central locations like New York to more than 20% in Honolulu.

*Keywords*: competition, markups, market power, concentration, misallocation, economic geography, multi-establishment firms, gravity.

*JEL classifications*: D4, F1, L1, O4, R1.

*Acknowledgments:* We particularly thank Mayara Felix, Jake Zhao, Sebastian Heise and Michael Waugh for their detailed comments. We also thank audiences at the 2026 SITE Workshop on the Causes and Consequences of Misallocation: Theory and Empirics, the 2026 Minnesota Workshop on Macroeconomic Theory, the 2026 FRB St Louis Conference on Firm Dynamics and Macroeconomics, the 2026 WESEAMS, 2026 ASSA meetings, 2025 WAMS, 2024 PHBS-Shenzhen workshop, 2023 OzMac workshop, 2023 FRB Dallas Conference on International Economics, and seminar participants at HKUST, HKBU, PHBS-Shenzhen, and the University of Nottingham–Ningbo.

*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?

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 as given, each firm chooses the cost-minimizing way to use its establishments. This pins down each firm's marginal cost of supplying each destination. Given these 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 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, despite the rich geography of these models, 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 tradeability, production is geographically concentrated, and firms range from single plants to large multi-establishment enterprises selling nationwide. We show that spatial frictions are quantitatively significant 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 trade flows from the Commodity Flow Survey (CFS) 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 other words, these frictions play a key role in determining the spatial distribution of production and sales. Intuitively, we find that production concentration is larger and more dispersed than sales concentration, with the gap being greatest in low-gravity sectors where goods are easily shipped to multiple destination markets.[1] Although our benchmark calibration lies much closer to free trade than to autarky, the remaining spatial frictions are still enough to matter substantially for competition.

Our main finding is that the spatial frictions in the US economy place substantial limits on competition, leading to a quantitatively significant amplification of the aggregate losses due to market power. In high-gravity sectors, local firms face less competition, leading to higher markups and higher productivity losses due to misallocation. We show that the *reduction in competition* provided by spatial frictions is equivalent to removing a randomly chosen 90% of firms from the economy. An otherwise equivalent model *without geography* — calibrated to match the same national concentration facts as our benchmark — has the same amount of competition, as measured by the aggregate markup and sales concentration, as our benchmark model *with geography* only after we remove 90% of its firms.

The intuition for this result is as follows. There are on average 600 firms in each 6-digit NAICS sector. In a standard calibration of a model without geography, this means that there are already enough large firms competing with each other that competitive outcomes are relatively insensitive to changes in the number of firms. We show that, without geography, this insensitivity is pervasive for a very wide range of the number of firms — doubling or halving the number of firms has negligible effects on national concentration,

[1] 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). This ordering of concentration is consistent with Benkard, Yurukoglu and Zhang (2026) who study concentration in finely disaggregated consumer survey data.

on the aggregate markup, and on the productivity losses due to misallocation. It takes a drastic reduction in the number of firms, down to about 60 firms per 6-digit NAICS sector, for the model without geography to replicate the reduction in competition that is generated by the spatial frictions we estimate in US data. With these frictions, most sales in many locations are accounted for by just a few firms, especially in high-gravity sectors.

Our finding that spatial frictions substantially limit competition — and that reducing these frictions would deliver large pro-competitive gains — stands in sharp contrast to a class of benchmark models of endogenously variable markups, such as Bernard, Eaton, Jensen and Kortum (2003), Atkeson and Burstein (2008), Edmond, Midrigan and Xu (2015), and Arkolakis, Costinot, Donaldson and Rodríguez-Clare (2019). These models span a range of assumptions about technology, market structure and conduct, but share the implication that changes in trade costs or in the number of competing firms generate modest, if any, movements in markups when (i) firms are heterogeneous in productivity, and (ii) the number of competitors is large. Since the typical narrowly-defined US manufacturing industry has on average 600 firms, this class of models would predict that pro-competitive effects in the US should be small. By contrast, in our model the extent of spatial frictions determines which firms actually compete head-to-head in any given destination market. Spatial frictions can leave a location served by only a small number of *effective* competitors, even when the same industry has many firms nationally. Reducing trade costs therefore expands the set of effective competitors faced by dominant firms in each destination market, lowering markups and reducing the misallocation they generate.

Accounting for geography and spatial frictions leads to substantially higher welfare costs of markups. We measure these 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 5.7% in consumption-equivalent terms, considerably higher than the 3.7% we find in the model without geography. The welfare costs are also very unevenly distributed, ranging from about 1% in greater New York to more than 20% in Honolulu. Abstracting from geography and spatial frictions not only substantially understates the average welfare costs of markups, but also prevents the model from capturing the large variation in the burden of market power across locations.

We show that our model is also consistent with key trends in concentration. Our model generates endogenously diverging trends in national concentration and local concentration, as in Rossi-Hansberg, Sarte and Trachter (2021), when trade costs fall. Our model interprets these 'diverging trends' as *convergence* towards a better-integrated national market — as trade costs fall from their benchmark levels, national sales concentration rises while local sales concentration falls. In this sense, national sales concentration may rise even as markets become more competitive, markups fall, and aggregate productivity rises. Trends in national concentration may be a poor guide to trends in aggregate market power.

Finally, we ask to what extent worker mobility can mitigate the welfare costs of markups.

Our benchmark 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 are almost identical.

**Intranational trade frictions.** A large body of evidence documents that trade within the US is geographically segmented — see e.g., Wolf (2000), Hillberry and Hummels (2008), and Martínez-San Román, Mateo-Mantecón and Sainz-González (2017). We take this evidence as a starting point and ask what it implies for competition and market power. Ramondo, Rodríguez-Clare and Saborío-Rodríguez (2016) quantify the importance of intranational trade frictions in an Eaton and Kortum (2002) model with perfect competition.

**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 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 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 but unlike us they abstract from multi-location firms. 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-location firms for the spatial distribution of markups.

**Trends in concentration.** This paper contributes to the extensive literature on US trends in concentration 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. A central debate in this literature concerns the divergence between national and local trends: Rossi-Hansberg, Sarte and Trachter (2021), using NETS data, find that local concentration has been declining even as national concentration has risen, while subsequent work using Census data finds mixed or opposing results (e.g., Autor, Patterson and Van Reenen, 2023; Smith and Ocampo, 2024). As argued by Benkard, Yurukoglu and Zhang (2026), Census-based measures classify economic activity by *production* rather than by *consumption*, and consumer access to substitutes is what ultimately determines producer market power. Using finely disaggregated consumer survey data, Benkard et al. and Neiman and Vavra (2023) both find decreasing local sales concentration.

**Multi-establishment production.** This paper also contributes to the recent literature on the 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), Hsieh and Rossi-Hansberg (2023), and many others. Most closely related is Oberfield, Rossi-Hansberg, Sarte and Trachter (2024), who study the endogenous spatial distribution of production in a model with constant markups, whereas we take the geographic footprint of production as given and ask what it implies for the endogenous spatial distribution of markups.

# 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 $k$.

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

**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*

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

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 $\Pi_j$ 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.

**Profits and the aggregate markup.** Under this assumption, expenditure in each location is $P_j C_j = (1+\bar{\pi}) W_j E_j L_j$ and it is then straightforward to show that

$$1+\bar{\pi} = \bar{\mu} \tag{7}$$

where $\bar{\mu}$ denotes the *aggregate markup*, the sales-weighted harmonic average of the location-level markups $\bar{\mu}_k$ defined below. The constant $\bar{\pi}$ is therefore the economy's aggregate profit-to-labor-income ratio, and the common profit share of income is $\bar{\pi}/(1+\bar{\pi}) = 1 - 1/\bar{\mu}$.

**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

$$c_{ijk}\,(s) = \left(\frac{\tau_{jk}(s) p_{ijk}(s)}{P_{ik}(s)}\right)^{-\eta} \underbrace{\left(\frac{P_{ik}(s)}{P_k(s)}\right)^{-\gamma} \left(\frac{P_k(s)}{P_k}\right)^{-\theta} C_k}_{=C_{ik}(s)} \tag{8}$$

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

$$P_k = \left(\int_0^1 P_k(s)^{1-\theta}\, ds\right)^{\frac{1}{1-\theta}}, \qquad P_k(s) = \left(\sum_{i=1}^{N(s)} P_{ik}(s)^{1-\gamma}\right)^{\frac{1}{1-\gamma}} \tag{9}$$

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

$$P_{ik}(s) = \left(\sum_{j=1}^{J} (\tau_{jk}(s) p_{ijk}(s))^{1-\eta}\right)^{\frac{1}{1-\eta}} \tag{10}$$

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

$$y_{ijk}(s) = z_{ij}(s) l_{ijk}(s) \tag{11}$$

**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

$$y_{ijk}(s) = \tau_{jk}(s)c_{ijk}(s) \tag{12}$$

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

$$\frac{W_j}{z_{ij}(s)} \tag{13}$$

**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

$$\Pi_{ik}(s) = \sum_{j=1}^{J} \left( p_{ijk}(s) - \frac{W_j}{z_{ij}(s)} \right) y_{ijk}(s) \tag{14}$$

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 $k$

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

subject to the firm-level price index (10). The Lagrangian for this problem can be written

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

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

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

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

$$p_{ijk}(s) = \mu_{ik}(s) \, \frac{W_j}{z_{ij}(s)}, \qquad \mu_{ik}(s) = \frac{\eta}{\eta - 1} \left( \frac{P_{ik}(s)}{\lambda_{ik}(s)} \right) \tag{18}$$

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 (10) and eliminating the multiplier gives

$$P_{ik}(s) = \mu_{ik}(s) \cdot \mathsf{MC}_{ik}(s) \tag{19}$$

where

$$\mathsf{MC}_{ik}(s) = \left( \sum_{j=1}^{J} \left( \frac{\tau_{jk}(s) W_j}{z_{ij}(s)} \right)^{1-\eta} \right)^{\frac{1}{1-\eta}} \tag{20}$$

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 (8), (14) and (19) to write the firm's profits from destination $k$

$$\begin{aligned} \Pi_{ik}(s) &= \left( P_{ik}(s) - \mathsf{MC}_{ik}(s) \right) C_{ik}(s) \\ &= \left( P_{ik}(s) - \mathsf{MC}_{ik}(s) \right) \left( \frac{P_{ik}(s)}{P_k(s)} \right)^{-\gamma} \left( \frac{P_k(s)}{P_k} \right)^{-\theta} C_k \end{aligned} \tag{21}$$

with each firm internalizing the effect of their price $P_{ik}(s)$ on the sector-level price index $P_k(s)$ in (9). 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

$$\mu_{ik}(s) = \frac{\varepsilon_{ik}(s)}{\varepsilon_{ik}(s) - 1} \tag{22}$$

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

$$\varepsilon_{ik}(s) = \left(\omega_{ik}(s)\frac{1}{\theta} + (1 - \omega_{ik}(s))\frac{1}{\gamma}\right)^{-1} \tag{23}$$

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

$$\omega_{ik}(s) := \frac{P_{ik}(s)C_{ik}(s)}{\sum_{i=1}^{N(s)} P_{ik}(s)C_{ik}(s)} = \frac{P_{ik}(s)^{1-\gamma}}{\sum_{i=1}^{N(s)} P_{ik}(s)^{1-\gamma}} \tag{24}$$

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 $k$. 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 $\gamma$ 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 $\theta$ 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 (22) and (23) to write the inverse markup as a linear function of the sales share

$$\frac{1}{\mu_{ik}(s)} = 1 - \frac{1}{\varepsilon_{ik}(s)} = \frac{\gamma - 1}{\gamma} - \left(\frac{1}{\theta} - \frac{1}{\gamma}\right)\omega_{ik}(s) \tag{25}$$

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 (24) to get

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

Here we see that, conditional on other firms' markups, each firm's market share is *strictly decreasing* in its markup. Together, equations (25) and (26) 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 (26) 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

$$\frac{1}{\mu_{ik}(s)} = \frac{\gamma - 1}{\gamma} - \left(\frac{1}{\theta} - \frac{1}{\gamma}\right) \frac{(\mu_{ik}(s)\mathsf{MC}_{ik}(s))^{1-\gamma}}{\sum_{i=1}^{N(s)}(\mu_{ik}(s)\mathsf{MC}_{ik}(s))^{1-\gamma}} \tag{27}$$

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

$$E_j L_j = \int_0^1 \sum_{k=1}^{J} \sum_{i=1}^{N(s)} l_{ijk}(s)\, ds \tag{28}$$

**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. See Appendix B for further details.

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

$$\bar{c}_{jk}(s) = \left(\sum_{i=1}^{N(s)} c_{ijk}(s)^{\frac{\gamma-1}{\gamma}}\right)^{\frac{\gamma}{\gamma-1}} \tag{29}$$

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

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

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

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

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

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

where

$$\omega_{ijk}(s) = \left(\frac{p_{ijk}(s)}{\bar{p}_{jk}(s)}\right)^{1-\gamma} = \left(\frac{\mu_{ik}(s)}{z_{ij}(s)}\,\frac{\bar{z}_{jk}(s)}{\bar{\mu}_{jk}(s)}\right)^{1-\gamma} \tag{33}$$

[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.

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, in efficiency units, *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

$$\bar{Z}_k(s) = \left( \sum_{j=1}^{J} \left( \frac{\bar{\mu}_{jk}(s)}{\bar{\mu}_k(s)} \right)^{-\gamma} \left( \frac{\bar{z}_{jk}(s)}{\tau_{jk}(s)} \right)^{\gamma-1} \left( \frac{W_j}{\bar{W}_k(s)} \right)^{-\gamma} \right)^{\frac{1}{\gamma-1}} \tag{34}$$

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 (31) 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 $j$.

**Aggregating across sectors.** Finally, we aggregate across sectors to obtain location-level productivity, markups, and wages. Let $\bar{\mu}_k$ denote the sales-weighted harmonic average of the sector-level markups $\bar{\mu}_k(s)$, and let $\bar{L}_k$ denote the total labor used across all source locations to produce for destination $k$. Then location-level productivity $\bar{Z}_k := C_k/\bar{L}_k$ can be written

$$\bar{Z}_k = \left( \int_0^1 \left( \frac{\bar{\mu}_k(s)}{\bar{\mu}_k} \right)^{-\theta} \bar{Z}_k(s)^{\theta-1} \left( \frac{\bar{W}_k(s)}{\bar{W}_k} \right)^{-\theta} ds \right)^{\frac{1}{\theta-1}} \tag{35}$$

where $\bar{W}_k$ is the location-level wage index implicitly defined by $P_k = \bar{\mu}_k \bar{W}_k/\bar{Z}_k$. The aggregate markup $\bar{\mu}$ is then given by the sales-weighted harmonic average of the location-level markups $\bar{\mu}_k$. Combining this with our assumption that profits are paid in proportion to labor income also gives $\bar{\mu} = 1 + \bar{\pi}$, as shown in Appendix B.

Overall we see that markup dispersion reduces aggregate productivity along three distinct dimensions. Dispersion *across firms* at a given source $j$ directly reduces the productivity $\bar{z}_{jk}(s)$ at each node of the productivity network, as in equation (31). For any given network of $\bar{z}_{jk}(s)$, dispersion *across source locations* reduces sector-level productivity $\bar{Z}_k(s)$ at each destination $k$, as in equation (34). And for any given $\bar{Z}_k(s)$, dispersion *across sectors* reduces location-level productivity $\bar{Z}_k$, as in equation (35). In this sense, the spatial dispersion in markups creates *endogenous misallocation*, across firms, across source locations, and across sectors. We quantify each of these three dimensions below.

# 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, as discussed in Appendix A. 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 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 manufacturing employment:

| | | | |
|---|---|---|---|
| 1 | Los Angeles-Riverside-Orange County | $14.6\%$ | of 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 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 $n_i(s) = \sum_j \mathbf{1}\{j \in \mathcal{J}_i(s)\}$ denote the number of locations where firm $i$ has establishments. We take this set $\mathcal{J}_i(s)$ exactly from the data, as Figure 1 illustrates for two sectors, *semi-conductors & electronic components* and *ready-mix concrete*.

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

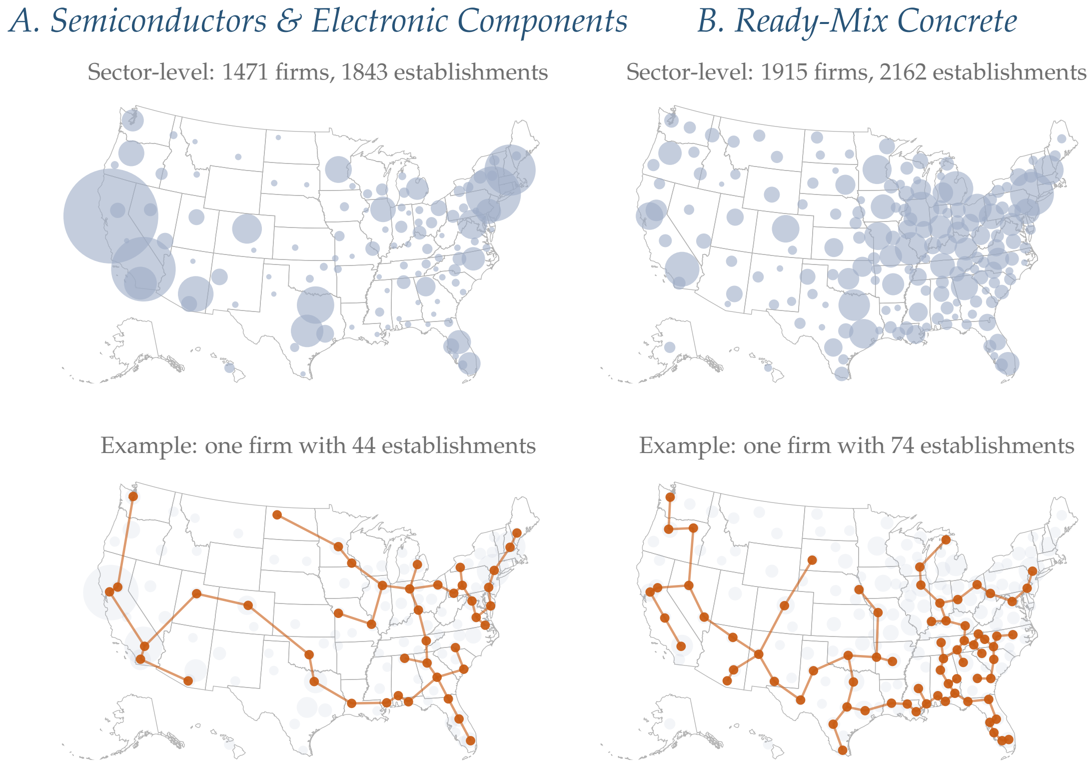
$$z_{ij}(s) = z_i(s) \cdot \mathbf{1}\{j \in \mathcal{J}_i(s)\} \tag{36}$$

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

$$\mathcal{C}(u, v) = \exp\left(-\left[(-\ln u)^{\frac{1}{1-\rho}} + (-\ln v)^{\frac{1}{1-\rho}}\right]^{1-\rho}\right) = \text{Gumbel Copula}(\rho) \tag{37}$$

Figure 1: Exact geographic footprints of firms and sectors




We match the exact patterns of geographic dispersion in locations at the NAICS 6-digit sector, firm- and establishment-levels. The top row illustrates the sector-level dispersion for *semiconductors & electronic components* (sector 334413) and *ready-mix concrete* (327320). In the top row, each dot is an Economic Area (EA), with area proportional to the number of establishments the sector operates there. The bottom row chooses a real firm from each sector to illustrate the firm- and establishment-level footprints that we match. Each firm's establishments are placed at the EA locations $\mathcal{J}_i(s)$ where the firm operates in the data and joined by a minimum spanning tree.

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 $n$ 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 1,123 firms that have $n = 10$ establishments. In the model, we can place the establishments of each of these 1,123 firms into the exact EAs 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 distribution of firm-level productivity conditional on $n = 10$. We then randomly sample 1,123 productivity draws with replacement from this conditional distribution and assign them to the 1,123 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 Pareto tail $\xi$ and the Gumbel rank correlation parameter $\rho$, as discussed further below.[4]

**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 $d_{jk}$

$$\ln \tau_{jk}(s) = \delta(s) \ln d_{jk} \tag{38}$$

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 national concentration from the US Census of Manufactures and the operations of multi-establishment firms from NETS. 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.[5]

[4]For the Gumbel copula in (37), 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 $\rho$ as the Gumbel correlation parameter.

[5]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

Table 1: Parameterization

| Parameter | | Value | Target |
|---|---|---|---|
| *Assigned values* | | | |
| Elas. subs. across sectors | $\theta$ | 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 | $\xi$ | 10.34 | National concentration |
| Gumbel rank correlation | $\rho$ | 0.81 | Employment share of multi-estab firms |
| Trade cost wrt distance | $\delta(s)$ | | Gravity coeff. $\beta(s)$ for 3-digit NAICS |

The elasticity of substitution across sectors $\theta$ and the elasticity of substitution within sectors $\gamma = \eta$ are assigned so that the model matches the sector-level relationship between inverse markups and sales concentration given in equation (39). The remaining parameters are calibrated internally by the simulated method of moments targeting the moments indicated in Table 2. The sector-specific elasticities of trade costs with respect to distance $\delta(s)$ are calibrated so that the model reproduces gravity coefficients $\beta(s)$ estimated at the 3-digit NAICS level from the Commodity Flow Survey (CFS), as in equation (40).

For any given sector $s$, this relationship can be obtained by multiplying both sides of equation (25) by market shares $\omega_{ik}(s)$ and then aggregating over firms and locations to get

$$\frac{1}{\bar{\mu}(s)} = \frac{\gamma - 1}{\gamma} - \left(\frac{1}{\theta} - \frac{1}{\gamma}\right) \overline{\text{HHI}}(s) \tag{39}$$

where $\overline{\text{HHI}}(s)$ denotes the expenditure-weighted average across location-level $\text{HHI}_k(s)$.

**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 $\xi$, (ii) measures of the employment share of multi-establishment firms to pin down the Gumbel correlation parameter $\rho$, 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.

of concentration in 4-digit US Census of Manufactures data. For their preferred specifications, they obtain estimates of the across-sector elasticity of substitution $\theta$ between 1.15 and 1.35 and estimates of the within-sector elasticity $\gamma$ between 7 and 13. We set $\theta = 1.25$ and $\gamma = 10$ as the rough midpoints of these ranges.

(i) NATIONAL SALES CONCENTRATION. We target the average national top-4 sales share and sales HHI. In the US Census of Manufactures, the average national top-4 sales share for 6-digit NAICS sectors is 42% while the average national HHI is 0.10.

(ii) MULTI-ESTABLISHMENT FIRMS. In NETS data, 3% of firms are multi-establishment. 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

$$\ln(\bar{p}_{jk}(s)\bar{y}_{jk}(s)) = \alpha_j(s) + \alpha_k(s) + \beta(s)\ln d_{jk} + \epsilon_{jk}(s) \tag{40}$$

where $\alpha_j(s), \alpha_k(s)$ denote sector-specific source and destination fixed effects. We estimate these gravity regressions using shipment-level microdata from the 2017 CFS for each 3-digit NAICS manufacturing sector. The geographic units in the CFS are origin and destination *regions*, most of which are individual US states, along with a small number of multi-state metropolitan areas. We aggregate weighted shipment values by region pair and sector and measure $d_{jk}$ as the great-circle distance between region centroids. Since the CFS regions are coarser than our 170 EAs, we estimate $\beta(s)$ at the CFS region level and then require that the model reproduce these same coefficients when we run the analogous regressions on simulated flows. Our estimated slope coefficients $\beta(s)$, reported in Appendix A.2, measure how sensitive trade flows are to geographical distance. Goods that are more easily tradeable, such as *computers & electronics* (sector 334), have estimated $\beta(s)$ that are small in magnitude. Goods that are less easily tradeable, such as *wood* (321) and *petroleum & coal products* (324), 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

$$\bar{p}_{jk}(s)\bar{y}_{jk}(s) = \sum_{i=1}^{N(s)} p_{ijk}(s)y_{ijk}(s). \tag{41}$$

To be consistent with our empirical gravity regressions from the CFS, we aggregate these simulated shipment flows to a 3-digit cluster of sectors and choose the parameters $\delta(s)$ in our specification (38) so that the estimated $\beta(s)$ in the model gravity regressions match their empirical counterparts from (40).

We report our internally calibrated parameters governing the productivity distribution and the operations of multi-establishment firms in Table 1. Jointly with our other parameters, our model matches the data on national sales concentration with a Pareto tail

Table 2: Model fit

| Moments [targeted] | Data | Model |
|---|---|---|
| *National concentration* | | |
| Top 4 sales share | 0.42 | 0.44 |
| HHI sales | 0.10 | 0.10 |
| *Local concentration* | | |
| HHI production | 0.36 | 0.37 |
| Top 4 sales share | | 0.58 |
| HHI sales | | 0.15 |
| *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 |

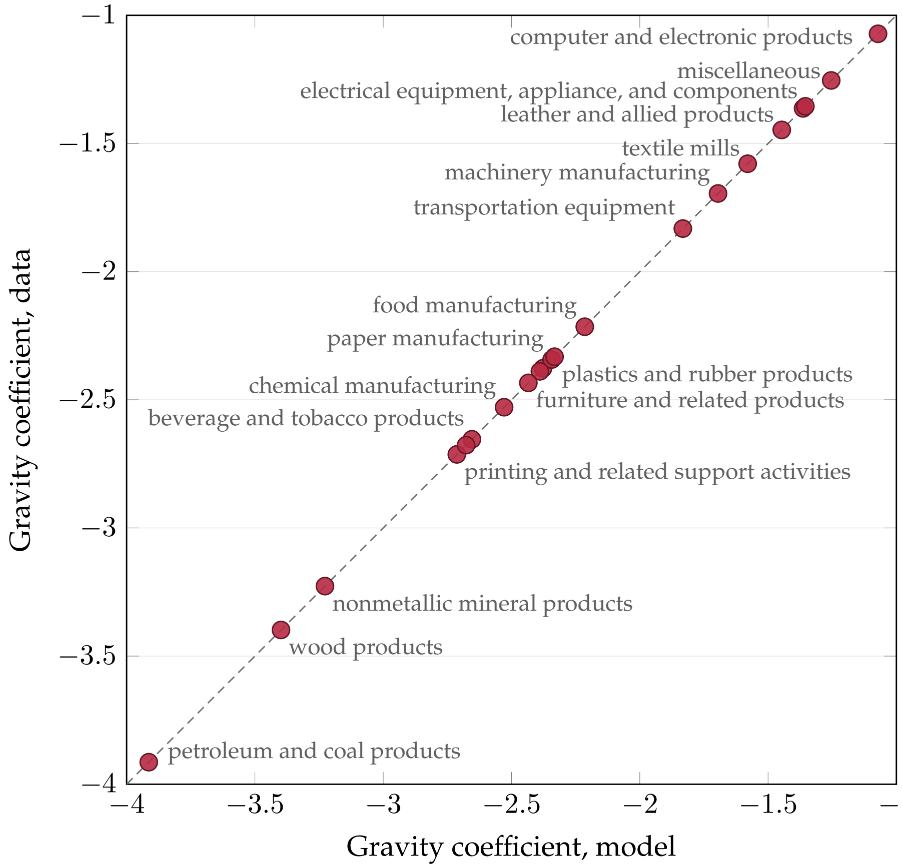
Targeted and untargeted moments in the data and their model counterparts. Moments highlighted in red are targeted by the simulated method of moments. All others are untargeted and serve as model validation. Local sales concentration has no data counterpart because it is not directly observed. The gravity coefficients $\beta(s)$ that pin down the trade cost elasticities $\delta(s)$ are reported separately in Figure 2.

$\xi = 10.34$, implying much 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 $\rho = 0.81$ between a firm's productivity fixed effect $z_i(s)$ and its number of establishments $n_i(s)$.

**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 the magnitude of 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 sales share of multi-establishment firms.

Figure 2: Gravity coefficients $\beta(s)$ in model and data




Each point is one 3-digit manufacturing sector, plotting the gravity coefficient $\beta(s)$ estimated from the Commodity Flow Survey (CFS) against its model counterpart. Our model exactly reproduces the sector-level gravity effects that pin down the magnitude of spatial trade frictions. Sectors with large, negative $\beta(s)$ have *high gravity* and are less tradeable than *low gravity* sectors.

# 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.

## 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 in the data, as indicated by the blank spaces in Table 2. But given that our model does a good job of reproducing national sales concentration and local production concentration, it seems natural to use the model to infer the amount of local sales concentration, i.e., to fill in these blanks.

Intuitively, our model implies local sales concentration lies between the national sales concentration we targeted in our calibration and local production concentration. Averaging across locations weighted by employment:

| **National Sales** | | **Local Sales** | | **Local Production** |
|---|---|---|---|---|
| $\text{HHI} = 0.10$ | $<$ | $\text{HHI} = 0.15$ | $<$ | $\text{HHI} = 0.37$ |
| *least concentrated* | | | | *most concentrated* |

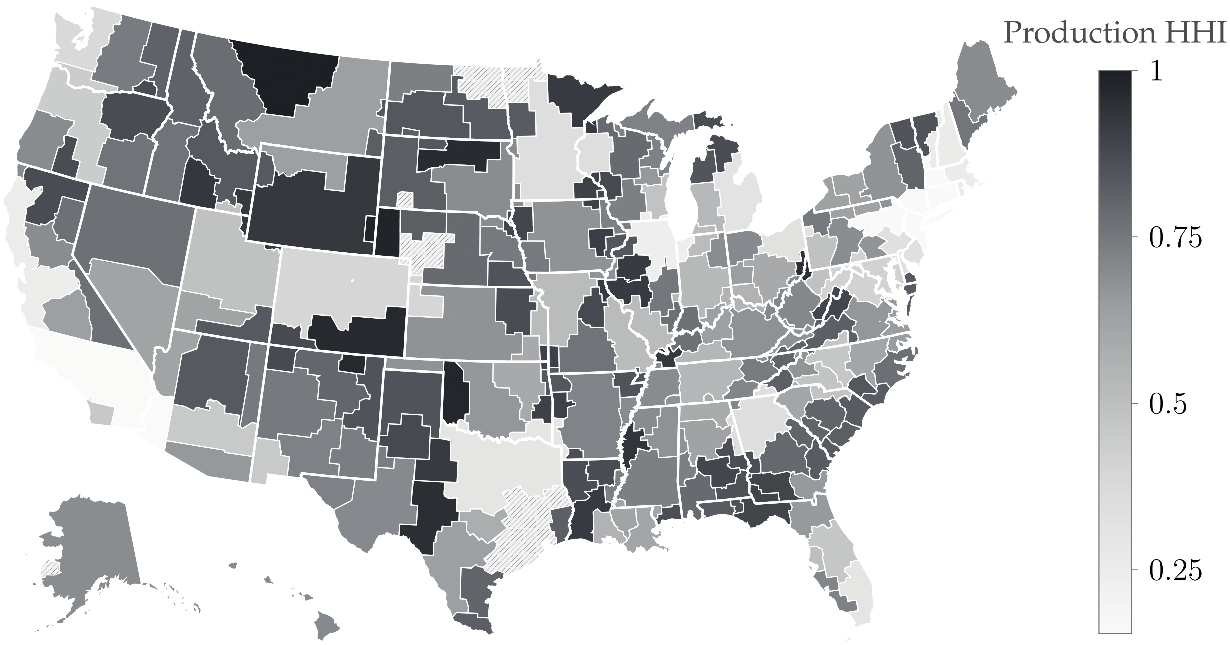
Since the inverse of the HHI corresponds to the number of equally-sized firms, to interpret these concentration statistics more intuitively, it is as if there are on average $10$ equally-sized firms nationally, about 6-7 equally-sized firms selling locally, and just under 3 equally-sized firms producing 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. 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.

## *4.2 Sales concentration vs. production concentration*

Our model predicts that local sales concentration is both lower and less dispersed than local production concentration. To assess this, Figure 3 reports average local production concentration (as measured by HHIs) across the 170 EAs in our data. These range from around $0.15$ to nearly $1$, indicating areas where most goods are produced by a single firm. By contrast, average local sales concentration (again measured by HHIs) in our model ranges from around $0.12$ to $0.23$. 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 one of the lowest production concentrations but the second-highest sales concentration.

**Spatial correlation of sales vs. production.** Figure 4 reports the scatter of local production concentration against local sales concentration in our model for the 170 EAs. If local production concentration were a good predictor of local sales concentration, we would expect the scatter to cluster tightly around the $45°$-line. Instead the scatter is close to

Figure 3: Local production concentration (HHI)

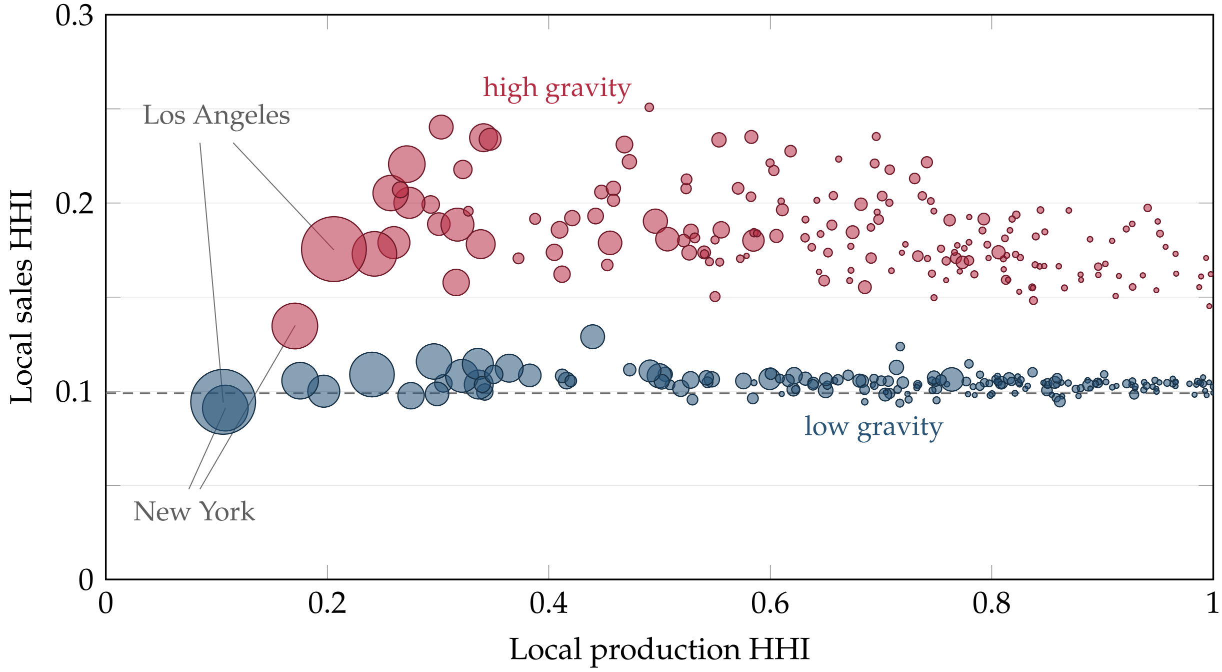



Average local production concentration, measured by the Herfindahl–Hirschman index (HHI), across the 170 Economic Areas (EAs) in our data. Production HHI ranges from about 0.15 to nearly 1, the latter indicating areas where most goods are produced by a single firm. County colors show the HHI of the EA containing each county. Hatched counties are not matched to an EA.

horizontal. Local production concentration, of the kind we can readily measure from data on shipments, is simply not informative about the local sales concentration that matters for competition and market power. So what does explain that variation? In Appendix D we consider a number of measures of market *contestability* and show how these help account for differences in local sales concentration and markups across locations.

**Spatial frictions and concentration.** The same figure also highlights the role of tradeability. 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* or *wood* products), and low gravity sectors, with low frictions that are much less costly to ship (e.g., *computer & electronic* products), shown by the color of each point. On average, high gravity sectors in red have higher local sales concentration than low gravity sectors in blue: trade costs can substantially limit competition, especially for smaller and more geographically remote locations. High gravity sectors also have sales concentration that is considerably more variable across locations. For low gravity sectors, by contrast, local sales HHIs cluster tightly just above 0.1 — just above the national sales HHI — so that these goods are traded in what amounts to a single national market.

Figure 4: Local sales concentration vs. local production concentration




Local production HHI against local sales HHI. Each Economic Area (EA) contributes two points, one averaging over its high-gravity (less tradeable) sectors and the other over its low-gravity (more tradeable) sectors — shown in red and blue respectively. Both points are then sized by the EA's total employment. If production concentration predicted sales concentration, points would cluster on the 45-degree line. Instead the scatter is close to horizontal. High-gravity points have both higher and more variable sales concentration. Low-gravity points cluster just above the national sales HHI of about 0.1.

## 4.3 *Markups and sales concentration*

In our model, a location's sales-weighted harmonic average markup $\bar{\mu}_k$ is in one-to-one correspondence with the location's sales-weighted arithmetic average HHI.[6] Because of this, the spatial distribution of markups is exactly the spatial distribution of local sales concentration. Figure 5 shows this common map, with the left axis reporting the average local sales HHI and the right axis reporting the corresponding average markup $\bar{\mu}_k$.

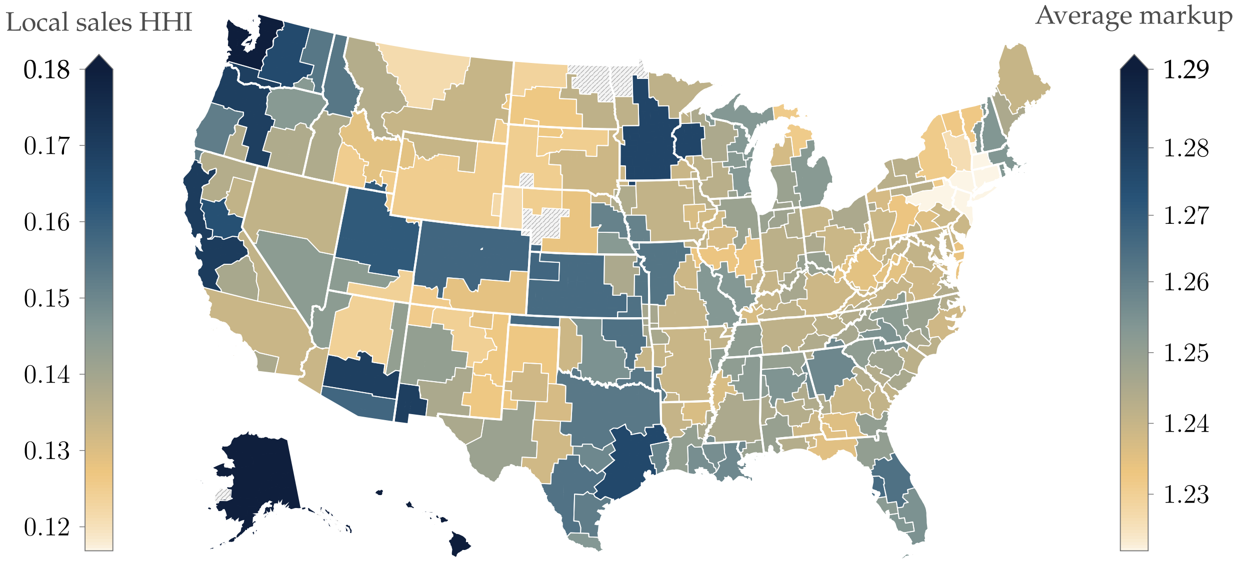
Across the 170 EAs in our data, average markups $\bar{\mu}_k$ range from a low of 1.22 in the greater New York area to highs of 1.29 in Anchorage, 1.31 in greater Seattle, and 1.35 in Honolulu. These location-specific averages $\bar{\mu}_k$ mask extensive variation in markups $\mu_{ik}(s)$ across sectors $s$ 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

[6] Similar to (39), this follows by aggregating (25) over firms and sectors to get the location-level relationship

$$\frac{1}{\bar{\mu}_k} = \frac{\gamma - 1}{\gamma} - \left(\frac{1}{\theta} - \frac{1}{\gamma}\right) \overline{\mathrm{HHI}}_k$$

between average markups $\bar{\mu}_k$ 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.

Figure 5: The spatial distribution of markups




Local sales concentration (sales-weighted average HHI) across the 170 Economic Areas (EAs). Because a location's average markup $\bar{\mu}_k$ is in one-to-one correspondence with its average sales HHI, they have the same geography. The left axis reports the average local sales HHI and the right axis reports the corresponding average markup $\bar{\mu}_k$. For visual clarity, the color scale is capped at an HHI of 0.18 (markup 1.29). Two EAs exceed this cap, greater Seattle (markup 1.31) and Honolulu (markup 1.35). County colors show the value of the EA containing each county. Hatched counties are not matched to an EA.

markup is 1.23, and the top percentile of markups is 1.62. The overall aggregate markup in our model is $\bar{\mu} = 1.26$.

# 5 Quantitative Importance of Spatial Frictions

In this section we present our main quantitative results on the aggregate effects of spatial frictions. We quantify these effects in two complementary ways. First, we ask how much we would need to reduce the number of firms in an otherwise equivalent economy *without geography* to reproduce the level of competition we see in our benchmark. The answer is stark: it takes a 90% reduction in the number of firms to reproduce the reduction in competition created by our benchmark spatial frictions. Second, we ask how the economy changes as we move along the spectrum of spatial frictions from intranational 'autarky' to 'free trade'. Our benchmark sits much closer to the free trade end of the spectrum, but even so the spatial frictions still deliver a quantitatively significant reduction in competition. We also use these results to show that our model can generate *endogenously divergent* trends in national and local concentration in response to falling trade costs, and to document the spatial distribution of markup and consumption changes in response to changes in trade costs. We show that the smallest, poorest, most remote locations generally gain more than the largest, richest, most economically central locations.

Table 3: Markups with and without geography

| Across sectors | Benchmark model | No geography |
|---|---|---|
| *Markup distribution* | | |
| 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 |
| *Concentration* (HHI) | | |
| National sales | 0.10 | 0.10 |
| Local sales | 0.15 | 0.10 |
| Local production | 0.37 | 0.10 |

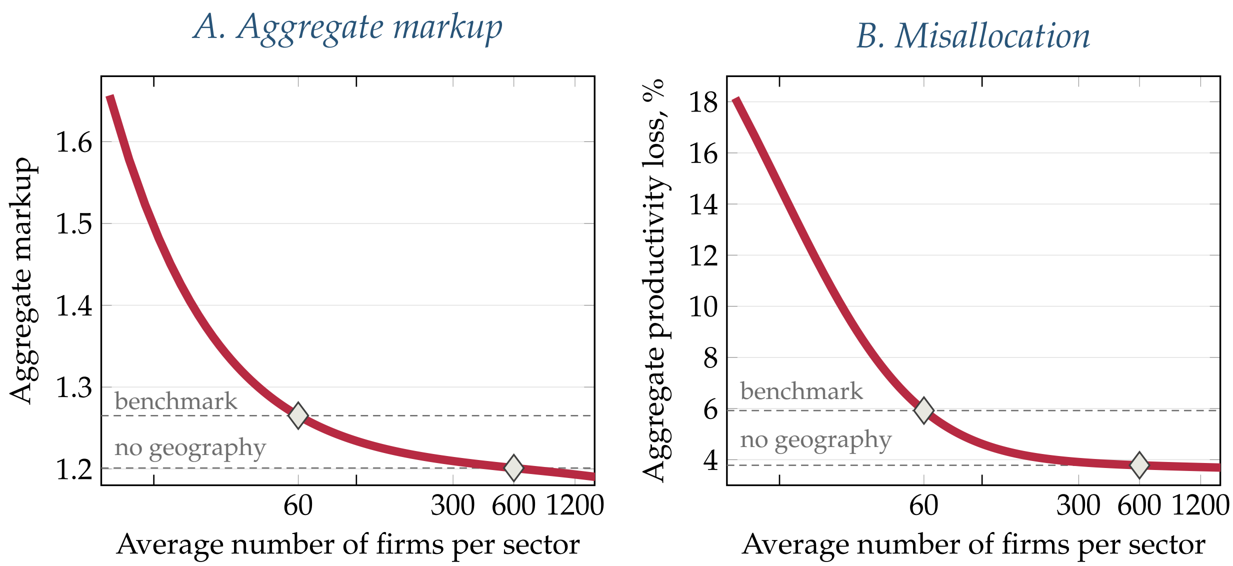
The distribution of markups across 6-digit NAICS sectors and concentration (HHI) in the benchmark model and in the no-geography model. The no-geography model abstracts from spatial frictions entirely but is calibrated to match the same national concentration facts as the benchmark. Without geography, local and national concentration coincide.

## 5.1 Quantifying the reduction in competition

One way to quantify the effect of spatial frictions on competition is to ask how much we would need to reduce the number of firms in an otherwise equivalent model *without geography* to reproduce the level of competition we see in our benchmark.

**The 'no geography' model.** To make this comparison, we construct a 'no geography' model that abstracts from spatial frictions altogether, with perfectly integrated product and labor markets, but that is calibrated to match the same national concentration facts as our benchmark. Table 3 reports the markup distribution and concentration facts for this no-geography model alongside our benchmark. Markups in the no-geography model are both lower and less dispersed than in our benchmark, with the aggregate markup falling from 1.26 to 1.19. To put this in context, the aggregate markup of 1.19 in the no-geography model is even lower than the average markups our benchmark implies for New York (1.22) and Los Angeles (1.24). Abstracting from geography is like assuming the competitive conditions of New York or Los Angeles (or better) prevail *everywhere* in the country.

Figure 6: Competition and the number of firms in the no-geography model



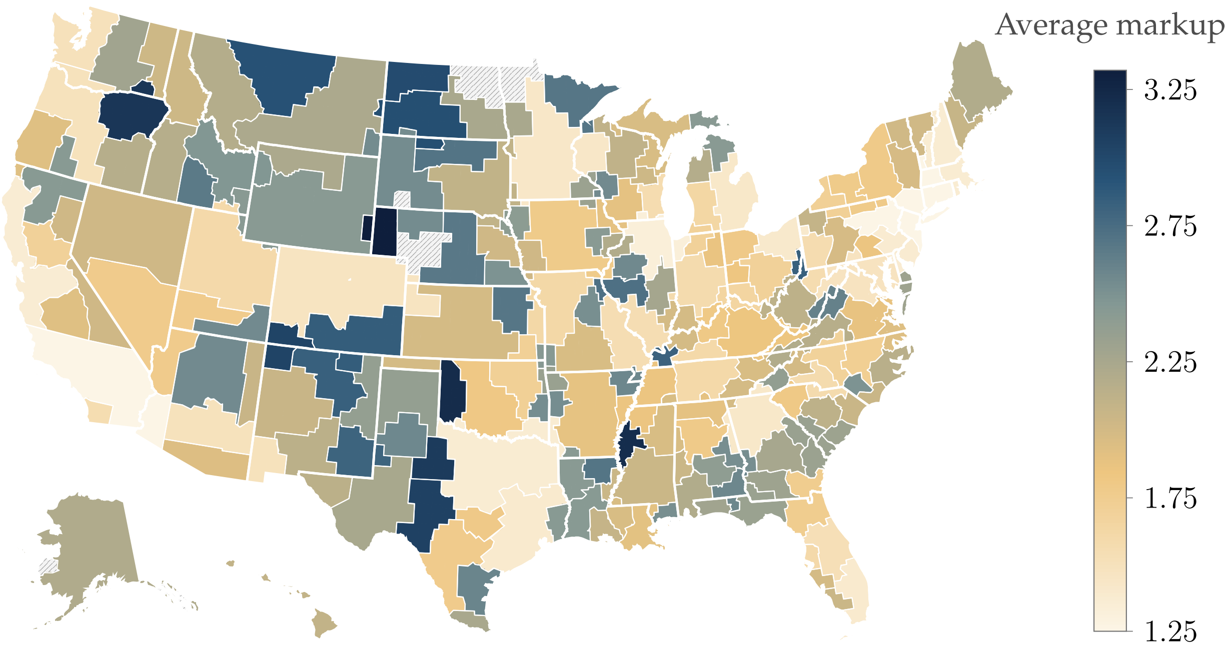

Panel A shows the aggregate markup as a function of the average number of firms per 6-digit NAICS sector in the *no-geography* model, on a log scale. Panel B shows the corresponding aggregate productivity losses due to misallocation, in percent. The marked diamonds are our benchmark calibration with an average of approximately 600 firms per sector and the much smaller count, approximately 60 firms per sector, at which the no-geography model reproduces the amount of competition in our benchmark model *with geography*.

**Changing the number of firms.** We report the results of this comparison in Figure 6, which shows how the aggregate markup and the productivity losses due to misallocation vary with the average number of firms per sector in the *no-geography* model. The diamonds mark an average of 600 firms per sector, as in our benchmark, and the average number of firms at which the no-geography model reproduces the level of competition in our benchmark. Two features stand out. First, competitive outcomes are relatively insensitive to the number of firms over a wide range. Doubling or halving the number of firms — from 600 to 1200, or from 600 to 300 — has small effects on the aggregate markup or on the productivity losses due to misallocation. When the number of firms is already large, changes in this number have little effect on competition. Second, reducing the number of firms below about 100 raises markups and misallocation sharply. The level required to replicate the level of competition in our benchmark with geography is about 60 firms per sector — roughly a 90% reduction from the benchmark. In this sense, the reduction in competition generated by US spatial frictions is equivalent to removing a randomly chosen 90% of firms from each sector of the no-geography economy.

## 5.2 *From autarky to free trade*

A second way to quantify the effect of spatial frictions on competition is to trace out the spectrum of outcomes from 'intranational autarky', where trade costs are so high as to

Figure 7: Markups in autarky



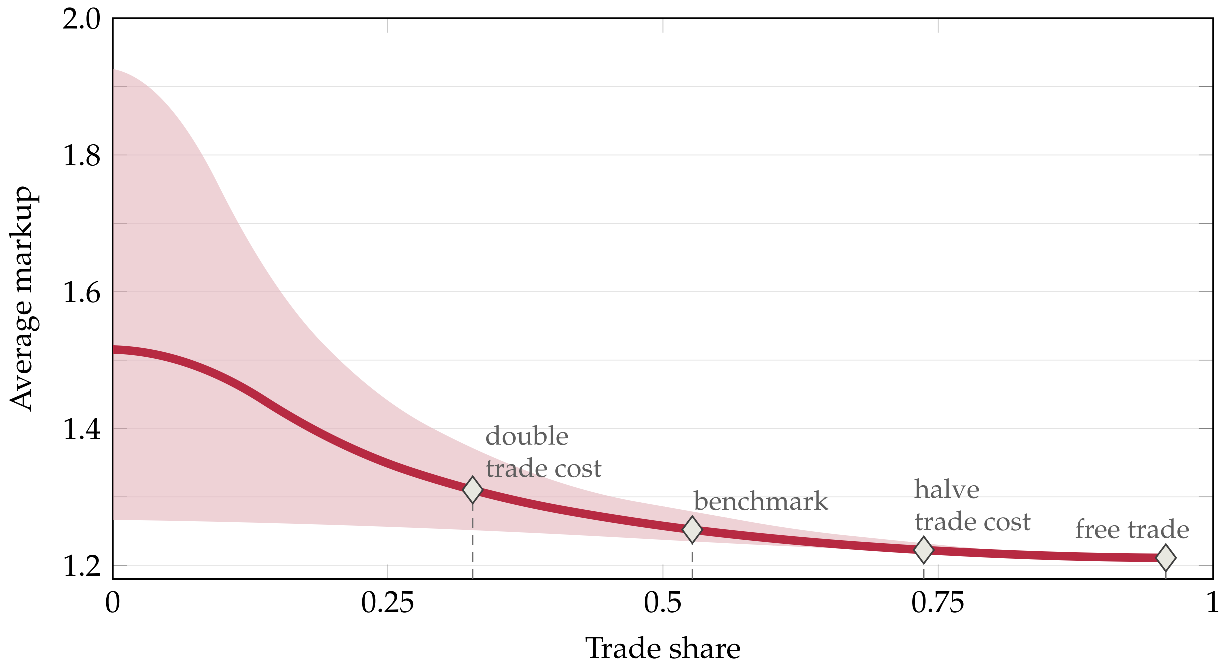

Average markup $\bar{\mu}_k$ across the 170 Economic Areas (EAs) under intranational autarky, where trade costs are high enough to prevent any trade between locations. Markups are far higher and far more dispersed than in the benchmark model, ranging from 1.26 to 3.32. Without trade, local sales concentration mirrors local production concentration so what matters is how much competition there is among home producers. County colors show the value of the EA containing each county. Hatched counties are not matched to an EA.

prevent any trade between locations, to 'free trade', where there are no trade costs at all.

**Autarky.** Under intranational autarky, with $\tau_{jk}(s) = +\infty$, markups are much larger and much more variable than in our benchmark model. Figure 7 reports our model's implications for average markups $\bar{\mu}_k$ across the 170 EAs in our model for this extreme case. Without trade, local sales concentration mirrors local production concentration location by location, so locations with low home production end up with much higher sales concentration — and hence much higher markups — than in our benchmark. The most extreme case is Scottsbluff, whose average markup rises from 1.23 in our benchmark to 3.32 under autarky, the largest increase of any location. In our benchmark Scottsbluff has almost no home production of its own and instead relies on imports from a large number of competing sources, which discipline markups, so without trade it has nothing to fall back on. By contrast Honolulu has the *highest* markup in our benchmark model, 1.35, but rises to only 2.09 under autarky. What matters under autarky is not how remote a location is, but how much competition there is among its home producers.

The lack of trade matters much less for locations with deep home production. New York's average markup is 1.26 under autarky as opposed to 1.22 in our benchmark, the smallest increase of any location — New York's low markups are driven by its deep,

Figure 8: Markups from autarky to free trade




Employment-weighted average markup as a function of the employment-weighted trade share, as spatial frictions vary from autarky (left) to free trade (right). The shaded band gives the 10th to 90th percentiles of the distribution of average markups across locations. Diamonds mark four points along the spectrum: double the trade costs, our benchmark trade costs, half the trade costs, and free trade.

competitive home market conditions, not by import competition. Seattle is a bit in between, with an average markup of 1.50 under autarky as opposed to 1.31 in our benchmark. Seattle is also large, but its home production is considerably shallower than New York's, so shutting down trade matters more for Seattle than for New York — though still far less than for locations with almost no home production of their own.

**Free trade.** Under free trade, $\tau_{jk}(s) = 1$, there is no geographic segmentation in product markets. Firms in each sector compete in a single national market and set the same firm-specific markup in every destination. Average markups are therefore the same in every location, $\bar{\mu}_k = \bar{\mu}$, and this common level works out to be $\bar{\mu} = 1.21$ — close to but not exactly equal to the 1.19 of the no-geography model, the small difference reflecting the fact that the free-trade economy still has local labor markets and differences in wages across locations. 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.

**Between the extremes.** Figure 8 traces out the full spectrum between these two extremes. In this figure we report the employment-weighted average markup as a function of the employment-weighted trade share, along with the 10th and 90th percentiles of the distribution of average markups across locations. Three features stand out. First, the average markup is more responsive to increases in integration near autarky and becomes less responsive when the extent of integration is already high. Second, the spatial dispersion in markups is high when spatial frictions are high and decreases substantially as the extent of integration increases. Most of the reduction in spatial dispersion is through a compression of the upper tail of the distribution. Near autarky, the 90th percentile markups are roughly 1.5 times the 10th percentile markups. Near free trade, the entire distribution has been compressed to a single point. Third, our benchmark calibration is much closer to the free-trade end of the spectrum. But even here the remaining spatial frictions are enough to generate the quantitatively significant effects on competition, as shown in Section 5.1.

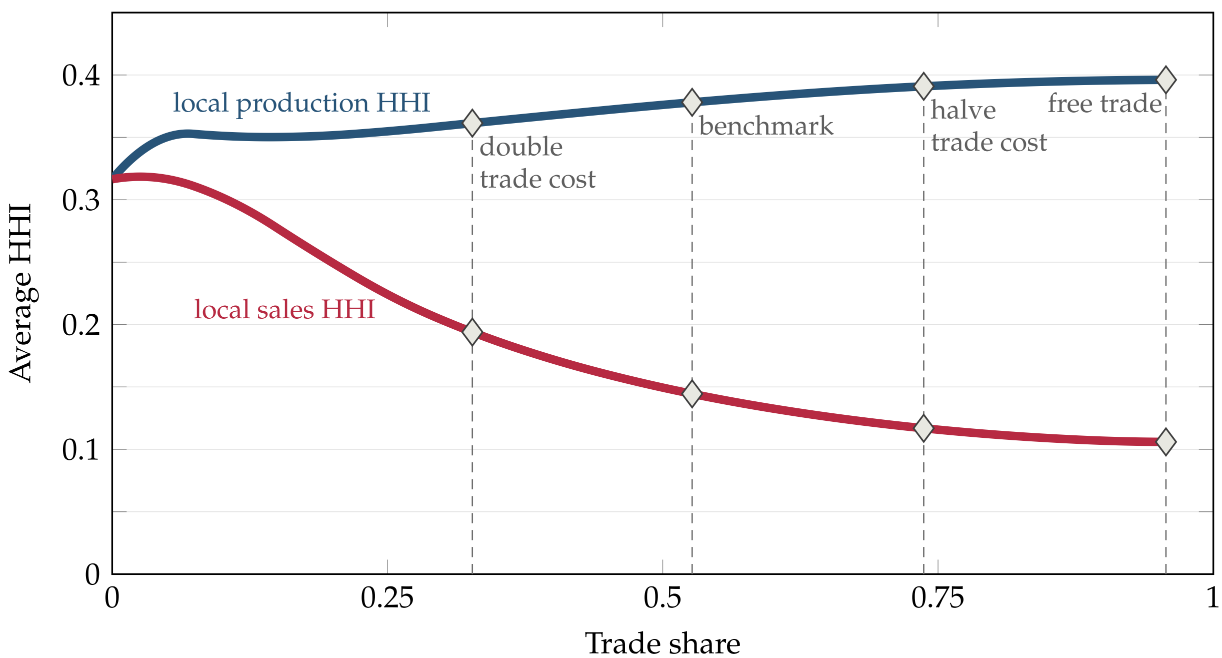
## *5.3 Divergence between local and national concentration*

Recent empirical research has documented significantly different trends in national sales concentration and local sales concentration. National sales concentration, along with 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 concentration emerge naturally in our model when trade costs are falling.

**Concentration from autarky to free trade.** Figure 9 reports the employment-weighted local sales HHI and local production HHI as a function of the employment-weighted trade share. In autarky, all goods are sold where they are produced so local sales and local production concentration coincide. As trade integration rises, local production concentration rises, as the most productive firms expand by accessing more distant markets, while local sales concentration falls, as consumers in each location gain access to a wider set of competitive suppliers. In the free trade limit, the local sales HHI converges to the national sales HHI.

**Reduction in trade costs.** To illustrate these effects more specifically, Table 4 reports outcomes for five points: autarky, 20% higher trade costs than our benchmark, our benchmark, 20% lower trade costs than our benchmark, and free trade. We choose the 20% reduction to match Coşar, Osotimehin and Popov (2024), who find, using interregional trade data, a 15 to 20% decrease in manufacturing distance elasticities from 1963 to 2017. Moving forward from 1963-style trade costs ($+20\%$) to our benchmark, the average national top-4 sales share rises modestly from 0.43 to 0.44 while the average local top-4 sales share falls from 0.61 to 0.58. A reduction in trade costs endogenously drives national sales concentration and

Figure 9: Concentration from autarky to free trade




Employment-weighted local production HHI and local sales HHI as functions of the employment-weighted trade share, as spatial frictions vary from autarky (left) to free trade (right). As trade integration rises, local production concentration rises while local sales concentration falls; in the free-trade limit the local sales HHI converges to the national sales HHI. Diamonds mark four points along the spectrum: double the trade costs, our benchmark trade costs, half the trade costs, and free trade.

local sales concentration in opposite directions. In response to lower trade costs, the most productive firms expand by accessing more distant markets. This increases national sales concentration and 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 decrease in trade costs are better understood as a kind of *convergence* towards a better-integrated national market — as trade costs fall from their benchmark levels, national sales concentration rises while local sales concentration falls. In the limit as intranational trade costs disappear, the distinction between national and local sales concentration disappears too. This can be seen in the last column of Table 4, which shows that in this free trade limit the national concentration moments exactly equal the local concentration moments.

Table 4: Effects of changes in trade costs

| | Autarky | +20% | Benchmark | −20% | Free trade |
|---|---|---|---|---|---|
| *Markup distribution* | | | | | |
| 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 |
| *Concentration* (HHI) | | | | | |
| National sales | – | 0.10 | 0.10 | 0.10 | 0.11 |
| Local sales | 0.31 | 0.16 | 0.15 | 0.13 | 0.11 |
| Local production | 0.31 | 0.36 | 0.37 | 0.38 | 0.40 |

The distribution of markups across 6-digit NAICS sectors and concentration (HHI) for five levels of trade costs along the spectrum from autarky to free trade, including trade costs ±20% relative to our benchmark model. National sales concentration is undefined under autarky. National sales concentration and local sales concentration are identical under free trade.

### 5.4 *Implications for markups and consumption*

We have seen that a reduction in intranational trade costs can lead to an increase in *national* sales concentration. We now show that nonetheless this reduction makes markets more competitive and increases consumption per worker in most locations. 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 actually sell their goods.

**Changes in the markup distribution.** Table 4 reports the effects of changes in intranational trade costs on the sector-level markup distribution and the aggregate markup. Since markups are determined in large part by the amount of competition in destination markets, and local sales concentration falls as trade costs fall, it is unsurprising that reductions in trade costs lead to both lower markups and lower markup dispersion — and hence lower productivity losses due to misallocation. Moving from 1963-style trade costs (+20%) to our benchmark, the aggregate markup falls from 1.27 to 1.26 while markup dispersion narrows. Further reductions in trade costs deliver further reductions in the aggregate markup and markup dispersion. The aggregate movements are modest — the aggregate markup falls by about 0.8% in response to a 20% reduction in trade costs from 1963 to our benchmark. But as we will now see, these modest aggregate changes mask considerable heterogeneity in markup changes across locations.

**Spatial heterogeneity in markup changes.** There is considerable spatial heterogeneity in how much markups respond to a given reduction in trade costs. In response to the 20% reduction in trade costs from our benchmark, the markup decreases range 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. The markup changes in more remote locations are up to around four times as large as the markup changes in more central locations. Consistent with our earlier 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.

**Spatial heterogeneity in consumption gains.** The spatial heterogeneity in consumption gains is even larger. In response to a 20% reduction in trade costs, consumption per worker rises by around 1.6% in greater New York but by nearly 16% in remote parts of North and South Dakota.[7] That is, there are ten-fold differences in consumption gains across locations, compared with four-fold differences in markup changes. The largest consumption gains are in remote locations, with more modest gains in coastal and large metro areas.

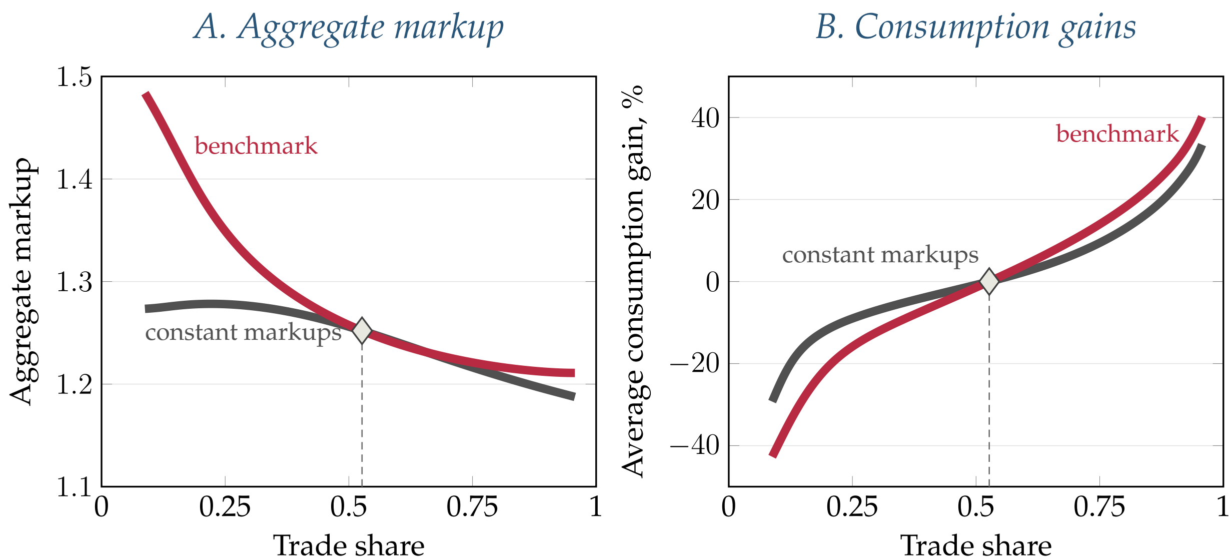
## 5.5 *Why endogenously variable markups matter*

In our model, markups vary across firms, sectors and locations because of differences in competition. But the amount of competition itself depends on more structural conditions, including trade costs and geography. When trade costs change, markups adjust. How much does this endogenous markup adjustment matter? To quantify this, Figure 10 reports our benchmark aggregate markup and employment-weighted average consumption gains compared to a counterfactual economy where each firm's markup is held constant at its benchmark level as we vary trade costs from near autarky to near free trade. The difference between the benchmark response to changing trade costs and the counterfactual response measures the additional effects of endogenously variable markups.

Panel A shows that the endogenous markup response to trade costs is large, especially as the economy gets close to autarky. For our benchmark economy, the aggregate markup rises from 1.26 at its initial level to 1.52 in autarky. In the counterfactual the aggregate markup responds much less and is close to flat, changing only because of changes

[7]Since labor $L_k$ for each location is fixed in our benchmark model, the change in final consumption $C_k$ is equivalent to the change in consumption per worker $C_k/L_k$.

Figure 10: Endogenous markups and the gains from trade




Panel A shows the aggregate markup as a function of the trade share, from high trade costs near autarky to low trade costs near free trade. Panel B shows the corresponding employment-weighted average consumption gains, in percent, measured relative to our benchmark. In each panel the red line is our benchmark economy, in which markups respond endogenously to competition, and the grey line is a counterfactual in which each firm's markup is held fixed at its benchmark level as trade costs vary. The marked diamonds are our benchmark calibration, at which the consumption gain is zero by construction.

in composition. The differences are more muted in the other direction, moving from the benchmark towards freer trade, because our benchmark already features extensive competition. Towards free trade, the aggregate markup falls by slightly more in the counterfactual — but this reflects sales reallocating towards the firms that have low markups in the benchmark, not reallocation towards the most productive firms. Panel B shows that the constant-markup counterfactual *understates the losses* from increases in trade costs relative to the benchmark and *understates the gains* from decreases in trade costs. Moving towards autarky, allowing markups to respond deepens the consumption losses — the benchmark loss exceeds the constant-markup loss by about 13 percentage points. Moving towards free trade, allowing markups to respond adds to the gains — the benchmark consumption gain exceeds the constant-markup gain by about 7 percentage points. In this sense, abstracting from endogenous markup responses leads to a quantitatively significant understatement of the welfare consequences of changing trade costs.

The consumption gains here depend on both 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 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, 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 5.7% in consumption-equivalent terms, and vary considerably across locations, from less than 1% 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 components: (i) *across firms* $i$ within source locations $j$, (ii) *across source locations* $j$ within destination $k$, and (iii) *across sectors* $s$ within destination $k$. We find that the across-firm component is the largest, accounting for over half of total misallocation, and is remarkably stable across locations. The across-source and across-sector components account for the bulk of 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.

**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 (18) we get

$$\chi_{ik}(s) \cdot p_{ijk}(s) = \mu_{ik}(s) \cdot \frac{W_j}{z_{ij}(s)} \tag{42}$$

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 5 reports our main results for the welfare costs of markups in our benchmark model. On average, the representative consumer would gain 5.7% in consumption-equivalent terms from eliminating markups. When we compute the costs of markups in our model without geography, as in Section 5.1 above, we find that the representative consumer would gain 3.7% in consumption-equivalent terms from eliminating markups. Again we see that abstracting from spatial frictions leads to a quantitatively significant understatement of the welfare costs of markups. Just as

Table 5: Consumption gains from eliminating markups, %

| Across Locations | Benchmark | Free trade |
|---|---|---|
| p01 | 0.9 | 2.0 |
| p10 | 3.6 | 2.9 |
| p25 | 3.9 | 3.8 |
| p50 | 5.6 | 4.5 |
| p75 | 6.9 | 5.7 |
| p90 | 9.1 | 6.1 |
| p99 | 14.5 | 7.6 |
| Average | 5.7 | 4.6 |

The distribution of consumption gains from eliminating markups across the 170 Economic Areas (EAs), in percent, in the benchmark model and under free trade. Gains are measured in consumption-equivalent terms. The distribution is wider in the benchmark, where spatial frictions make the welfare cost of markups much larger in some locations than others. Under free trade the gains are lower on average and less dispersed because markups are lower and less dispersed.

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.** Overall we find that there is extensive geographic variation in the welfare costs of markups, with lower costs in larger, more economically central locations like greater New York, for instance — just 0.9% — and higher costs in more remote locations, reaching as much as 21.5% in Honolulu. 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 and (ii) locations which have higher initial trade shares have larger costs of 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.

## 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 across firms, across locations, and across sectors. We now use our aggregation results from Section 2.2 to quantify these components of misallocation.

**Efficient productivity.** If markup variation is eliminated then the bilateral productivity network $\bar{z}_{jk}(s)$ from equation (31) above simplifies to the familiar CES index

$$\bar{z}_j^*(s) = \left(\sum_{i=1}^{N(s)} z_{ij}(s)^{\gamma-1}\right)^{\frac{1}{\gamma-1}} \tag{43}$$

This is the *efficient* productivity level at each source $j$ in sector $s$. 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 $j$ we can compute the sector-level productivity that would prevail at each destination $k$ if markup variation is eliminated

$$\bar{Z}_k^*(s) = \left(\sum_{j=1}^{J} \left(\frac{\bar{z}_j^*(s)}{\tau_{jk}(s)}\right)^{\gamma-1} \left(\frac{W_j}{\bar{W}_k(s)}\right)^{-\gamma}\right)^{\frac{1}{\gamma-1}} \tag{44}$$

Importantly, we perform this calculation *holding wages fixed* at their benchmark levels $W_j$. As discussed in Appendix C, 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 levels

$$\bar{Z}_k^* = \left(\int_0^1 \bar{Z}_k^*(s)^{\theta-1} \left(\frac{\bar{W}_k(s)}{\bar{W}_k}\right)^{-\theta} ds\right)^{\frac{1}{\theta-1}} \tag{45}$$

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

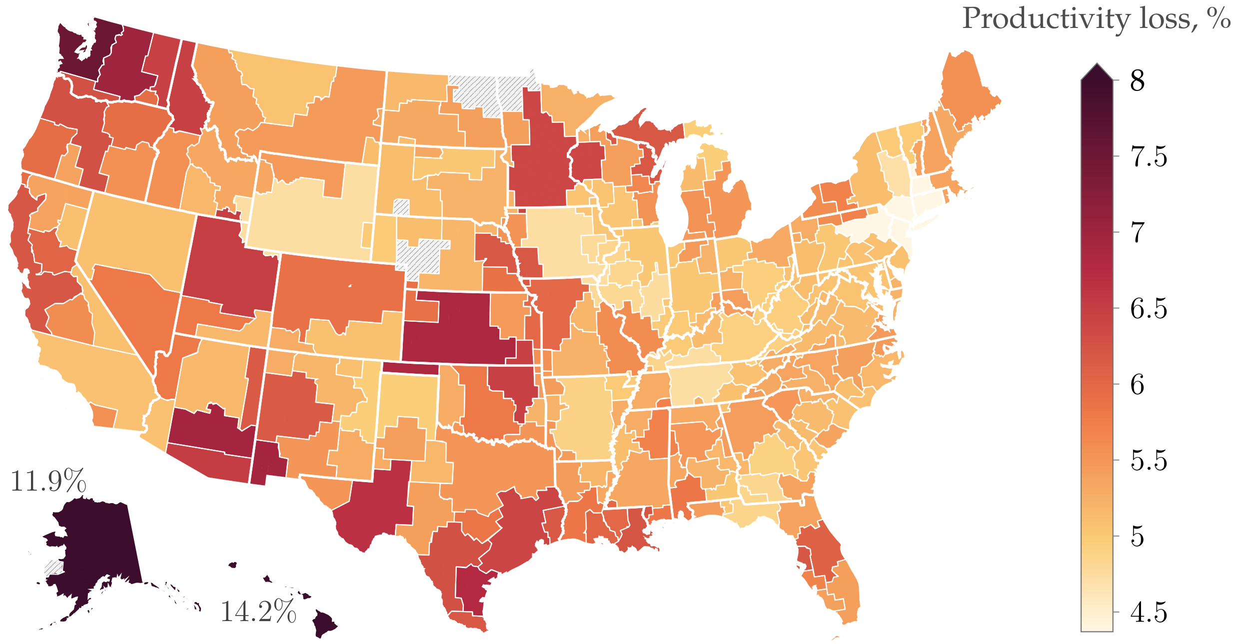
$$\hat{Z}_k(s) := \frac{\bar{Z}_k(s)}{\bar{Z}_k^*(s)} \leq 1 \tag{46}$$

and similarly measure overall misallocation at each destination $k$ by

$$\hat{Z}_k := \frac{\bar{Z}_k}{\bar{Z}_k^*} \leq 1 \tag{47}$$

Note that because we hold relative wages fixed in computing $\bar{Z}_k^*$, this measure of misallocation does not correspond exactly, location by location, to the consumption gains reported in Table 5 — the latter are general equilibrium outcomes in which relative wages adjust, while $\hat{Z}_k$ isolates the direct effect of markup dispersion at fixed wages.

Figure 11: The spatial distribution of misallocation

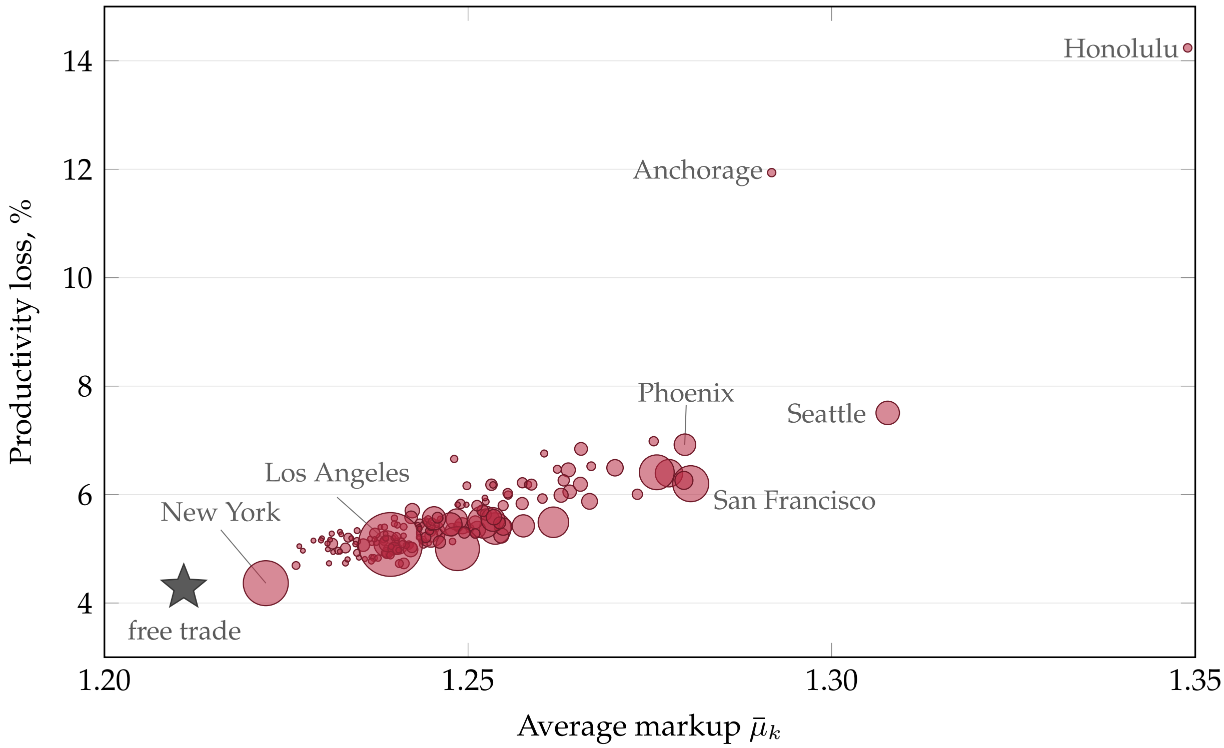



Productivity losses from misallocation, $1 - \hat{Z}_k$ as defined in equation (47), across the 170 Economic Areas (EAs). Losses are largest in more remote locations and smallest in the largest, most central markets. For visual clarity, the color scale is capped at 8%; two EAs exceed this cap, Anchorage and Honolulu, and each is labelled with its actual value directly on the map. County colors show the value of the EA containing each county. Hatched counties are not matched to an EA.

**Spatial heterogeneity in misallocation.** Figure 11 reports the full geographic variation in the productivity losses due to misallocation across locations. These losses range from 4.4% in greater New York to 11.9% in Anchorage and 14.2% in Honolulu, which are genuine outliers — across all other locations, misallocation is never more than 8%. Even excluding these outliers, misallocation varies by a factor of roughly 2 across locations, considerably more than the variation in average markups across locations, which range from 1.22 in greater New York to 1.35 in Honolulu. The same spatial frictions which lead to higher average markups in some locations give rise to greater misallocation too, as shown in Figure 12. The correlation between the two is 0.84, rising to 0.85 if we exclude the outliers Anchorage and Honolulu.

**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 6 reports the decomposition for a selection of illustrative locations. The across-firm component is generally the largest, averaging 3.3% across locations, and is relatively stable geographically, ranging from 2.6% in Anchorage to 3.7% in Seattle among these locations and from 2.5%

## Figure 12: Productivity losses from misallocation




Productivity losses from misallocation, $1-\hat{Z}_k$, in percent, against the average markup, $\bar{\mu}_k$, for each of the 170 Economic Areas (EAs) with area proportional to EA employment. Losses range from 4.4% in greater New York to 14.2% in Honolulu and are strongly correlated with average markups. The star marks the free-trade limit, where markups and misallocation collapse to a single point.

## Table 6: Decomposition of productivity losses from misallocation

| | Across firms | Across sources | Across sectors | Total |
|---|---|---|---|---|
| New York | 3.1 | 0.8 | 0.5 | 4.4 |
| Los Angeles | 3.4 | 0.9 | 0.8 | 5.1 |
| Minneapolis | 3.5 | 2.0 | 1.0 | 6.4 |
| Phoenix | 3.5 | 2.3 | 1.2 | 6.9 |
| Seattle | 3.7 | 2.4 | 1.6 | 7.5 |
| Anchorage | 2.6 | 3.7 | 5.8 | 11.9 |
| Honolulu | 3.0 | 4.4 | 6.9 | 14.2 |
| Average | 3.3 | 1.6 | 0.7 | 5.5 |

Productivity losses from misallocation, $1-\hat{Z}_k$, in percent, for a selection of Economic Areas (EAs). These losses are decomposed into three components: *across-firms*, capturing misallocation across firms within a source location; *across sources* serving a given destination; and *across sectors* within a destination. The average is employment-weighted and taken across all 170 EAs. As shown in Appendix C.3, the three components sum to the total up to an interaction term ($<0.1$pp) between the across-firms and across-sources components.

to 3.8% across all EAs. By contrast the across-source and across-sector components are generally smaller, 1.6% and 0.7% respectively, but much more variable across locations. Across all EAs the across-source component varies by a factor of 5 and the across-sector component by a factor of more than 40. These two components 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 3.3% 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.

## 6.3 *Aggregating welfare costs across locations*

The welfare costs of markups are very unevenly distributed across locations, ranging from 0.9% in greater New York to 21.5% in Honolulu. How should we think about aggregating these costs across locations? One approach would be to assume a *social welfare function* and along with it a set of weights for each location. In recent work, Baqaee and Burstein (2025) propose a measure of aggregate welfare that avoids taking a stand on these weights. Applied to our welfare costs of markups calculation, their measure can be obtained as follows. Let $C_k$ denote consumption at location $k$ in our benchmark economy with markups and let $C_k^*$ denote *any allocation without markups that is feasible given the economy's resources*, allowing lump-sum transfers across locations. Then define

$$A^* := \max \{ \, A > 0 \; : \; \text{there exists } C_k^* \geq A\, C_k \; \text{ for all } k \, \} \tag{48}$$

That is, $A^*$ is the largest *common* consumption-equivalent gain that could be delivered to every location simultaneously. Equivalently, $1 - 1/A^*$ is the fraction of the economy's resources that could be saved while still leaving every location weakly better off. We compute $A^*$ by solving jointly for the wage vector and the scalar $A^*$ that support the no-markup allocation, $C_k^* = A^*\, C_k$, subject to labor market clearing for each location.

**Aggregate welfare costs.** For our benchmark economy we find that $A^* = 1.058$. We could eliminate markups and make *every* location better off by 5.8% in consumption-equivalent terms. Equivalently, the amount of aggregate productivity lost due to the markup distortions, $1 - 1/A^*$, is 5.5%. Interestingly, the common welfare cost of 5.8% is quite close to the 5.7% employment-weighted average cost reported in Table 5 above. The possibility of lump-sum transfers across locations adds only about 0.1 percentage points. In this sense, our overall welfare costs are not an artifact of how locations are weighted. Moreover, the aggregate productivity loss $1 - 1/A^*$ of 5.5% is the same, to one decimal place, as the employment-weighted average misallocation reported in Table 6 above.

These two measures are closely related. Under our assumption on the distribution of profits, each location's expenditure share equals its labor income share, and the compensated allocation must satisfy aggregate budget balance without profits. Together these imply that, to a first-order approximation, $1 - 1/A^*$ is a weighted average of the location-level misallocation losses $1 - \hat{Z}_k$ with weights that differ from employment shares only because of markup dispersion. Because the locations with the largest misallocation losses have very low employment shares, this average is robust to alternative weighting schemes — employment shares, expenditure shares, and markup-adjusted shares all give an overall productivity loss of between 5.51% and 5.53%.

# 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 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 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 an 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

$$\boldsymbol{v} = (v_1, v_2, ..., v_J) \tag{49}$$

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

$$\text{Prob}[v_j \leq v] = \exp(-v^{-\sigma}), \qquad \sigma > 0 \tag{50}$$

with tail parameter $\sigma$. 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

$$P_j c_j(\boldsymbol{v}) = (1+\bar{\pi}) W_j E_j \tag{51}$$

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

$$u_j(\boldsymbol{v}) = A_j v_j(\boldsymbol{v}) c_j(\boldsymbol{v}) = A_j v_j(\boldsymbol{v})(1+\bar{\pi})\frac{W_j}{P_j} E_j \tag{52}$$

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

$$u(\boldsymbol{v}) = \max_{j=1,\ldots,J} u_j(\boldsymbol{v}) \tag{53}$$

**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

$$L_j = \frac{\left(A_j E_j \frac{W_j}{P_j}\right)^\sigma}{\sum_k \left(A_k E_k \frac{W_k}{P_k}\right)^\sigma} \bar{L} \tag{54}$$

In short we have a labor supply curve, increasing in the real wage $W_j/P_j$ for each location $j$, with elasticity given by the Fréchet tail parameter $\sigma$. Profit income per location cancels out because of our assumption that locations receive profit income in proportion to labor income. A higher $\sigma$ 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

$$E_j L_j = \frac{\left(A_j E_j \frac{W_j}{P_j}\right)^\sigma}{\sum_k \left(A_k E_k \frac{W_k}{P_k}\right)^\sigma} \bar{L} = \int_0^1 \sum_{k=1}^{J} \sum_{i=1}^{N(s)} l_{ijk}(s)\, ds \tag{55}$$

**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

Table 7: Eliminating markups: labor mobility

| | Immobile | Mobile | | |
|---|---|---|---|---|
| Across locations | $C_j/L_j$ | $C_j/L_j$ | $C_j$ | $L_j$ |
| p10 | 3.6 | 3.7 | −0.4 | −3.9 |
| p25 | 3.9 | 4.1 | 0.9 | −3.1 |
| p50 | 5.6 | 5.6 | 5.4 | −0.3 |
| p75 | 6.9 | 6.9 | 9.3 | 2.0 |
| p90 | 9.1 | 8.9 | 15.6 | 6.1 |
| Average | 5.7 | 5.8 | 6.0 | 0.0 |

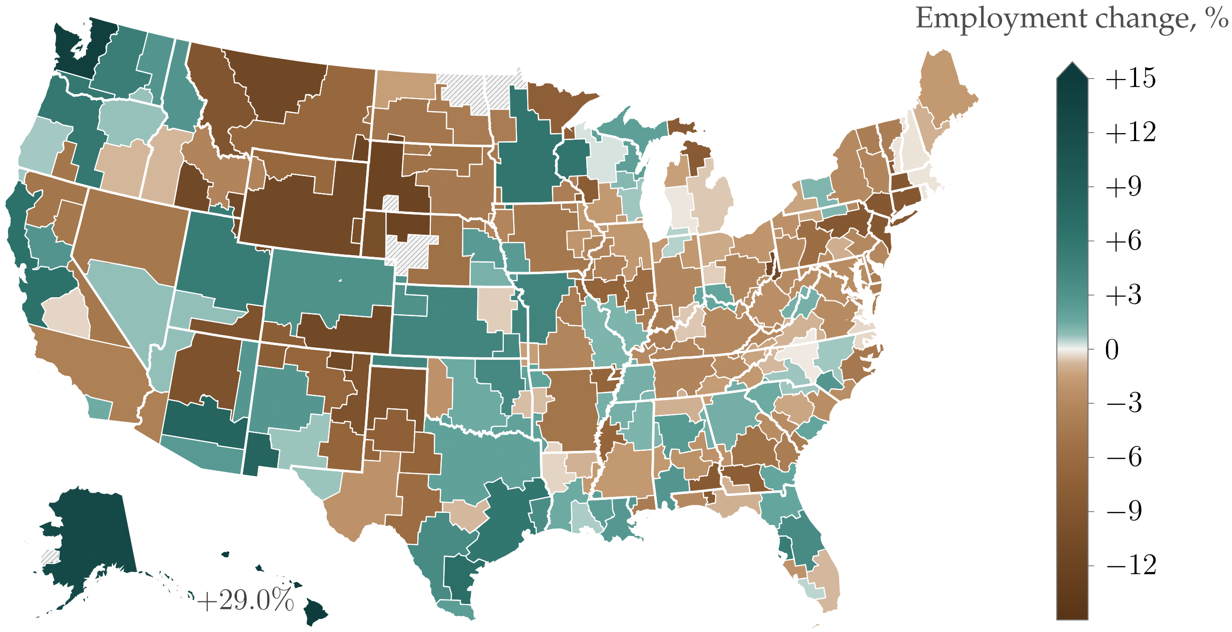
The distribution of consumption gains from eliminating markups across the 170 Economic Areas (EAs), in percent, in the benchmark model with immobile labor and in the model with labor mobility. For the model with labor mobility we report gains in consumption per worker $C_j/L_j$, in aggregate consumption $C_j$, and the change in employment $L_j$. Percentiles and averages are weighted by benchmark employment $L_j$. Total employment $\bar{L}$ is fixed, so the average employment change is zero by construction and these flows redistribute workers across locations.

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 7 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 the percentage changes in consumption per worker, $C_j/L_j$, in aggregate consumption, $C_j$, and in employment $L_j$. 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 the model with mobile labor and 5.7% in 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 even smaller effects.

That said, the model with labor mobility implies more substantial differences in the changes in aggregate consumption and employment across locations. For example, the p75 gain in consumption per worker is the same, 6.9%, in both models. But with mobile labor, that 6.9% gain corresponds to a 9.3% gain in aggregate consumption with a 2.0% increase in employment. Likewise, the p90 gain in consumption per worker of 8.9% with mobile labor corresponds to a large 15.6% gain in aggregate consumption with a 6.1% increase in employment. The locations which gain the most from the elimination of markups see

Figure 13: Percentage labor flows from eliminating markups




Percentage change in employment across the 170 Economic Areas (EAs) from eliminating markups, in the model with mobile labor. Changes range from an 11.9% decline in Rapid City to a 29.0% increase in Honolulu. Locations gaining workers are shown in teal and those losing workers in brown. Aggregate labor $\bar{L}$ is fixed, so these flows redistribute workers across locations and net to zero. The locations that gain most from eliminating markups receive large inflows, drawing on labor from locations that shrink. County colors show the value of the EA containing each county. Hatched counties are not matched to an EA.

large increases in employment: employment rises by 29.0% in Honolulu, 14.5% in Seattle and 12.6% in Anchorage. Intuitively, there are large labor inflows to locations which are growing substantially. At the other end of the spectrum, the p10 gain in consumption per worker of 3.7% with mobile labor masks a 0.4% *decrease* in aggregate consumption and a 3.9% decrease in employment. Employment falls by 8.6% in greater New York and by 11.9% in Rapid City, the largest decline of any location. Locations at this end of the distribution 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 13 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 market power in a quantitative model with multi-establishment firms, oligopolistic competition, and endogenously variable markups. We calibrate our model to match US manufacturing data across 170 Economic Areas, placing each firm's establishments exactly where they appear in the data. We find that spatial frictions have large effects on competition. The reduction in effective competition due to spatial frictions is equivalent to removing a randomly chosen 90% of firms from each sector of an otherwise identical economy without geography. The average welfare costs of markups are large, 5.7% in consumption-equivalent terms, considerably higher than the 3.7% implied by an economy without geography, and very unevenly distributed — ranging from about 1% in greater New York to more than 20% in Honolulu.

Our analysis provides a constructive response to a well-known concern in the industrial organization literature, namely that concentration data are a poor guide to market power. Berry, Gaynor and Scott Morton (2019) summarize these concerns, arguing forcefully that markets are not observable and that geography is one of the reasons why — a sector with many small firms may be either a sector where those firms compete directly or one where each firm dominates separate geographic locations. Similarly, Eeckhout (2021) argues that, since the boundaries of a market cannot be observed, market structure has to be inferred through the lens of an economic model rather than read off a classification. Our analysis is in this spirit. We take each firm's establishment locations from the data and discipline spatial frictions using sector-level gravity estimates. Rather than assuming some classification of 'markets' — sector $\times$ location cells, say — we let the model determine, endogenously, who competes with whom in each location, across the entire macroeconomy.

The model gives us a measure of *local sales* concentration, the concentration of the goods that consumers in a given location actually buy. This is a different object from the *local production* concentration that can be computed from establishment-level data. In our model, local sales concentration is an excellent guide to market power, mapping one-to-one into markups. But local production concentration is much higher, varies enormously across locations, and tells us almost nothing about local sales concentration — and hence provides a genuinely *poor* guide to market power.

To keep our analysis tractable we have abstracted from a number of potentially important considerations. First, our analysis is confined to competition between domestic producers. But imports are a substantial share of US expenditure on manufacturing and foreign firms compete with domestic firms in the same destination markets. Because our framework treats sources symmetrically, foreign producers can, in principle, be incorporated as additional sources with their own trade costs — trade costs that need not be aligned with distance from domestic production centers.

Second, our analysis takes each firm's establishment locations as given, allowing us to represent the geographic richness of US economic activity, but at the cost of holding firm locations *fixed* in our counterfactuals. Does this matter? On the one hand, lower trade costs would allow a given firm to serve more markets from a smaller set of production locations, reducing the number of local producers. On the other, those same lower trade costs would expose each destination to a wider set of competing firms across the country.

These forces pull in offsetting directions. A natural direction for future research is to make the extensive margin endogenous and to quantify these effects.

Third, we study manufacturing, where the shipments data needed to calibrate trade costs are available. By treating manufactured goods as final consumption, we abstract from the compounding of markup distortions through production networks, as emphasized by Baqaee and Farhi (2020), networks which in our context would have both technological and geographic dimensions. More broadly, some services are highly geographically segmented, more so than any manufactured good, while other services are highly tradeable. Incorporating services would also require dealing with the vastly larger number of services firms relative to manufacturing and the computational challenges that poses.

Finally, we represent US geography with 170 Economic Areas and measure spatial frictions as a log-linear function of great-circle distance between their centroids. Both are approximations. Since we assume perfect integration *within* areas as large as greater Los Angeles and New York, finer spatial units would, if anything, increase measured segmentation and strengthen our conclusions. And physical distance is at best a proxy for the cost of moving goods along optimal routes over multi-modal transportation networks.

While these are all natural directions for future research, our analysis here is deliberately focused on a single basic mechanism. What disciplines a firm's market power is the competition it faces in the markets where it sells, and, in a country as large as the United States, for most goods those markets are not national. Once competition is measured where it actually happens, the burden of market power is both larger and more unevenly distributed than national accounting would suggest — with low burdens in large, well-connected metro areas where firms from everywhere compete for the same customers, and much heavier burdens in more distant places where goods arrive from a handful of suppliers. To abstract from geography is to assume that the competitive conditions of the largest, best-connected areas prevail everywhere. But for most of the country, they do not.

# 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.

**Arkolakis, Costas, Arnaud Costinot, Dave Donaldson, and Andrés Rodríguez-Clare**, "The Elusive Pro-Competitive Effects of Trade," *Review of Economic Studies*, January 2019, *86* (1), 46–80.

**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 R. and Ariel Burstein**, "Aggregate Productivity with Heterogeneous Agents," August 2025. NBER Working Paper 34176.

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

**Barnatchez, Keith, Leland D. Crane, and Ryan A. Decker**, "An Assessment of the National Establishment Time Series (NETS) Database," November 2017. Finance and Economics Discussion Series 2017-110, Board of Governors of the Federal Reserve System.

**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.

**Bernard, Andrew B., Jonathan Eaton, J. Bradford Jensen, and Samuel Kortum**, "Plants and Productivity in International Trade," *American Economic Review*, September 2003, *93* (4), 1268–1290.

**Berry, Steven, Martin Gaynor, and Fiona Scott Morton**, "Do Increasing Markups Matter? Lessons from Empirical Industrial Organization," *Journal of Economic Perspectives*, Summer 2019, *33* (3), 44–68.

**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.

**Eaton, Jonathan and Samuel Kortum**, "Technology, Geography, and Trade," *Econometrica*, September 2002, *70* (5), 1741–1779.

**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.

**Eeckhout, Jan**, "Comment on "Diverging Trends in National and Local Concentration"," *NBER Macroeconomics Annual*, 2021, *35*, 151–166.

**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.

**Martínez-San Román, Valeriano, Ingrid Mateo-Mantecón, and Rubén Sainz-González**, "Intra-national Home Bias: New Evidence from the United States Commodity Flow Survey," *Economics Letters*, February 2017, *151*, 4–9.

**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.

**Oberfield, Ezra, Esteban Rossi-Hansberg, Pierre-Daniel Sarte, and Nicholas Trachter**, "Plants in Space," *Journal of Political Economy*, March 2024, *132* (3), 867–909.

**Ramondo, Natalia, Andrés Rodríguez-Clare, and Milagro Saborío-Rodríguez**, "Trade, Domestic Frictions, and Scale Effects," *American Economic Review*, October 2016, *106* (10), 3159–3184.

**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*, 2021, 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



## A Data

In this appendix we document how we use four datasets to parameterize the model. We use Bureau of Economic Analysis (BEA) Economic Areas to define the model's physical geography and distances between locations, use the Commodity Flow Survey (CFS) to quantify sector-specific trade costs, use the National Establishment Time Series (NETS) to set which locations each firm operates in, and use County Business Patterns (CBP) to measure location-level employment and wages.

### *A.1 BEA: geography*

**Locations.** Our model locations are BEA Economic Areas (EAs). We assign each US county to an EA using the BEA county-to-EA concordance based on the 1999 EA definitions. We aggregate all county-level data to the EA level. We exclude EAs covering American Samoa, Guam and the Northern Mariana Islands, the Gulf of Mexico, Puerto Rico, and the US Virgin Islands. We include only EAs with positive manufacturing employment in the CBP. Two EAs, Grand Forks and North Platte, are excluded on this basis. Both report manufacturing establishments but no manufacturing employment in the CBP. This leaves us with 170 EAs.

**Distances.** We measure the distance $d_{jk}$ between any two EAs as the distance in kilometers between their geographic centroids, where each EA's centroid is computed from the boundaries of its constituent counties. 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 set all within-EA distances to one, so home production faces no trade cost.

## *A.2 CFS: trade costs and gravity*

**Trade costs.** Sector-specific trade costs are log-linear in physical distance

$$\ln \tau_{jk}(s) = \delta(s) \ln d_{jk} \tag{A.1}$$

where $d_{jk}$ is the normalized EA distance constructed above, common to all sectors, and $\delta(s)$ is a sector-specific elasticity that we calibrate.

**Trade cost elasticities.** 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 by indirect inference, so that the coefficients of gravity regressions run on simulated model data match the corresponding coefficients estimated from the data.

**CFS shipments.** We estimate sector-specific distance elasticities based on public-use microdata from the 2017 CFS. The unit of observation in the raw data is an individual shipment. For each shipment we observe a 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 use only the 2017 cross section, we do not deflate these flows. Origin-destination pairs with no shipments do not enter the aggregated data and are therefore absent from the regressions. We do not apply any additional trimming or winsorization to these flows.

**CFS locations and distances.** In our gravity regressions, the notion of geography is the CFS origin and destination *region*, which is coarser than the EAs used in the model. For most shipments these regions coincide with individual US *states*, though in a few cases 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 $d_{jk}^{\text{cfs}}$ as the great-circle distance in kilometers between the centroids of regions $j$ and $k$. Shipments with the same origin and destination region do not enter the gravity regressions.

Table A.1: Sector-level gravity coefficients

| NAICS | Sector | $\widehat{\beta}(s)$ | Standard error |
|---|---|---|---|
| 311 | Food | −2.215 | 0.095 |
| 312 | Beverage and tobacco products | −2.654 | 0.138 |
| 313 | Textile mills | −1.579 | 0.139 |
| 314 | Textile product mills | −2.377 | 0.124 |
| 315 | Apparel | −1.363 | 0.118 |
| 316 | Leather and allied products | −1.447 | 0.133 |
| 321 | Wood products | −3.398 | 0.135 |
| 322 | Paper | −2.343 | 0.120 |
| 323 | Printing and related support activities | −2.713 | 0.116 |
| 324 | Petroleum and coal products | −3.914 | 0.194 |
| 325 | Chemicals | −2.529 | 0.109 |
| 326 | Plastics and rubber products | −2.332 | 0.099 |
| 327 | Nonmetallic mineral products | −3.227 | 0.131 |
| 331 | Primary metals | −2.389 | 0.130 |
| 332 | Fabricated metal products | −2.677 | 0.103 |
| 333 | Machinery | −1.695 | 0.102 |
| 334 | Computer and electronic products | −1.072 | 0.109 |
| 335 | Electrical equipment and appliances | −1.355 | 0.109 |
| 336 | Transportation equipment | −1.832 | 0.136 |
| 337 | Furniture and related products | −2.434 | 0.107 |
| 339 | Miscellaneous manufacturing | −1.254 | 0.102 |

Estimated distance elasticities $\widehat{\beta}(s)$ from sector-level gravity regressions on 2017 Commodity Flow Survey shipments, as described in Appendix A.2. Each regression includes origin and destination fixed effects. All coefficients are significant at the 1% level. Goods with more negative coefficients are less easily traded across space. These estimates are the targets used to calibrate the trade cost elasticities $\delta(s)$. Figure 2 compares these estimates with their model counterparts.

**CFS gravity regressions.** For each three-digit manufacturing sector $s$, we estimate a standard gravity equation of the form

$$\ln x_{jk}^{\mathrm{cfs}}(s) = \alpha_j(s) + \alpha_k(s) + \beta(s) \ln d_{jk}^{\mathrm{cfs}} + \varepsilon_{jk}(s) \tag{A.2}$$

where $x_{jk}^{\mathrm{cfs}}(s)$ is the shipment value from origin region $j$ to destination region $k$ in sector $s$ and $d_{jk}^{\mathrm{cfs}}$ is the bilateral distance between regions $j$ and $k$ constructed above. Conceptually, the shipment value in the data corresponds to $\bar{p}_{jk}(s)\bar{y}_{jk}(s)$ in the model, as given in equation (41). The origin and destination fixed effects $\alpha_j(s)$ and $\alpha_k(s)$ absorb all location-specific determinants of trade. The slopes $\beta(s)$ discipline the distance-sensitivity of iceberg trade costs by sector in our quantitative model. We report our estimates of these gravity coefficients in Table A.1.

## A.3 NETS: firm-establishment operations

We match the geographic distribution of firm-establishment operations over the 170 EAs. We use a 2014 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 Federal Information Processing Series (FIPS) code, six-digit NAICS codes, employment, and total sales. We impose the following filters in constructing our sample of manufacturing establishments:

- We drop establishments located in US 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 (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 EA using the county-to-EA mapping described in Appendix A.1 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. Producing locations are also highly concentrated across EAs: the 10th, 50th, and 90th percentiles of the EA distribution are 158, 582, and 2,926 producing locations, respectively. The five largest EAs together host 67,344 producing locations (27.8% of the total) and the largest 25 EAs host 149,247 (61.6%).

### A.4 CBP: employment and wages by location

We construct EA-level employment and wages from CBP data on manufacturing employment and wage payments. We use 2012 CBP data on county-level manufacturing employment and manufacturing wage bills mapped to EAs using the county-to-EA mapping described in Appendix A.1 above. This gives us total manufacturing employment and total manufacturing wage payments for each EA. From the EA-level manufacturing employment and wage bill we then construct:

- The employment level $L_k$ in location $k$, given by CBP manufacturing employment, normalized so that average employment across locations is one.
- The average wage per worker $W_k^{\text{cbp}}$ in location $k$, given by the CBP manufacturing wage bill divided by CBP manufacturing employment in that EA, again normalized so that the average wage per worker across locations is one.

In the model, $L_k$ is the number of workers in location $k$. Each worker at that location supplies $E_k$ *efficiency units* of labor and is paid the wage $W_k$ per efficiency unit, so that a worker in location $k$ earns $W_k E_k$. We choose efficiency units $E_k$ so that, for our benchmark calibration, the model wage bill $W_k E_k L_k$ matches the observed CBP wage bill location-by-location, i.e., so that $W_k E_k = W_k^{\text{cbp}}$. Beyond dropping EAs with no manufacturing activity, we do not trim or winsorize wages or employment when constructing EA-level aggregates.

Across the 170 EAs, 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.

## B Computation

The quantitative model has 363 sectors, 219,365 firms, and matches the geographic footprint of firms derived from 270,141 establishments over 170 locations. Each firm can serve every destination from the locations in which it operates. Thus a candidate wage vector changes marginal costs and potentially about 37 million firm-destination markups, which in turn change trade flows, price indices, and labor demand in all 170 locations.

We solve the model in two steps, starting with (i) an *inner solve*, which finds firm-level markups $\mu_{ik}(s)$ conditional on location-level wages $W_k$, and then (ii) an *outer solve* which finds location-level wages $W_k$ by solving a system of local labor market clearing conditions (up to a choice of numeraire). The inner solve is straightforward. The outer solve involves the solution of a 170-dimensional system of nonlinear equations via Newton methods. In practice, we obtain a good initial condition for the Newton procedure by first solving an *auxiliary model* where aggregate profits are distributed on an equal per-worker basis. With this uniform distribution of profits, the outer solve reduces to a system of homogeneous linear equations. The relative wages which solve this system of linear equations are then used to initialize the full nonlinear Newton procedure for our benchmark model where profits are distributed in proportion to a location's wage income.

## B.1 Inner solve: firm-level markups

For the inner solve we take as given location-level wages $W_k$ and can thus construct the firm-and-destination specific marginal costs $\mathsf{MC}_{ik}(s)$ from (20) above.

**Preliminaries.** Let

$$a_{ik}(s) := \mathsf{MC}_{ik}(s)^{1-\gamma} \tag{B.1}$$

denote a firm's *effective competitiveness*, inversely related to marginal cost since $\gamma > 1$, and note that we can then write a firm's market share from (26) as

$$\omega_{ik}(s) = \frac{\mu_{ik}(s)^{1-\gamma}\, a_{ik}(s)}{\sum_{i=1}^{N(s)} \mu_{ik}(s)^{1-\gamma}\, a_{ik}(s)} = \mu_{ik}(s)^{1-\gamma} \frac{a_{ik}(s)}{D_k(s)} \tag{B.2}$$

where $D_k(s)$ denotes the destination-specific denominator

$$D_k(s) = \sum_{i=1}^{N(s)} \mu_{ik}(s)^{1-\gamma}\, a_{ik}(s) \tag{B.3}$$

**Solving for firm-level markups.** Now let $x_{ik}(s) := a_{ik}(s)/D_k(s)$ denote firm-level effective competitiveness relative to this common denominator so that we can write the condition (25) linking a firm's sales share and its markup

$$\frac{1}{\mu_{ik}(s)} = \frac{\gamma - 1}{\gamma} - \left(\frac{1}{\theta} - \frac{1}{\gamma}\right) \mu_{ik}(s)^{1-\gamma}\; x_{ik}(s) \tag{B.4}$$

Let $\mu(x)$ denote the markup function mapping $x_{ik}(s)$ to $\mu_{ik}(s)$ implied by (B.4) and note that this function is strictly increasing, since $\gamma > \theta > 1$, and that its inverse $x(\mu)$ is available in closed form. In practice we compute $\mu(x)$ once on a dense grid and interpolate.

Given this markup function $\mu(x)$ we then pin down the denominator terms $D_k(s)$ by solving, for each destination $k$ and sector $s$, the scalar equation

$$\sum_{i=1}^{N(s)} \mu\left(\frac{a_{ik}(s)}{D_k(s)}\right)^{1-\gamma} \left(\frac{a_{ik}(s)}{D_k(s)}\right) = 1 \tag{B.5}$$

which simply imposes that sales shares sum to one for each $k$ and $s$. With these values for $D_k(s)$ in hand we then have $x_{ik}(s) = a_{ik}(s)/D_k(s)$ and the corresponding markups $\mu_{ik}(s) = \mu(x_{ik}(s))$ for each firm $i$, destination $k$, and sector $s$. Note that solving for $D_k(s)$ is separable across locations $k$ and sectors $s$ and depends only on the number of firms $N(s)$ and the collection of $a_{ik}(s)$ *within* a given $k, s$ cell. Interactions across locations are implicitly summarized by the wages $W_k$ that this inner solve takes as given.

## B.2 Outer solve: local labor market clearing

In the outer solve, we find the location-specific wages $W_k$ that clear local labor markets. More precisely, since the system of labor market clearing conditions is homogeneous of degree zero in wages, we find $J - 1 = 169$ *relative wages*, together with one additional unknown that depends on how aggregate profits are distributed across locations. To solve this 170-dimensional system of nonlinear equations with Newton methods, we first solve an *auxiliary model* where aggregate profits are distributed *uniformly* — on an equal per-worker basis.

**Labor market clearing.** Let $\Lambda_{jk}$ denote the share of expenditure at destination $k$ that is paid as labor income in source location $j$

$$\Lambda_{jk} := \int_0^1 \frac{\tau_{jk}(s)\, W_j}{\bar{z}_{jk}(s)\, P_k(s)} \left(\frac{P_k(s)}{P_k}\right)^{1-\theta} ds \tag{B.6}$$

where $\tau_{jk}(s)W_j/\bar{z}_{jk}(s)$ is the unit cost of delivering the bilateral composite from $j$ to $k$ and $(P_k(s)/P_k)^{1-\theta}$ is the aggregate share of expenditure at destination $k$ spent on sector $s$. Adding up labor demand across firms and sectors, the labor market clearing condition in each location $j$ can then be written

$$W_j E_j L_j = \sum_{k=1}^{J} \Lambda_{jk}\, P_k C_k \tag{B.7}$$

which equates the value of labor supplied in location $j$ to the value of labor demanded from $j$ by all destinations $k$.

**Auxiliary model: uniform profit rule.** In our benchmark model, profits in each location are proportional to local labor income, $\Pi_k = \bar{\pi} W_k E_k L_k$ for some constant $\bar{\pi} \geq 0$, to be determined. In our *auxiliary model*, profits are instead distributed on an equal per-worker basis, $\Pi_k = \hat{\pi} L_k$ for some constant $\hat{\pi} \geq 0$, so that expenditure in location $k$ is $P_k C_k = (W_k E_k + \hat{\pi}) L_k$. Then let $\bar{\mu}_k$ denote the aggregate markup in location $k$. Aggregate profits are $\sum_k (1 - \bar{\mu}_k^{-1}) P_k C_k$, so aggregate budget balance implies

$$\hat{\pi} = \frac{\sum_{k=1}^{J} \left(1 - \bar{\mu}_k^{-1}\right) W_k E_k L_k}{\sum_{k=1}^{J} \bar{\mu}_k^{-1} L_k} \tag{B.8}$$

**Auxiliary model: linear system.** The advantage of this profit rule is computational. Since the denominator of (B.8) does not involve wages, the solution for $\hat{\pi}$ is *linear* in $\boldsymbol{W}$. Because of this, we can substitute (B.8) and the budget constraint into (B.7) and divide through by $E_j L_j$, to write the labor market clearing conditions as a homogeneous system of linear equations

$$\boldsymbol{B}\boldsymbol{W} = \boldsymbol{0}, \qquad \boldsymbol{W} = (W_1, \ldots, W_J)^\top \tag{B.9}$$

where the elements of the $J \times J$ matrix $\boldsymbol{B}$ are given by

$$b_{jk} := \delta_{jk} - \Lambda_{jk} \frac{E_k L_k}{E_j L_j} - \phi_j \left(1 - \bar{\mu}_k^{-1}\right) \frac{E_k L_k}{E_j L_j}, \qquad \phi_j := \frac{\sum_{k=1}^{J} \Lambda_{jk} L_k}{\sum_{k=1}^{J} \bar{\mu}_k^{-1} L_k} \tag{B.10}$$

and $\delta_{jk}$ is the Kronecker delta. By Walras' Law $\boldsymbol{B}$ has rank $J-1$, so its null space is one-dimensional and pins down relative wages. We compute the null vector of $\boldsymbol{B}$, normalize the wage numeraire, then return to the inner solve, recomputing marginal costs, solving the scalar $D_k(s)$ problems, and constructing a new matrix $\boldsymbol{B}$. Iterating between these two inexpensive steps produces an equilibrium wage vector for the auxiliary economy.

**Benchmark model: proportional profit rule.** For our benchmark profit rule, $\Pi_k = \bar{\pi} W_k E_k L_k$, expenditure in location $k$ is $P_k C_k = (1 + \bar{\pi}) W_k E_k L_k$ and every location has the same profit share of income. Aggregate budget balance now implies

$$\bar{\pi} = \frac{\sum_{k=1}^{J} \left(1 - \bar{\mu}_k^{-1}\right) W_k E_k L_k}{\sum_{k=1}^{J} \bar{\mu}_k^{-1} W_k E_k L_k} \tag{B.11}$$

Although similar in appearance to (B.8), the denominator now involves wages $W_k$ and because of this we are left with a genuinely nonlinear system of equations.

**Benchmark model: nonlinear system.** Substituting the budget constraint into (B.7), local labor market clearing can be written

$$W_j E_j L_j = (1 + \bar{\pi}) \sum_{k=1}^{J} \Lambda_{jk} W_k E_k L_k, \qquad j = 1, \ldots, J \tag{B.12}$$

By Walras' Law one of these conditions is redundant. The equilibrium system therefore consists of $J-1=169$ independent labor market clearing conditions in *relative wages* — plus one additional condition, (B.11), which determines $\bar{\pi}$. All together, then, we have a 170-dimensional nonlinear system.

We initialize at the auxiliary solution and solve with a Newton-type method. Each function evaluation requires a complete inner solve: we recompute the marginal costs $\mathsf{MC}_{ik}(s)$, solve the scalar root-finding problems for $D_k(s)$ for each of the $170 \times 363 = 61{,}710$ destination $k$ sector $s$ cells, recover market shares and markups, and then recompute price indices, the bilateral shares $\Lambda_{jk}$, and labor demand in every location.

**Efficiency units and counterfactuals.** The location-specific efficiency endowments $E_k$ are determined jointly with the benchmark equilibrium, updated at each iteration as $E_k = W_k^{\text{cbp}}/W_k$ so that the model reproduces the observed manufacturing wage bill in each location, as described in Appendix A.4. In all subsequent counterfactuals we hold these $E_k$ fixed at their benchmark values and solve the same system for the new equilibrium wages, initializing at the benchmark wage vector.

**Aggregate markup.** Also note that from equation (B.11) we can recover the result given in Section 2 above, that $1+\bar{\pi}$ is the aggregate markup $\bar{\mu}$. To see this, add $\sum_k \bar{\mu}_k^{-1} W_k E_k L_k$ to both the numerator and the denominator of (B.11) to get

$$1+\bar{\pi} = \frac{\sum_{k=1}^{J} W_k E_k L_k}{\sum_{k=1}^{J} \bar{\mu}_k^{-1} W_k E_k L_k} \tag{B.13}$$

Under our benchmark profit rule, expenditure is $P_k C_k = (1+\bar{\pi}) W_k E_k L_k$, so we can write

$$1+\bar{\pi} = \frac{\sum_{k=1}^{J} P_k C_k}{\sum_{k=1}^{J} \bar{\mu}_k^{-1} P_k C_k} = \left( \sum_{k=1}^{J} \bar{\mu}_k^{-1} \, \frac{P_k C_k}{\sum_k P_k C_k} \right)^{-1} = \bar{\mu} \tag{B.14}$$

since $\bar{\mu}$ is the sales-weighted harmonic average of the location-level markups $\bar{\mu}_k$.

# C Aggregate Productivity and Misallocation

In this appendix we provide further details on our measures of aggregate productivity and misallocation and show how to decompose the losses due to misallocation into components capturing markup dispersion across firms, across source locations, and across sectors.

## *C.1 Aggregate productivity*

**How markup dispersion reduces productivity.** As shown in Section 2.2, markup dispersion reduces productivity in three distinct layers. First, for each given sector $s$, source location $j$ and destination $k$, dispersion in firm-level markups $\mu_{ik}(s)$ reduces the node-level productivities $\bar{z}_{jk}(s)$ that make up the *productivity network*

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

Second, for each given sector $s$ and destination $k$, dispersion across $j$ in node-level markups $\bar{\mu}_{jk}(s)$ around $\bar{\mu}_k(s)$ reduces sector-level productivity

$$\bar{Z}_k(s) = \left( \sum_{j=1}^{J} \left( \frac{\bar{\mu}_{jk}(s)}{\bar{\mu}_k(s)} \right)^{-\gamma} \left( \frac{\bar{z}_{jk}(s)}{\tau_{jk}(s)} \right)^{\gamma-1} \left( \frac{W_j}{\bar{W}_k(s)} \right)^{-\gamma} \right)^{\frac{1}{\gamma-1}} \tag{C.2}$$

Third, for each given destination $k$, dispersion in sector-level markups $\bar{\mu}_k(s)$ around the location-level markup $\bar{\mu}_k$ reduces destination-level productivity

$$\bar{Z}_k = \left( \int_0^1 \left( \frac{\bar{\mu}_k(s)}{\bar{\mu}_k} \right)^{-\theta} \bar{Z}_k(s)^{\theta-1} \left( \frac{\bar{W}_k(s)}{\bar{W}_k} \right)^{-\theta} ds \right)^{\frac{1}{\theta-1}} \tag{C.3}$$

Below we give a decomposition of the losses due to misallocation that switches off each of these layers in turn.

**Interpreting these productivity indexes.** Recall from Section 2.2 that the wage indexes $\bar{W}_k(s)$ and $\bar{W}_k$ are the average cost of the labor used to produce for destination $k$, satisfying $\bar{W}_k(s)\bar{L}_k(s) = \sum_j W_j \bar{l}_{jk}(s)$ with $\bar{L}_k(s) = \sum_j \bar{l}_{jk}(s)$, and $\bar{W}_k$ implicitly defined by $P_k = \bar{\mu}_k \bar{W}_k / \bar{Z}_k$. With this notation we can write

$$C_k(s) = \bar{Z}_k(s)\bar{L}_k(s), \qquad \text{and} \qquad C_k = \bar{Z}_k \bar{L}_k \tag{C.4}$$

But note that the labor index $\bar{L}_k$ is *not* the exogenous labor supply $L_k$ at location $k$, for two reasons. First, $\bar{L}_k$ is *all* the labor used to produce for destination $k$, not the labor supplied at location $k$ — only under autarky would these coincide. Second, the labor index $\bar{L}_k$ is in *efficiency units*, so that, since all labor is used to produce for some destination, $\sum_k \bar{L}_k = \sum_k E_k L_k$. In short, $\bar{Z}_k$ is output per efficiency unit of labor used to produce for $k$, regardless of where that labor is located.

**Efficient productivity.** If markup variation is eliminated, then the node-level productivity index (C.1) simplifies to the familiar CES index $\bar{z}_j^*(s)$ reported in equation (43). This efficient level of productivity is independent of the destination $k$ and independent of wages. Building up from these efficient nodes, $\bar{z}_j^*(s)$, the efficient sector-level and location-level productivity indexes are obtained by setting $\bar{\mu}_{jk}(s)/\bar{\mu}_k = 1$ in the sector-level index (C.2) and setting $\bar{\mu}_k/\bar{\mu}_k = 1$ in the location-level index (C.3), both evaluated at the wages

$W_j^*$ for this efficient economy. We write these indexes

$$\bar{Z}_k^{\text{efficient}}(s) = \left( \sum_{j=1}^{J} \left( \frac{\bar{z}_j^*(s)}{\tau_{jk}(s)} \right)^{\gamma-1} \left( \frac{W_j^*}{\bar{W}_k^*(s)} \right)^{-\gamma} \right)^{\frac{1}{\gamma-1}} \tag{C.5}$$

and

$$\bar{Z}_k^{\text{efficient}} = \left( \int_0^1 \bar{Z}_k^{\text{efficient}}(s)^{\theta-1} \left( \frac{\bar{W}_k^*(s)}{\bar{W}_k^*} \right)^{-\theta} ds \right)^{\frac{1}{\theta-1}} \tag{C.6}$$

Eliminating markup dispersion has both (i) a *direct effect* on productivity through the elimination of the three wedges $\mu_{ik}(s)/\bar{\mu}_{jk}(s)$, $\bar{\mu}_{jk}(s)/\bar{\mu}_k(s)$, and $\bar{\mu}_k(s)/\bar{\mu}_k$, and (ii) an *indirect effect* on productivity through the changes in relative wages from their benchmark levels $W_j/\bar{W}_k(s)$ and $\bar{W}_k(s)/\bar{W}_k$ to their efficient levels $W_j^*/\bar{W}_k^*(s)$ and $\bar{W}_k^*(s)/\bar{W}_k^*$. In our benchmark model, labor supply is exogenous at each location. But the elimination of markup variation induces changes in labor demand at the firm, sector, and location levels which in turn induces changes in equilibrium wages across locations.

**Isolating the direct effect.** To isolate the direct effect of markup dispersion on productivity we consider a 'partial equilibrium' case where we eliminate markup dispersion but evaluate outcomes at the benchmark relative wages $W_j/\bar{W}_k(s)$ and $\bar{W}_k(s)/\bar{W}_k$, giving

$$\bar{Z}_k^{\text{partial}}(s) = \left( \sum_{j=1}^{J} \left( \frac{\bar{z}_j^*(s)}{\tau_{jk}(s)} \right)^{\gamma-1} \left( \frac{W_j}{\bar{W}_k(s)} \right)^{-\gamma} \right)^{\frac{1}{\gamma-1}} \tag{C.7}$$

and

$$\bar{Z}_k^{\text{partial}} = \left( \int_0^1 \bar{Z}_k^{\text{partial}}(s)^{\theta-1} \left( \frac{\bar{W}_k(s)}{\bar{W}_k} \right)^{-\theta} ds \right)^{\frac{1}{\theta-1}} \tag{C.8}$$

where, because they are independent of wages, the node-level productivities $\bar{z}_j^*(s)$ are the same as in the fully efficient economy.

## *C.2 Misallocation*

Our preferred measure of misallocation isolates the direct effect of markup dispersion on productivity abstracting from relative wage changes. To that end we define

$$\hat{Z}_k(s) := \frac{\bar{Z}_k(s)}{\bar{Z}_k^{\text{partial}}(s)} \leq 1 \tag{C.9}$$

and

$$\hat{Z}_k := \frac{\bar{Z}_k}{\bar{Z}_k^{\text{partial}}} \leq 1 \tag{C.10}$$

where $\bar{Z}_k(s)$ and $\bar{Z}_k$ are the sector-level and location-level productivity indexes in the benchmark equilibrium with markup dispersion, given in (C.2)-(C.3).

**Why we hold relative wages fixed.** Does holding relative wages fixed matter? For some locations, yes. Anchorage, for example, switches from having the second most misallocation to having the least misallocation when ranked relative to full efficiency. For Anchorage, the relative wage movements implied by the reallocation of labor demand across the US economy swamp the direct productivity losses due to markup dispersion. That said, overall our preferred measure of misallocation implies similar average losses to the full general equilibrium calculations. For example, as reported in Section 6.3, across locations the employment-weighted average of $1-\hat{Z}_k$, is $5.51\%$, the same, to two decimal places, as the aggregate resource saving $1-1/A^* = 5.51\%$ implied by (48), which is a full general equilibrium object in which wages adjust to clear all 170 local labor markets.

## C.3 Misallocation decomposition

In this section we decompose the losses due to misallocation into components capturing markup dispersion across firms, across source locations, and across sectors.

**Multiplicative representation.** We first write sector-level misallocation

$$\hat{Z}_k(s) := \frac{\bar{Z}_k(s)}{\bar{Z}_k^{\text{partial}}(s)} = \underbrace{\frac{\bar{Z}_k(s)}{\bar{Z}_k^{\text{inter}}(s)}}_{=:\hat{A}_k(s)} \times \underbrace{\frac{\bar{Z}_k^{\text{inter}}(s)}{\bar{Z}_k^{\text{partial}}(s)}}_{=:\hat{B}_k(s)} \leq 1 \tag{C.11}$$

and define the *intermediate index*

$$\bar{Z}_k^{\text{inter}}(s) := \left(\sum_{j=1}^{J} \left(\frac{\bar{\mu}_{jk}(s)}{\bar{\mu}_k(s)}\right)^{-\gamma} \left(\frac{\bar{z}_j^*(s)}{\tau_{jk}(s)}\right)^{\gamma-1} \left(\frac{W_j}{\bar{W}_k(s)}\right)^{-\gamma}\right)^{\frac{1}{\gamma-1}} \tag{C.12}$$

which is the same as the sector-level index (C.2) except that the distorted productivity nodes $\bar{z}_{jk}$ are replaced by the (destination-independent) efficient productivity nodes $\bar{z}_j^*$.

The multiplicative representation $\hat{Z}_k(s) = \hat{A}_k(s)\hat{B}_k(s)$ decomposes sector-level misallocation into a first term $\hat{A}_k(s) := \bar{Z}_k(s)/\bar{Z}_k^{\text{inter}}(s)$ which isolates markup dispersion across firms $i$, within source locations $j$ and a second term $\hat{B}_k(s) := \bar{Z}_k^{\text{inter}}(s)/\bar{Z}_k^{\text{partial}}(s)$ which isolates markup dispersion across source locations $j$ serving destination $k$. We then need to aggregate across sectors $s$ to obtain our full three-way decomposition.

Table C.1: Distribution of productivity losses from misallocation

| | Across firms | Across sources | Across sectors | Total |
|---|---|---|---|---|
| p01 | 2.8 | 0.8 | 0.3 | 4.4 |
| p10 | 2.9 | 0.9 | 0.5 | 5.0 |
| p25 | 3.1 | 1.3 | 0.5 | 5.1 |
| p50 | 3.2 | 1.6 | 0.7 | 5.4 |
| p75 | 3.5 | 1.9 | 0.8 | 5.8 |
| p90 | 3.7 | 2.2 | 1.1 | 6.4 |
| p99 | 3.8 | 2.6 | 2.8 | 8.8 |
| Average | 3.3 | 1.6 | 0.7 | 5.5 |

The distribution of productivity losses from misallocation across the 170 Economic Areas (EAs), in percent, decomposed into three components defined in Appendix C.3. Percentiles and averages are weighted by employment. The three components sum to the total up to a small interaction term between the across-firms and across-sources components, which never exceeds 0.1 percentage points.

**Aggregating across sectors.** Let $\omega_k(s)$ denote the share of destination $k$ expenditure on sector $s$ and define the average *losses* implied by $\hat{A}_k(s)$ and $\hat{B}_k(s)$ according to

$$\hat{a}_k := \int_0^1 \big(1 - \hat{A}_k(s)\big)\, \omega_k(s)\, ds, \qquad \hat{b}_k := \int_0^1 \big(1 - \hat{B}_k(s)\big)\, \omega_k(s)\ ds \tag{C.13}$$

We then define the across-sector component

$$\hat{c}_k := \big(1 - \hat{Z}_k\big) - \int_0^1 \big(1 - \hat{Z}_k(s)\big)\, \omega_k(s)\, ds \tag{C.14}$$

With this definition, $\hat{c}_k$ is the difference between the location-level loss $1 - \hat{Z}_k$ and the average of the sector-level losses $1 - \hat{Z}_k(s)$ within that location. Adding up and collecting terms we can then write the losses due to misallocation as

$$1 - \hat{Z}_k = \hat{a}_k + \hat{b}_k + \hat{c}_k + \hat{d}_k \tag{C.15}$$

where the residual term $\hat{d}_k$ is given by interactions between the across-firm and across-source components averaged across sectors

$$\hat{d}_k = -\int_0^1 (1 - \hat{A}_k(s))(1 - \hat{B}_k(s))\, \omega_k(s)\, ds \tag{C.16}$$

In practice this residual is small. Across all 170 locations, $\max_k |\hat{d}_k| \leq 0.00084$, less than 0.1 percentage points.

**Distribution across locations** We report the across-firm component $\hat{a}_k$, the across-source location component $\hat{b}_k$, and the across-sector component $\hat{c}_k$ for a selection of locations in Table 6 above. We further report the distribution of these three components across all 170 locations in Table C.1. The across-firm component $\hat{a}_k$ varies relatively little across locations, ranging from 2.8% to 3.8% between the first and ninety-ninth percentiles. The across-source component $\hat{b}_k$ and the across-sector component $\hat{c}_k$ are much more dispersed. The right tail of the distribution of total losses is thin, only Anchorage and Honolulu have losses above 8%. Since both of these locations are low employment shares, the employment-weighted ninety-ninth percentile total loss, 8.8%, is substantially below the maximum loss, 14.2%.

# D Geographic Variation in Concentration and Markups

As shown in Section 4, local production concentration is a poor guide to local sales concentration and markups. So what does help explain these outcomes? Table D.1 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 D.1 reports for each selected location $k$ employment $L_k$, the average markup, $\bar{\mu}_k$, 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 $k$, so that home supply costs are $\mathsf{C}_{kk}(s) = \mathsf{C}_k(s)$ since $\tau_{kk}(s) = 1$. Then let $\mathsf{B}_k(s)$ denote the costs of the *best* outside source for $k$

$$\mathsf{B}_k(s) := \min_{j \neq k} \mathsf{C}_{jk}(s) \tag{D.1}$$

and let $j_k^*(s)$ denote the source location that achieves this minimum. For each location $k$ we then compute and report in Table D.1 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 now review outcomes for a number of illustrative 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 $\bar{\mu}_k = 1.22$ 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.

**San Francisco.** The greater San Francisco area is also large and productive, but in our model has higher sales concentration, with a sales HHI of 0.17, and higher markups, $\bar{\mu}_k = 1.28$. The import share is again relatively low, 0.26. What really makes San Francisco different to New York in our model is that domestic production is more concentrated and faces even less effective competition from outside sources. On average the best outside source has delivered costs 0.39 log points higher, compared to New York's 0.23. On average there are even fewer cost-realistic outside options, 0.143, some half that of New York, 0.284. Overall, greater San Francisco's higher sales concentration and markups reflect both more domestic production concentration and a lack of effective competition from imports.

**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 $\bar{\mu}_k = 1.31$ 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.

**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 $\bar{\mu}_k = 1.23$ 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.

**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 $\bar{\mu}_k = 1.35$, 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 five 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 contrasts between New York, San Francisco and Seattle, or between Scottsbluff and Honolulu, make clear, the finer details of each location's geography and production structure still play an important role in determining their individual location-level outcomes.

## Table D.1: Characteristics of selected locations

| | | Outcomes | | Home supply | | | Outside sources | | Best gap | |
|---|---|---|---|---|---|---|---|---|---|---|
| Location | $L_k$ | $\bar{\mu}_k$ | HHI | Coverage | Plants | Top 1 | Imports | Contest | Trade | Comp. |
| Los Angeles | 24.9 | 1.24 | 0.13 | 0.98 | 56.79 | 0.505 | 0.22 | 0.22 | 0.277 | 0.043 |
| New York | 12.3 | 1.22 | 0.12 | 0.98 | 61.02 | 0.511 | 0.36 | 0.28 | 0.270 | −0.044 |
| San Francisco | 7.84 | 1.28 | 0.17 | 0.95 | 24.98 | 0.598 | 0.26 | 0.14 | 0.306 | 0.084 |
| Minneapolis | 4.48 | 1.28 | 0.17 | 0.93 | 14.71 | 0.650 | 0.39 | 0.60 | 0.473 | −0.091 |
| Seattle | 3.24 | 1.31 | 0.19 | 0.91 | 10.21 | 0.657 | 0.37 | 0.41 | 0.426 | −0.049 |
| Phoenix | 2.71 | 1.28 | 0.17 | 0.86 | 8.61 | 0.680 | 0.50 | 1.11 | 0.372 | −0.107 |
| San Antonio | 0.641 | 1.26 | 0.16 | 0.76 | 3.85 | 0.766 | 0.77 | 5.11 | 0.336 | −0.342 |
| Honolulu | 0.298 | 1.35 | 0.23 | 0.55 | 1.74 | 0.786 | 0.62 | 0.92 | 0.983 | −0.249 |
| Anchorage | 0.289 | 1.29 | 0.18 | 0.36 | 0.99 | 0.782 | 0.77 | 0.34 | 0.906 | −0.240 |
| Redding | 0.0266 | 1.24 | 0.14 | 0.27 | 0.52 | 0.856 | 0.97 | 17.43 | 0.367 | −0.642 |
| Minot | 0.00995 | 1.23 | 0.12 | 0.15 | 0.24 | 0.897 | 0.97 | 20.07 | 0.498 | −0.560 |
| Scottsbluff | 0.00206 | 1.23 | 0.12 | 0.13 | 0.18 | 0.935 | 1.00 | 47.63 | 0.442 | −0.916 |

Characteristics of twelve selected Economic Areas (EAs), ordered by employment $L_k$, measured relative to the cross-EA average. Outcomes: $\bar{\mu}_k$ is the average markup and HHI the average local sales Herfindahl index. Home supply: Coverage is the fraction of sectors with home production, Plants the average number of production plants per sector, and Top 1 the average cost share of the lowest-cost home plant. Outside sources: Imports is average import penetration and Contest the average number of outside origins with delivered costs within 10% of home delivered cost. Best gap: Trade and Comp. are the two components of the log delivered cost advantage of the best outside source relative to home, $\ln \tau_{j^*k}$ and $\ln \mathsf{C}_{j^*} - \ln \mathsf{C}_k$, which sum to $\ln \mathsf{B}_k - \ln \mathsf{C}_k$. All statistics are averages across sectors, weighted by sales except Coverage, which is weighted by expenditure. Full EA names: Los Angeles-Riverside-Orange County, CA-AZ; New York-North New Jersey-Long Island, NY-NJ-CT-PA-MA-VT; San Francisco-Oakland-San Jose, CA; Minneapolis-St. Paul, MN-WI-IA; Seattle-Tacoma-Bremerton, WA; Phoenix-Mesa, AZ-NM; San Antonio, TX; Honolulu, HI; Anchorage, AK; Redding, CA-OR; Minot, ND; Scottsbluff, NE-WY.