Cognitive Hubs and Spatial Redistribution*

Esteban Rossi-Hansberg
University of Chicago

Pierre-Daniel Sarte
Federal Reserve Bank of Richmond

Felipe Schwartzman
Federal Reserve Bank of Richmond

December 15, 2023

Abstract

In the US, cognitive non-routine (CNR) occupations are disproportionately and increasingly represented in large cities. To study the allocation of workers across cities, we propose a quantitative spatial equilibrium model with multiple industries employing CNR and non-CNR workers. Productivity is city-industry-occupation specific and, as we estimate, partly determined by externalities that depend on local occupation shares and total employment. An optimal policy that benefits workers equally, incentivizes the formation of cognitive hubs in large cities. It also creates higher overall activity in small cities, greater industrial specialization in both the largest and smallest cities, and greater diversification in medium-sized cities.

1 Introduction

"*Most of what we know we learn from other people (...) most of it we get for free.*"
Robert E. Lucas Jr.[1]

Workers capable of doing the complicated cognitive non-routine (CNR) tasks required in a modern economy are scarce. Acquiring the expertise to work as an engineer, doctor, manager, lawyer, computer scientist, or researcher requires sustained effort and investments as well as individual ability. Where and how these workers are employed has significant implications for the overall efficiency and welfare of an economy. The marginal productivity of workers depends on opportunities to learn from and collaborate with similar worker types located nearby. Large cities with a large

*rossihansberg@uchicago.edu, pierre.sarte@rich.frb.org, felipe.schwartzman@rich.frb.org. We thank the editor and three anonymous referees for suggestions and comments that have greatly improved the paper. We also thank David Albouy, Arnaud Costinot, Morris Davis, Jonathan Dingel, Cecile Gaubert, and Edward Glaeser for useful comments and discussions, as well as participants in various seminars and workshops. Finally, we thank Nick Garvey, Mike Finnegan, Xiang Zhang, Daniel Ober-Reynolds, and Jackson Evert for excellent research assistance. The views expressed here are those of the authors and do not reflect those of the Federal Reserve Bank of Richmond or the Federal Reserve System.

[1] Lucas (1988), page 38.

fraction of workers in these scarce occupations offer learning and collaboration opportunities that enhance the productivity of other similar workers. However, the local abundance of these workers can also lower their marginal product. The equilibrium interaction of these forces then determines the spatial distribution of workers and, relatedly, the industrial composition of cities. Can the economy allocate scarce occupations in a way that improves the lives of all workers? Should optimal spatial policy reinforce or mitigate the currently polarized pattern of occupations across space? Our aim in this paper is to study the allocation of occupations and industries across cities in the US and to characterize the optimal spatial allocation and the policies to implement it.

The need for spatial policy follows directly from the presence of urban externalities. Externalities that enhance worker productivity in larger cities have been discussed, analyzed, and measured at least since Marshall (1920).[2] It is natural to hypothesize that production externalities depend on the occupational composition of cities. After all, cognitive non-routine work often requires close interactions between knowledgeable workers. Moreover, the industrial composition of cities makes this relationship between externalities and occupational composition particularly pertinent as CNR-intensive industries have become increasingly common, especially in large coastal cities. Estimating these externalities for different worker types and clarifying how they interact with local industrial structure forms a central part of our analysis.

The detailed quantitative assessment of the optimal spatial policy we propose requires a number of contributions. These fall along four main dimensions:

First, we develop a spatial equilibrium model with multiple industries and two occupations (CNR and non-CNR) as well as occupation-specific externalities. Multiple industries, input-output linkages, and costly trade are all key features of the environment since the demand for different occupations depends on the occupational intensity of the specific industries in each location. In addition, the industrial structure in each location depends on the supply of workers, their productivity, and the local availability of inputs produced by other industries. The framework also includes heterogeneous preferences for locations that act as a form of migration costs. Crucially, it features production externalities in each city that depend on the share of workers of different types and its total workforce.

Second, we derive the optimal spatial policy by obtaining the efficient allocation in this setup. We choose to study the efficient allocation that benefits both occupations equally. Implementing this allocation requires us to characterize particular transfers specific to locations and occupations.

Third, the details of optimal transfers require that we quantify the model and estimate the parameters that determine the endogenous component of city-industry-occupation specific productivity. Thus, we first recover productivity across locations, industries, and occupations such that the equilibrium of our model matches observed data. We then parameterize the relationship between productivity and the occupational composition and size of cities and estimate this equation using an instrumental variables approach. As proposed in the empirical literature (e.g., Card (2001) and Moretti (2004a)), we use past migration flows of particular immigrant groups and the location

[2] See Duranton and Puga (2004) for a review of the literature on agglomeration externalities.

of land-grant colleges as instruments. We also present results using model-implied instrumental variables. Our strategy yields robust results that corroborate previous work, though, as a contribution, estimated here directly from productivity measures recovered from the general equilibrium framework we use to study optimal policy. Our findings imply that the productivity of CNR and non-CNR workers depends similarly on city size. However, the productivity of CNR workers also depends strongly and significantly on the share of CNR workers. We find less evidence that the productivity of non-CNR workers depends on the composition of occupations.

Finally, informed by these findings, we compute the optimal allocation in the US economy and discuss its implementation using specific policy tools. We also highlight important quantitative aspects of this allocation through various counterfactual exercises.

Our findings advocate for a new approach to spatial policy. They indicate that the spatial allocation of workers and industries may be improved by reducing the size of large CNR-intensive cities while, at the same time, increasing their fraction of CNR workers. These 'cognitive hubs' take advantage of scarce CNR workers in the economy by clustering them to maximize externalities. Such clustering is further enhanced by the fact that CNR-intensive industries purchase disproportionately more of their inputs from other CNR-intensive industries. By contrast, in a counterfactual economy with homogeneous industry linkages, the optimal policy prescribes instead increasing the CNR population of some small cities as well. In summary, both the nature of externalities and of industrial linkages are essential to our result. Because of the externality, it is optimal to have workers spatially sort, but because of linkages, it is optimal for those cognitive hubs to be located in large cities.

The externalities associated with CNR workers that we estimate are substantial. In equilibrium, the average social value of CNR workers is 79 percent larger than their private value. However, the industrial makeup of cities, as well as their location, does impose limits on the creation of cognitive hubs in the optimal allocation. Some large cities, such as Miami or Las Vegas, remain largely populated by non-CNR workers since they are particularly productive in industries that employ CNR workers less intensively. Trade costs also play an important role in the optimal allocation, dispersing cognitive hubs across the country rather than concentrating them in a few coastal cities.

In order to increase the share of CNR workers while alleviating congestion in larger cities, the optimal policy provides non-CNR workers with an incentive to live in smaller cities with relatively more non-CNR-intensive industries. The end result is that the optimal policy helps the smallest cities grow in size by playing to their strengths and expanding industries in which a large share of their employment already resides. The corresponding growth of smaller cities also makes it possible for them to sustain more employment in non-tradable industries such as retail and construction, as well as in non-CNR-intensive industries such as accommodation. Hence, contrary to some previous literature and much of the public discourse, the economics of the problem suggest that with the appropriate transfers, small industrial cities in the US should strive to expand non-CNR-intensive industries and not try to become the next San Jose.

Naturally, implementing the optimal allocation requires a number of transfers and taxes that

depend on the location and occupation of workers. To internalize the effects of externalities, workers need to be incentivized to move to the cities where they are most productive. Thus, occupation and city-specific transfers are positively correlated with city size for CNR workers and negatively correlated with city size and the CNR share for non-CNR workers. Those city-occupation-specific transfers are supplemented by large base transfers, ensuring that non-CNR workers benefit equally from the optimal policy. In our analysis, the base transfers to non-CNR workers amount to $16,887 (in 2015 dollars), while CNR workers, who earn substantially more, end up paying a base transfer of $15,276. Finally, as described in Fajgelbaum and Gaubert (2020), spatial efficiency requires a flat wage tax on all individuals to correct for differences in the marginal utility of consumption generated by heterogeneous preferences for location. Therefore, once all incentives are taken into account, non-CNR workers receive on average $2,025, while CNR workers contribute on average $3,324, though there is considerable variation in net transfers across cities. Importantly, the incentive for CNR workers to earn more by locating in larger cities stands in contrast to typical phase-out features of actual transfer policies that impose an implicit tax on high earners.

A comparison of the current spatial equilibrium to that in 1980 reveals that the spatial allocation of workers has historically evolved towards that implied by the optimal policy (with current fundamentals). Specifically, since the 1980s, the relative number of CNR workers has not only increased nationally but also become increasingly concentrated in CNR-intensive hubs, many of which are large cities. This formation of cognitive hubs has taken place in parallel with a well-documented increase in wage inequality across space and occupations. Our quantitative framework implies that both processes were linked through local occupation-specific externalities. Our analysis also indicates that absent these spillovers, the spatial polarization of workers would have been greatly mitigated, and the welfare gains received by CNR workers would have been smaller than those of non-CNR workers.

Relationship to the Literature A substantial literature has pointed to increasing spatial concentration of skilled workers (Berry and Glaeser (2005), Diamond (2016), and Giannone (2017)), as well as increasing wage inequality across space and within cities (Baum-Snow and Pavan (2013), and Autor (2019)), with the skill premium increasing the most in large cities. Our paper speaks to the optimal policy reaction to those trends.

We focus on production externalities as a key driving force behind those spatial patterns. The estimation of those externalities is a central theme in urban and spatial economics. A robust finding is the existence of a relationship between city size and productivity (see Melo et al. (2009) for a meta-analysis). While we allow for such agglomeration externalities, our main focus is on externalities tied to the occupational composition of cities. This focus aligns with empirical evidence by Ellison et al. (2010) whereby industries with similar occupational make-up tend to be spatially proximate. Given the correlation between occupational types and skill levels, our findings of strong spillovers stemming from the occupational composition of cities mirror findings by Moretti (2004a; 2004b) regarding the local external effects of human capital.

There has been ample research on the extent of spatial misallocation in the US economy and the degree to which it corresponds to heterogeneity in taxation policy (or its local incidence), zoning laws, or other unspecified sources of distortions. Examples of papers in that vein are Albouy (2009), Desmet and Rossi-Hansberg (2013), Ossa (2015), Fajgelbaum et al. (2018), Colas and Hutchinson (2017), Hsieh and Moretti (2019) and, most recently Herkenhoff et al. (2018).

Our paper sheds light on place-based policies in that it highlights the optimal endogenous expansion of different industries in different locations. A summary of the related literature can be found in Neumark and Simpson (2015). Rather than evaluating exogenous policies, we derive the optimal allocation in a quantitative spatial model with local externalities. Our derivation of optimal policy thus generalizes that of Fajgelbaum and Gaubert (2020) in an environment with production linkages and where trade is differentially costly across industries. Introducing these observable features of the economy changes the implied optimal policy implications in an essential way. Instead of promoting concentrations of CNR workers in small and large cities, as in Fajgelbaum and Gaubert (2020), it now promotes the creation of 'cognitive hubs' in today's largest cities.[3].

We integrate industrial, occupational, and spatial heterogeneity in a single coherent framework. Other recent work that has emphasized the joint distribution of industrial and skill composition within the US are Hendricks (2011) and Brinkman (2014). As in Caliendo et al. (2017), we allow for trade costs, thus capturing an explicitly spatial dimension, but add to that framework by also allowing for occupational heterogeneity and local production externalities. Finally, on a more technical note, our paper adds to the rapidly expanding 'quantitative spatial economics' literature that uses general equilibrium models to address issues related to international, regional, and urban economics (see Redding and Rossi-Hansberg (2017) for a review of this approach).

The rest of the paper is organized as follows. Section 2 presents our multi-industry spatial model with occupation-specific externalities within cities. Section 3 quantifies the model, including our estimation of the externality parameters. It also discusses the role of externalities in the equilibrium allocation. Section 4 presents the optimal policy, and Section 5 its implementation. Section 6 presents exercises that probe the role of industrial linkages and trade costs. Section 7 provides a decomposition of the impact of fundamental changes in the national CNR employment share and in technology across sectors and cities between 1980 and the recent data. Section 8 concludes. We relegate many of the model's and quantification's details, additional robustness exercises, and counterfactuals to the Appendix.

2 A Spatial Model with Multiple Industries and Occupations

The economy has N cities and J sectors. We denote a particular city by $n \in \{1, ..., N\}$ and a particular sector by $j \in \{1, ..., J\}$. Individuals are endowed with an occupational type and cannot switch types. There are two occupational types, denoted by $k \in \{\text{CNR}, \text{non-CNR}\}$, with aggregate number of workers L^k per type (total employment in occupation k aggregated across industries and

[3] Eeckhout and Guner (2015) and Albouy et al. (2019) discuss the optimal distribution of city sizes.

cities).[4]

Firms use multiple types of labor but in potentially different proportions depending on the industry and the city. Aggregate land and structures in location n are denoted by H_n. Labor of all types moves freely across locations and sectors, while structures are location-specific. Some sectors are tradable, while others have prohibitive trade costs.

Quantities and prices in the economy may be associated with industries, cities, or occupations. For notational convenience, we denote aggregates across a given dimension by omitting the corresponding index. Thus, for example, L_n^{kj} is the number of workers employed in occupation k in industry j in city n, $L_n^k = \sum_j L_n^{kj}$ represents the number of workers employed in occupation k in city n, $L^k = \sum_n L_n^k$ represents all workers in occupation k, and $L = \sum_k L^k$ is simply total employment.

2.1 Workers

Workers are endowed with an occupation (their type) and differ in how much they value living in different cities. These idiosyncratic preferences are captured by a vector, $\mathbf{a} = \{a_1, a_2, ..., a_N\}$, where each entry, a_n, shifts an individual's utility from living in city n.

Sector $j \in \{1, ..., J\}$ goods consumed in city n are purchased at prices P_n^j. Workers in occupation k and city n supply one unit of labor inelastically and earn wages w_n^k. Conditional on living in city n, the problem of a worker with amenity vector $\mathbf{a}$ employed in occupation k is

$$v_n^k(\mathbf{a}) \equiv \max_{\left\{C_n^{kj}(\mathbf{a})\right\}_{j=1}^J} a_n A_n^k \prod_j \left(C_n^{kj}(\mathbf{a})\right)^{\alpha^j} \text{ subject to } \sum_j P_n^j C_n^{kj}(\mathbf{a}) = w_n^k + \chi^k \equiv I_n^k,$$

where $\sum_j \alpha^j = 1$. $C_n^{kj}(\mathbf{a})$ is the consumption by workers with amenity draws $\mathbf{a}$ of good j in city n and P_n^j is the price of that good. Here, α_j represents sector j's consumption share, and A_n^k is a city-specific exogenous component of utility common to all individuals in occupation k living in city n. As in Caliendo et al. (2017), the term χ^k represents the return per household on a national portfolio of land and structures from all cities, $\chi^k = b^k \sum_{n'} r_{n'} H_{n'}/L^k$, where r_n is the rental rate on land and structures in that city, and b^k denotes the share of the national portfolio accruing to workers in occupation k. In what follows, we assume that b^k is determined such that, in equilibrium, different worker types receive a share of the national portfolio proportional to their share of wages in the total wage bill, so that $b^k = \sum_n w_n^k L_n^k / \sum_{k',n'} w_{n'}^{k'} L_{n'}^{k'}$.

Given homothetic preferences, cost-minimization implies that one can aggregate those sectoral prices into a price index, P_n, so that $C_n^k(\mathbf{a}) = I_n^k/P_n$. Given that income I_n^k and prices do not depend on $\mathbf{a}$, all workers in a given occupation living in a given city choose the same consumption basket. It follows that we can write $C_n^{kj}(\mathbf{a}) = C_n^{kj}\ \forall \mathbf{a}$. Therefore, the value of locating in a particular city, n, for an individual employed in a given occupation, k, with a specific idiosyncratic

[4] We do not model occupational choice in this paper, though Burstein et al. (2019) provide a potential framework along those lines.

preference vector, $\mathbf{a}$, is $v_n^k(\mathbf{a}) = a_n A_n^k \frac{w_n^k + \chi^k}{P_n} \equiv a_n v_n^k$, where v_n^k is the part of the value function common to all workers in a given occupation and city. In equilibrium, workers move to the location where they receive the highest utility so that $v^k(\mathbf{a}) = \max_n v_n^k(\mathbf{a})$. We denote by Ψ the joint *cdf* for the elements of $\mathbf{a}$ across workers in occupation k. We take Ψ to be a Fréchet distribution with shape parameter ν. One can then show that the fraction of workers in each city is

$$\frac{L_n^k}{L^k} = \Pr\left(v_n^k(\mathbf{a}) > \max_{n' \neq n} v_{n'}^k(\mathbf{a})\right) = \frac{\left(v_n^k\right)^\nu}{\sum_{n'}\left(v_{n'}^k\right)^\nu}. \tag{1}$$

2.2 Firms

There are two types of firms: those producing intermediate goods and those producing final goods. Following Caliendo and Parro (2015), intermediate goods are traded across locations and are aggregated into final goods that are locally consumed or used as material inputs.[5]

Tradable intermediate varieties: A continuum of varieties of intermediate goods aggregates into a finite number of final goods corresponding to J sectors. Varieties of intermediate goods are characterized by the sector in which they are produced, and by a vector of city-specific productivity parameters, $\mathbf{z} = \{z_1, z_2, ..., z_N\}$, with each element, z_n, scaling the productivity of firms in city n producing that variety. Idiosyncratic productivity draws, $\mathbf{z}$, arise from a Fréchet distribution with shape parameter θ. Draws are independent across goods, sectors, and locations. We denote the joint CDF of variety-specific productivity parameters across locations by Φ.

Within each sector, j, production of intermediate varieties is Cobb-Douglas in materials produced in different sectors, land and structures, and a CES aggregator of different labor types. They use a production technology given by

$$q_n^j(\mathbf{z}) = z_n^j \left[\left(\sum_k \left(\lambda_n^{kj} L_n^{kj}(\mathbf{z})\right)^{\frac{\varepsilon-1}{\epsilon}}\right)^{\frac{\epsilon}{\epsilon-1}(1-\beta_n^j)} \left(H_n^j(\mathbf{z})\right)^{\beta_n^j}\right]^{\gamma_n^j} \prod_{j'} \left(M_n^{j'j}(\mathbf{z})\right)^{\gamma_n^{j'j}},$$

where for each city n, industry j, variety $\mathbf{z}$, $q_n^j(\mathbf{z})$ is the quantity produced, $L_n^{kj}(\mathbf{z})$ is the use of labor of type k, $H_n^j(\mathbf{z})$ is the use of land and structures, $M_n^{j'j}(\mathbf{z})$ are the uses of materials produced by each sector j', λ_n^{kj} are labor-augmenting productivity shifters, γ_n^j is the value-added share of gross output, β_n^j the share of housing and structures of value-added, $\gamma_n^{j'j}$ is the input share of gross output sourced from sector j', ε is the elasticity of substitution between different types of labor.

The implied unit cost of producing a variety $\mathbf{z}$ in industry j in city n is

$$x_n^j(\mathbf{z}) = \frac{x_n^j}{z_n^j} = B_n^j \left\{ r_n^{\beta_n^j} \left[\sum_{k=1}^K \left(\frac{w_n^k}{\lambda_n^{kj}}\right)^{1-\epsilon}\right]^{\frac{1-\beta_n^j}{1-\epsilon}} \right\}^{\gamma_n^j} \prod_{j'=1}^J \left(P_n^{j'}\right)^{\gamma_n^{j'j}}, \tag{2}$$

[5] This classification of goods in intermediate and final is a modeling convention that does not preclude trade of goods in particular industries.

where in each city n and industry j, x_n^j is a unit cost index, and B_n^j is a normalizing constant.[6] Given constant returns to scale, firms earn zero profits, which implies a unit price for variety $\mathbf{z}$ in industry j, location n, equal to its unit cost, $x_n^j(\mathbf{z})$.

The terms λ_n^{kj} are labor-augmenting productivity shifters specific to each city, industry, and occupation. Importantly, we allow λ_n^{kj} to reflect externalities that depend on the composition of the labor force. We assume that occupational spillovers have the same labor-augmenting effect across sectors,

$$\ln \lambda_n^{kj} = \tau^{R,k} \ln \left(\frac{L_n^k}{L_n} \right) + \tau^{L,k} \ln (L_n) + \ln \tilde{\lambda}_n^{kj}, \tag{3}$$

where $\tilde{\lambda}_n^{kj}$ is an exogenously determined component of technology. Consistent with this formulation, Ellison et al. (2010), find that, aside from natural advantages and production linkages, occupational complementarities are the main source of industrial co-location.

Equation (3) allows for an agglomeration effect that depends on overall city size, through $\tau^{L,k}$, and an additional effect related to the ratio of a worker's own type to all workers, through $\tau^{R,k}$ (i.e., the share of each worker type). With four elasticity parameters to be estimated, the specification has enough degrees of freedom to allow for a full set of cross-elasticities between the two types of labor and their respective productivities.[7] Moreover, the elasticity of productivity with respect to the concentration of a given worker of type k in city n is given by

$$\frac{\partial \ln \lambda_n^{kj}}{\partial \ln L_n^k} = \tau^{R,k} + \left(\tau^{L,k} - \tau^{R,k} \right) \frac{L_n^k}{L_n}. \tag{4}$$

The marginal effect of workers in occupation k on their own productivity is always positive, but when $\tau^{L,k} < \tau^{R,k}$, it is smaller in cities where those workers are already more concentrated.[8] In addition, there are cross-occupational effects. Specifically, for $k \neq k'$, we have that

$$\frac{\partial \ln \lambda_n^{kj}}{\partial \ln L_n^{k'}} = \left(\tau^{L,k} - \tau^{R,k} \right) \frac{L_n^{k'}}{L_n}, \tag{5}$$

so that cross-occupational externalities are negative when $\tau^{L,k} < \tau^{R,k}$. In other words, worker productivity in a given occupation declines with each additional worker in an alternative occupation, a form of congestion effect.

Final Goods: Intermediate varieties are assembled into final goods sold in the city where they are produced. Final goods producers operating in city n, sector j, purchase a quantity $Q_{nn'}^j(\mathbf{z})$ of

[6] In particular, $B_n^j = \left[\left(1 - \beta_n^j\right)^{\beta_n^j - 1} \left(\beta_n^j\right)^{-\beta_n^j} \right]^{\gamma_n^j} \left[\prod_{j'} \left(\gamma_n^{j'j} \right)^{-\gamma_n^{j'j}} \right] \left(\gamma_n^j\right)^{-\gamma_n^j}$

[7] In a companion paper (Rossi-Hansberg et al. (2021)), we estimate a more flexible model that allows for industry-specific spillovers by letting τ^R and τ^L vary across industries (so we estimate $\tau^{R,j}$ and $\tau^{L,j}$). See Section 7.2 of the Appendix.

[8] Note also that the specification can be readily extended to additional occupations as long as we restrict all cross-occupation effects to take place directly or through total population.

the variety of intermediate goods indexed by $\mathbf{z}$ from city n'. They produce the quantity Q_n^j using a technology given by

$$Q_n^j = \left[\int \left[\sum_{n'} Q_{nn'}^j(\mathbf{z}) \right]^{\frac{\eta-1}{\eta}} d\Phi(\mathbf{z}) \right]^{\frac{\eta}{\eta-1}},$$

where η is the elasticity of substitution between varieties of intermediate goods.

Transport costs are of the iceberg form. Namely one unit of any intermediate good in sector j shipped from region n' to region n requires producing $\kappa_{nn'}^j \geq 1$ units in the origin n'. Therefore, producers of final goods in each sector solve

$$\max_{Q_{nn'}^j(\mathbf{z})} P_n^j Q_n^j - \sum_{n'} \int \kappa_{nn'}^j x_{n'}^j(\mathbf{z}) Q_{nn'}^j(\mathbf{z})\, d\Phi(\mathbf{z}),$$

subject to $Q_{nn'}^j(\mathbf{z}) \geq 0$, where P_n^j is the price of the final good in sector j, city n. Intermediate goods produced by non-tradable sectors cannot be shipped between cities, that is, $\kappa_{nn'}^j = \infty$ for $n \neq n'$. Final goods producers make zero profits, so that the price of good j in city n, P_n^j, ends up a function of $x_{n'}^j$, the cost index for corresponding intermediate goods in all origin cities n'.[9]

Final goods firms purchase intermediate goods from the location in which they can acquire them most cheaply, allowing for transportation costs. We denote by X_n^j total expenditures on final goods j by city n which, given competitive markets, must equate to the value of final goods in that sector, $X_n^j = P_n^j Q_n^j$.[10] Following Eaton and Kortum (2002), the share of intermediate inputs imported from location n' for the production of final good j in city n is

$$\pi_{nn'}^j = \frac{\left[\kappa_{nn'}^j x_{n'}^j\right]^{-\theta}}{\sum_{n''=1}^N \left[\kappa_{nn''}^j x_{n''}^j\right]^{-\theta}}, \tag{6}$$

where x_n^j is a cost index associated with the production of varieties in sector j and city n. In equation (2), this cost index reflects wages, material input prices, and labor-augmenting productivity shifters (see above). In mapping the model to the data, we also allow for non-tradable sectors in which case $\pi_{nn}^j = 1$.

2.3 Market Clearing

Within each city n, the number of workers employed in occupation k must equal the number of those workers who choose to live in that city. In addition, final goods are used either as consumption goods or as materials for the production of local varieties. See Section 1 of the Appendix for the formal expressions governing market clearing and other equilibrium conditions.

[9] Specifically, $P_n^j = \Gamma(\xi)^{\frac{1}{1-\eta}} \left[\sum_{n'=1}^N \left[\kappa_{nn'}^j x_{n'}^j\right]^{-\theta}\right]^{-\frac{1}{\theta}}$.

[10] Zero profits in the final goods sector implies that total expenditures on intermediate goods in any sector must equal the cost of inputs in that sector.

3 Quantifying the Model

Conditional on observed wages, rents, and employment, the economic environment described above provides a framework for recovering productivity measures specific to each occupation, industry, and city. Put another way, the model implies specific productivity measures that deliver observed wages, rents, and employment as equilibrium outcomes of the model. Furthermore, the associated model inversion then implies amenities for each city and occupation that rationalize observed location choices in the data. We provide an overview of the model inversion below, with the specifics of the quantification method presented in Sections 2.1 and 2.2 of the Appendix.

3.1 Data

To characterize employment and wages, we distinguish between two large occupational groups over the period 2011-2015: those that are intensive in cognitive non-routine (CNR) tasks and all others (non-CNR). Specifically, we follow Jaimovich and Siu (2018) and define CNR occupations to include occupations with SOC-2 classifications 11 to 29 and non-CNR occupations to include the remainder of SOC-2 classifications. This classification builds on the observation by Acemoglu and Autor (2011) that one can best assess wage inequality trends through a task-based lens.

Our division of workers by occupation differs somewhat from the spatial economics literature that has more often emphasized college attainment. We favor the occupational classification since it is closer to the theoretical foundation of productivity spillovers, emphasizing learning from others in similar tasks. In addition, this distinction is not merely theoretical. In Section 7.3 of the Appendix, we document that while the majority of CNR workers do have college degrees, a large fraction do not. Importantly, CNR workers without college degrees may include managers whose share of total occupations is particularly large in smaller cities. These non-college-educated managers likely confer productivity spillovers onto other CNR workers, arguably more so than college-educated workers in non-CNR occupations.[11]

The model is also mapped into 22 industries and 382 Core Based Statistical Areas (CBSAs). This partition allows us to capture important elements of industrial heterogeneity (see Section 6.1 below) while still keeping the model tractable from a computational standpoint.[12] Two industries are non-tradable, in that their output is produced and used locally. The two non-tradable sectors include real-estate services, and a composite sector comprising retail, construction, and utilities. Tradable industries include 10 manufacturing sectors and 10 service sectors.

We obtain data pertaining to the wage and occupational composition of each industry in each

[11] In Section 7.3 of the Appendix, we carry out our analysis using college attainment groups rather than occupational groups. Our findings are qualitatively similar to those presented in the main text though somewhat attenuated. This is consistent with the view that college attainment does not fully capture the extent of worker externalities. In section 7.4 of the same Appendix, we further experiment with narrower group definitions that single out CNR workers with professional degrees. This only reinforces our findings but at the cost of potentially abstracting from important sources of spillovers. Further breaking up occupational categories, while interesting, becomes impractical given how the increased noise associated with finer categories affects our instruments.

[12] With 382 cities and 22 industries, solving the model requires repeatedly inverting a square matrix with more than 10,000 rows.

city from the American Community Survey (ACS). To address the effect of sorting on wage data, we control for observable individual characteristics including education, gender, race, and other attributes.[13] Any remaining variation in our occupational wage data, therefore, primarily reflects differences stemming from location.[14] The Census provides measures of total employment, L_n^j, from the County Business Patterns (CBP) that better match Bureau of Economic Analysis (BEA) industry-level counts than ACS totals. We thus combine total employment from the CBP with ACS data on employment shares to obtain L_n^{kj}. Finally, we use rent data included in the Regional Price Parities produced by the BEA. Those are, in turn, produced using ACS data on rents paid after adjusting for individual home characteristics.

The main challenge in processing the data lies in ensuring consistency between different sources as well as with national accounting identities. Thus, we now provide a summary of this process while a detailed account of the data and the specific procedures adopted to harmonize different sources are given, respectively, in Sections 9 and 2 of the Appendix.

3.2 Share Parameters

We assume that share parameters in the tradable sectors are invariant across cities. We discipline material input share parameters, $\gamma^{j'j}$, and consumption shares, α^j, using an average of the 2011 to 2015 Bureau of Economic Analysis (BEA) Use Tables. Because we restrict our attention to CBSAs, we make adjustments to those tables to ensure that nationwide output and expenditures are equal for each sector (as in the model). In particular, given our focus on production technology, we adjust consumption shares as needed so that aggregate identities hold (see Section 2.1 of the Appendix).[15]

In addition, mapping the data into model parameters requires taking a stance on the origins of gross operational income (i.e., essentially income from equipment, land, and structures). This is in part because the model is static and omits capital accumulation. Thus, we attribute equipment income to materials (in that equipment fully depreciates in a static setting where the period lasts indefinitely). The remaining component of gross operational income is attributed to land and structures. Since the model does not differentiate between land and structures that are rented versus owned, we take all land and structures to be operated by firms in the real estate sector. These firms then sell their services to other sectors and to final consumers.[16] This approach implies that the share of land and structures used in production, β_n^j, is zero in all sectors except for real estate.

[13] Specifically, we include as control variables education, potential experience, race, gender, English proficiency, number of years in the US, marital status, having had a child in the last year, citizenship status, and veteran status.

[14] The possibility remains that observed wage differences reflect sorting on unobservable variables. Recent work from de la Roca and Puga (2017) leveraging Spanish administrative data offers evidence against this possibility. In particular, they find that sorting on worker fixed effects accounts for only a small fraction of cross-city differences in wages once one allows for the dynamic effects of location on individual wages. In the context of our static model, those dynamic effects are interpreted as stemming from city-specific productivity.

[15] Our adjusted consumption shares remain highly correlated with the raw consumption shares in the data.

[16] These assignments do not affect aggregate operational surplus (net of equipment investment), aggregate labor compensation, or aggregate value added (net of equipment investment).

3.3 Elasticity Parameters

Idiosyncratic preferences pertaining to city-specific amenities, a_n, are drawn independently from a Fréchet distribution across cities parameterized by the shape parameter, ν, which captures the extent of preference heterogeneity across workers employed in each occupation. Higher values of ν imply less heterogeneity across workers with preferences for locations that are more similar. We set the distribution of amenity types so that the average elasticity of employment with respect to real wages in our model matches the estimate from Fajgelbaum et al. (2018). This implies $\nu = 2.02$.

We set the Fréchet parameter governing the trade elasticity, θ, to 10. Estimates of trade elasticities range from 3 to 17 in the literature (see Footnote 44 in Caliendo and Parro (2015), as well as section 4.2 in Head and Mayer (2014) for comprehensive summaries of estimates). While θ has been estimated for various levels of disaggregation, estimates vary somewhat widely for a given sector or commodity across studies. For example, while Caliendo and Parro (2015) estimate an elasticity of 7.99 for basic metals and 4.75 for chemicals, Feenstra et al. (2018) estimate elasticities of 1.16 and 1.46, respectively, for these two categories. In addition, this uncertainty is further compounded by the fact that trade elasticities relevant for trade between countries may be less relevant for trade between regions or cities.

We treat transportation costs in non-tradable sectors ('real estate,' and 'retail, construction, and utilities') as infinite. For tradable sectors, we follow Anderson et al. (2014) and assume that trade costs increase with distance. Specifically, in order to ship one unit of good to city n from city n', $\kappa^j_{nn'} = (d_{nn'})^{t^j}$ units of the good need to be produced in city n', with $d_{nn'}$ denoting the distance between city n and city n' in miles. The within-city distance is set to 20 miles. The parameter t^j is industry-specific. For commodities, we directly estimate t^j from the Commodity Flow Survey synthetic microdata using standard gravity regressions based on model trade shares. In tradable services, we use the values obtained by Anderson et al. (2014) using Canadian data.[17]

Finally, Ciccone and Peri (2005) summarize estimates for the elasticity of substitution between skilled and unskilled labor in the literature as ranging between 1.36 and 2. Card (2001) estimates the elasticity of substitution between occupations to be closer to 10. We adopt $\epsilon = 2$ as a benchmark.

3.4 Obtaining Equilibrium Measures of Productivity

In the environment above, data on wages, rents, and employment may be used to recover measures of productivity by occupation, industry, and location. In particular, observed production allocations in space require a particular alignment of unit costs across sectors and regions. Thus, if computer and electronic manufacturing is heavily concentrated in San Jose, for example, the model requires that households and firms prefer purchasing these goods from San Jose to producing them locally. We formalize this result in Lemma 1 below.[18]

[17] We do not need to calibrate η, the elasticity of substitution between varieties, since in equilibrium this only scales prices P^j_n by a fixed amount for each industry j. Therefore, they do not affect spatial allocations and are absorbed by the normalization in Step 6 of the model inversion algorithm described in Section 2.2 of the Appendix.

[18] The proof of Lemma 1 is the set of steps provided in the algorithm used to compute the model inversion (steps 2 through 9) in Section 2.2 of the Appendix.

Lemma 1. *Conditional on data on wages by city and occupation (w_n^k), employment by city, occupation and industry (L_n^{kj}), and prices in non-tradable sectors (P_n^j for j such that $\kappa_{nn'}^j = \infty$ if $n \neq n'$), as well as consumption and input share parameters ($\alpha^j, \beta_n^j, \gamma_n^{j'j}$), the parameters determining the distribution of productivities (θ) and amenities (ν), and trade costs ($\kappa_{nn'}^j$), the model determines for each city n, sector j, and occupation k:*

(1) Values for gross output, $\sum_{n'} \pi_{n'n}^j X_{n'}^j$ and total expenditures X_n^j.

(2) Unit production costs, x_n^j, and price indices, P_n^j, given $\sum_{n'} \pi_{n'n}^j X_{n'}^j$ and X_n^j derived in (1).

(3) Productivity shifters, λ_n^{kj}, given x_n^j and P_n^j's derived in (2).

(4) City-specific average amenity shifters for each occupation, A_n^k, given gross output derived in (1) and prices derived in (2).

Item (1) of Lemma 1 follows from Cobb-Douglas technology and preferences. Given appropriate share parameters, we can recover for all sectors j and locations n the value of gross output, $\sum_{n'} \pi_{nn'}^j X_{n'}^j$, and expenditures, X_n^j (where recall that $\pi_{nn'}^j$ is the share city n''s expenditures on good j sourced from city n).

In tradable sectors gross output and expenditures measured in step (1) do not have to match within each city, leading to sectoral trade imbalances within each location. Recall from equation (6) that trade shares are determined by differences in unit production costs x_n^j, given trade costs and elasticities. Therefore, Item (2) in Lemma 1 backs out values of x_n^j required by the model to match the trade imbalances obtained from (1). Finally, the unit costs x_n^j can be aggregated into local final goods prices, P_n^j (see Footnote 9). One cannot use this same algorithm for nontradable sectors, since they do not feature trade deficits by construction. For those sectors, unit costs are simply equal to the nontradable prices in the data.[19]

Item (3) recovers city-industry-occupation specific productivity, λ_n^{kj}. It uses the unit costs and final good prices obtained in Item (2) together with equation (2) and the observed wage-bill share of each worker type. The equation implies that productivity λ_n^{kj} is high whenever unit costs x_n^j are low, or whenever wages in that occupation are high. In other words, when observed wages of a given labor type are high, the model requires that the corresponding labor type be particularly productive to justify its use in production.

Finally, item (4) uses equation (1). Given local income and local prices obtained in items (1) and (2), one can solve for real consumption associated with each location and occupation. Then, conditional on real consumption and the observed distribution of workers in space, L_n^k, one can obtain amenity shifters A_n^k.

Lemma 1 imposes additional constraints on the parameters for non-tradable goods sectors. By definition, all output in those sectors is consumed where it is produced. However, the model

[19] We match the price of real estate services to residential rent prices (which the BEA calculates using a hedonic adjustment to rent prices reported in the ACS) and choose the final goods price of retail, construction, and utilities to match the relative price of services reported by the BEA.

inversion implied by Lemma 1 does not constrain gross expenditures to exactly exhaust gross output in all cities. Spurious trade deficits or surpluses, therefore, can arise in non-traded goods or services. To rule out such counterfactual outcomes, we choose the share of land and structures of value-added, β_n^{RE}, and the value-added share of gross output in retail, construction, and utilities, γ_n^{RCU}, to ensure that in non-tradable sectors, gross output equals expenditures in every city.[20]

3.5 Comparison with Prior Empirical Work and Model Validation

Lemma 1 allows us to obtain tradable prices, total factor productivity, and amenity measures in a way that is consistent with the general equilibrium framework underlying our analysis of optimal policy. However, independent measurements of those variables based on alternative empirical approaches also exist. Thus, a brief comparison of our findings with prior work is given below.

Using Nielsen home-scanned data on tradable goods bought in grocery stores, Handbury and Weinstein (2014) find that tradable consumer prices decrease with city size and a corresponding elasticity of -0.011.[21] Table 1 below summarizes the relationship between prices and city size obtained in our model inversion. Similar to Handbury and Weinstein (2014), our general equilibrium framework reveals that prices decrease with city size and, in fact, do so across all tradable sectors, not just those bought in grocery stores. The average elasticity of prices with respect to city size implied by our model inversion is -0.012. In the food and beverage sector, our model inversion reveals an elasticity of -0.010, virtually identical to that for the grocery products studied by Handbury and Weinstein (2014). Remarkably, our model inversion is able to replicate these negative elasticities without direct observations on prices.

We can also verify that the model is able to generate a reasonable ranking of industry-level TFP across cities. Table 1 below shows the city in which TFP (recovered from the model inversion) is highest across different industries. The results largely conform to intuition. For example, productivity in computers and electronic equipment is highest in San Jose, CA; Anchorage stands out for oil, chemicals, and nonmetallic minerals; and Seattle for motor vehicles (which include aircraft). It is also interesting to note that the two largest cities in the country, New York and Los Angeles, rank as the top city in several sectors. New York dominates in TFP in most service sectors while Los Angeles stands out in several manufacturing sectors.

In Section 2.3 of the Appendix, we further verify that amenities pertinent to CNR workers are relatively higher in cities where CNR workers are disproportionately located, in line with Diamond (2016). Moreover, our model inversion also implies that TFP rises with both city size and the CNR share according to magnitudes similar to those found in work summarized, for example, by Melo et al. (2009) and Moretti (2004b).

[20] See Section 2.2 of the Appendix, step 3, for the exact implementation of this constraint.
[21] See Handbury and Weinstein (2014) Table 5, column 9.

Table 1: Elasticities of Final Goods Prices, P_n^j, w.r.t. L_n and Top TFPs

Sector	Elasticity	Highest TFP City
Food and Beverage	−0.010	San Francisco, CA
Textiles	−0.022	Los Angeles, CA
Wood, Paper, and Printing	−0.011	Minneapolis, MN
Oil, Chemicals, and Nonmetallic Minerals	−0.016	Anchorage, AK
Metals	−0.014	Los Angeles, CA
Machinery	−0.005	Houston, TX
Computer and Electronic	−0.013	San Jose, CA
Electrical Equipment	−0.003	Los Angeles, CA
Motor Vehicles (Air, Cars, and Rail)	−0.009	Seattle, WA
Furniture and Fixtures	−0.009	Los Angles, CA
Miscellaneous Manufacturing	−0.013	Los Angles, CA
Wholesale Trade	−0.007	New York, NY
Transportation and Storage	−0.007	New York, NY
Professional and Business Services	−0.018	San Jose, CA
Other	−0.011	Los Angeles, CA
Communication	−0.003	New York, NY
Finance and Insurance	−0.011	New York, NY
Education	−0.024	New York, NY
Health	−0.030	New York, NY
Accommodation	−0.012	San Francisco, CA

Elasticity column is the coefficients from univariate OLS regression of final-goods prices ($\ln xP_n^j$) on $\ln L_n$. Within the context of the model, TFP is measured by $\ln TFP_n^j = \ln \frac{\sum_{n'} \pi_{n'n} X_{n'}^j}{P_n^j} - \gamma_n^j \beta_n^j \ln H_n^j - \gamma_n^j (1-\beta_n^j) \sum_k \delta^{kj} \ln L_n^{kj} - \sum_{j'} \gamma_n^{j'j} \ln M_n^{j'j}$

3.6 Estimating Production Externalities by Worker Type

Thus far, we have described how city-industry-occupation-specific productivity, λ_n^{kj}, is recovered to be consistent with observed data on wages and sectoral employment. Importantly, Lemma 1 provides us with overall equilibrium productivity, λ_n^{kj}, but not with how it is partitioned between its endogenous and exogenous components in equation (3), $\tau^{R,k} \ln\left(\frac{L_n^k}{L_n}\right) + \tau^{L,k} \ln(L_n)$ and $\ln \tilde{\lambda}_n^{kj}$ respectively. Therefore, given the productivity measures obtained through the model inversion, we now turn to estimating the relationship between productivity, λ_n^{kj}, and both the scale and composition of population across cities.

In any OLS estimation of equation (3), coefficients are potentially biased since workers may choose to live in cities where they are relatively more productive. This would induce a correlation between the exogenous component of worker productivity, $\tilde{\lambda}_n^{kj}$, and the share of each type of worker in a given city. Moreover, the estimates might be biased if there were omitted variables correlated with both $\tilde{\lambda}_n^{kj}$ and the occupational ratio.

To address omitted variable bias, we specify $\tilde{\lambda}_n^k$ such that equation (3) becomes

$$\ln \lambda_n^{kj} = \tau^{R,k} \ln\left(\frac{L_n^k}{L_n}\right) + \tau^{L,k} \ln(L_n) + a_Z^k Z_n^j + d^{kj} + u_n^{kj},$$

where Z_n^j is a vector of observable city-industry characteristics, d^{kj} denotes a set of industry dummies, and u_n^{kj} captures unobserved city-specific sources of natural advantages in the production of

sector j goods with workers of type k.[22]

The vector Z_n^j includes the following controls: (i) dummies for 9 census divisions interacted with industry dummies; these aim to absorb geographical and historical components that jointly determine amenities and productivity in different places, (ii) geographic amenities constructed by the United States Department of Agriculture (USDA) that include measures of climate, topography, and water area; these controls account for the possibility that whatever geographic characteristics lead workers to choose certain cities also potentially influence their productivity, (iii) the share of manufacturing workers in 1920 which aims to extract long-standing factors that may have influenced the industrial composition of particular locations, and (iv) controls for demographic characteristics of different cities including racial composition, gender split, the fraction of immigrant population, and age structure. Together, these controls narrow down the identification of externality coefficients to rely only on the extent to which more productive but otherwise similar cities attract individuals of a particular demographic makeup.[23]

Columns 1 and 2 of Table 2 report the coefficients from an OLS estimation of equation (3) with all the controls. The estimation also allows for two-way clustered standard errors by city and industry and produces coefficients that are positive and significant. They indicate that the presence of other workers of the same occupational group enhances individual productivity. The coefficients also indicate the presence of congestion effects, $\tau^{L,k} < \tau^{R,k}$, in which case equation (5) implies that cross-occupational externalities are negative.

Instrumenting for Employment Levels and Composition

In order to isolate the residual simultaneity between exogenous productivity variations and labor allocations, we resort to variants of instruments proposed in the literature. Specifically, we follow Ciccone and Hall (1996) and use population from a century ago to capture historical determinants of the current population. We also follow Card (2001) and Moretti (2004a) and use the variation in the early immigrant population and the presence of land-grant colleges to capture historical determinants of skill composition across cities. A detailed discussion of the specific instruments is provided in Section 2.5 of the Appendix.

Table 2 shows estimation results with instrumental variables (IVs) and all the controls. Columns 3 and 4 show the corresponding two-stage least-squares estimates. Those are similar to the OLS estimates, being well within one standard error from one another. To evaluate the strength of the instrumental variables, we follow the procedure in Sanderson and Windmeijer (2016) and obtain separate first-stage F statistics for each of the endogenous variables.[24] Since these F-statistics are

[22] When estimating (3), we observe that u_n^{kj} also incorporates measurement error in x_n^j's, which passes through to $\ln \lambda_n^{kj}$ multiplied by $1/\gamma_n^j$. For that reason, we multiply both sides of (3) by γ_n^j before estimating.

[23] Results from OLS regressions with different sets of controls are described in Section 2.4 of the Appendix, table A5. The census division dummies are: 1. New England, 2. Mid-Atlantic, 3. East North Central, 4. West North Central, 5. South Atlantic, 6. East South Central, 8. Mountain and 9. Pacific. Geographic controls include average temperature for January and July, hours of sunlight in January, humidity in July from 1941 to 1970, variations in topography, and percent of water area. Demographic controls are, by city, the percent female, black, Hispanic, and percent in the age bins 16-25 and 26-65 (observations related to the younger-than-16 population are dropped from the sample, and the age bin 66+ is omitted from the regression).

[24] This follows the intuition laid out by Angrist and Pischke (2008) that requires strong IVs to predict the two

Table 2: Externality Parameter Estimate

		OLS		2SLS		CUE	
		(1)	(2)	(3)	(4)	(5)	(6)
Variable	Coefficient	CNR	non-CNR	CNR	non-CNR	CNR	non-CNR
$\ln(\frac{L_n^k}{L_n})$	τ^R	0.889***	0.702***	1.177***	0.263	1.304***	0.835*
		(0.12)	(0.22)	(0.38)	(0.51)	(0.38)	(0.51)
$\ln(L_n)$	τ^L	0.386***	0.322***	0.334***	0.284***	0.349***	0.357***
		(0.05)	(0.04)	(0.06)	(0.05)	(0.06)	(0.04)
Observations		7,460	7,460	7,460	7,460	7,460	7,460
R^2		0.416	0.317	0.413	0.312	0.411	0.313
J Test P-Value				0.375	0.115	0.384	0.122
K.P. F				3.912	5.425	3.912	5.425
S.W.F. L_n^k Share				5.975	8.369	5.975	8.369
S.W.F. L_n				5.997	8.587	5.997	8.587

Regressions estimate equation (3) multiplied by γ_n^j to correct for heteroskedasticity in measurement errors (see Footnote 22). The dependent variable is $\ln \lambda_n^{kj}$ obtained from the model inversion procedure described in text. Standard errors in parentheses, clustered two-ways by city and by industry. ***p<0.01, ** p<0.05, * p<0.1

less than 10, the associated estimates may be subject to bias and have incorrect standard errors. The IV literature then recommends using limited information maximum likelihood (LIML) estimators. Columns 5 and 6 of Table 2, therefore, carry out an estimation using a continuously updated GMM estimator (GMM-CUE), similar to a limited information maximum likelihood estimator but which allows for clustered and heteroskedastic standard errors. The Stock and Yogo (2005) critical values in the LIML model for a p-value of 5 percent or better is 6.46, at or close to our obtained values.

These instruments, inspired by previous studies, exploit the idea that after allowing for controls, all long-term local productivity effects of either historical immigration enclaves, land-grant college location, or historical population derive from their repercussions on the current population and its occupational composition. To the degree that historical factors affect current productivity through other channels, or that the exogenous component of productivity has persisted for more than a century, the instruments we use may nevertheless be called into question. Furthermore, we also assume that the spillover parameters, τ^L and τ^R, are the same across cities. However, given that we have more instruments than we have parameters to estimate, we can check the implied J-statistic for the joint orthogonality of our instruments with respect to the exogenous component of productivity. We find that the J-statistic is indeed not significantly different from zero at the 5 percent critical level. To further address potential misspecification concerns with our approach, we leverage the model in two ways.

First, to verify that our empirical strategy indeed identifies external effects, we carry out the same regressions but on data generated by a counterfactual allocation where we set $\tau^{R,k} = \tau^{L,k} = 0$ for all occupations k. The results are presented in Table A6 in Section 2.6 of the Appendix. They confirm that the OLS estimates capturing the effect of occupation shares on productivity are biased

endogenous variables independently from one another.

downward (the coefficients are now negative), whereas the estimates related to external productivity effects driven by the overall population exhibit little bias. Importantly, the exercise also shows that our IVs successfully eliminate most of those biases, especially in the case of GMM-CUE estimates.

Second, we carry out an estimation exercise using IVs implied by the model. Recall that, in our framework, the size and composition of the population in different locations are determined simultaneously by local productivity, amenities, input-output linkages, and trade costs. Thus, we construct a counterfactual allocation where, for each industry and occupation, we set productivity to be fully exogenous and equal to the cross-city averages of productivity parameters, λ_n^{kj}. We then use as instruments the resulting counterfactual employment shares and totals conditional on fixed exogenous productivity and no endogenous productivity differences. The estimates in Section 2.7 of the Appendix, Table A7, confirm the main findings. Specifically, the effect of population composition is larger than that of city size, and the worker composition effect is larger for CNR workers than for non-CNRs.

3.7 The Role of Externalities

We adopt the GMM-CUE coefficients in the last column of Table 2 as our benchmark. These coefficients imply that cross-occupational externalities are negative for CNR workers in that $\tau^{L,CNR}$ - $\tau^{R,CNR}$ in equation (5) is significantly negative. Hence, non-CNR workers create negative congestion effects for other CNR workers. In contrast, the difference between $\tau^{L,\text{non-CNR}}$ and $\tau^{R,\text{non-CNR}}$ is half as large as for CNR workers while the estimate for $\tau^{R,\text{non-CNR}}$ is only marginally significant. Thus, evidence of congestion effects from CNR workers to non-CNR workers is weaker.

The coefficients in the first two lines of Table 2 capture the spillover effects from city occupation composition (L_n^k/L_n), and city size, (L_n), on the productivity of each worker type. We can also use the estimated coefficient to calculate the effects of spillovers on TFP in each sector j and city n, given by[25]

$$\ln TFP_n^j \simeq \sum_k \delta^{kj}\gamma^j \ln \lambda_n^{kj}$$

where δ^{kj} is the wage-bill share of workers in occupation k and γ^j is the labor share of gross output.

Since the average labor share across sectors is close to 0.3, and $\tau^{L,CNR} \simeq \tau^{L,non-CNR} \simeq 0.35$ the implied population size effects on tradable TFP are close to 0.1, i.e., a 10 percentage point increase in city size raises TFP by 1 percentage point in that city, in line with prior estimates.[26]

Importantly, we find material externality effects on worker productivity stemming from local occupational composition. Specifically, moving from a city with a CNR share of 0.35, close to the national average, to 0.45 (while keeping total population constant) would increase the productivity of CNR workers by close to 30% and reduce that of non-CNR workers by 15%. The net effect on

[25] See Section 2.3.3 of the Appendix for the derivation.

[26] The wage impact of externalities may, in principle be strengthened by capital mobility between cities. Within our static model, this mechanism operates through greater imports of materials. Nevertheless, it remains that a substantial component of capital takes the form of non-mobile land and structures.

TFP is the weighted average of those effects with weights given by the wage bill share of each type of worker (close to 0.5) and then multiplied by the labor share of gross output (0.3). The net effect on tradable TFP is then close to 2%.

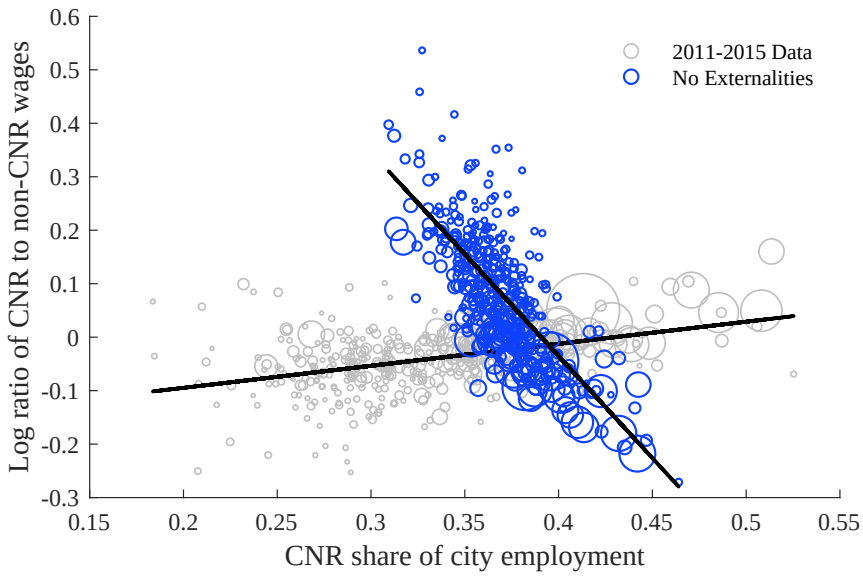

Figure 1: Occupational share and wage premium - no externalities

Counterfactual values obtained from assuming no externalities ($\tau^{R,k} = \tau^{L,k} = 0$), while keeping the exogenous part of productivity as originally quantified (blue markers). Grey markers correspond to equilibrium values. The equilibrium correlation is 0.4, and the counterfactual correlation is -0.74.

Externalities from city size and occupational composition have important implications for occupational wage premia. The gray markers in Figure 1 show that wages of workers employed in Cognitive Non-Routine (CNR) occupations are relatively high both in larger cities (captured by the size of scatter-plot markers) and cities where those workers are more abundant. In particular, the correlation between occupational wage premia and CNR shares suggests that differences in occupational wage premia are, to a large degree driven by differences *in the relative demand for CNR workers.* Absent externalities (i.e., set $\tau^{K,k} = \tau^{L,k} = 0$ for all k), the abundance of CNR workers decreases their marginal product and therefore their relative wage.

4 Optimal Allocation

We now describe the optimal allocation and the policies that implement it. We start by defining social preferences and setting up the planner's problem.

4.1 The Planner's Problem

The planner's problem recognizes that workers in each occupation can freely move across cities. Under this assumption, and given multiplicative amenities drawn from a Fréchet distribution, the expected utility of a worker of type k is given by

$$v^k = \Gamma\left(\frac{\nu-1}{\nu}\right)\left(\sum_n \left(A_n^k C_n^k\right)^\nu\right)^{\frac{1}{\nu}}.$$

Then, if ϕ^k denotes welfare weights for each occupation, we can postulate the generalized social welfare function as

$$\mathcal{W} = \sum_k \phi^k U\left[\Gamma\left(\frac{\nu-1}{\nu}\right)\left(\sum_{n=1}^N \left(A_n^k C_n^k\right)^\nu\right)^{\frac{1}{\nu}}\right] L^k, \tag{7}$$

where U is increasing and concave. This generalized social welfare function nests the leading cases of a utilitarian planner, in which U is linear, and the limit in which U becomes infinitely concave so that $\mathcal{W}$ approximates the max-min welfare function of a Rawlsian planner. The planner maximizes the social welfare function (7) subject to the availability of labor in each occupation within each city (1), as well as the resource utilization constraints on labor, land, structures, final goods, and intermediate goods.[27] The key difference between the optimal and equilibrium allocations stems from a wedge between the private and social marginal products of labor due to labor augmenting externalities. Lemma 2 characterizes this wedge.

Lemma 2. *Let Δ_n^k denote the wedge between the private and social marginal value of a worker in occupation k in city n. Then,*

$$\Delta_n^k = w_n^k \frac{\partial \ln \lambda_n^k}{\partial \ln L_n^k} + w_n^{k'} \frac{L_n^{k'}}{L_n^k} \frac{\partial \ln \lambda_n^{k'}}{\partial \ln L_n^k}, \tag{8}$$

where $\frac{\partial \ln \lambda_n^k}{\partial \ln L_n^k}$ and $\frac{\partial \ln \lambda_n^{k'}}{\partial \ln L_n^k}$ are given by equations (4) and (5).

The wedge in Lemma 2 points to the distortions that the planner is seeking to correct. It increases with the elasticity of worker productivity with respect to the number of workers in occupation k, $\partial \ln \lambda_n^{k'} / \partial \ln L_n^k$, in each occupation. Moreover, the contribution of this elasticity in a given occupation and city varies with the proportion of workers in that occupation within that city, $L_n^{k'}/L_n^k$, as well as their private marginal product, namely their wage, w_n^k.

The estimated externality parameters in Table 2 imply that occupation k's own elasticity is positive, $\partial \ln \lambda_n^k / \partial \ln L_n^k > 0$, while its cross-elasticity with other occupations is negative, $\partial \ln \lambda_n^{k'} / \partial \ln L_n^k < 0$. It follows that the wedge, Δ_n^k, for any occupation k increases with its wage, w_n^k, and decreases with the wage of occupation k', $w_n^{k'}$. Hence, all else equal, the planner seeks to increase the concentration of workers of a given occupation in places where those workers are most productive (as reflected in their wages). Conversely, the planner seeks to lower the concentration of workers of a

[27] See Sections 3 and 4 of the Appendix for a formal statement of the planner's problem and its solution.

given type in locations where that type is less productive. The end result is an increase in spatial polarization.

Given these wedges, the optimal policy is then most intuitively framed in terms of a set of taxes and subsidies that incentivize workers to move to cities where their spillovers are larger. Put another way, the planner internalizes the wedge between the private and social marginal productivity of workers. At the same time, a utilitarian planner also attempts to balance gains between different types of workers. Proposition 1 provides an exact characterization of this spatial policy.

Proposition 1. *If the planner's problem is globally concave, the optimal allocation can be achieved by a set of city-invariant labor taxes, t_L^k, city-specific labor subsidies, Δ_n^k, and transfers, R^k, such that*

$$P_n C_n^k = (1 - t_L^k)(w_n^k + \Delta_n^k) + \chi^k + R^k,$$

where $t_L^k = \frac{1}{1+\nu}$, and R^k is such that

$$\phi^k U'(v^k) v^k L^k = \sum_n P_n C_n^k L_n^k.$$

The proposition generalizes a key insight in Fajgelbaum and Gaubert (2020) to a multi-industry environment. Because spillover elasticities are only occupation-specific, one need not keep track of sectoral employment shares in order to determine the optimal subsidy. Therefore, even with multiple industries, the optimal policy collapses to the special case in which occupational shares and wages become sufficient statistics.[28] However, unlike Fajgelbaum and Gaubert (2020), because the difference between workers' private and social marginal value depends on industry composition, the optimal allocation depends in important ways on local industry-specific productivity and production linkages. The global concavity condition on $\mathcal{W}$ in equation (7) generally depends on the concavity of U in the planner's objective. In our numerical results, we assume that U is sufficiently concave to guarantee that Proposition 1 holds.

The resulting planner's solution tells us that households' consumption differs from that implied by their budget constraint in two ways. First, the planner's solution depends on the social marginal product of labor, given by $w_n^k + \Delta_n^k$, rather than its private counterpart. Second, in the planner's solution, consumption increases less than one-for-one with the (social) marginal product of labor. This emerges because workers have to trade off amenities against consumption when choosing a location. In contrast, the planner can allow some workers to live in high-amenity locations while enjoying the high consumption produced by those working in high-productivity locations. Because the planner cannot distinguish between workers with different amenity draws, it trades off this spatial smoothing of consumption against the incentives to have workers live in high-productivity

[28] In Section 7.2 of the Appendix, we relax the assumption of industry invariant elasticities and describe how one would account for such heterogeneity. More generally, when spillover elasticities are industry-specific, the optimal subsidy requires that the planner keep track of sectoral employment shares but not the details of intersectoral linkages.

places.[29]

In the optimal allocation, workers also receive transfers (R^k in the implementation equation presented in Proposition 1) that are dependent on worker type but not on where they live. The planner adjusts those to ensure that welfare for different worker types increases in a way that is consistent with the weights ϕ_k.[30]

4.2 The Value of Social Wedges Across Cities and Occupations

The wedge between the social and private marginal product of labor, Δ_n^k in equation (8), may be computed for each city and each occupation. Figure 2 shows the differences between the wedges corresponding to CNR and non-CNR workers. These differences illustrate how much total consumption increases by switching one type of worker for the other across cities, over and above any wage differences. The wedge difference is positive in all cities, with the population average difference being fairly large at \$63,182 dollars per worker, larger than the average wage difference between CNR and non-CNR workers. This sizeable average difference in wedges follows in part from the relative scarcity of CNR workers. Reallocating these workers more productively across space then also leads to a sizable efficiency gain. Furthermore, this gain suggests that education and migration policies that create and attract CNR workers can potentially have high social value. Here, however, we take the supply of CNR and non-CNR workers as given.

The correlation between the wedge differential illustrated in Figure 2 and the CNR share is 0.36. The analogous correlation with the log population is 0.72. Externalities from CNR workers appear to be particularly large in coastal areas close to New York and in California, as well as Houston and Dallas. They appear less pronounced in non-CNR intensive Florida and, more broadly, in the South and Mid-West (except, modestly, in Chicago).

5 Quantifying the Optimal Allocation

In computing the optimal allocation, we set the Pareto weights, ϕ^k, such that gains under the planner's solution are proportionately equal for both types of workers. Figure 3 shows the percentage change in employment in the optimal allocation relative to the equilibrium allocation for CNR workers. The figure shows that the optimal allocation features a greater concentration of CNR workers in larger cities relative to the decentralized equilibrium, thereby intensifying the spatial polarization of occupations. This rise in spatial polarization follows in part from the spillover coefficient estimates in Section 3.6. Those underscored that CNR workers become more productive when

[29] See Davis and Gregory (2020) for a critique of this argument. Fajgelbaum and Gaubert (2020) show that heterogeneity in preferences induces the same optimal tax as an isoelastic negative spillover in amenities.

[30] As such, because the planner's problem only features real variables, transfers depend on how the returns to the national portfolio of land and structures are shared, χ^k. We should note that the system of taxes and transfers we describe is just one way of implementing the optimal allocation. For given χ^k's, the planner chooses transfers, R^k's, to achieve a particular real allocation of consumption summarized by $\phi^k U'(v^k) v^k L^k = \sum_n P_n C_n^k L_n^k$ in Proposition 1. Alternative distributions of the returns to land and structures, therefore, would imply alternative transfers, R^k's, consistent with the same real optimal allocations. As an example, one could imagine a real estate tax on all rental income that is redistributed in such a way that leaves optimal choices unchanged.

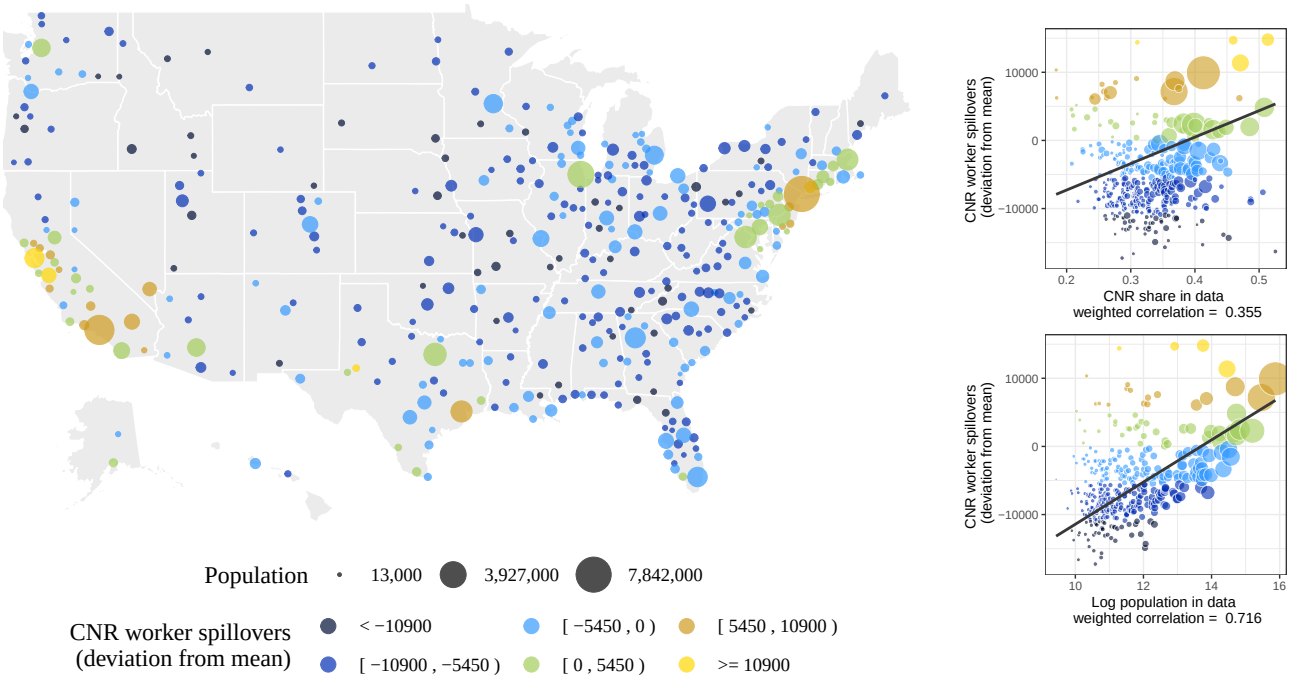

Figure 2: $\Delta_n^{\text{CNR}} - \Delta_n^{\text{non-CNR}}$

Δ_n^{CNR} captures the wedge between the social and the private marginal product of labor for CNR workers, as described in Lemma 2 and $\Delta_n^{\text{non-CNR}}$ for non-CNRs. The figure depicts deviation from the employment weighted average ($63,182) Each marker in the map refers to a CBSA. Marker sizes are proportional to total employment in each city.

clustered with other CNR workers. As discussed below, the degree of heterogeneity in industrial composition as well as linkages between industries also play a key role in driving the optimal policy.

Figure 3 shows that increases in CNR workers under the optimal allocation are particularly large in cities like New York, San Francisco, or San Jose where the difference in wedges between social and private marginal products of labor is especially large. These cities, together with other relatively large agglomerations including Chicago, Dallas, Houston, and Los Angeles that are somewhat less CNR intensive, become cognitive hubs under the optimal allocation. More generally, as shown in the scatterplots of Figure 3, the optimal policy creates cognitive hubs in larger cities that are already CNR abundant under the decentralized equilibrium. Because trade is costly, cities that gain CNR workers are somewhat uniformly distributed in space according to overall economic activity, including in places such as Denver, Minneapolis, and Atlanta. They constitute cognitive hubs in that they absorb CNR workers and are now surrounded by smaller cities with more non-CNR workers.[31]

Figure 4 shows the corresponding figure for non-CNR workers. It illustrates that while the planner generally chooses to incentivize non-CNR workers to move to smaller cities, a few large cities do nevertheless become more non-CNR abundant under the optimal allocation. This is the

[31] See Section 6.2 for an exercise showing how trade costs matter for optimal policy.

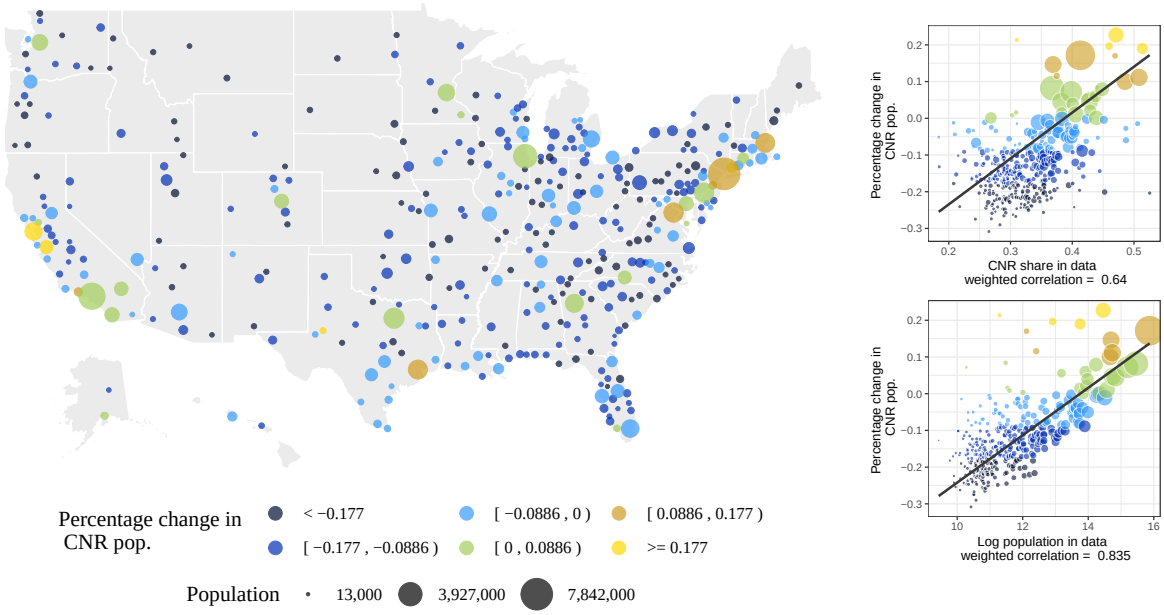

Figure 3: L_n^{CNR} (percentage change from data equilibrium)

Percentage change in employment of CNR workers between equilibrium and optimal values. Each marker in the map refers to a CBSA. Marker sizes are proportional to total equilibrium employment in each city.

case for Miami, Las Vegas, Phoenix, and San Antonio among others, where the wedge differential is not as large in the decentralized equilibrium. These cities become new non-CNR centers. They specialize in non-CNR-intensive industries such as accommodation and retail and grow in size due to a large inflow of non-CNR workers.

While the share of CNR workers increases in large cities under the optimal allocation, Figure 5 also shows that these cities lose overall population. In contrast, smaller cities increase in size. New cognitive hubs, therefore, emerge alongside growing small and non-CNR-abundant cities. Put another way, the city size distribution evens out under the optimal allocation. This feature recognizes that while the productivity of CNR workers increases with the number of those workers, congestion also increases with city size. Furthermore, heterogeneous location preferences imply that attracting CNR workers to a given city becomes increasingly difficult on the margin.

5.1 The Effects of Optimal Policy on Industrial Specialization

Along with cities becoming more even in size under the optimal allocation, we observe that both small and large cities generally increase their degree of industrial specialization while medium-size cities tend to move towards greater industrial diversification. Figure 6 highlights this pattern using changes in the Gini coefficient associated with the distribution of wage bill shares across industries. The figure illustrates how these changes vary with the share of CNR employment. The relationship is U-shaped: in the efficient allocation, cities with low and high CNR worker shares become more specialized. In contrast, cities with intermediate CNR shares exhibit zero

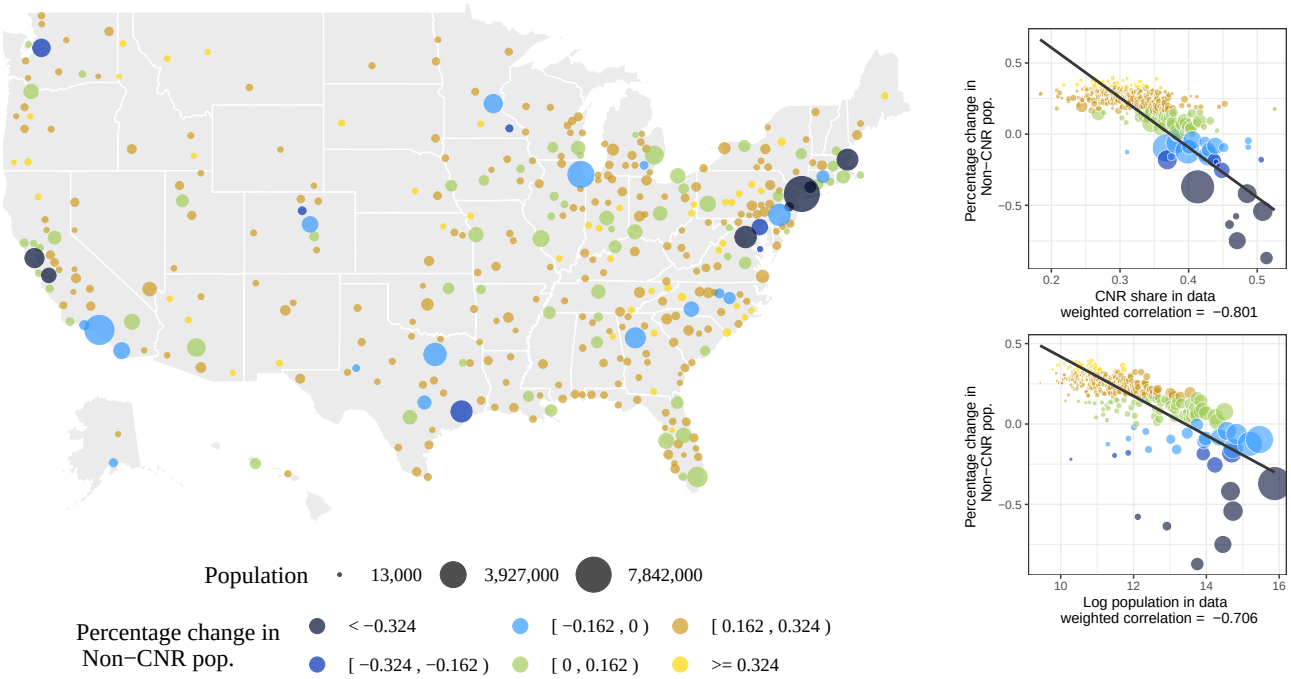

Figure 4: $L_n^{\text{non-CNR}}$ (percentage change from data equilibrium)

Percentage change in employment of non-CNR workers between equilibrium and optimal values. Each marker in the map refers to a CBSA. Marker sizes are proportional to total equilibrium employment in each city.

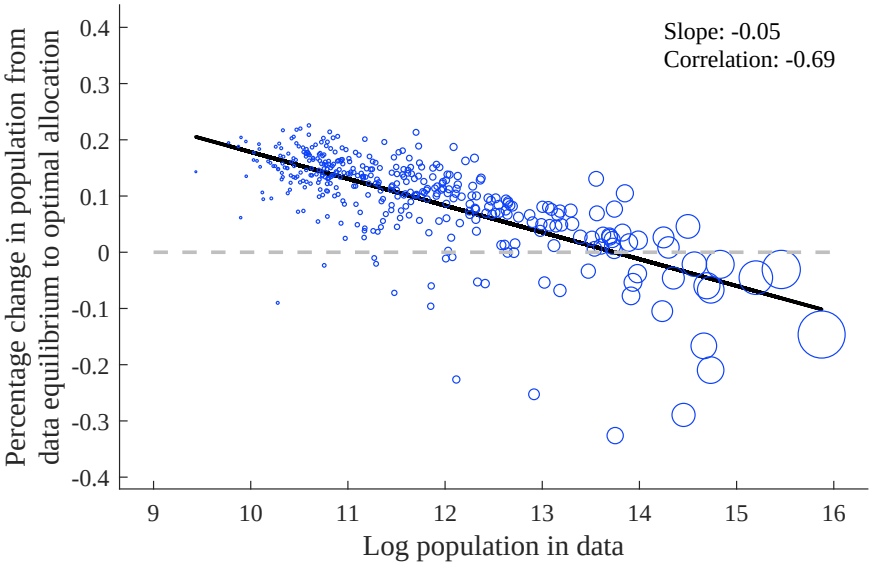

Figure 5: Percentage change in L_n from equilibrium to optimal

Note: Each observation refers to a CBSA. Marker sizes are proportional to total employment.

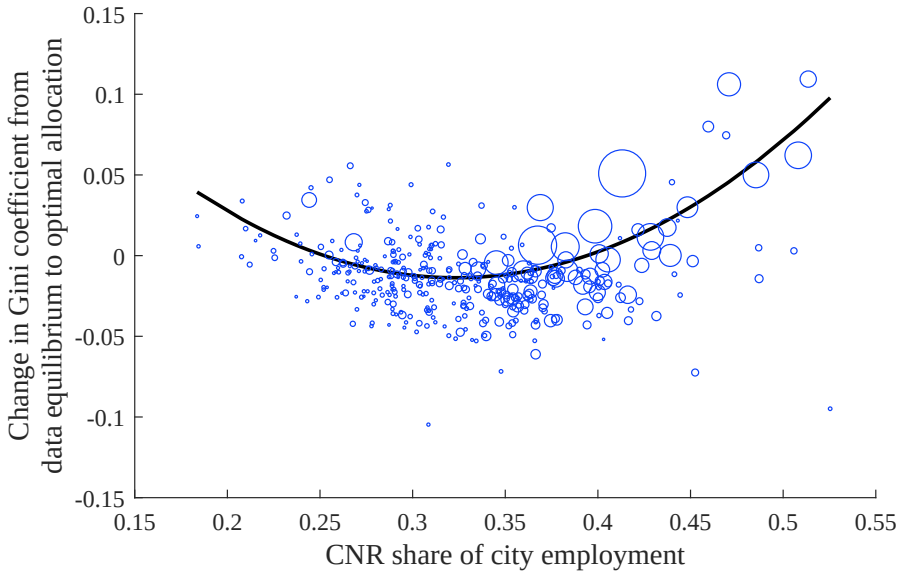

Figure 6: Changes in the Gini coefficient between the data and optimal allocation.

Note: Each observation refers to a CBSA. Marker sizes are proportional to total employment. The solid-black line is a cubic fit on the data. The Gini is constructed using the Lorenz curves depicting within city wage bill and industry rank.

or negative changes in Gini coefficients, indicating no change or greater industrial diversification in those locations. This finding emerges because concentrating occupations is more valuable in cities whose industries employ a specific occupation intensively. The planner's solution, therefore, prescribes further expanding industries in cities that have a more extreme skill mix. Moreover, these cities tend to be at either end of the size distribution so that the U-shaped relationship shown in Figure 6 also holds, though somewhat attenuated, with respect to population (see Figure A16 in Section 8 of the Appendix).

As examples of how efficient allocations change the industrial composition landscape, we highlight two cities at either end of the CNR share distribution. At one end, San Francisco, CA, with close to 2 million workers, sees its share of CNR workers increase from 47 percent to 81 percent. This change reflects an increase in specialization summarized by a 0.1 change in the industrial Gini coefficient. It captures in part an increase of 15 percentage points in the employment share of San Francisco's top industry, professional and business services, and a 12 percentage point increase in that industry's wage bill share. At the other end, Harrisonburg, VA, with only 50,126 workers, sees instead its share of non-CNR workers increase from 73 percent to 83 percent. This change stems from the planner emphasizing employment in the industries in which Harrisonburg's non-CNR workers are already intensively employed; hence, the industrial Gini coefficient increases by 0.04 under the optimal allocation. The employment share in Harrisonburg's top industry, retail, construction, and utilities, increases by 2 percentage points while its wage bill sees a 3 percentage point rise.

In the middle of the city size distribution, changes in the industrial Gini coefficient associated

with wage bill shares across industries are generally close to zero or negative. In other words, many of those medium-sized cities either stay with or diversify their industrial mix. For example, Scranton, PA, a city of around 234,000 workers, sees its Gini coefficient fall by around 0.05 as its employment spreads out across more industries. The wage bill share of its top industry, namely health, declines by 7.5 percentage points while the wage bill share of its new largest industry, retail, construction, and utilities, increases by 2.1 percentage points.

5.2 Taxes and Transfers: Implementing the Optimal Allocation

Implementing the optimal allocation involves transfers specific to each occupation and location. With those transfers, total consumption is equal to total income in every city. Importantly, transfers are optimal and serve two key functions. First, they incentivize CNR workers to live in cognitive hubs and non-CNR workers to further take advantage of the cheaper cost of living in smaller towns. Second, they guarantee that relative welfare gains are the same across occupations.

To achieve equal welfare gains across occupations, the optimal transfer scheme in Proposition 1 has two components. One is related to wedges and incentivizes agents to go to the 'right' location, $\frac{\nu}{1+\nu}\Delta_n^k - \frac{1}{1+\nu}w_n^k$. The other is a fixed transfer by occupation, R^k. This fixed transfer guarantees that all workers obtain equal gains from moving to the planner's allocation. This fixed transfer amounts to a payment of \$15,255 by CNR workers that finances a positive transfer of \$16,872 to all non-CNR workers. Put another way, the optimal transfer scheme requires in part a redistribution between occupations that CNR workers willingly accept in exchange for the formation of the cognitive hubs where they can thrive.

The optimal transfers, $\frac{\nu}{1+\nu}\Delta_n^k - \frac{1}{1+\nu}w_n^k$, incentivize CNR workers to move to larger cities where they tend to already concentrate, and for non-CNR workers to move to smaller cities already specializing in non-CNR-intensive industries. Because externalities are occupation-specific, this reallocation yields larger CNR productivity increases in larger agglomerations. This raises CNR wages and mitigates any need for large net transfers to those workers.

Optimal net transfers, after incentives are taken into account, are given by $R^k + \frac{\nu}{1+\nu}\Delta_n^k - \frac{1}{1+\nu}w_n^k$ which are depicted in Figures 7 and 8.[32] The net average contribution of CNR workers is \$3,324. Observe that CNR workers do not need to be particularly incentivized to stay in the large CNR-abundant cities. In fact, as Figure 7 shows, transfers decrease slightly with the CNR share and city size. The increase in productivity and, therefore, wages that result from enhanced externalities in cognitive hubs is sufficient to attract those workers. CNR workers are net recipients of transfers in only 8 percent of locations but net payers in 92 percent of locations. On the whole, net payments from CNR workers range from \$312 (in the 10th percentile of the population) to \$4,433 (in the 90th percentile).

Once incentive-based transfers are taken into account, non-CNR workers receive on average a net transfer of \$2,025, ranging from \$1,824 (in the 10th percentile of the population) to as much as \$10,531 (in the 90th percentile). Because cognitive hubs offer a high wage to non-CNR workers,

[32] More specifically, net transfers in Proposition 1 are also such that $R^k + \frac{\nu}{1+\nu}\Delta_n^k - \frac{1}{1+\nu}w_n^k = P_n C_n^k - w_n^k - \chi^k$.

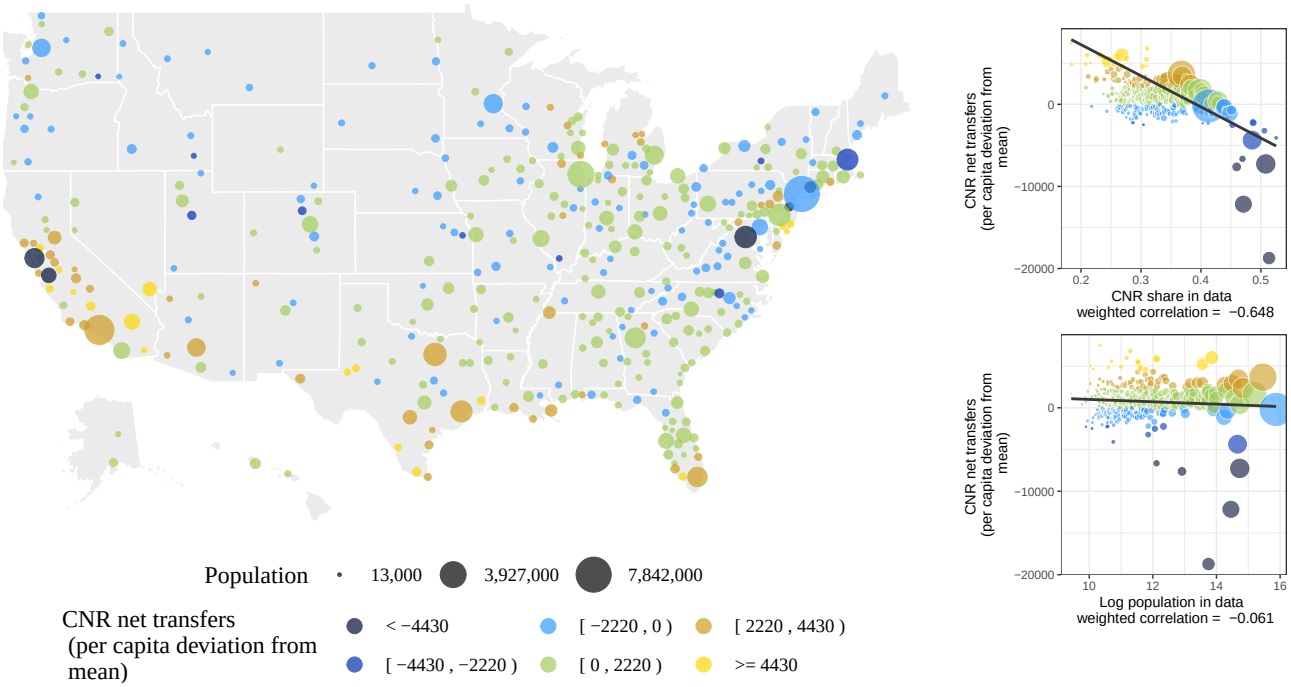

Figure 7: Optimal transfers to CNR workers (per CNR worker)

Note: Optimal transfers per CNR workers are defined as the difference in the optimal allocation between the value consumed and the income they would receive given optimal wages and rents but absent the transfers ($P_n C_n^{\text{CNR}} - w_n^{\text{CNR}} - \chi^k$). The figure depicts deviations from the employment-weighted average ($ -3,316). Each marker in the map refers to a CBSA. Marker sizes are proportional to total equilibrium employment in each city.

those workers also accept a smaller net transfer to discourage other non-CNR types from joining them. This observation accounts for the wide range of non-CNR transfers and reduces the average net burden on CNR workers.

The overall gain in welfare from implementing the optimal allocation amounts to 0.59 percent of GDP for workers in both occupations. As recognized by Kline and Moretti (2013), one can have large local externalities and no aggregate gains if spillovers are iso-elastic, because the gains from inflows in some locations cancel the losses from outflows in others. In our multisector model with occupation specific externalities a similar logic moderates, but does not eliminate, welfare gains. These welfare gains are larger than those found in recent work by Bartelme et al. (2019), 0.4 percent of GDP, in a study of optimal industrial policy in a multi-country trade model. One further important factor moderating welfare gains is that the observed equilibrium allocation is already fairly polarized, so standard envelope theorem arguments imply that one may not have large gains even if the changes in allocation are substantial. In fact, in the next section, we explore features that have led to this fairly polarized state starting from the prevailing conditions in 1980. Before we do so, we benchmark the optimal spatial policy with existing government policies and possible extensions.

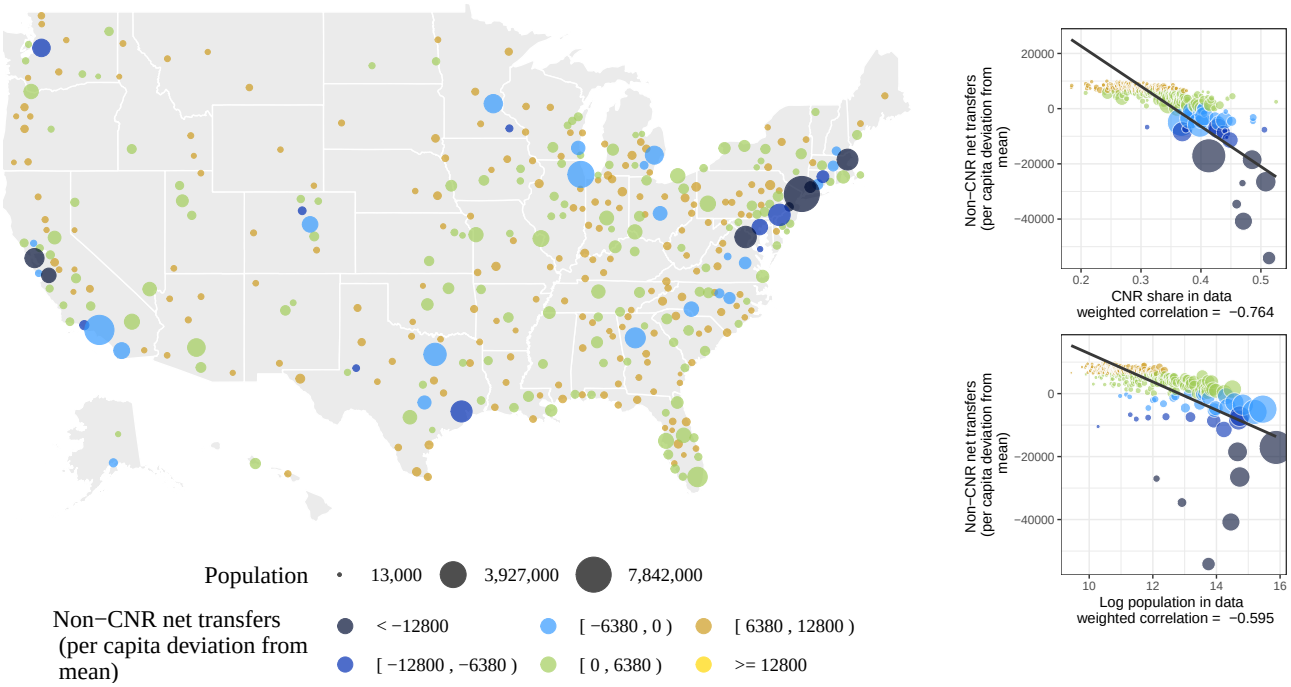

Figure 8: Optimal transfers to non-CNR workers (per non-CNR worker)

Note: Optimal transfers per non-CNR workers are defined as the difference in the optimal allocation between the value consumed and the income they would receive given optimal wages and rents but absent the transfers ($P_n C_n^{\text{non-CNR}} - w_n^{\text{non-CNR}} - \chi^k$). The figure depicts deviations from the employment-weighted average ($2,020). Each marker in the map refers to a CBSA. Marker sizes are proportional to total equilibrium employment in each city.

Benchmarking to existing policies

The size of the recent COVID-19 pandemic assistance programs suggests a practical benchmark regarding the US government's capacity to transfer resources to individuals. These programs also constitute an interesting example of how transfer policies are implemented in practice. In the period after March 2020, a household consisting of two adults and two children between the ages of 6 and 17 and earning less than $75,000 dollars per adult received $6,200 per adult, or $12,400 per household, from the three Economic Impact Payments (the "checks") and child tax credit programs. This compares with an average transfer of about $2,000 per non-CNR worker under the optimal policy reported above. The benchmark is potentially even larger given other concurrent proposals to make some of these programs permanent, such as the narrowly defeated child tax credit program.[33]

[33]The main distinction between our proposed optimal policy and public policies that have been or are under consideration is their dependence on income. In particular, the latter policies feature phase-outs that serve as a common money-saving device in programs aiming to reduce poverty. When applied at somewhat high income levels, these amount to an effective tax on higher-earning individuals. Even before the COVID pandemic, phase-outs at such relatively high income levels were already in place for higher-education provisions such as the Student Loan Interest Deduction (phasing out in 2020 for incomes above $ 82,350) or the American Opportunity Tax Credit (phasing out at $80,000). In contrast, optimal transfers in this study are higher for CNR workers living in higher-wage locations. Those and other similar phase-out provisions are provided on the tax policy center website, https://www.taxpolicycenter.org/briefing-book/how-do-phaseouts-tax-provisions-affect-taxpayers

Additional Considerations

It is worth emphasizing that the optimal allocation may not account for the external effects of population composition on amenities (see Diamond (2016)). In particular, high-earning workers may enjoy living next to other high-earning workers for reasons other than productivity spillovers. If so, spatial sorting would only be more pronounced both in equilibrium and under our optimal policy. We explore the role of amenity externalities in Section 6.3 of the Appendix.[34]

One might also consider alternative non-log-linear effects of worker externalities. This feature may be especially relevant in cities that, per the optimal policy, end up with high shares of one or the other occupation. However, equation (4) implies that $\lim_{L_n^k/L_n \to 0} \frac{\partial \log \lambda_n^k}{\partial L_n^k} \to \tau^R$ whereas $\lim_{L_n^k/L_n \to 1} \frac{\partial \log \lambda_n^k}{\partial L_n^k} \to \tau^L$. Hence, given our findings that $\tau^L < \tau^R$, as the share of CNR workers increases in a given city, their spillover to other CNR workers declines from τ^R to τ^L. Thus, our functional form assumption for externalities implies that the planner stays away from extreme values of CNR or non-CNR shares of employment. In that sense, our functional form assumption favors a comparatively conservative policy prescription.

Finally, the optimal allocation is calculated in a static model that allows for moving costs in the form of heterogeneous preferences for location. These preferences, in turn, help determine the size of transfers needed to compensate workers for moving to an otherwise less preferred city under the optimal policy. However, the model abstracts from repeated moving costs that would necessarily arise in a dynamic setting as workers transition from place to place. Moreover, these transitions may each be associated with capital gains and losses on real estate holdings that are challenging to track in a spatial setting but may nevertheless ease or hinder moving decisions. For now, we leave these considerations to future research.

6 Probing the Model Features

The model economy includes a rich array of features including, for example, production linkages, industry heterogeneity, and sector-specific trade costs. We now present a set of exercises that probe what role those features play in determining key aspects of optimal policy.

6.1 Industrial Linkages and Cognitive Hubs

The presence of industrial production linkages is, in fact, essential for the formation of cognitive hubs under the optimal policy and absent from previous work. This is the case despite the optimal tax formula not appearing at first glance to depend on the industrial composition of different cities. The reason is that the arguments of the optimal tax formula are endogenous. Therefore, any change in any aspect of industry heterogeneity that leads to a change in wages and employment will also lead to a change in optimal taxes.

[34] Moreover, we believe that allowing for endogenous amenities as measured by Diamond (2016) could lead to some "double counting" of externalities since some of the amenity externalities that she documents are also reflected in productivity (for example the density measures in her "retail index" and quality of jobs (measured by patents per capita in her "job index").

To illustrate the role of industrial composition in the optimal mixing of worker types across cities, we carry out a counterfactual exercise that eliminates heterogeneity in linkages between sectors and makes them all symmetric. We then characterize the difference between the equilibrium and optimal allocations under this counterfactual parameterization.

Observe first that our multi-sector, multi-region, economy does not admit a representative aggregate sector irrespective of parameter values. In other words, there is no aggregation theorem that applies in our model. Nonetheless, we can explore the role of sectoral heterogeneity by studying a symmetric version of the economic environment where all sectors matter equally as providers and purchasers of inputs to and from other sectors. That is, we treat all sectors (other than real estate, which is not traded) identically. Thus, we set value-added shares and the share of real estate services to be the same for all $J-1$ sectors and divide the remaining gross output equally across inputs purchased from those same $J-1$ sectors.[35]

The results from this exercise are depicted in Figure 9. Doing away with sectoral heterogeneity yields an optimal policy that implies a U-shaped relationship between changes in CNR population and city size. Namely, with uniform sectoral linkages, the optimal policy leads to "cognitive hubs" in both the smallest and the largest cities. This finding also emerges in Fajgelbaum and Gaubert (2020), who derive the optimal policy in a framework without industry heterogeneity. Importantly, they make clear that the qualitative nature of optimal allocations should depend crucially on the degree of sectoral heterogeneity in the economic environment and on the nature of cross-sectoral linkages. The map in Figure 9 shows in red the set of cities where the percentage change in the number of CNR workers in the optimal allocation is 20% larger without than with heterogeneous industrial linkages (based on the difference between gray and blue markers for each city in the scatterplot). Namely, these are the set of cities that we would mistakenly designate as cognitive hubs if we ignore the industrial linkages in the US. Clearly, this includes many very small cities in the Midwest and on the East and West coasts.

What drives the U-shaped relationship in this counterfactual exercise? The exogenous component of productivity in some CNR-intensive sectors, such as health and education, declines with city size while in other CNR-intensive sectors, such as communications and financial services, it increases with city size. Therefore, absent linkages, the planner has good reasons to have some small cities specialize to a greater degree in health and education, thus also concentrating some CNR workers in those cities while letting larger cities specialize in communication or financial services.

Heterogeneous linkages across industries overturn this mechanism for small cities. In particular, CNR-intensive industries often use materials from other CNR-intensive industries. The planner, therefore, aims to keep these industries together in large cities to take full advantage of externalities among CNR workers. Heterogeneous linkages work towards keeping CNR-intensive sectors close to one another in the optimal policy.

[35] Said differently, these exercises do away with heterogeneity in production linkages across all sectors other than real estate. Specifically, we set γ_n^j $\forall j$, other than real estate, equal to the value-added weighted average of γ_n^j for each n. Similarly, for each n, we set $\gamma_n^{j'j} = \frac{1-\gamma_n^j-\gamma_n^{\text{real estate},j}}{J-1}$ $\forall j, j'$.

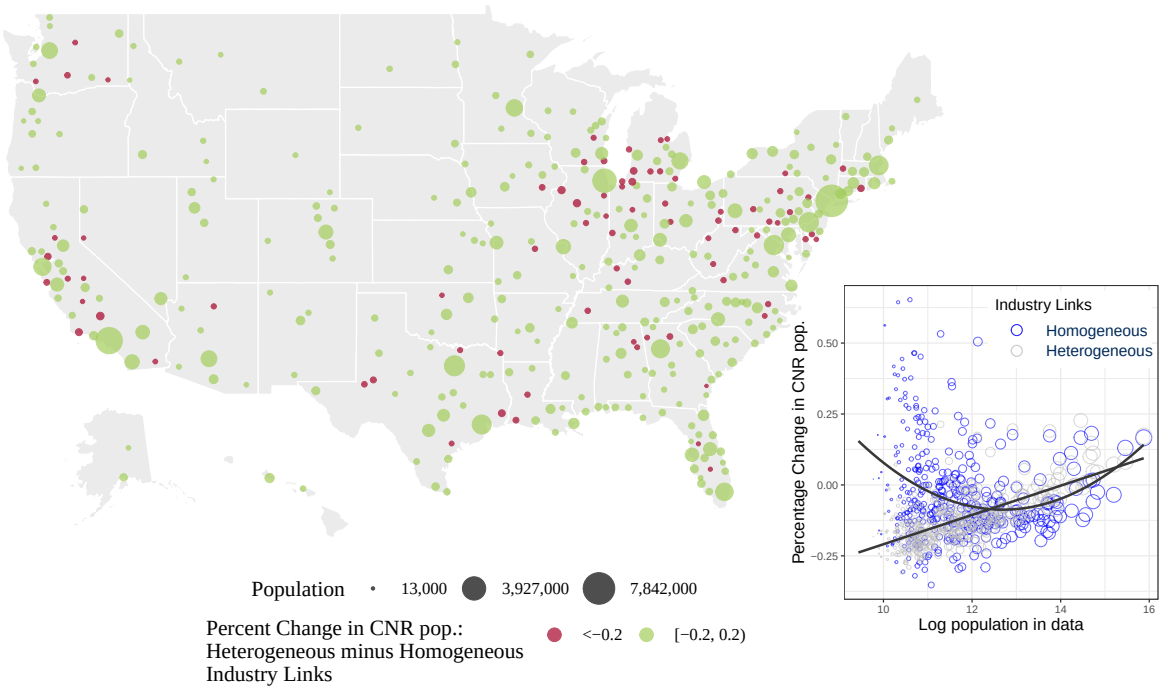

Figure 9: Optimal Policy in Baseline vs Counterfactual with Homogeneous Industry Links

Note: Comparison between counterfactual equilibrium and optimal allocations. Each observation refers to a CBSA. Markers on the map refer to the difference in the percentage change in the CNR population implied by the optimal policy in the baseline with heterogeneous linkages and the change prescribed by the one in the counterfactual with homogeneous industry links. The solid lines in the scatterplot are either quadratic or linear fits to the data. Marker sizes are proportional to total employment.

6.2 Transportation Costs

To clarify the role of transportation costs in the optimal allocation, we now present an exercise where we set trade costs in all sectors (other than housing) to zero. As before, we compare the optimal allocation given this counterfactual parameterization with the corresponding equilibrium allocation. The results are presented in Figure 10. Trade costs play a key role in the geographic distribution of cognitive hubs. Without trade costs, optimal policy tends to concentrate CNR workers mostly in coastal cities such as New York, Washington D.C., or San Francisco, as well as smaller nearby towns. As a result, the formation of inland regional hubs is simultaneously reduced in places like Chicago, Minneapolis, or Dallas.

More generally, eliminating trade costs does not materially affect how the optimal reallocation correlates with city size.[36] The reason involves two offsetting forces. On the one hand, absent trade costs, the optimal policy allows for a large increase in CNR intensity in some smaller cities that can now easily trade their output with other locations. On the other hand, eliminating trade costs also allows CNR-intensive sectors to further concentrate in locations where they are already productive, namely larger cities where CNR labor is already relatively abundant and thus cheaper. Crucially, this push towards a higher concentration of CNR workers in those cities is then further reinforced

[36] See Figure A5 in Section 6.2 of the Appendix

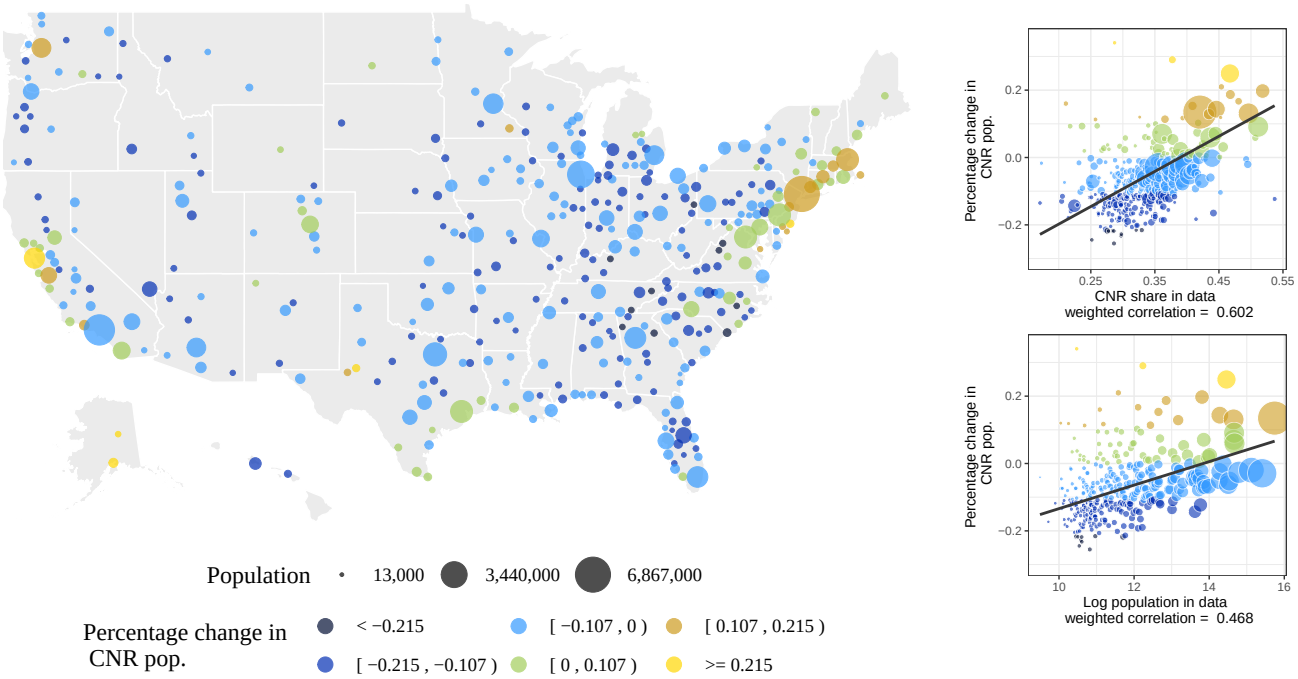

Figure 10: Change in L_n^{CNR} from no trade costs equilibrium to optimal allocation

Percentage change in employment of CNR workers between equilibrium and optimal values. Each marker in the map refers to a CBSA. Marker sizes are proportional to total equilibrium employment in each city. In the scatterplots, two outlier observations are not pictured for clarity. These are Atlantic City, NJ, and Midland, TX.

through externalities. The net effect of these two countervailing forces is to leave our benchmark findings relatively unchanged.

7 The Formation of Cognitive Hubs after 1980

The US economy has evolved towards the formation of cognitive hubs at least since the 1980s. Quantifying our model to 1980 data and comparing the results to our baseline period allows us to study this phenomenon in detail. In particular, a series of counterfactual exercises makes the case for the key role of externalities and the formation of cognitive hubs in generating widespread gains across the US economy. At the same time, these exercises also highlight limits on these gains implied by restrictive housing supply regulations implemented in these same hubs.

To compare the spatial structure of the economy in 1980 to that in 2011-15, we abstract from aggregate technology and population trends. Thus, we first build a 'Baseline' 1980 economy comparable to that in 2011-15 with respect to overall population, input shares, and aggregate productivity levels, but similar to 1980 in location and occupational-specific characteristics.[37] In particular, the Baseline economy includes neither location nor occupation-specific changes in productivity across industries relative to the original 1980 allocation. It further leaves amenities unchanged relative to

[37] For details on the construction of this Baseline economy and other counterfactuals, see Section 5 of the Appendix.

that year. We then study the role of these components by considering their implications one at a time.

In this Baseline 1980 economy, the US population is concentrated in cities with close to an average CNR share of employment for that year. In the 2011-15 data, however, population is more dispersed around a now larger average. More concretely, in 1980 about 2% of the US population lived in cities with CNR share of employment more than 10 percentage points away from the national average, but in 2011-15 12.5% did. In comparison, the distribution of CNR workers implied by the 2011-2015 planner's solution calculated above implies that 42% of the US population would live in such cities. Together, these observations underscore that the concentration of CNR workers increased in the direction implied by today's optimal allocation.

Table 3: Welfare Comparison, Relative to Baseline

	CNR	non-CNR	CNR-to-non-CNR Ratio
Full Model			
ratio of CNR to non-CNR welfare in Baseline		*1.91*	
1. Baseline + change in occ. shares in employment	1.045	1.058	0.987
2. (1) + change in local technology (ex real estate)	1.066	1.079	0.988
3. (2) + change in real estate productivity	1.042	1.060	0.983
4. (3) + change in amenities (2011-15 parameters)	1.039	1.061	0.980
5. Optimal Allocation	1.045	1.067	0.980
6. 2011-15 parameters minus change in real estate productivity	1.054	1.070	0.985
7. Optimal Allocation with parameters in (6)	1.064	1.081	0.985
Model without externalities			
ratio of CNR to non-CNR welfare in Baseline		*3.77*	
8. Baseline + change in occ. shares in employment	0.858	1.110	0.773
9. (8) + change in local technology (ex real estate)	0.852	1.116	0.763
10. (9) + change in real estate productivity	0.860	1.129	0.762
11. (10) + change in amenities (2011-15 parameters)	0.861	1.129	0.762
12. Optimal Allocation	0.863	1.132	0.762
13. 2011-15 parameters minus change in real estate productivity	0.846	1.109	0.763
14. Optimal Allocation with parameters in (13)	0.849	1.112	0.763

To better understand this spatial evolution of workers, we carry out a series of counterfactual exercises that clarify the importance of different forces in driving national trends. Table 3 shows the cumulative effects of these exercises on the welfare of CNR and non-CNR workers. The columns depict welfare levels relative to the Baseline 1980 economy for each occupation. Given the quantitative spatial model we developed, one can also examine the role of local spillovers by carrying out

the same exercises absent externalities. Findings from a world without externalities are shown in the lower panel (i.e., where the externality elasticity parameters are set to zero).

Starting from the Baseline 1980 economy, Rows 1 and 8 in Table 3 change the composition of employment to 2011-15 levels, with more CNR workers and fewer non-CNR workers. Absent externalities, the model implies that CNR workers experience significant losses as they become relatively more abundant. Not surprisingly, the increase in the supply of CNR workers implies a reduction in their relative wages. The welfare of CNR workers then ends up 14 percent below its baseline counterfactual. Conversely, the welfare of non-CNR workers rises by 11 percent relative to its baseline. In sharp contrast, when allowing for externalities among workers, both CNR and non-CNR workers end up gaining about the same as the new CNR workers generate positive productivity spillovers and occupations become more polarized in space.

Rows 2 and 9 add exogenous changes to local technology (other than for the real estate sector studied separately) over and above those implied by average national trends. The implied gains are larger in cities that had high CNR shares in 1980.[38] The sustained productivity growth experienced by CNR-intensive sectors in large cities is documented and explored in detail by Duranton and Puga (2005) and Eckert et al. (2020), among others. These studies underscore how information technology enhanced the productivity of, respectively, CNR-intensive service industries and, within all industries, CNR-intensive management functions. As such, exogenous changes to local technology capture, for example, the fact that computers and electronics became particularly more productive in San Jose while finance became particularly more productive in New York. These location-specific technological changes interact with externalities to increase the welfare of both types of workers.

Rows 3 and 10 add changes to productivity in real estate. This exercise embodies the effects of two different underlying processes. On the one hand, real estate productivity increased to a greater extent in fast-growing cities, with the stock of housing rising to accommodate growing populations. On the other hand, as emphasized in Glaeser and Gyourko (2018), Herkenhoff et al. (2018), and Hsieh and Moretti (2019), housing regulations impeded development in some very productive areas. The table shows that the net effect of these two forces would have been positive in the absence of externalities. Housing development would have then accommodated the increasing population in many fast-growing locations. However, this net effect becomes negative once externalities are taken into account since housing productivity lagged in CNR-intensive cities.[39] These findings indicate that housing restrictions are likely not a productive avenue for creating cognitive hubs since they tend to hinder growth in precisely the locations most amenable to such an evolution.[40]

To complete the historical decompositions, Rows 4 and 11 add changes in amenities. In particular, Row 4 corresponds to the 2011-15 equilibrium allocation. Changes in the spatial distribution of amenities appear to add little to total welfare.

[38] The population-weighted correlation between CNR shares and the increase in the exogenous part of technology was 0.395.

[39] The population-weighted correlation between CNR share and increase in housing productivity was -0.36.

[40] Homothetic preferences imply that housing restrictions are unlikely to have had a material effect on local sorting.

In considering the results presented in Table 3, it is important to bear in mind that the specific sequence in which the changes between 1980 and 2011-15 are added can have an effect on the results. We chose to present a sequence that is intuitive to us, but the main findings highlighted above are robust to other sequences.

Aside from these historical decompositions, we explore optimal allocations under two scenarios. First, Rows 5 and 12 correspond to the optimal allocation described in Section 4 with and without externalities. The second scenario assesses the role of housing policy in hindering optimal policy. In particular, we compute the optimal policy under the assumption that housing productivity was distributed as in 1980 (the corresponding equilibrium counterfactual welfare is presented in Rows 6 and 13 and the corresponding optimal allocation in Rows 7 and 14). We find that the increment in welfare is more than 50 percent larger in the latter scenario than when starting from the actual equilibrium. In other words, the optimal policy is considerably less effective than it might otherwise be because of housing supply restrictions that have slowed down growth in cognitive hubs.

8 Concluding Remarks

Since the 80s the US economy has experienced increased skill and occupational polarization across space. Large cities increasingly have more highly educated CNR workers who earn more. In contrast, many medium and small cities have suffered an exodus of skilled workers and experienced persistent population declines. These trends, amplified by local externalities, were also associated with a rise in income inequality between occupations. This growing gap between top and small-sized cities has motivated policymakers and city governments to advocate policies to attract CNR workers to smaller towns in order to reverse their fortunes. Our analysis shows that these efforts could be counterproductive.

Our analysis underscores that while CNR workers are extremely useful, they are also scarce. Furthermore, their productivity is tremendously enhanced by living with other CNR workers. So, attracting them to smaller towns with more mixed populations represents an under-utilization of resources. CNR workers are too valuable for society to be used in this way. A better policy is to reinforce existing trends and let them concentrate in cognitive hubs while incentivizing non-CNR workers to move and help smaller cities grow. Of course, some non-CNR workers are always needed in "cognitive hubs" because of the imperfect substitutability of occupations in production. The result would be smaller, more CNR intensive, "cognitive hubs" in some of today's largest cities. We show that the resulting migration of non-CNR workers that allows small towns to grow may be implemented with a baseline transfer to non-CNR workers, reminiscent of a universal basic income, and a set of occupation-location specific transfers. Overall, CNR workers would transfer resources to non-CNR workers to generate equal welfare gains.

Our findings suggest that efforts to stop the spatial polarization of occupations are misguided. In fact, encouraging it further can yield benefits for everyone when accompanied by the necessary transfers. Implementing these transfers, however, is critical. Otherwise, cognitive hubs might use

other indirect means of pushing out non-CNR workers such as, for example, housing supply constraints, zoning restrictions, or a lack of investment in transportation networks to aid commuting. Such efforts can generate occupational polarization across space *without* generating Pareto gains for all workers.[41] Implementing the necessary transfers would not only help avoid those inefficient policies and benefit CNR workers, but it would also improve the welfare of non-CNR workers and the many small and medium-sized cities where they would end up living and working.

Our analysis abstracts from the role that spatial polarization might have on human capital formation. In principle, the migration of CNR workers towards cognitive hubs may be detrimental to smaller cities in a setting where the learning technology also features meaningful externalities from CNR workers as in Lhuillier (2023) or Crews (2023). At the same time, however, the transfers that CNR workers are able and willing to make to non-CNR workers, given the productivity gains they experience from living in cognitive hubs, might naturally be invested in education and other training in the smaller cities. These transfers, if directed properly, have the potential to ameliorate, or even reverse, the conceivably negative effects of spatial polarization on human capital.

References

Acemoglu, D. and D. Autor (2011). Skills, tasks and technologies: Implications for employment and earnings. In *Handbook of labor economics*, Volume 4, pp. 1043–1171. Elsevier. 3.1

Albouy, D. (2009). The unequal geographic burden of federal taxation. *Journal of Political Economy 117*(4), 635–667. 1

Albouy, D., K. Behrens, F. Robert-Nicoud, and N. Seegert (2019). The optimal distribution of population across cities. *Journal of Urban Economics 110*, 102–113. 3

Anderson, J. E., C. A. Milot, and Y. V. Yotov (2014). How much does geography deflect services trade? canadian answers. *International Economic Review 55*(3), 791–818. 3.3

Angrist, J. D. and J.-S. Pischke (2008). *Mostly harmless econometrics: An empiricist's companion.* Princeton university press. 24

Autor, D. H. (2019). Work of the past, work of the future. In *Richard T. Ely Lecture given at Annual Meeting of American Economics Association, Atlanta, GA, January*, Volume 4. 1

Bartelme, D., A. Costinot, D. Donaldson, and R.-C. Andres (2019). The textbook case for industrial policy: Theory meets data. Technical report, CEPR Discussion Papers. 5.2

Baum-Snow, N. and R. Pavan (2013). Inequality and city size. *Review of Economics and Statistics 95*(5), 1535–1548. 1

[41] The welfare gains from the optimal policy in our setup are small relative to the gains found by Hsieh and Moretti (2019) in part because our exercise does not allow for changes in the supply of land and structures.

Berry, C. R. and E. L. Glaeser (2005). The divergence of human capital levels across cities. *Papers in regional science 84*(3), 407–444. 1

Brinkman, J. (2014). The supply and demand of skilled workers in cities and the role of industry composition. 1

Burstein, A., E. Morales, and J. Vogel (2019). Changes in between-group inequality: computers, occupations, and international trade. *American Economic Journal: Macroeconomics 11*(2), 348–400. 4

Caliendo, L. and F. Parro (2015). Estimates of the trade and welfare effects of nafta. *The Review of Economic Studies 82*(1), 1–44. 2.2, 3.3

Caliendo, L., F. Parro, E. Rossi-Hansberg, and P.-D. Sarte (2017). The impact of regional and sectoral productivity changes on the us economy. *The Review of economic studies 85*(4), 2042–2096. 1, 2.1

Card, D. (2001). Immigrant inflows, native outflows, and the local labor market impacts of higher immigration. *Journal of Labor Economics 19*(1), 22–64. 1, 3.3, 3.6

Ciccone, A. and R. E. Hall (1996). Productivity and the density of economic activity. *The American Economic Review 86*(1), 54. 3.6

Ciccone, A. and G. Peri (2005). Long-run substitutability between more and less educated workers: evidence from us states, 1950–1990. *Review of Economics and statistics 87*(4), 652–663. 3.3

Colas, M. and K. Hutchinson (2017). Heterogeneous workers and federal income taxes in a spatial equilibrium. Technical report. 1

Crews, L. G. (2023). A dynamic spatial knowledge economy. 8

Davis, M. A. and J. Gregory (2020). Place-based redistribution in simple location-choice models. 29

de la Roca, J. and D. Puga (2017). Learning by working in big cities. *The Review of Economic Studies 84*(1), 106–142. 14

Desmet, K. and E. Rossi-Hansberg (2013). Urban accounting and welfare. *American Economic Review 103*(6), 2296–2327. 1

Diamond, R. (2016). The determinants and welfare implications of us workers' diverging location choices by skill: 1980-2000. *American Economic Review 106*(3), 479–524. 1, 3.5, 5.2, 34

Duranton, G. and D. Puga (2004). Micro-foundations of urban agglomeration economies. In *Handbook of regional and urban economics*, Volume 4, pp. 2063–2117. Elsevier. 2

Duranton, G. and D. Puga (2005). From sectoral to functional urban specialisation. *Journal of urban Economics 57*(2), 343–370. 7

Eaton, J. and S. Kortum (2002). Technology, geography, and trade. *Econometrica 70*(5), 1741–1779. 2.2

Eckert, F., S. Ganapati, and C. Walsh (2020). Skilled scalable services: The new urban bias in economic growth. Technical report, CESifo Working Paper. 7

Eeckhout, J. and N. Guner (2015). Optimal spatial taxation: Are big cities too small? 3

Ellison, G., E. L. Glaeser, and W. R. Kerr (2010). What causes industry agglomeration? evidence from coagglomeration patterns. *American Economic Review 100*(3), 1195–1213. 1, 2.2

Fajgelbaum, P. and C. Gaubert (2020). Optimal spatial policies, geography, and sorting. *The Quarterly Journal of Economics 135*(2), 959–1036. 1, 1, 4.1, 29, 6.1

Fajgelbaum, P. D., E. Morales, J. C. Suárez Serrato, and O. Zidar (2018). State taxes and spatial misallocation. *The Review of Economic Studies 86*(1), 333–376. 1, 3.3

Feenstra, R. C., P. Luck, M. Obstfeld, and K. N. Russ (2018). In search of the armington elasticity. *Review of Economics and Statistics 100*(1), 135–150. 3.3

Giannone, E. (2017). Skilled-biased technical change and regional convergence. 1

Glaeser, E. and J. Gyourko (2018, February). The economic implications of housing supply. *Journal of Economic Perspectives 32*(1), 3–30. 7

Handbury, J. and D. E. Weinstein (2014). Goods prices and availability in cities. *The Review of Economic Studies 82*(1), 258–296. 3.5, 21

Head, K. and T. Mayer (2014). Gravity equations: Workhorse, toolkit, and cookbook. In *Handbook of international economics*, Volume 4, pp. 131–195. Elsevier. 3.3

Hendricks, L. (2011). The skill composition of us cities. *International Economic Review 52*(1), 1–32. 1

Herkenhoff, K. F., L. E. Ohanian, and E. C. Prescott (2018). Tarnishing the golden and empire states: Land-use restrictions and the us economic slowdown. *Journal of Monetary Economics 93*, 89–109. 1, 7

Hsieh, C.-T. and E. Moretti (2019). Housing constraints and spatial misallocation. *American Economic Journal: Macroeconomics*. 1, 7, 41

Jaimovich, N. and H. E. Siu (2018). The trend is the cycle: Job polarization and jobless recoveries. *The Review of Economics and Statistics*. 3.1

Kline, P. and E. Moretti (2013). Local economic development, agglomeration economies, and the big push: 100 years of evidence from the tennessee valley authority. *The Quarterly Journal of Economics 129*(1), 275–331. 5.2

Lhuillier, H. (2023). Should i stay or should i grow? 8

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

Marshall, A. (1920). *Principles of Economics (Revised ed.)*. London: Macmillan; reprinted by Prometheus Books. 1

Melo, P. C., D. J. Graham, and R. B. Noland (2009). A meta-analysis of estimates of urban agglomeration economies. *Regional science and urban Economics 39*(3), 332–342. 1, 3.5

Moretti, E. (2004a). Estimating the social return to higher education: evidence from longitudinal and repeated cross-sectional data. *Journal of econometrics 121*(1-2), 175–212. 1, 1, 3.6

Moretti, E. (2004b). Workers' education, spillovers, and productivity: evidence from plant-level production functions. *American Economic Review 94*(3), 656–690. 1, 3.5

Neumark, D. and H. Simpson (2015). Place-based policies. In *Handbook of regional and urban economics*, Volume 5, pp. 1197–1287. Elsevier. 1

Ossa, R. (2015). A quantitative analysis of subsidy competition in the us. Technical report, National Bureau of Economic Research. 1

Redding, S. J. and E. Rossi-Hansberg (2017). Quantitative spatial economics. *Annual Review of Economics 9*, 21–58. 1

Rossi-Hansberg, E., P.-D. Sarte, and F. Schwartzman (2021). Local industrial policy and sectoral hubs. In *AEA Papers and Proceedings*, Volume 111, pp. 526–31. 7

Sanderson, E. and F. Windmeijer (2016). A weak instrument f-test in linear iv models with multiple endogenous variables. *Journal of Econometrics 190*(2), 212–221. 3.6

Stock, J. H. and M. Yogo (2005). Testing for weak instruments in linear iv regression. *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, 80. 3.6