

# Supplemental Appendix to Cognitive Hubs and Spatial Redistribution\*

Esteban Rossi-Hansberg      Pierre-Daniel Sarte      Felipe Schwartzman

February 4, 2026

## Contents

<b>1</b>	<b>Model Details</b>	<b>3</b>
1.1	Household Decisions . . . . .	3
1.2	Firms . . . . .	4
1.2.1	Intermediate Goods Producers . . . . .	4
1.2.2	Final Goods . . . . .	5
1.2.3	Derivation of Prices . . . . .	6
1.2.4	Trade Shares . . . . .	6
1.3	Market Clearing and Aggregation at the Industry and City Level . . . . .	7
1.4	Definition of Equilibrium . . . . .	9
1.5	Aggregate Equilibrium . . . . .	10
<b>2</b>	<b>Quantifying the Model and Model Inversion</b>	<b>11</b>
2.1	Details of City-Invariant Parameters . . . . .	12
2.2	Model Inversion and Granular Parameters . . . . .	14
2.3	Comparisons to Previous Work and Model Validation . . . . .	21
2.3.1	Wages . . . . .	21
2.3.2	Amenities . . . . .	22
2.3.3	Total Factor Productivity . . . . .	23
2.4	Controls for Determinants of Productivity . . . . .	26
2.5	Instrumenting for Employment . . . . .	28
2.6	A check on Instruments: Estimates in Counterfactual Without Externalities . . . . .	29
2.7	Model-Implied IV . . . . .	30

---

\*Rossi-Hansberg: University of Chicago (email: [rossihansberg@uchicago.edu](mailto:rossihansberg@uchicago.edu)), Sarte: Federal Reserve Bank of Richmond (email: [pierre.sarte@rich.frb.org](mailto:pierre.sarte@rich.frb.org)), Schwartzman: Federal Reserve Bank of Richmond (email: [felipe.schwartzman@rich.frb.org](mailto:felipe.schwartzman@rich.frb.org)).

<b>3</b>	<b>The Planner’s Problem</b>	<b>30</b>
3.1	Solving the Planner’s Problem . . . . .	32
<b>4</b>	<b>Characterization of the Planner’s Solution</b>	<b>34</b>
4.1	The Social and Private Marginal Value of Workers of Type $k$ in City $n$ (Proof of Lemma 1) . . . . .	35
4.2	Implementation (Proof of Proposition 1) . . . . .	36
4.3	Robustness: Policy Without a Linear Labor Tax . . . . .	38
<b>5</b>	<b>Quantifying the Model for 1980 and Counterfactual Exercises</b>	<b>39</b>
5.1	Quantifying the Model for 1980 . . . . .	39
5.2	Details of Counterfactual Exercises . . . . .	40
<b>6</b>	<b>Counterfactuals</b>	<b>40</b>
6.1	Homogeneous Linkages . . . . .	40
6.2	No Trade Costs . . . . .	41
6.3	A Counterfactual Economy Without Endogenous Amenities . . . . .	42
<b>7</b>	<b>Robustness</b>	<b>44</b>
7.1	Stress-Testing Externality Parameters . . . . .	44
7.2	Industry-Specific Externality Parameters . . . . .	46
7.3	College Attainment vs. Occupation . . . . .	48
7.4	Restricting CNRs to Graduate Degree Holders . . . . .	49
<b>8</b>	<b>Additional Figures and Tables</b>	<b>53</b>
<b>9</b>	<b>Data</b>	<b>54</b>
9.1	Definitions and Data Sources . . . . .	54
9.2	Employment . . . . .	57
9.2.1	MSA-Industry Employment ( $L_n^j$ ) . . . . .	57
9.2.2	Occupation Split by MSA-Industry . . . . .	58
9.3	Wages . . . . .	60
9.3.1	Accounting for non-wage compensation . . . . .	60
9.3.2	Mincerian Regression . . . . .	61
9.3.3	Adjusted Wages . . . . .	61
9.4	Industry Interactions . . . . .	61
9.5	Trade Costs . . . . .	63
9.6	Prices . . . . .	64
9.7	Joint distribution of prices, $P_n^j$ . . . . .	64
9.8	Industry Level Prices . . . . .	65
9.9	Crosswalks . . . . .	66

9.9.1	Geographic Crosswalks . . . . .	66
9.9.2	Industry Crosswalks . . . . .	66
9.9.3	Occupation Crosswalks . . . . .	68

<b>References</b>	<b>70</b>
-------------------	-----------

# 1 Model Details

## 1.1 Household Decisions

In a given occupation, all households living in the same city choose the same consumption basket. It follows that  $C_n^{kj}(\mathbf{a}) = C_n^{kj}$  for all  $\mathbf{a}$ . Moreover, the demand for good  $j$  by workers in occupation  $k$  living in city  $n$  is

$$C_n^{kj} = \alpha^j \frac{P_n}{P_n^j} C_n^k \tag{A-1}$$

where  $P_n = \prod_{j=1}^J \left( \frac{P_n^j}{\alpha^j} \right)^{\alpha^j}$  is the ideal price index in city  $n$  and  $P_n^j$  is the price of good  $j$  in city  $n$ .

Agents move freely across cities. The value,  $v_n^k(\mathbf{a})$ , of locating in a particular city  $n$  for an individual employed in occupation  $k$ , with idiosyncratic preference vector  $\mathbf{a}$  is

$$v_n^k(\mathbf{a}) = \frac{a_n A_n^k I_n^k}{P_n} = a_n A_n^k C_n^k,$$

where recall that

$$I_n^k = w_n^k + \chi^k. \tag{A-2}$$

In equilibrium, workers move to the location where they receive the highest utility so that

$$v^k(\mathbf{a}) = \max_n v_n^k(\mathbf{a}),$$

where  $v^k(\mathbf{a})$  now denotes the equilibrium utility of an individual in occupation  $k$  with amenity vector  $\mathbf{a}$ . Our benchmark case assumes that  $a_n$  is drawn from a Fréchet distribution. Draws are independent across cities. We denote by  $\Psi$  the joint *cdf* for the elements of  $\mathbf{a}$  across workers in a given occupation, with

$$\Psi(\mathbf{a}) = \exp \left\{ - \sum_n (a_n)^{-\nu} \right\},$$

where the shape parameter  $\nu$  reflects the extent of preference heterogeneity across workers. Higher values of  $\nu$  imply less heterogeneity, with all workers ordering cities in the same way when  $\nu \rightarrow \infty$ .

Assuming that workers of different types can freely move between cities, the average utility of

a worker of type  $k$  is given by

$$v^k = \Gamma\left(\frac{\nu-1}{\nu}\right) \left(\sum_n (A_n^k C_n^k)^\nu\right)^{\frac{1}{\nu}}, \quad (\text{A-3})$$

where  $\Gamma(\cdot)$  is the Gamma function.

The assumption of a Fréchet distribution for idiosyncratic amenity parameters implies closed form expressions for the fraction of workers in each city:

$$L_n^k = \Pr\left(v_n^k(\mathbf{a}) > \max_{n' \neq n} v_{n'}^k(\mathbf{a})\right) = \frac{(A_n^k C_n^k)^\nu}{\sum_{n'} (A_{n'}^k C_{n'}^k)^\nu} L^k. \quad (\text{A-4})$$

Combining this equation with equation (A-4) describing labor supply yields an expression relating the value of each occupational type to consumption and employment in particular locations:

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

## 1.2 Firms

### 1.2.1 Intermediate Goods Producers

Cost minimization implies that input demand satisfies:

$$\frac{r_n H_n^j(\mathbf{z})}{x_n^j(\mathbf{z}) q_n^j(\mathbf{z})} = \gamma_n^j \beta_n^j, \quad (\text{A-5})$$

$$\frac{w_n^k L_n^{kj}(\mathbf{z})}{x_n^j(\mathbf{z}) q_n^j(\mathbf{z})} = \frac{\left(\frac{w_n^k}{\lambda_n^{kj}}\right)^{1-\epsilon}}{\sum_{k'} \left(\frac{w_n^{k'}}{\lambda_n^{k'j}}\right)^{1-\epsilon}} \gamma_n^j (1 - \beta_n^j), \quad (\text{A-6})$$

$$\frac{P_n^{j'} M_n^{j'j}(\mathbf{z})}{x_n^j(\mathbf{z}) q_n^j(\mathbf{z})} = \gamma_n^{j'j}, \quad (\text{A-7})$$

where  $x_n^j(\mathbf{z})$  is the Lagrange multiplier which in this case reflects the unit cost of production,  $r_n$  is local rents per unit of land. We can solve for  $x_n^j(\mathbf{z})$  by substituting optimal factor choices into the production function,

$$x_n^j(\mathbf{z}) \equiv \frac{x_n^j}{z_n} = \frac{B_n^j}{z_n} \left\{ \frac{r_n^{\beta_n^j}}{Z_n^j} \left[ \sum_k \left(\frac{w_n^k}{\lambda_n^{kj}}\right)^{1-\epsilon} \right]^{\frac{1-\beta_n^j}{1-\epsilon}} \right\}^{\gamma_n^j} \prod_{j'=1}^J (P_n^{j'})^{\gamma_n^{j'j}} \quad (\text{A-8})$$

where  $x_n^j$  is a city and industry specific unit cost index such that

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

Given constant returns to scale and competitive intermediate goods markets, a firm produces positive but finite amounts of a variety only if its price is equal to its unit production cost,

$$p_n^j(\mathbf{z}) = x_n^j(\mathbf{z}) = \frac{x_n^j}{z_n}. \quad (\text{A-9})$$

### 1.2.2 Final Goods

Let  $Q_n^j(\mathbf{z}) = \sum_{n'} Q_{nn'}^j(\mathbf{z})$  denote the total amount of intermediate goods of variety  $\mathbf{z}$  purchased from different cities by a final goods producer in city  $n$ , sector  $j$ . Given that intermediate goods of a given variety produced in different cities are perfect substitutes, final goods producers purchase varieties only from cities that offer the lowest unit cost,

$$Q_{nn'}^j(\mathbf{z}) = \begin{cases} Q_n^j(\mathbf{z}) & \text{if } \kappa_{nn'}^j p_{n'}^j(\mathbf{z}) < \min_{n'' \neq n'} \kappa_{nn''}^j p_{n''}^j(\mathbf{z}), \\ 0 & \text{otherwise} \end{cases},$$

where we abstract from the case where  $\kappa_{nn'}^j p_{n'}^j(\mathbf{z}) = \min_{n'' \neq n'} \kappa_{nn''}^j p_{n''}^j(\mathbf{z})$  since, given the distributional assumption on  $\mathbf{z}$ , this event only occurs on a set of measure zero.

Denote by  $P_n^j(\mathbf{z})$  the unit cost paid by a final good producer in city  $n$  and sector  $j$  for a particular variety whose vector of productivity draws is  $\mathbf{z}$ . Given that final goods firms only purchase intermediate goods from the lowest cost supplier,

$$P_n^j(\mathbf{z}) = \min_{n'} \left\{ \kappa_{nn'}^j p_{n'}^j(\mathbf{z}) \right\} = \min_{n'} \left\{ \frac{\kappa_{nn'}^j x_{n'}^j}{z_{n'}} \right\}. \quad (\text{A-10})$$

For non-tradable intermediate goods, firms must buy those goods locally, so that if  $j$  is non-tradable,

$$P_n^j(\mathbf{z}) = \frac{x_n^j}{z_n}. \quad (\text{A-11})$$

Then, the demand function for intermediate goods of variety  $\mathbf{z}$  in industry  $j$  and city  $n$  is given by

$$Q_n^j(\mathbf{z}) = \left( \frac{P_n^j(\mathbf{z})}{\tilde{P}_n^j} \right)^{-\eta} Q_n^j, \quad (\text{A-12})$$

where  $\tilde{P}_n^j$  the ideal cost index for final goods produced in sector  $j$  in city  $n$ ,

$$\tilde{P}_n^j = \left[ \int P_n^j(\mathbf{z})^{1-\eta} d\Phi(\mathbf{z}) \right]^{\frac{1}{1-\eta}}. \quad (\text{A-13})$$

Since the production function for final goods is constant returns to scale, and the market for

final goods is competitive, a final goods firm produces positive but finite quantities of a final good if its price is equal to its cost index, that is if  $P_n^j = \tilde{P}_n^j$ .

### 1.2.3 Derivation of Prices

We follow [Eaton and Kortum \(2002\)](#) in solving for the distribution of prices. Given this distribution and zero profits for final goods producers, when sector  $j$  is tradable, the price of final goods in sector  $j$  in region  $n$  solves

$$(P_n^j)^{1-\eta} = \int P_n^j(\mathbf{z})^{1-\eta} d\Phi(\mathbf{z}) d\mathbf{z},$$

which is the expected value of the random variable  $P_n^j(\mathbf{z})^{1-\eta}$ .

Let  $P_{nn'}^j(\mathbf{z}) = \frac{\kappa_{nn'}^j x_{n'}^j}{z_{n'}}$  denote the unit cost of a variety indexed by  $\mathbf{z}$  produced in city  $n'$  and sold in  $n$ . Following the steps described in [Caliendo et al. \(2017\)](#), we have that

$$\Pr \left[ P_{nn'}^j(\mathbf{z}) \leq p \right] = 1 - e^{-\omega_{nn'}^j p^\theta}$$

where  $\omega_{nn'}^j = \left[ \kappa_{nn'}^j x_{n'}^j \right]^{-\theta_j}$ . The price of variety  $\mathbf{z}$  in city  $n$  and industry  $j$ ,  $P_n^j(\mathbf{z})$ , is the minimum across  $P_{nn'}^j(\mathbf{z})$ . Its *cdf* is,

$$\Pr \left[ P_n^j(\mathbf{z}) \leq p \right] = 1 - e^{-\Omega_n^j p^\theta},$$

where  $\Omega_n^j = \sum_{n'} \omega_{nn'}^j = \sum_{n'} \left[ \kappa_{nn'}^j x_{n'}^j \right]^{-\theta}$  ( $\Omega_n^j$  does not depend on  $n'$  because we are integrating out the city dimension).

Let  $F_{P_n^j}(p)$  denote the *cdf* of  $P_n^j(\mathbf{z})$ ,  $\Pr \left[ P_n^j(\mathbf{z}) \leq p \right]$ . Then, its associated *pdf*, denoted  $f_{P_n^j}(p)$ , is  $\Omega_n^j \theta p^{\theta-1} e^{-\Omega_n^j p^\theta}$ . As in [Caliendo et al. \(2017\)](#), we have that

$$P_n^j = \Gamma(\xi)^{\frac{1}{1-\eta}} (\Omega_n^j)^{-\frac{1}{\theta}},$$

where  $\Gamma(\xi)$  is the Gamma function evaluated at  $\xi = 1 + \frac{1-\eta}{\theta}$ . The price of goods in tradable sector  $j$  may then also be expressed as

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

In a given non-tradable sector  $j$ ,  $\kappa_{nn'}^j = \infty \forall n' \neq n$ , so that the equation reduces to

$$P_n^j = \Gamma(\xi)^{\frac{1}{1-\eta}} x_n^j.$$

### 1.2.4 Trade Shares

Let  $X_n^j$  denote *total expenditures* on final goods  $j$  by city  $n$ , which must equal of the value of final goods in that sector,  $X_n^j = P_n^j Q_n^j$ . Recall that because of zero profits in the final goods sector,

total expenditures on intermediate goods in a given sector are then also equal to the cost of inputs used in that sector, so that  $P_n^j Q_n^j = \int P_n^j(\mathbf{z}) Q_n^j(\mathbf{z}) d\Phi(\mathbf{z})$ . Let  $X_{nn'}^j = \int \kappa_{nn'}^j P_{n'}^j(\mathbf{z}) Q_{nn'}^j(\mathbf{z}) d\Phi(\mathbf{z})$  denote the value spent by city  $n$  on intermediate goods of sector  $j$  produced in city  $n'$ . Further, let  $\pi_{nn'}^j$  denote the share of city  $n$ 's expenditures on sector  $j$  goods purchased from region  $n'$ . Then,

$$\pi_{nn'}^j = \frac{X_{nn'}^j}{X_n^j}.$$

Observe that, since there is a continuum of varieties of intermediate goods, the fraction of goods that firms in city  $n$  purchase from firms in city  $n'$  is given by

$$\tilde{\pi}_{nn'}^j \equiv \Pr \left[ P_{nn'}^j(\mathbf{z}) \leq \min_{n'' \neq n'} \left\{ P_{nn''}^j(\mathbf{z}) \right\} \right].$$

Following the steps described in Caliendo et al. (2017), we have that

$$\begin{aligned} \tilde{\pi}_{nn'}^j &= \frac{\omega_{nn'}^j}{\Omega_n^j} \\ &= \frac{\left[ \kappa_{nn'}^j x_{n'}^j \right]^{-\theta}}{\sum_{n''=1}^N \left[ \kappa_{nn''}^j x_{n''}^j \right]^{-\theta}} \end{aligned}$$

We can verify that  $\tilde{\pi}_{nn'}^j = \pi_{nn'}^j$ , that is, the share of goods that firms in city  $n$  purchase from city  $n'$  is equal to the share of the *value* of goods produced in city  $n'$  in the bundle purchased by firms in city  $n$  (see Eaton and Kortum (2002), Footnote 17). Observe also that  $\sum_{n'=1}^N \left[ \kappa_{nn'}^j x_{n'}^j \right]^{-\theta} = \left( P_n^j \right)^{-\theta} \Gamma(\xi)^{\frac{\eta}{1-\eta}}$ . Therefore, we may alternatively write the trade share  $\pi_{nn'}^j$  as

$$\pi_{nn'}^j = \frac{X_{nn'}^j}{X_n^j} = \left[ \frac{\kappa_{nn'}^j x_{n'}^j \Gamma(\xi)^{\frac{1}{1-\eta}}}{P_n^j} \right]^{-\theta}$$

In non-tradable sectors,  $\pi_{nn}^j = 1$ .

### 1.3 Market Clearing and Aggregation at the Industry and City Level

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. Put alternatively,

$$\sum_j \int L_n^{kj}(\mathbf{z}) d\Phi(\mathbf{z}) = \int \zeta_n^k(\mathbf{a}) d\Psi(\mathbf{a}), \quad \forall n = 1, \dots, N, k = 1, \dots, K. \quad (\text{A-14})$$

where  $\zeta_n^k(\mathbf{a}) \in \{0, 1\}$  denotes the location choice of households as a function of their type. Market clearing for land and structures in each region imply that

$$\sum_j \int H_n^j(\mathbf{z}) d\Phi(\mathbf{z}) = H_n, n = 1, \dots, N. \quad (\text{A-15})$$

Final goods market clearing implies that

$$\sum_k L_n^k C_n^{kj} + \sum_{j'} \int M_n^{jj'}(\mathbf{z}) d\Phi(\mathbf{z}) = Q_n^j. \quad (\text{A-16})$$

Finally, intermediate goods market clearing requires that

$$q_n^j(\mathbf{z}) = \sum_{n'} \kappa_{n'n}^j Q_{n'n}^j(\mathbf{z}). \quad (\text{A-17})$$

Given the labor supply equation (A-4) and the definition  $L_n^{kj} = \int L_n^{kj}(\mathbf{z}) d\Phi(\mathbf{z})$ , the labor market clearing equation (A-14) may be rewritten as

$$\sum_j L_n^{kj} = L^k \frac{(A_n^k C_n^k)^\nu}{\sum_{n'} (A_{n'}^k C_{n'}^k)^\nu}, \forall n = 1, \dots, N, k = 1, \dots, K.$$

Given the definition  $H_n^j = \int H_n^j(\mathbf{z}) d\Phi(\mathbf{z})$ , the market clearing equation for structures in each city (A-15) may be rewritten as

$$\sum_j H_n^j = H_n, n = 1, \dots, N.$$

Given our definition of total final expenditures,  $X_n^j = P_n^j Q_n^j$ , and the demand function for consumption goods of sector  $j$  (A-1), the market clearing condition for final goods in each city  $n$  and sector  $j$  (A-16) may be expressed in terms of sectoral and city aggregates,

$$\sum_k L_n^k (\alpha^j P_n C_n^k) + P_n^j \sum_{j'} M_n^{jj'} = X_n^j.$$

Finally, given that  $\pi_{n'n}^j X_{n'}^j = X_{n'n}^j = \int p_n^j(\mathbf{z}) \kappa_{n'n}^j Q_{n'n}^j(\mathbf{z}) d\Phi(\mathbf{z})$ , the market clearing condition for intermediate inputs (A-17) may be rewritten in terms of sectoral city aggregates as

$$\underbrace{\int p_n^j(\mathbf{z}) q_n^j(\mathbf{z}) d\Phi(\mathbf{z})}_{\text{Total value of intermediate goods produced in city } n} = \sum_{n'} \pi_{n'n}^j X_{n'}^j,$$

where  $\sum_{n'} \pi_{n'n}^j X_{n'}^j$  is the total value of expenditures across all cities spent on intermediate goods produced in city  $n$ .

We can use this last aggregation relationship to obtain aggregate factor input demand equations as follows,

$$\begin{aligned}
w_n^k L_n^{kj} &= \gamma_n^j (1 - \beta_n^j) \frac{\left(\frac{w_n^k}{\lambda_n^{kj}}\right)^{1-\epsilon}}{\sum_{k'=1}^K \left(\frac{w_n^{k'}}{\lambda_n^{k'j}}\right)^{1-\epsilon}} \sum_{n'} \left(\pi_{n'n}^j X_{n'}^j\right), \\
r_n H_n^j &= \gamma_n^j \beta_n^j \sum_{n'} \left(\pi_{n'n}^j X_{n'}^j\right), \\
P_n^{j'} M_n^{jj'} &= \gamma_n^{j'} \sum_{n'} \left(\pi_{n'n}^j X_{n'}^j\right).
\end{aligned}$$

Finally, combining these factor demand equations yields the aggregate production function,

$$\sum_{n'} \pi_{n'n}^j X_{n'}^j = x_n^j \left[ \left( \sum_k \left( \lambda_n^{kj} L_n^{kj} \right)^{1-\frac{1}{\epsilon}} \right)^{\frac{\epsilon}{\epsilon-1} (1-\beta_n^j)} (H_n^j)^{\beta_n^j} \right]^{\gamma_n^j} \prod_{j'} \left( M_n^{j'j} \right)^{\gamma_n^{j'j}}.$$

#### 1.4 Definition of Equilibrium

Equilibrium for this system of cities is given by a set of final goods prices  $P_n^j$ , wages in different occupations,  $w_n^k$ , rental rates,  $r_n$ , intermediate goods prices paid by final goods producers,  $P_n^j(\mathbf{z})$ , intermediate goods prices received by intermediate goods producers,  $p_n^j(\mathbf{z})$ , consumption choices,  $C_n^{kj}$ , intermediate input choices,  $Q_n^j(\mathbf{z})$ , intermediate input production,  $q_n^j(\mathbf{z})$ , demand for materials,  $M_n^{jj'}(\mathbf{z})$ , labor demand,  $L_n^{kj}(\mathbf{z})$ , demand for structures,  $H_n(\mathbf{z})$ , and location decisions,  $\zeta_n^k(\mathbf{a})$ , such that:

i) Workers choose consumption of each final good optimally, as implied by equation (A-1) and the budget constraint,  $\sum_j P_n^j C_n^{kj} = P_n C_n^k = I_n^k$ , where  $P_n = \prod_j \left(\frac{P_n^j}{\alpha^j}\right)^{\alpha^j}$  and  $I_n^k$  is given by equation (A-2).

ii) Workers choose optimally where to live as implied by equation (A-4).

iii) Intermediate input producers choose their demand for materials, labor and structures optimally (as implied by factor demand equations (A-5), (A-6) and (A-7)), and produce positive but finite amounts only if (A-9) holds, where  $x_n^j$  in that equation is given by (A-8).

iv) Final goods producers choose the origin of intermediate inputs optimally, implying that a producer in city  $n$  and industry  $j$  imports a variety  $\mathbf{z}$  from city  $n'$  if and only if  $\kappa_{nn'}^j p_n^j(\mathbf{z}) = \min_{n''} \left\{ \kappa_{nn''}^j p_{n''}^j(\mathbf{z}) \right\}$ . The price that they pay for intermediate goods satisfies (A-10) if the good is tradable and (A-11) if it is non-tradable.

v) Final goods producers choose their intermediate input use optimally according to (A-12) and produce positive but finite amounts only if (A-13) holds.

vi) Market clearing conditions for employment (equation A-14), land and structures (equation A-15), final goods (equation A-16), and intermediate goods (equation A-17) hold.

## 1.5 Aggregate Equilibrium

At the aggregate level, equilibrium is given by values for the prices  $P_n$ ,  $P_n^j$ ,  $x_n^j$ ,  $r_n$ ,  $w_n^k$ , aggregate quantities  $C_n^k$ ,  $L_n^{kj}$ ,  $H_n^j$ ,  $M_n^{j'j}$ , expenditures,  $X_n^j$ , and expenditure shares,  $\pi_{nn'}^j$ , that satisfy the following equations

$$\sum_{k,j'} L_n^{kj'} (\alpha^j P_n C_n^k) + \sum_{j'} P_n^j M_n^{j'j} = X_n^j \quad (NJ \text{ eqs.}) \quad (\text{A-18})$$

$$L_n^k = \sum_j L_n^{kj} = \frac{(A_n^k C_n^k)^\nu}{\sum_{n'} (A_{n'}^k C_{n'}^k)^\nu} L^k \quad (NK \text{ eqs.}) \quad (\text{A-19})$$

$$\sum_j H_n^j = H_n \quad (N \text{ eqs.}) \quad (\text{A-20})$$

$$P_n = \prod_j \left( \frac{P_n^j}{\alpha^j} \right)^{\alpha^j} \quad (N \text{ eqs.}) \quad (\text{A-21})$$

$$P_n C_n^k = w_n^k + b^k \frac{\sum_{n'} r_{n'} H_{n'}}{L^k} \quad \text{where } b^k = \frac{\sum_n w_n^k L_n^k}{\sum_{n,k'} w_n^{k'} L_n^{k'}} \quad (NK \text{ eqs.}) \quad (\text{A-22})$$

$$w_n^k L_n^{kj} = \frac{\left( \frac{w_n^k}{\lambda_n^{kj}(\mathbf{L}_n)} \right)^{1-\epsilon}}{\sum_{k'} \left( \frac{w_n^{k'}}{\lambda_n^{k'j}(\mathbf{L}_n)} \right)^{1-\epsilon}} \gamma_n^j (1 - \beta_n^j) \sum_{n'} \pi_{n'n}^j X_{n'}^j \quad (NKJ \text{ eqs.}) \quad (\text{A-23})$$

$$r_n H_n^j = \gamma_n^j \beta_n^j \sum_{n'} \pi_{n'n}^j X_{n'}^j \quad (NJ \text{ eqs.}) \quad (\text{A-24})$$

$$P_n^{j'} M_n^{j'j} = \gamma_n^{j'j} \sum_{n'} \pi_{n'n}^j X_{n'}^j \quad (NJ^2 \text{ eqs.}) \quad (\text{A-25})$$

$$P_n^j = \begin{cases} \Gamma(\xi)^{\frac{1}{1-\eta}} \left( \sum_{n'} [\kappa_{nn'}^j x_{n'}^j]^{-\theta} \right)^{-\frac{1}{\theta}} & \text{if } j \text{ is tradable} \\ \Gamma(\xi)^{\frac{1}{1-\eta}} x_n^j & \text{if } j \text{ is non-tradable} \end{cases} \quad (NJ \text{ eqs.}) \quad (\text{A-26})$$

$$\begin{aligned} & \sum_{n'} \pi_{n'n}^j X_{n'}^j \\ = & x_n^j \left[ \left( \sum_k \left( \lambda_n^{kj}(\mathbf{L}_n) L_n^{kj} \right)^{1-\frac{1}{\epsilon}} \right)^{\frac{\epsilon}{\epsilon-1} (1-\beta_n^j)} (H_n^j)^{\beta_n^j} \right]^{\gamma_n^j} \prod_{j'} (M_n^{j'j})^{\gamma_n^{j'j}} \quad (NJ \text{ eqs.}) \quad (\text{A-27}) \end{aligned}$$

$$\pi_{nn'}^j = \frac{\left[ \kappa_{nn'}^j x_{n'}^j \right]^{-\theta}}{\sum_{n''} \left[ \kappa_{nn''}^j x_{n''}^j \right]^{-\theta}} \quad (N^2 J \text{ eqs.}) \quad (\text{A-28})$$

This system of equations comprises  $2N + 2NK + 4NJ + NKJ + NJ^2 + N^2J$  equations in the same number of unknowns.

By substituting equation (A-25) into equation (A-18), adding over all industries ( $j$ ) and all cities ( $n$ ) and rearranging, we arrive at the National Accounting identity stating that aggregate value added is equal to aggregate consumption expenditures in the economy,

$$\sum_{n,k,j} L_n^{kj} P_n C_n^k = \sum_{n,j} \gamma_n^j X_n^j \quad (\text{A-29})$$

At the same time, multiplying both sides of equation (A-22) by  $L_n^k$ , adding over city ( $n$ ) and occupation ( $k$ ), and substituting out  $w_n^k$  and  $r_n H_n$  using equations (A-20), (A-23) and (A-24), yields the same national accounting identity. The fact that we can arrive at that same identity by manipulating different sets of equations implies that there is one redundant equation in the system, leading to one too many unknowns relative to the number of equations. The presence of a redundant equation is a feature of Walrasian systems. In order to pin down the price level, therefore, we need to amend the system with an additional equation defining the numeraire. Specifically, we set :

$$\sum_{n,j} \omega_n \ln(P_n) = \ln(\bar{P}), \quad (\text{A-30})$$

where  $\omega_n$  are a set of weights. When computing counterfactuals, we set those weights to be proportional to local nominal consumption:  $\omega_n \propto P_n \sum_k C_n^k L_n^k$ . Finally, observe that if we substitute the factor demand equations (A-23), (A-24), (A-25) into (A-27), we obtain the expression for the unit cost index,

$$x_n^j = B_n^j \left\{ r_n^{\beta_n^j} \left[ \sum_{k=1}^K \left( \frac{w_n^k}{\lambda_n^{kj}} \right)^{1-\epsilon} \right]^{\frac{1-\beta_n^j}{1-\epsilon}} \right\}^{\gamma_n^j} \prod_{j'=1}^J \left( P_n^{j'} \right)^{\gamma_n^{j'j}}. \quad (\text{A-31})$$

## 2 Quantifying the Model and Model Inversion

We now provide additional detail on how we quantify the model. The set of parameters needed to quantify our framework fall into broadly two types: i) parameters that are constant across cities (but may vary across occupations and/or industries) and that are directly available from statistical agencies, or that may be chosen to match national or citywide averages, and ii) parameters that vary at a more granular level and require using all of the model's equations (i.e. by way of model inversion) to match data that vary across cities, industries, and occupations.

## 2.1 Details of City-Invariant Parameters

**Input use shares in gross output**  $(\gamma_n^j, \gamma_n^{jj'})$  : To obtain an initial calibration for these share parameters, we use an average of the 2011 to 2015 BEA Use Tables, each adjusted by the same year's total gross output. The Use Table divides the value of the output in each sector  $j$ ,  $\sum_{n',n} \pi_{n'n}^j X_{n'}^j = \sum_n X_n^j$ , into the value of input purchases from other sectors  $j'$ ,  $\sum_{n,j'} P_n^{j'} M_n^{j'j}$ , labor compensation,  $\sum_{n,k} w_n^k L_n^{kj}$ , operational surplus,  $\sum_n r_n H_n^j$ , and taxes on production and imports,  $-\sum_n s_n^j X_n^j$ ,

$$\sum_n X_n^j = \sum_{n,j'} P_n^{j'} M_n^{j'j} + \sum_{n,k} w_n^k L_n^{kj} + \sum_n r_n H_n^j - \sum_n s_n^j X_n^j. \quad (\text{A-32})$$

Input purchases from other sectors are separated into purchases from domestic producers and purchases from international producers. Since the model does not allow for foreign trade, we adjust the Use Table by deducting purchases from international producers from the input purchases and, for accounting consistency, from the definition of gross output for the sector.

The model is static, so when mapping aggregate sectoral data into model parameters we need to take a stance on how gross operating surplus is distributed. A first observation is that equipment investment should be regarded in our static setting as equivalent to materials purchase. In particular, we follow Caliendo et al. (2018) and, for all sectors, we augment material purchases to include the purchases of equipment. Specifically, we subtract from the operational surplus of each sector 17 percent of their value added and then add the same value back to materials.<sup>1</sup> This 17 percent value is estimated by Greenwood et al. (1997) as the equipment share in output. We then pro-rate the equipment share of value added to different materials in proportion to their use within each sector.

Second, we interpret the remaining gross operational surplus as income from real estate ownership. The Use Tables treat the income from firm-owned plants differently from rent payments. The income from the former is a residual, included in the gross operating surplus of each sector, whereas rents are accounted as purchases of real estate services. Since the model does not differentiate between owned and rented property, for consistency we adopt the convention that all land and structures are managed by firms in the real estate sector, which then sell their services to other sectors. Accordingly, for all sectors other than real estate, we reassign the gross operating surplus remaining, after deducting equipment investment, to purchases from the real estate sector. These surpluses are in turn added to the gross operating surplus of real estate.

It follows that, for all sectors  $j$  other than real estate,

$$\sum_n r_n H_n^j = 0.$$

and in each of those sectors,

---

<sup>1</sup>When gross operation surplus amounts to less than 17 percent of value added, the entire operational surplus is deducted.

$$\begin{aligned}
P_n^{\text{real estate}} M_n^{\text{real estate},j} &= \text{Purchases from real estate by } j \\
&+ \text{Operational Surplus of } j \\
&- \text{Equipment Investment by } j.
\end{aligned}$$

In contrast, in the real estate sector,

$$\begin{aligned}
\sum_n r_n H_n^{\text{real estate}} &= \text{Total Operational Surplus across all } j \\
&- \text{Total Equipment Investment across all } j.
\end{aligned}$$

One can verify that those reassignments do not affect aggregate operational surplus (net of equipment investment), aggregate labor compensation, and aggregate value added (net of equipment investment).

We assume that tradable sectors have a  $\gamma_n^j = \gamma^j$ , constant across cities and similarly for  $\gamma_n^{j'}$ 's. The two non-tradable sectors have city specific parameters. Given these adjustments to the Use Table, the share parameters for tradable sectors follow immediately,

$$\gamma^j = \frac{\sum_{n,k} w_n^k L_n^{kj} + \sum_n r_n H_n^j}{\sum_n (1 + s_n^j) X_n^j}, \quad \gamma^{j'} = \frac{\sum_n P_n^{j'} M_n^{j'j}}{\sum_n (1 + s_n^j) X_n^j}. \quad (\text{A-33})$$

Furthermore, we have that, for the non-tradable sectors,

$$\gamma_n^j = \frac{\sum_k w_n^k L_n^{kj} + r_n H_n^j}{(1 + s_n^j) X_n^j}, \quad (\text{A-34})$$

where  $s_n^j$  is an ad-valorem subsidy for city  $n$ , sector  $j$ , which we introduce to account for the fact that part of the sectoral value added calculated by the BEA is in fact paid out in indirect taxes. Finally, since we do not observe use of materials by individual sectors in each city, we assume that, the proportions of materials used in each city by nontradable sectors if fixed at the national averages  $\frac{\gamma_n^{j'}}{1 - \gamma_n^j}$  is the same for all cities and satisfies equals  $\frac{\sum_n P_n^{j'} M_n^{j'j}}{\sum_{n,j'} P_n^{j'} M_n^{j'j}}$

$$\frac{\gamma_n^{j'}}{1 - \gamma_n^j} = \hat{\gamma}^{j'} = \frac{\sum_n P_n^{j'} M_n^{j'j}}{\sum_{n,j'} P_n^{j'} M_n^{j'j}}$$

The calibration of  $\gamma_n^j$  and, therefore of  $\gamma_n^{j'}$ , will require choosing additional parameters as described below but consistent with the above equations.

**Trade costs:** We assign trade costs to be a log-linear function of distance, that is,

$$\kappa_{nn'}^j = (d_{nn'})^{t^j}$$

where  $\kappa_{n,n'}$  is the amount of goods that need to be shipped from location  $n'$  in order for one unit of the good to be available in location  $n$ , and  $d_{n,n'}$  is the distance (in miles) between the two locations.

From equation (A-28) we can write

$$\log(\pi_{nn'}^j) = -\theta t^j \log(d_{nn'}) + c_n + c_{n'} \quad (\text{A-35})$$

where  $c_n$  and  $c_{n'}$  are  $n$  and  $n'$  location-specific factors

$$c_{n'} = -\theta \log(x_{n'}^j)$$

and

$$c_n = -\log \left( \sum_{n''} \left[ \kappa_{nn''}^j x_{n''}^j \right]^{-\theta} \right)$$

We assign trade costs to industries using three different methods. First, we assign two industries (retail, construction and utilities, and real estate) to be non-tradable. Second, we use the estimates in Table 1 of [Anderson et al. \(2014\)](#) to obtain gravity coefficients for services. Third, we rely on equation (A-35) to obtain the gravity coefficients. In order to do this, we use the 2012 Commodity Flow Survey Public Use Microdata File. We add up shipment values by industry, origin and destination and then, for each industry, we regress the log of those averages on log of average shipment distance between each origin and destination.

The gravity coefficients used are summarized in Table A-1, below.

## 2.2 Model Inversion and Granular Parameters

From the ACS, we obtain data pertaining to  $w_n^k$ , and  $\frac{L_n^{kj}}{\sum_{k'} L_n^{k'j}}$ . The spatial distribution of CNR shares ( $L_n^k/L_n$ ) is depicted in Figure A-1 below. The Census County Business Patterns (CBP) provide us measures of total employment  $\sum_{k'} L_n^{k'j}$  that better match BEA industry-level counts, which we combine with the ACS data to obtain  $L_n^{kj}$ . From the BEA, we obtain regional price parity (RPP) indices for each city, disaggregated into goods, services and rents. As explained below, we use the level of rents and the relative price of goods and services, providing us with  $2(N - 1)$  additional restrictions (we deduct 2 since prices in any given city are only defined relative to those in other cities). Furthermore, we can use the BEA Use Tables to calculate the national share of income from land and structures in the production of real estate, providing us with one additional equation. Lastly, as we explain in more detail below, we can apply a normalization to each sector, for a total of  $J$  normalizations.

The data plus normalizations above impose  $NK + NKJ + 2N + J - 1$  independent restrictions that allows us to solve for  $NK$  values for amenity parameters,  $A_n^k$ ,  $NKJ$  scaling factors in produc-

Table A-1: Estimated Gravity coefficients ( $-\theta^j$ )

Industry	Gravity Coefficient	Source
Retail, Construction and Utilities	$-\infty$	Non-tradable
Food and Beverage	-1.24	own estimate
Textiles	-0.88	own estimate
Wood, Paper, and Printing	-1.36	own estimate
Oil, Chemicals, and Nonmetallic Minerals	-1.32	own estimate
Metals	-1.20	own estimate
Machinery	-0.81	own estimate
Computer and Electric	-0.77	own estimate
Electrical Equipment	-0.64	own estimate
Motor Vehicles (Air, Cars, and Rail)	-0.90	own estimate
Furniture and Fixtures	-1.18	own estimate
Miscellaneous Manufacturing	-0.83	own estimate
Wholesale Trade	-0.56	Anderson et al. (2014)
Transportation and Storage	-0.62	Anderson et al. (2014)
Professional and Business Services	-0.93	Anderson et al. (2014)
Other	-0.72	Anderson et al. (2014)
Communication	-0.30	Anderson et al. (2014)
Finance and Insurance	-0.68	Anderson et al. (2014)
Real Estate	$-\infty$	Non-tradable
Education	-1.01	Anderson et al. (2014)
Health	-1.42	Anderson et al. (2014)
Accommodation	-0.93	Anderson et al. (2014)

See text for own estimation details. Coefficients from Anderson et al. (2014) are extracted from Table 1.

tion,  $T_n^{kj} \equiv \left(H_n^j\right)^{\gamma_n^j \beta_n^j} \left(\lambda_n^{kj}\right)^{\gamma_n^j (1-\beta_n^j)}$ ,  $(N-1)$  shares of non-residential structures in value added in the real estate sector,  $\beta_n^{\text{real estate}}$ ,  $(N-1)$  shares of value added in non-tradable output, and  $J-1$  independent values for consumption share parameters,  $\alpha^j$ .<sup>2</sup>

The steps below describe the model inversion.

1. *Computing consumption shares,  $\alpha^j$ .* We first add up equation (A-18) across  $n$  and  $j$ . We then use the factor demand equations (A-23) and (A-24) to obtain  $\gamma_n^j X_n^j = \left(\sum_n r_n H_n^j + \sum_{n,k} w_n^k L_n^{kj}\right)$ , and the national accounting identity (A-29) to substitute out  $X_n^j$ 's and  $C_n^k$ 's from the aggregated equation (A-18):

$$\alpha^j \sum_{j'} \left( \sum_n r_n H_n^{j'} + \sum_{n,k} w_n^k L_n^{kj'} \right) + \sum_{n,j'} P_n^j M_n^{jj'} = \sum_n r_n H_n^j + \sum_{n,k} w_n^k L_n^{kj} + \sum_{n,j'} P_n^{j'} M_n^{j'j}.$$

The ACS does not provide data on  $r_n H_n^j$ ,  $\sum_{j'} P_n^j M_n^{jj'}$  or their sum across cities. While the BEA provides data on sectoral aggregates, those cover the whole country as opposed to only MSA's.

<sup>2</sup>Furthermore, additional restrictions imposed on  $\alpha^j$  and  $\beta^j$ , specifically that  $\alpha^j \in [0, 1]$ , and  $\beta^j \in [0, 1]$ , imply some overidentifying restrictions.

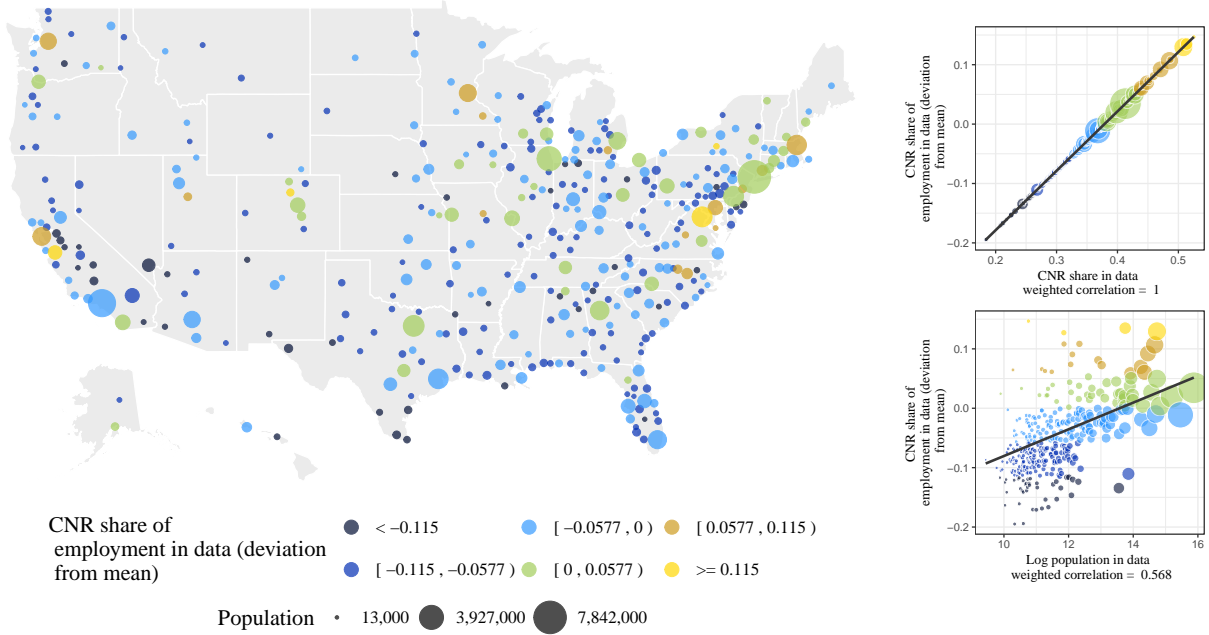


Figure A-1: CNR share from 2011-2015

Each marker in the map refers to a CBSA. Marker sizes are proportional to total equilibrium employment in each city.

Thus, we rely instead on ratios,  $\frac{\sum_n r_n H_n^j}{\sum_{n,k} w_n^k L_n^{kj}}$  and  $\frac{\sum_{n,j'} P_n^{j'} M_n^{j'j'}}{\sum_{n,k} w_n^k L_n^{kj}}$  obtained from the Use Tables, which we can then combine with data on  $\sum_{n,k} w_n^k L_n^{kj}$  and the above equation. Specifically,

$$\begin{aligned} & \alpha^j \sum_{n,j',k} w_n^k L_n^{kj'} \left( \frac{\sum_n r_n H_n^{j'}}{\sum_{n,k} w_n^k L_n^{kj'}} + 1 \right) + \sum_{j',n,k} \frac{\sum_{n,j'} P_n^{j'} M_n^{j'j'}}{\sum_{n,k} w_n^k L_n^{kj'}} w_n^k L_n^{kj'} \\ &= \left( \sum_{n,k} w_n^k L_n^{kj} \right) \times \left( \frac{\sum_n r_n H_n^j}{\sum_{n,k} w_n^k L_n^{kj}} + \frac{\sum_{n,j'} P_n^{j'} M_n^{j'j'}}{\sum_{n,k} w_n^k L_n^{kj}} + 1 \right), \end{aligned}$$

or

$$\begin{aligned} & \alpha^j \sum_{n,j',k} w_n^k L_n^{kj'} \left( \varrho_H^{j'} + 1 \right) + \sum_{j',n,k} w_n^k L_n^{kj'} \varrho_M^{j'j'} \\ &= \sum_{n,k} w_n^k L_n^{kj} \left( \varrho_H^j + \varrho_M^{j'j} + 1 \right), \end{aligned}$$

where  $\varrho_H^j$  and  $\varrho_M^{j'j}$  denote, respectively, the ratio of national aggregate rental income and the ratio

of national aggregate material inputs usage from sector  $j'$  to national aggregate wage income in sector  $j$  which are consistent with the Use Tables. The  $J$  equations above can be solved for  $J$  values of  $\alpha^j$ . One can verify that any value of  $\alpha^j$  obtained from those equations will satisfy  $\sum_j \alpha^j = 1$ . One complicating factor is that in each sector  $j$ ,  $\alpha^j$  must live in  $[0, 1]$ . However, because of measurement inconsistencies between ACS and BEA data, the procedure generates negative values of  $\alpha^j$  in three out of 22 sectors. One of those sectors (“Oil, Chemicals, and Nonmetallic Minerals”) indeed has much of its employment located outside of urban areas. We use information from the Use Tables to calibrate  $\alpha^j$  in that sector, setting it equal to 5.57 percent. The other two sectors (“Wood, Paper, and Printing”, and “Metals”) are to a large degree producers of inputs for other industries, so that we set their consumption shares to 0. To ensure that all equations hold while satisfying those restrictions, we allow  $\varrho_M^{j'j}$ 's to deviate somewhat from those obtained from the Use Tables. This requires adjusting  $\gamma^j$  and  $\gamma^{j'j}$  for the tradable sectors, since those satisfy  $\gamma^j = \frac{1+\varrho_H^j}{1+\varrho_H^j+\sum^{j'} \varrho_M^{j'j}}$ ,  $\gamma^{j'j} = \frac{\varrho_M^{j'j}}{1+\varrho_H^j+\sum^{j'} \varrho_M^{j'j}}$ .

2. *Expressing gross output and rental income from each sector and city as functions of share parameters and wage bills.* Using the labor demand equations (A-23), we obtain

$$\sum_{n'} \pi_{n'n}^j X_{n'}^j = \frac{\sum_k w_n^k L_n^{kj}}{(1 - \beta_n^j) \gamma_n^j}. \quad (\text{A-36})$$

In the non-tradable sectors,  $\pi_{nn}^j = 1$  and  $\pi_{n'n}^j = 0$  for  $n' \neq n$  so that

$$X_n^j = \frac{\sum_k w_n^k L_n^{kj}}{(1 - \beta_n^j) \gamma_n^j}.$$

For all sectors other than real estate, we have that  $\beta_n^j = 0$ , so that  $r_n H_n^j = 0$ . For the real estate sector, we have from the first-order conditions in that sector that

$$r_n H_n^{\text{real estate}} = \frac{\beta_n^{\text{real estate}}}{1 - \beta_n^{\text{real estate}}} \sum_k w_n^k L_n^{k, \text{real estate}}.$$

Since real estate services are the only sector with positive rental income, this is also equal to the total rental income in each city.

3. *Computing the shares of land and structures in value added for the real estate sector,  $\beta_n^{\text{real estate}}$ .* We use equations (A-22) to substitute out  $P_n C_n^k$  in equations (A-18). We then apply the relationships from equation (A-36) to substitute out gross output in (A-23) to (A-25), and use the resulting equations to substitute out factor demands in (A-18). Given that in the non-tradable sectors (“real estate” and “retail, construction, and utilities”), expenditures are equal to gross output, this implies that, for  $j \in \{\text{“real estate”, “retail, construction, and utilities”}\}$ , we have that

$$\begin{aligned}
& \frac{1}{1 - \beta_n^j} \frac{\sum_k w_n^k L_n^{kj}}{\gamma_n^j} \\
&= \alpha^j \sum_k w_n^k L_n^k \\
&+ \alpha^j \sum_k \frac{L_n^k}{L^k} b^k \sum_{n'} \left( \frac{\beta_{n'}^{\text{real estate}}}{1 - \beta_{n'}^{\text{real estate}}} \sum_{k'} w_{n'}^{k'} L_{n'}^{k', \text{real estate}} \right) \\
&+ \sum_{j'} \frac{1 - \gamma_n^{j'}}{\gamma_n^{j'} (1 - \beta_n^{j'})} \hat{\gamma}^{jj'} \sum_k w_n^k L_n^{kj'},
\end{aligned}$$

where we are using the fact that  $\beta_n^j = 0$  for all sectors other than real estate. Given that we have two non-tradable sectors, this is a system of  $2N$  equations, in  $N$  values for  $\gamma_n^j$  and  $N$  values of  $\beta_n^{\text{real estate}}$ .

4. *Computing individual values for nominal expenditures,  $X_n^j$ , in tradable sectors.* We use equations (A-22) to substitute out  $P_n C_n^k$  in equations (A-18). We then apply the relationships from equation (A-36) to substitute out gross output in (A-23) to (A-25), and use the resulting equations to substitute out factor demands in (A-18). In the tradable sectors, this gives us

$$\begin{aligned}
X_n^j &= \alpha^j \sum_k w_n^k L_n^k \\
&+ \alpha^j \sum_k \frac{L_n^k}{L^k} b^k \sum_{n'} \left( \frac{\beta_{n'}^{\text{real estate}}}{1 - \beta_{n'}^{\text{real estate}}} \sum_{k'} w_{n'}^{k'} L_{n'}^{k'j'} \right) \\
&+ \sum_{j'} \gamma_n^{jj'} \frac{\sum_k w_n^k L_n^{kj'}}{(1 - \beta_n^{j'}) \gamma_n^{j'}}
\end{aligned}$$

Given values for  $\beta_n^j$  from the previous step, values for  $X_n^j$  are then immediately determined from the data.

5. *Computing relative cost indices for tradable goods,  $\tilde{x}_n^j$ .* For  $N(J - 2)$  tradable sectors (all but “real estate,” as well as “retail, construction, and utilities”), we now solve for  $(N - 1)(J - 2)$  values of the cost index,  $\frac{x_n^j}{\sum_{n'} x_{n'}^j}$ , for each  $j \in \{1, \dots, J\}$  from the system of  $(N - 1)(J - 2)$  independent equations,

$$\frac{\sum_k w_n^k L_n^{kj}}{(1 - \beta_n^j) \gamma_n^j} = \sum_{n'=1}^N \pi_{n'n}^j(\mathbf{x}^j) X_{n'}^j,$$

where  $\mathbf{x}^j = \{x_1^j, \dots, x_N^j\}$  is the vector of unit production costs. This system comprises only  $(N - 1)(J - 2)$  independent equations since, for each  $j$ , adding up the right-hand-side and

left-hand-side over  $n$  gives the same result on both sides irrespective of  $\mathbf{x}^j$ . At the same time  $\pi_{n'n}^j(\mathbf{x}^j)$  is homogeneous of degree 0 in  $\mathbf{x}^j$  for each  $j$  in equation (A-28), so that we can still solve for the ratio,  $\tilde{x}_n^j \equiv \frac{x_n^j}{\sum_{n'} x_{n'}^j}$ .<sup>3</sup>

6. *Computing relative tradable consumer prices,  $\tilde{P}_n^j$ , in every sector and city.* Substituting  $\tilde{x}_n^j$  from the previous step into equation (A-26) and rearranging, we have that for the tradable sectors,

$$P_n^j = \Gamma(\xi)^{\frac{1}{1-\eta}} \left( \sum_{n'} [\kappa_{nn'}^j \tilde{x}_n^j]^{-\theta} \right)^{-\frac{1}{\theta}} \times \sum_{n'} x_{n'}^j,$$

which gives a system of  $N(J-2)$  equations. We can thus determine

$$\Xi_P^j \equiv \Gamma(\xi)^{\frac{1}{1-\eta}} \frac{\sum_{n'} x_{n'}^j}{P^j} = \left[ \sum \varpi_n^j \left( \sum_{n'} [\kappa_{nn'}^j \tilde{x}_n^j]^{-\theta} \right)^{-\frac{1}{\theta}} \right]^{-1}$$

for each  $j$  by imposing  $\sum_n \varpi_n^j P_n^j = P^j$ , where  $\varpi_n^j$  are model-consistent expenditure weights given by  $X_n^j / \sum_{n'} X_{n'}^j$  obtained in step 4. We may then obtain for all tradable  $j$ 's

$$\tilde{P}_n^j \equiv \frac{P_n^j}{P^j} = \Xi_P^j \left( \sum_{n'} [\kappa_{nn'}^j \tilde{x}_n^j]^{-\theta} \right)^{-\frac{1}{\theta}}.$$

Note that data on  $P^j$  is only available in changes from a base period. Thus, we define the base period to be 2011-2015, our benchmark period, and set  $P^j = 1$  in all sectors in that period. For the remainder of the analysis, therefore,  $\tilde{P}_n^j = P_n^j$ .

7. *Computing non-tradable consumer prices.* In the non-tradable sectors, we have that  $P_n^j = \Gamma(\xi)^{\frac{1}{1-\eta}} x_n^j$  for all  $n$  and  $j$ , and for those sectors, we determine prices based on data from the Regional Price Parity (RPP) indices calculated by the BEA. We directly obtain values for  $\Gamma(\xi)^{\frac{1}{1-\eta}} x_n^{\text{real estate}} \equiv P_n^{\text{real estate}}$  from the RPP estimates of the price of real estate services in different cities. For the other non-tradables (“retail, construction, and utilities”), we choose  $P_n^{\text{retail, etc.}} = \Gamma(\xi)^{\frac{1}{1-\eta}} x_n^{\text{retail, etc.}}$  so that the price of services (other than real estate) relative to tradable goods in the model matches its counterpart in the RPP. To carry out this calculation, observe that the price index for services can be defined by  $P_n^{\text{Services}} = \prod_{j \in \text{Services}} \left( \frac{\sum_{j' \in \text{Services}} \alpha^{j'} P_n^{j'}}{\alpha_j} \right)^{\frac{\alpha^j}{\sum_{j' \in \text{Services}} \alpha^{j'}}$  where the service sectors include retail, etc., wholesale trade, transportation and storage, professional and business services, other, communication, finance and insurance, education, health, and accommodation. The price

---

<sup>3</sup>Numerically, the system is easier to solve for  $\frac{(x_n^j)^{\theta j}}{\sum_{n'} (x_{n'}^j)^{\theta j}}$  from which we can easily obtain values for  $x_n^j$ 's.

index for goods can be defined analogously where the goods sector includes all remaining sectors other than real estate.

8. *Computing firm productivity in different sectors,  $j$ , and cities,  $n$ , associated with occupation  $k$ ,  $\lambda_n^{kj}$ .* From equations (A-23), we have that

$$w_n^k L_n^{kj} = \frac{\left(\frac{w_n^k}{\lambda_n^{kj}}\right)^{1-\epsilon}}{\sum_{k'} \left(\frac{w_n^{k'}}{\lambda_n^{k'j}}\right)^{1-\epsilon}} \sum_k w_n^k L_n^{kj},$$

which can rewrite as

$$w_n^k L_n^{kj} = \frac{\left(\frac{\sum_{k'} \lambda_n^{k'j}}{\lambda_n^{kj}} w_n^k\right)^{1-\epsilon}}{\sum_{k'} \left(\frac{\sum_{k'} \lambda_n^{k'j}}{\lambda_n^{k'j}} w_n^{k'}\right)^{1-\epsilon}} \sum_k w_n^k L_n^{kj}.$$

Thus, for each city  $n$  and industry  $j$ , we can use  $K - 1$  of those equations to solve for  $K - 1$  ratios,  $\tilde{\lambda}_n^{kj} = \frac{\lambda_n^{kj}}{\sum_{k'} \lambda_n^{k'j}}$ . With some rearrangement, those can be written as

$$\tilde{\lambda}_n^{kj} = \frac{(w_n^k)^{\frac{\epsilon}{\epsilon-1}} (L_n^{kj})^{\frac{1}{\epsilon-1}}}{\sum_{k'} (w_n^{k'})^{\frac{\epsilon}{\epsilon-1}} (L_n^{k'j})^{\frac{1}{\epsilon-1}}}.$$

From equations (A-31) (obtained by substituting the factor demand equations (A-23), (A-24) and (A-25) into equations (A-27)), and the value for  $r_n H_n$  obtained in step 2, we obtain

$$\begin{aligned} & H_n^{\gamma_n^j \beta_n^j} \left( \sum_{k'} \lambda_n^{k'j} \right)^{\gamma_n^j (1-\beta_n^j)} \Gamma(\xi)^{-\frac{1}{1-\eta}} \tag{A-37} \\ &= \frac{B_n^j}{\tilde{x}_n^j \Xi_P^j} \left\{ \left( \sum_j \frac{\beta_n^j}{1-\beta_n^j} \sum_k w_n^k L_n^{kj} \right)^{\beta_n^j} \left[ \sum_{k=1}^K \left( \frac{w_n^k}{\tilde{\lambda}_n^{kj}} \right)^{1-\epsilon} \right]^{\frac{1-\beta_n^j}{1-\epsilon}} \right\}^{\gamma_n^j} \prod_{j'=1}^J (P_n^{j'})^{\gamma_n^{j'j}}, \end{aligned}$$

where we set  $\Gamma(\xi)^{-\frac{1}{1-\eta}} = 1$  since it is common to all sectors and cities and thus immaterial in any counterfactual exercise. Recall that the use of land and structures as inputs has been folded in the real estate sector that then sells real estate services to all other sectors (i.e.  $\beta_n^j = 0$  in all sectors but real estate). Then, multiplying both sides of equation (A-31) by the ratios  $(\tilde{\lambda}_n^{kj})^{\gamma_n^j}$  gives  $NK(J-1)$  values for the productivity of firms in different sectors,  $j$ , and cities,  $n$ , associated with occupation  $k$ ,  $T_n^{kj} \equiv \left( H_n^j \right)^{\gamma_n^j \beta_n^j} \left( \lambda_n^{kj} \right)^{\gamma_n^j (1-\beta_n^j)}$ , which, in the special case where one abstracts from differences in occupational composition across cities,

reproduces measured regional and sectoral productivity in Caliendo et al. (2018).

9. *Computing the idiosyncratic amenity distribution parameter  $\nu$  and amenity shifters  $A_n^k$  for each city  $n$  and occupation  $k$*

To compute  $\nu$  we match the estimate for local labor supply elasticity with respect to local wage estimated by Fajgelbaum, Serrato and Zidar (2018) of 1.14. In our setup, for any occupation  $k$  and city  $n$ , this elasticity is  $\nu \frac{w_n^k}{P_n C_n^k}$ . The average elasticity is equal to 1.14 if  $\nu = 2.017$ . Given  $\nu$ , we can now back-out amenities from the labor supply equation (A-4).<sup>4</sup>

## 2.3 Comparisons to Previous Work and Model Validation

We now show that the wage data used in our model inversion broadly conforms to patterns observed in previous literature. Moreover, our model-consistent TFP measures and tradable prices compare favorably with previous empirical work, but within a single general equilibrium framework that can also be used to guide optimal policy. There exists a large literature that has estimated and studied the role of agglomeration externalities. Much of this work has relied on a production function approach using measures of output and factor inputs to estimate Total Factor Productivity (TFP), or using labor productivity more directly, in exploring how productivity depends on the scale of city employment or its skill composition. There is also a literature that has sought to understand how tradable goods prices vary with city size. We add to those literatures by combining their results with a framework that can then be used to provide a quantitative assessment of optimal spatial policy.

### 2.3.1 Wages

Table A-2 compares the relationships between wages, employment, and employment composition across different cities highlighted in previous work relative to the data used in our model inversion. The first three rows of the table show regression coefficients of log wages for CNR workers, non-CNR workers, and the CNR wage premium, on different measures of city employment and employment composition. The subsequent rows show similar regression coefficients obtained in previous literature. The data we use implies relationships that are consistent with those in other work. In particular, all wages increase with city size, more so for skilled workers. A similar relationship holds for wages and city composition where proportionally more skilled cities exhibit higher wages for all workers, more so for skilled workers.<sup>5</sup>

---

<sup>4</sup>For given  $k$ , labor supply is homogeneous of degree zero in  $A_n^k$ . This implies that amenities are only determined up to an arbitrary, occupation-specific scaling constant, that is, we can change  $A_n^k$  to  $\tilde{A}_n^k$  so that  $\tilde{A}_n^k = m^k A_n^k$  without any observable implications.

<sup>5</sup>An exception is Moretti (2004a) who finds no statistically significant differences in the way that wages of college educated workers and non-college educated workers vary with employment composition across cities. Our findings, however, rely on a more recent time period where other work has found an increasingly pronounced relationship between skill and city size (see Baum-Snow and Pavan (2013)).

### 2.3.2 Amenities

We now turn to the occupation-specific amenities implied by the model inversion,  $A_n^k$ . The relationship between relative amenities for CNR and non-CNR workers against the size and composition of cities is depicted in Figure A-2. Our findings conform to [Diamond \(2016\)](#) in that cities with more CNR workers are also relatively more amenable to those same workers. At the same time, larger cities are relatively more amenable to CNR workers helping account for the concentration of CNR workers in large cities.

[Diamond \(2016\)](#) provides evidence for a causal impact of local population composition on amenities. In Appendix 6.3, we show the effect of filtering out the component of amenities that is endogenous to the local labor composition.<sup>6</sup> While suppressing those endogenous effects eliminates

<sup>6</sup>Here, we use the parameterization that [Fajgelbaum and Gaubert \(2020\)](#) obtain based on the estimates by [Diamond \(2016\)](#).

Table A-2: Wages, Employment, and City Composition

Dependent Variable	$\ln(L_n)$	$\ln\left(\frac{L_n^{\text{CNR}}}{L_n^{\text{non-CNR}}}\right)$	$\frac{L_n^{\text{CNR}}}{L_n}$
$\ln(w_n^{\text{CNR}})$	0.06 (0.004)	0.34 (0.02)	1.50 (0.09)
$\ln(w_n^{\text{non-CNR}})$	0.05 (0.002)	0.22 (0.02)	0.99 (0.07)
$\ln\left(\frac{w_n^{\text{CNR}}}{w_n^{\text{non-CNR}}}\right)$	0.02 (0.001)	0.12 (0.01)	0.51 (0.04)
Moretti HS <sup>1</sup>	—	—	0.85 (0.06)
Moretti Some College	—	—	0.86 (0.06)
Moretti College +	—	—	0.74 (0.06)
Roca & Puga wage log wage constant <sup>2</sup>	0.046 (0.008)	—	—
Diamond log college wage <sup>3</sup>	—	0.26 (0.11)	—
Diamond log non-college wage <sup>4</sup>	—	0.18 (0.01)	—
Baum-Snow et al. log wage, 2005-2007 <sup>5</sup>	0.065 (< 0.01)	—	—
Baum-Snow et al. log wage ratio <sup>6</sup>	0.029 (< 0.003)	—	—

1. [Moretti \(2004a\)](#) "Estimate the social return to higher education: evidence from longitudinal and repeated cross-sectional data", Table 5.

2. [de la Roca and Puga \(2017\)](#) "Learning by Working in Big Cities", Table 1.

3. [Diamond \(2016\)](#) "The Determinants and Welfare Implications of US Workers' Diverging Location Choices by Skill: 1980-2000", Figure 4.

4. [Diamond \(2016\)](#), Figure 3

5. [Baum-Snow et al. \(2018\)](#) "Why Has Urban Inequality Increased?", Table 1. Standard error reported as less than 0.01.

6. [Baum-Snow et al. \(2018\)](#), Table 2. Standard error reported as less than 0.003.

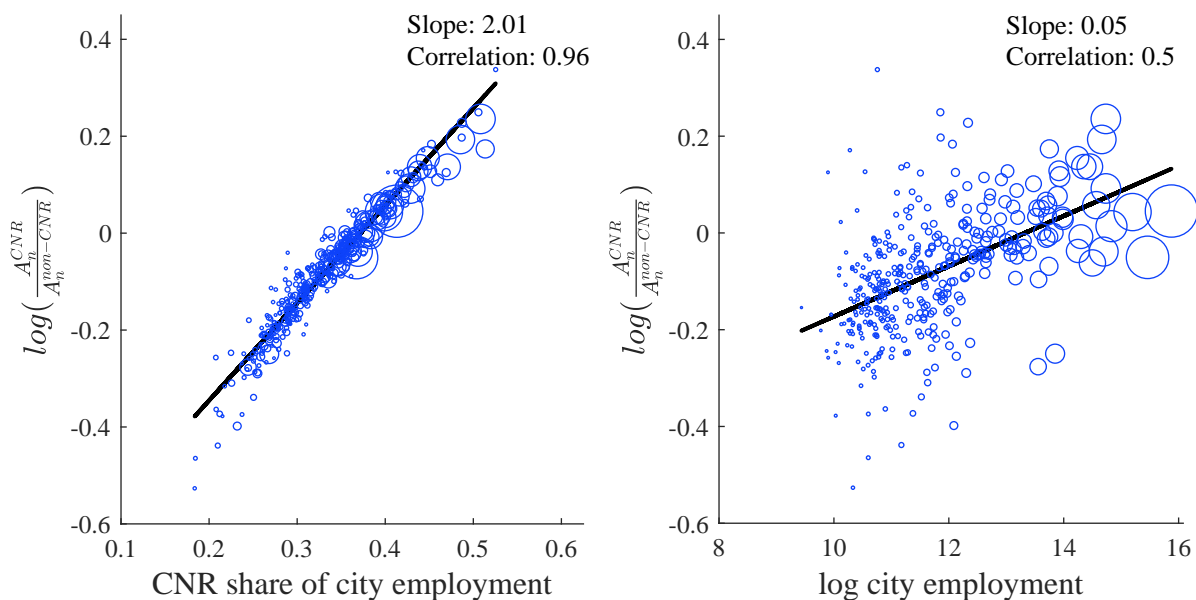


Figure A-2: Relative amenities and city size and composition

Ratio of occupational-specific amenity parameters for each city obtained from the model quantification against employment share of CNR workers and log total employment. Each observation refers to a CBSA. Marker sizes are proportional to total city employment.

the positive relationship between the CNR share and relative amenities, the relationship between relative residual amenities and city size becomes stronger. Intuitively, given the estimates in [Diamond \(2016\)](#), large non-CNR populations generate larger congestion effects on CNR workers than on non-CNR workers. The bottom line, therefore, is that our findings below regarding the optimality of concentrating CNR workers, computed without endogenous amenities, are if anything conservative.

### 2.3.3 Total Factor Productivity

A substantive literature in urban economics has addressed the relationship between productivity and city size (i.e. “agglomeration economies”), as well as that between productivity and employment composition. Baseline estimates of real Total Factor Productivity (TFP) typically rely on Cobb-Douglas production functions that allow for different types of labor to enter separately. Within the context of our model, we follow [Caliendo et al. \(2017\)](#) and express measured TFP as

$$\ln TFP_n^j = \ln \frac{\sum_{n'} \pi_{n'n} X_{n'}^j}{P_n^j} - \gamma_n^j \beta_n^j \ln H_n^j - \gamma_n^j (1 - \beta_n^j) \sum_k \delta^{kj} \ln L_n^{kj} - \sum_{j'} \gamma_n^{j'j} \ln M_n^{j'j}, \quad (\text{A-38})$$

where  $\delta^{kj}$  is the share of occupation  $k$  wages in sector  $j$ 's wage bill. In the model we have laid out, and up to a first-order approximation (abstracting from selection effects induced by trade), we have that for tradable sectors,

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

where  $(\lambda_n^{kj})^{\gamma^j}$  may thus be interpreted as the component of TFP in sector  $j$  and city  $n$  associated with occupation  $k$ .<sup>7</sup>

From equation (A-27),

$$\sum_{n'} \pi_{n'n}^j X_{n'}^j = x_n^j \left[ \left[ \left( \sum (\lambda_n^{kj} L_n^{kj})^{\frac{\epsilon-1}{\epsilon}} \right)^{\frac{\epsilon}{\epsilon-1}} \right]^{1-\beta_n^j} (H_n^j)^{\beta_n^j} \right]^{\gamma_n^j} \prod_{j'} (M_n^{j'j})^{\gamma_n^{j'j}}$$

Also, recall that we can write  $\frac{x_n^j}{P_n^j} = \frac{1}{\kappa_{nn}^j} (\pi_{nn}^j)^{-\frac{1}{\theta}}$ , which picks up the role of selection effects on productivity (see [Caliendo et al. \(2017\)](#)). We take a first order log-linear approximation of  $\sum_{n'} \pi_{n'n}^j X_{n'}^j$  around national averages to obtain

$$\begin{aligned} d \ln \left( \sum_{n'} \pi_{n'n}^j X_{n'}^j \right) &\simeq d \ln P_n^j - \frac{1}{\theta} d \ln \pi_{nn}^j + \gamma_n^j (1 - \beta_n^j) \sum_k \frac{(\lambda^{kj} L^{kj})^{\frac{\epsilon-1}{\epsilon}}}{\sum_{k'} (\lambda^{kj} L^{kj})^{\frac{\epsilon-1}{\epsilon}}} (d \ln L_n^{kj} + d \ln \lambda_n^{kj}) \\ &\quad + \gamma_n^j \beta_n^j d \ln H_n^j + \sum_{j'} \gamma_n^{j'j} d \ln M_n^{j'j}, \end{aligned}$$

where  $\frac{(\lambda^{kj} L^{kj})^{\frac{\epsilon-1}{\epsilon}}}{\sum_{k'} (\lambda^{kj} L^{kj})^{\frac{\epsilon-1}{\epsilon}}}$  is the national average of  $\frac{(\lambda_n^{kj} L_n^{kj})^{\frac{\epsilon-1}{\epsilon}}}{\sum_{k'} (\lambda_n^{kj} L_n^{kj})^{\frac{\epsilon-1}{\epsilon}}}$ . From manipulating equation (A-23), we can verify that:

$$\frac{(\lambda_n^{kj} L_n^{kj})^{\frac{\epsilon-1}{\epsilon}}}{\sum_{k'} (\lambda_n^{kj} L_n^{kj})^{\frac{\epsilon-1}{\epsilon}}} = \frac{w_n^k L_n^{kj}}{\sum_{k'} w_n^{k'} L_n^{k'j}}$$

If we log-linearize around a national weighted average across cities, where we weight individual cities by their wage bill, we have that

$$\frac{(\lambda^{kj} L^{kj})^{\frac{\epsilon-1}{\epsilon}}}{\sum_{k'} (\lambda^{kj} L^{kj})^{\frac{\epsilon-1}{\epsilon}}} = \sum_n \frac{w_n^k L_n^{kj}}{\sum_{k'} w_n^{k'} L_n^{k'j}} \times \frac{\sum_{k'} w_n^{k'} L_n^{k'j}}{\sum_{n',k'} w_{n'}^{k'} L_{n'}^{k'j}} = \frac{\sum_n w_n^k L_n^{kj}}{\sum_{n',k'} w_{n'}^{k'} L_{n'}^{k'j}} = \delta^{kj},$$

so that

---

<sup>7</sup>For non-tradable sectors, the city-specific share parameters make it challenging to compare this term across-cities. Furthermore, our data does not allow us to separate the productivity of the real-estate sector from the stock of housing.

$$d \ln \left( \sum_{n'} \pi_{n'n}^j X_{n'}^j \right) \simeq d \ln P_n^j - \frac{1}{\theta} d \ln \pi_{nn}^j + \gamma_n^j (1 - \beta_n^j) \sum_k \delta^{kj} \left( d \ln L_n^{kj} + d \ln \lambda_n^{kj} \right) \\ + \gamma_n^j \beta_n^j d \ln H_n^j + \sum_{j'} \gamma_n^{j'j} d \ln M_n^{j'j}$$

Comparing to the expression for TFP, it follows that, up to a first order approximation,

$$d \ln TFP_n^j \simeq -\frac{1}{\theta} d \ln \pi_{nn}^j + \gamma_n^j (1 - \beta_n^j) \sum_k \delta^{kj} \left( d \ln \lambda_n^{kj} \right)$$

which, abstracting from selection effects, reduces to

$$d \ln TFP_n^j = \gamma_n^j (1 - \beta_n^j) \sum_k \delta^{kj} \left( d \ln \lambda_n^{kj} \right)$$

Defining  $T_n^{kj} \equiv \left( \lambda_n^{kj} \right)^{(1-\beta_n^j)\gamma_n^j} \left( H_n^{kj} \right)^{\beta_n^j \gamma_n^j}$  it follows that for tradable sectors (in which case  $\beta_n^j = 0$  and  $\gamma_n^j = \gamma^j$  for all  $n$ ),

$$d \ln TFP_n^j = \sum_k \delta^{kj} d \ln T_n^{kj}$$

or

$$\ln TFP_n^j = \sum_k \delta^{kj} \ln T_n^{kj} + \text{constant independent of } n$$

For the purposes of comparing  $TFP_n^j$  across space, we can omit that constant. In the remainder of the paper, we let  $T_n^{kj} = \left( \lambda_n^{kj} \right)^{\gamma^j}$ .

Table A-3 compares estimates of productivity elasticities with respect to city size and employment composition obtained from our model inversion to those found in previous work. In particular, we report elasticities with respect to city size from the meta analysis carried out in [Melo et al. \(2009\)](#). As reported in Table A-3, all results point to a positive relationship between TFP and city size. Moreover, our findings fall within the range of reduced form estimates found in the literature, with the possible exception of services. However, as in previous literature, the elasticity of (tradable) services productivity with respect to city size is substantially larger than that of manufacturing. To the extent that regional prices are not readily available, the relationship between TFP and city size estimated in some of the existing literature captures variations in nominal TFP, that is  $\ln TFP_n^j + \ln P_n^j$ . In other words, while our model inversion produces measures of  $P_n^j$ , the absence of local price data can otherwise bias downward empirically estimated elasticities of TFP

Table A-3: Elasticities of TFP with respect to City Size and Employment Composition

	$\ln(L_n)$		$\frac{L_n^{CNR}}{L_n}$	
	Real	Nominal	Real	Nominal
Average <sup>1</sup>	0.04	0.03	0.81	0.63
Manufacturing Average	0.03	0.02	0.58	0.46
Tradable Services Average	0.05	0.04	1.08	0.84
Melo et. al. Economy <sup>2</sup>	0.03 (0.01)		—	
Melo et. al. Manufacturing	0.04 (0.01)		—	
Melo et. al. Services	0.15 (0.15)		—	
Moretti College Share <sup>3</sup> (Manufacturing)	—		0.84 (0.10)	

Average coefficients from univariate OLS regression of  $\ln TFP_n^j$  defined in equation (A-38) on  $\ln L_n$  and  $L_n^{CNR}/L_n$ .

1. Excludes non-tradables
2. [Melo et al. \(2009\)](#) "A Meta-analysis of estimates of urban agglomeration economies", Table 2. "By type of response variable" and "By industry group."
3. [Moretti \(2004b\)](#) "Workers' Education, Spillovers, and Productivity: Evidence from Plant-Level Production Functions", Table 2. College share in other industries, Cobb-Douglas production, 1992.

with respect to city size. Indeed, our findings indicate that elasticities of real TFP with respect to city-size are somewhat larger than those of nominal TFP.

We also compare our TFP regressions coefficients with respect to the share of CNR workers to those estimated by [Moretti \(2004b\)](#) using a panel of firms (we use the CNR share of employment, whereas he uses the college educated share of employment). Again, our TFP measures are consistent with semi-elasticities that are of the same sign and comparable in magnitude to those in [Moretti \(2004b\)](#). As before, the regression coefficients become larger when deflated by the model-consistent regional price index.

Table A-4 shows the coefficients in Table A-3 for tradable sectors disaggregated by industry. We find that the positive relationship between TFP and city size holds uniformly across all tradable sectors. In addition, we find that the semi-elasticity of TFP in Computer and Electronics with respect to employment composition across cities is more than twice as large as the average for manufacturing, replicating the finding by [Moretti \(2004b\)](#) for high-tech sectors.

## 2.4 Controls for Determinants of Productivity

OLS estimates or production externalities are potentially biased since workers of a given type may choose to live in cities where they are relatively most productive. This would induce a correlation between the exogenous component of worker productivity,  $\hat{T}_n^{kj}$ , and the share of each type of worker in a given city. Moreover, the estimates might be biased if there are omitted variables which are correlated with both  $\hat{T}_n^{kj}$  and the occupational ratio.

Table A-4: Sectoral Elasticities of TFP with respect to city size and employment composition

	$\ln(L_n)$		$\frac{L_n^{CNR}}{L_n}$	
	Real	Nominal	Real	Nominal
Food and Beverage	0.02	0.01	0.43	0.32
Textiles	0.03	0.01	0.29	0.25
Wood, Paper, and Printing	0.02	0.01	0.58	0.29
Oil, Chemicals, and Nonmetallic Minerals	0.05	0.03	0.93	0.80
Metals	0.02	0.01	0.49	0.27
Machinery	0.02	0.01	0.43	0.37
Computer and Electronic	0.06	0.04	1.48	1.22
Electrical Equipment	0.02	0.01	0.41	0.36
Motor Vehicles (Air, Cars, and Rail)	0.02	0.01	0.54	0.39
Furniture and Fixtures	0.02	0.01	0.18	0.28
Miscellaneous Manufacturing	0.03	0.02	0.64	0.50
Wholesale Trade	0.05	0.04	0.92	0.82
Transportation and Storage	0.03	0.03	0.60	0.50
Professional and Business Services	0.05	0.04	1.15	0.83
Other	0.06	0.05	1.07	0.95
Communication	0.05	0.04	0.10	0.94
Finance and Insurance	0.06	0.05	1.36	1.14
Education	0.08	0.05	1.63	1.10
Health	0.07	0.04	1.42	0.79
Accommodation	0.04	0.03	0.60	0.49

Coefficients from univariate OLS regression of  $\ln TFP_n^j$  defined in equation (A-38) on  $\ln L_n$  and  $L_n^{CNR}/L_n$ .

To help address the omitted variable bias, Table A-5 explores the effects of adding various controls to our basic OLS regression. Column 2 includes dummies for 9 census divisions interacted with industry dummies.<sup>8</sup> These should absorb many of the geographical and historical components that may jointly determine amenities and productivity in different places. Column 3 introduces geographic amenities constructed by the United States Department of Agriculture (USDA) that include measures of climate, topography and water area.<sup>9</sup> These controls allow for the possibility that the same geographic characteristics that may lead workers to choose certain cities may also influence their productivity. Column 4 introduces the share of manufacturing workers in 1920 as a control.<sup>10</sup> This aims to extract long standing factors that may influence the industrial composition in individual places. Finally, column 5 adds controls for demographic characteristics of different cities, including racial composition, gender split, the fraction of immigrant population, and age structure.<sup>11</sup> Together, these controls help narrow down the identification of the externality

<sup>8</sup>They are 1. New England, 2. Mid-Atlantic, 3. East North Central, 4. West North Central, 5. South Atlantic, 6. East South Central, 7. West SouthCentral, 8. Mountain and 9. Pacific.

<sup>9</sup>Geographic controls include average temperature for January and July, hours of sunlight in January, humidity in July from 1941 to 1970, variation in topography, and percent of water area.

<sup>10</sup>Just as with our labor force variables of interest, this and other controls are likewise interacted with the value added shares  $\gamma_n^j$ .

<sup>11</sup>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).

Table A-5: OLS Estimates

VARIABLES	(1)		(2)		(3)		(4)		(5)	
	CNR	non-CNR	CNR	non-CNR	CNR	non-CNR	CNR	non-CNR	CNR	non-CNR
$\ln(\frac{L_n^k}{L_n})$	0.82*** (0.13)	0.63*** (0.23)	0.81*** (0.12)	0.67*** (0.20)	0.84*** (0.12)	0.69*** (0.22)	0.83*** (0.12)	0.67*** (0.21)	0.89*** (0.12)	0.70*** (0.22)
$\ln(L_n)$	0.42*** (0.05)	0.35*** (0.04)	0.42*** (0.05)	0.34*** (0.04)	0.41*** (0.05)	0.35*** (0.04)	0.41*** (0.05)	0.34*** (0.04)	0.39*** (0.05)	0.32*** (0.04)
Jan. Temp					0.02 (0.03)	-0.06** (0.03)	0.01 (0.03)	-0.06** (0.02)	0.01 (0.03)	-0.04** (0.02)
Jan. Hrs Sun					0.08*** (0.01)	0.05*** (0.01)	0.07*** (0.01)	0.05*** (0.01)	0.07*** (0.02)	0.06*** (0.02)
July Temp					-0.08*** (0.03)	-0.03 (0.02)	-0.08*** (0.02)	-0.02 (0.02)	-0.07*** (0.02)	-0.04** (0.02)
July Humid					-0.02 (0.02)	-0.02 (0.03)	-0.02 (0.02)	-0.01 (0.03)	0.00 (0.02)	-0.03 (0.02)
Topography					-0.03* (0.01)	-0.03** (0.01)	-0.03* (0.01)	-0.02** (0.01)	-0.04*** (0.01)	-0.02 (0.01)
% Water Area					0.04*** (0.01)	0.00 (0.01)	0.04*** (0.01)	0.00 (0.01)	0.03*** (0.01)	0.01 (0.01)
% 1920 Mfg Workers							-0.00 (0.01)	0.01 (0.01)	-0.00 (0.01)	0.00 (0.01)
% female									-0.06*** (0.01)	-0.05*** (0.01)
% black									0.00 (0.02)	0.04*** (0.01)
% hispanic									0.03* (0.02)	0.00 (0.01)
% Age 16-25									0.01 (0.01)	0.04** (0.02)
% Age 26-65									0.04** (0.01)	0.07*** (0.02)
Industry FE	X	X	X	X	X	X	X	X	X	X
Census Division FE			X	X	X	X	X	X	X	X
Observations	7,640	7,640	7,640	7,640	7,560	7,560	7,460	7,460	7,460	7,460
R-squared	0.49	0.39	0.51	0.43	0.53	0.45	0.53	0.45	0.54	0.46

Regressions estimate equation (6) in main text multiplied by  $\gamma_n^j$  to correct for heteroskedasticity in measurement errors (see Footnote 15 in main text). Dependent variable is  $\ln \lambda_n^{kj}$  obtained from model inversion procedure. Individuals are classified by whether they are in CNR occupations and have more years of study than college or not. Standard errors in parentheses, clustered two-ways by city and by industry.

\*\*\*p<0.01, \*\* p<0.05, \* p<0.1

coefficients to the extent that more productive cities attract individuals of certain demographic make-up.

The point estimates of the coefficients on CNR workers change only slightly with the controls, while they increase the effect of labor market composition on non-CNR workers. These controls help extract exogenous sources of productivity variation that affect individual location decision.

## 2.5 Instrumenting for Employment

In order to isolate the residual simultaneity between exogenous productivity variation and labor allocation, we resort to variants of instruments proposed in the literature. Specifically, we follow

Ciccone and Hall (1996) and use population a century prior to our data period to capture historical determinants of current population, and we follow Card (2001) and Moretti (2004a), and use variation in early immigrant population and the presence of land-grant colleges to capture historical determinants of skill composition of cities. We now discuss the particular instruments in more detail.

**Population in 1920** Ciccone and Hall (1996) argue for the validity of historical variables as instruments under the assumption that, after allowing for the controls described above, original sources of agglomeration only affect current population patterns through the preferences of workers, and not through their effect on the residual component of productivity. This reasoning motivates using population almost one hundred years prior to our data period as an instrument and will also serve as motivation for the other instruments, described below.

**Irish immigration in 1920** Next, we use the fraction of Irish immigrants in the population of each city in 1920. This instrument is motivated by Card (2001), who uses the location of immigrant communities as an instrument for labor supply in different occupations. For our purposes, we focus on the location of Irish immigrants following evidence reviewed by Neal (1997), and further studied by Altonji et al. (2005), showing that attending catholic schools substantially increases the likelihood of completing high school and college education. We use as an instrument the fraction of Irish immigrants, rather than the overall catholic population, because Irish immigrants represented the first wave of catholic immigration to the U.S. and, therefore, historically were the first to invest in education. As additional validation for this instrument, we compile data on the current location of catholic colleges, and observe that MSAs in which catholic colleges are present had in 1920 more than three times the fraction of Irish immigrants as other locations.

**The Presence of land-grant colleges** Lastly, following Moretti (2004a), we also use as an instrument the presence of a land-grant college within the city. Land-grant colleges were established as a result of the Morrill Act of 1862, and extended in 1890, a federal act that sought to give states the opportunity to establish colleges in engineering and other sciences. Since the act is more than a century old, the presence of a land-grant college in the city is unlikely to be related to unobservable factors affecting productivity in different cities over our base period, 2011 – 2015. At the same time, as shown in Moretti (2004a), the presence of land-grant colleges is generally correlated with the composition of skills across cities.

## 2.6 A check on Instruments: Estimates in Counterfactual Without Externalities

Table A-6 below shows the results from carrying out the same estimation exercises as in Table 2 in the text using employment and productivity values obtained from a counterfactual equilibrium in which externality elasticities,  $\tau^{R,k}$  and  $\tau^{L,k}$ , are set to zero but all other model parameters are kept at their original levels.

Table A-6: Estimates with data generated by counterfactual without externalities

VARIABLES	(1)		(2)		(3)	
	<u>OLS</u>		<u>2SLS</u>		<u>CUE</u>	
	CNR	non-CNR	CNR	non-CNR	CNR	non-CNR
$\ln(\frac{L_n^k}{L_n})$	-1.00*** (0.29)	-0.19 (0.47)	-0.28 (0.73)	-0.93 (0.94)	-0.06 (0.72)	0.08 (0.93)
$\ln(L_n)$	0.02 (0.06)	-0.04 (0.04)	-0.03 (0.06)	-0.07 (0.04)	0.00 (0.06)	0.00 (0.04)
Observations	7,460	7,460	7,460	7,460	7,460	7,460
R-squared	0.01	0.01	0.00	-0.00	0.00	-0.00
J Test P-Value			0.38	0.12	0.39	0.122
K.P. F			7.30	7.36	7.30	7.36
S.W.F. $L_n^k$ Share			11.54	11.59	11.54	11.59
S.W.F. $L_n$			14.57	17.06	14.57	17.06

Robust standard errors in parentheses

\*\*\* p&lt;0.01, \*\* p&lt;0.05, \* p&lt;0.1

Regressions estimate equation (6) in main text multiplied by  $\gamma_n^j$  to correct for heteroskedasticity in measurement errors (see Footnote 15 in main text). Dependent variable is  $\ln \lambda_n^{kj}$  obtained from model inversion procedure. Individuals are classified by whether they are in CNR occupations and have more years of study than college or not. Standard errors in parentheses, clustered two-ways by city and by industry.

\*\*\*p<0.01, \*\* p<0.05, \* p<0.1

## 2.7 Model-Implied IV

In this exercise, we estimate externalities using an IV implied by the model. This is obtained by calculating the counter-factual allocation associated with an economy where, for any given industry/occupation category, productivity is constant across cities, and using the resulting counter-factual labor allocation as instruments.

This instrument will correct for a reverse causality problem since, by construction, there is no exogenous variation in productivity across cities. Table A-7 below shows the estimates. The F-statistics are very large, implying no need to explore GMM-CUE estimates. Moreover, the coefficients present the same general pattern as our baseline estimates: occupational externalities are generally stronger than those associated with total population.

## 3 The Planner's Problem

This section describes the solution to the planner's problem taking as given that workers in different occupations can freely choose in which city to live. Under this assumption, the expected utility of a worker of type  $k$  is given by equation (A-3). Given welfare weights for each occupation,  $\phi^k$ , the

Table A-7: Externality estimates with Model Implied IV's

VARIABLES	(1)		(2)	
	<u>OLS</u>		<u>2SLS</u>	
	CNR	non-CNR	CNR	non-CNR
$\ln\left(\frac{L_n^k}{L_n}\right)$	0.89*** (0.12)	0.70*** (0.22)	0.75*** (0.12)	0.45** (0.21)
$\ln(L_n)$	0.39*** (0.05)	0.32*** (0.04)	0.41*** (0.05)	0.31*** (0.04)
Observations	7,460	7,460	7,460	7,460
R-squared	0.42	0.32	0.42	0.32
K.P. F			600.2	595.5
S.W.F. $L_n^k$ Share			1705	1778
S.W.F. $L_n$			1187	1205

Regressions estimate equation (6) in main text multiplied by  $\gamma_n^j$  to correct for heteroskedasticity in measurement errors (see Footnote 15 in main text). Dependent variable is  $\ln \lambda_n^{kj}$  obtained from model inversion procedure. Individuals are classified by whether they are in CNR occupations and have more years of study than college or not. Standard errors in parentheses, clustered two-ways by city and by industry.

\*\*\*p<0.01, \*\* p<0.05, \* p<0.1

utilitarian planner then solves

$$\mathcal{W} = \sum_k \phi^k U \left[ \Gamma \left( \frac{\nu-1}{\nu} \right) \left( \sum_{n=1}^N (A_n^k C_n^k)^\nu \right)^{\frac{1}{\nu}} \right] L^k, \quad (\text{A-39})$$

where recall that  $C_n^k$  aggregates final goods from different sectors:

$$C_n^k = \prod_j (C_n^{kj})^{\alpha^j}. \quad (\text{A-40})$$

The planner maximizes (A-39) subject to the resource constraints for final goods,

$$\sum_k L_n^k C_n^{kj} + \sum_{j'} \int M_n^{jj'}(\mathbf{z}) d\Phi(\mathbf{z}) = \left( \int \left[ \sum_{n'} Q_{nn'}^j(\mathbf{z}) \right]^{\frac{\eta-1}{\eta}} d\Phi(\mathbf{z}) \right)^{\frac{\eta}{\eta-1}}, \quad (\text{A-41})$$

where  $Q_{nn'}^j(\mathbf{z})$  are the purchases of intermediate goods produced in city  $n'$  by final goods firms in city  $n$ , the resource constraints for intermediate goods of all varieties  $\mathbf{z}$  and industries  $j$  produced in all cities  $n$

$$\sum_{n'} Q_{n'n}^j(\mathbf{z}) \kappa_{n'n}^j = q_n^j(\mathbf{z}), \quad \forall \mathbf{z} \in \mathbb{R}_n^+, \quad (\text{A-42})$$

where

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

labor markets constraints in all locations,

$$\sum_j \int L_n^{kj}(\mathbf{z}) d\Phi(\mathbf{z}) = L_n^k, \quad (\text{A-43})$$

where labor supply in each city,  $L_n^k$ , satisfies

$$L_n^k = \frac{(A_n^k C_n^k)^\nu}{\sum_{n'} (A_{n'}^k C_{n'}^k)^\nu} L^k, \quad (\text{A-44})$$

the resource constraints in the use of land and structures,

$$\sum_j \int H_n^j(\mathbf{z}) d\Phi(\mathbf{z}) = H_n, \quad (\text{A-45})$$

as well as non-negativity constraints applying to both household consumption of different goods and input flows:

$$C_n^{kj} \geq 0 \text{ and } Q_{n'n}^j(\mathbf{z}) \geq 0.$$

From the resource constraint on local labor markets (A-43), and the labor supply condition (A-44), it follows immediately that national labor markets clear (i.e.,  $\sum_{n,j} \int L_n^{kj}(\mathbf{z}) d\Phi(\mathbf{z}) = L^k$ ).

### 3.1 Solving the Planner's Problem

We solve the Planner's problem for interior allocations, (i.e., where  $C_n^k$  and  $L_n^k$  are strictly greater than zero for all  $n$  and  $k$ ). For each city  $n$  and sector  $j$ , let  $P_n^j$  be the Lagrange multiplier corresponding to the final goods resource constraint in city  $n$ , sector  $j$  (A-41),  $\tilde{P}_n$  the multiplier corresponding to the aggregation of sectoral goods in each city (A-40), and  $\tilde{p}_n^j(\mathbf{z})$  the multiplier corresponding to the intermediate goods resource constraints (A-42). For each city  $n$  and occupation  $k$ , let  $w_n^k$  be the multiplier corresponding to regional labor market clearing (A-43),  $W_n^k$  the multiplier corresponding to the definitions of employment in each occupation and sector (A-44). Finally, for each city  $n$ , let  $r_n$  denote the multiplier corresponding to market clearing for structures (A-45).

The first-order conditions associated with the planner's problem are:

$$\partial C_n^{kj} : \tilde{P}_n \alpha^j \frac{C_n^k}{C_n^{kj}} = P_n^j L_n^k, \quad (\text{A-46})$$

which also defines an ideal price index,

$$P_n = \frac{\tilde{P}_n}{L_n^k} = \prod_j \left( \frac{P_n^j}{\alpha^j} \right)^{\alpha^j}. \quad (\text{A-47})$$

In addition,

$$\begin{aligned} \partial C_n^k &: \phi^k U' (v^k) v^k \frac{(A_n^k C_n^k)^\nu}{\sum_{n'} (A_{n'}^k C_{n'}^k)^\nu} \frac{1}{C_n^k} L^k \\ &= L_n^k P_n - \sum_{n'=1}^N \frac{\partial \zeta_{n'}^k(\mathbf{C}^k)}{\partial C_n^k} W_{n'}^k, \end{aligned} \quad (\text{A-48})$$

where

$$v^k = \Gamma \left( \frac{\nu - 1}{\nu} \right) \left( \sum_{n'} (A_{n'}^k C_{n'}^k)^\nu \right)^{\frac{1}{\nu}}$$

and

$$\frac{\partial \zeta_{n'}^k(\mathbf{C}^k)}{\partial C_n^k} = \begin{cases} \left( \frac{\nu}{C_n^k} \right) \left( 1 - \frac{L_n^k}{L^k} \right) L_n^k & \text{if } n' = n \\ - \left( \frac{\nu}{C_n^k} \right) \left( \frac{L_{n'}^k}{L^k} \right) L_n^k & \text{if } n' \neq n \end{cases}.$$

Also

$$\partial L_n^k : \sum_{j=1}^J P_n^j C_n^{kj} - \tilde{w}_n^k + W_n^k = 0. \quad (\text{A-49})$$

where

$$\begin{aligned} \tilde{w}_n^k &= w_n^k \\ &+ \sum_j \int \frac{\partial z_n \left[ H_n^j(\mathbf{z})^{\beta_n^j} \left[ \sum_{k''} \left( \lambda_n^{k''j}(\mathbf{L}_n) L_n^{k''j}(\mathbf{z}) \right)^{\frac{\epsilon-1}{\epsilon}} \right]^{\frac{\epsilon}{\epsilon-1} (1-\beta_n^j)} \right]^{\gamma_n^j} \prod_{j'=1}^J M_n^{j'j}(\mathbf{z})^{\gamma_n^{j'j}}}{\partial L_n^k} \tilde{p}_n^j(\mathbf{z}) d\mathbf{z} \end{aligned} \quad (\text{A-50})$$

denotes the total social marginal value of an extra worker of type  $k$  in city  $n$ . On the production side, efficient allocations dictate

$$\partial Q_{nn'}^j(\mathbf{z}) : \begin{cases} Q_{nn'}^j(\mathbf{z}) > 0 & \text{if } \kappa_{nn'}^j \tilde{p}_{n'}^j(\mathbf{z}) = P_n^j (Q_n^j)^{\frac{1}{\eta}} \left[ \sum_{n'=1}^N Q_{nn'}^j(\mathbf{z}) \right]^{-\frac{1}{\eta}} d\Phi(\mathbf{z}) \\ Q_{nn'}^j(\mathbf{z}) = 0 & \text{if } \kappa_{nn'}^j \tilde{p}_{n'}^j(\mathbf{z}) > P_n^j (Q_n^j)^{\frac{1}{\eta}} \left[ \sum_{n'=1}^N Q_{nn'}^j(\mathbf{z}) \right]^{-\frac{1}{\eta}} d\Phi(\mathbf{z}) \end{cases}. \quad (\text{A-51})$$

This last equation delivers efficient trade shares,  $\pi_{nn'}^j$ , and prices,  $P_n^j$ , using the usual Eaton and

Kortum derivations. In addition,

$$\partial L_n^{kj}(\mathbf{z}) : \gamma_n^j (1 - \beta_n^j) \frac{q_n^j(\mathbf{z})}{L_n^{kj}(\mathbf{z})} \frac{\left( \frac{w_n^k}{(\lambda_n^{kj}(\mathbf{L}_n))} \right)^{1-\epsilon}}{\sum_{k'} \left( \frac{w_n^{k'}}{(\lambda_n^{kj}(\mathbf{L}_n))} \right)^{1-\epsilon}} \tilde{p}_n^j(\mathbf{z}) = w_n^k d\Phi(\mathbf{z}), \quad (\text{A-52})$$

$$\partial H_n^j(\mathbf{z}) : \gamma_n^j \beta_n^j \frac{q_n^j(\mathbf{z})}{H_n^j(\mathbf{z})} \tilde{p}_n^j(\mathbf{z}) = r_n d\Phi(\mathbf{z}), \quad (\text{A-53})$$

$$\partial M_n^{j'j}(\mathbf{z}) : \gamma_n^{j'j} \frac{q_n^j(\mathbf{z})}{M_n^{j'j}(\mathbf{z})} \tilde{p}_n^j(\mathbf{z}) = P_n^{j'} d\Phi(\mathbf{z}). \quad (\text{A-54})$$

With the usual manipulations of these equations, one obtains

$$\tilde{p}_n^j(\mathbf{z}) \equiv p_n^j(\mathbf{z}) d\Phi(\mathbf{z}) = \frac{x_n^j d\Phi(\mathbf{z})}{z_n}, \quad (\text{A-55})$$

where

$$x_n^j = B_n^j \left[ r_n^{\beta_n^j} \left[ \sum_k \left( \frac{w_n^k}{(\lambda_n^{kj}(\mathbf{L}_n))} \right)^{1-\epsilon} \right]^{\frac{1-\beta_n^j}{1-\epsilon}} \right]^{\gamma_n^j} \prod_{j'} (P_n^{j'})^{\gamma_n^{j'j}}, \quad (\text{A-56})$$

and  $B_n^j$  is defined as above.

## 4 Characterization of the Planner's Solution

In the decentralized equilibrium, the budget constraint of a household of type  $k$  in city  $n$  satisfies

$$P_n C_n^k = w_n^k + \chi^k,$$

where  $\chi^k = b^k \frac{\sum_{n'} r_{n'} H_{n'}'}{\sum_{n',j} L_{n'}^{k,j}}$ . In contrast, we now show that the consumption of a household of type  $k$  in city  $n$  implied by the planner's solution satisfies

$$P_n C_n^k = \frac{\nu}{1+\nu} \tilde{w}_n^k + \chi^k + R^k,$$

and recall that  $\tilde{w}_n^k$  is the social marginal product of labor associated with occupation  $k$  in city  $n$ .

**Proof:**

Equation (A-48) may alternatively be expressed as

$$\phi^k U'(v^k) v^k \frac{L_n^k}{C_n^k} = L_n^k P_n - \left( \frac{\nu}{C_n^k} \right) L_n^k W_n^k + \sum_{n'=1}^N \left( \frac{\nu}{C_n^k} \right) \left( \frac{L_{n'}^k}{L^k} \right) L_n^k W_{n'}^k,$$

where  $v^k$  is defined in equation (A-3). Alternatively, we have that

$$\underbrace{\phi^k U'(v^k) v^k - \sum_{n'} \nu \left( \frac{L_{n'}^k}{L^k} \right) W_{n'}^k}_{(1+\nu)(\chi^k + R^k)} = P_n C_n^k - \nu W_n^k.$$

Substituting for  $W_n^k$  from (A-49) in this last expression gives

$$P_n C_n^k = \nu \left( \tilde{w}_n^k - P_n C_n^k \right) + (1 + \nu) (\chi^k + R^k)$$

or

$$P_n C_n^k = \frac{\nu}{1 + \nu} \tilde{w}_n^k + \chi^k + R^k. \quad (\text{A-57})$$

□

Observe that we can also use (A-49) to write  $\chi^k + R^k$  as a function of prices,  $\tilde{w}_n^k$ ,  $P_n$ , and consumption,  $C_n^k$ . In particular,

$$\chi^k + R^k = \frac{\phi^k U'(v^k) v^k}{1 + \nu} - \frac{\nu}{1 + \nu} \sum_{n'} \left( \frac{L_{n'}^k}{L^k} \right) \left( \tilde{w}_n^k - P_{n'} C_{n'}^k \right).$$

We can then obtain an expression for the total consumption expenditures of households of type  $k$  by adding (A-57) across cities  $n$ , with the expression for  $\chi^k + R^k$  substituted in,

$$\phi^k U'(v^k) v^k L^k = \sum_n P_n C_n^k L_n^k. \quad (\text{A-58})$$

Substituting out  $\phi^k U'(v^k) v^k$  back into the expression for  $\chi^k + R^k$  and rearranging, we obtain

$$\chi^k + R^k = \frac{\sum_n P_n C_n^k L_n^k}{L^k} - \sum_n \frac{\nu^k}{1 + \nu} \left( \frac{L_n^k}{L^k} \right) \tilde{w}_n^k.$$

Finally, note that  $\sum_{n,k} P_n C_n^k L_n^k = \sum_{n,k} (w_n^k L_n^k + r_n H_n)$ , so that

$$\sum_k L^k (\chi^k + R^k) = \sum_{n,k} \frac{1}{1 + \nu} w_n^k L_n^k - \sum_{n,k} \frac{\nu}{1 + \nu} \left( \tilde{w}_n^k - w_n^k \right) L_n^k + \sum_n r_n H_n \quad (\text{A-59})$$

The individual values for  $\chi^k$  are determined to be such that equation (A-59) is satisfied.

#### 4.1 The Social and Private Marginal Value of Workers of Type $k$ in City $n$ (Proof of Lemma 1)

We first prove a more general version of Lemma 1 that allows for industrial heterogeneity in spillover parameters:

**Lemma 1** (general version). *Let  $\Delta_n^k$  denote the wedge between the private and the social marginal value of a worker in occupation  $k$  in city  $n$ . Then*

$$\Delta_n^k = \sum_{k',j} w_n^{k'} \frac{L_n^{k'j}}{L_n^k} \frac{\partial \ln \lambda_n^{k'j}(\mathbf{L}_n)}{\partial \ln L_n^k}. \quad (\text{A-60})$$

Solving the derivative in the equation defining the social value of workers of type  $k$  in city  $n$  (A-50), we obtain

$$\begin{aligned} & \tilde{w}_n^k - w_n^k \\ = & \sum_j \int \frac{\partial z_n^j \left[ H_n^j(\mathbf{z})^{\beta_n^j} \left[ \sum_{k'} \left( \lambda_n^{k'j}(\mathbf{L}_n) L_n^{k'j}(\mathbf{z}) \right)^{1-\frac{1}{\epsilon}} \right]^{\frac{\epsilon}{1-\epsilon} (1-\beta_n^j)} \right]^{\gamma_n^j} \prod_{j'=1}^J M_n^{j'j}(\mathbf{z})^{\gamma_n^{j'j}}}{\partial L_n^k} p_n^j(\mathbf{z}) d\Phi(\mathbf{z}) \end{aligned}$$

where  $p_n^j(\mathbf{z}) d\Phi(\mathbf{z}) = \tilde{p}_n^j(\mathbf{z})$ . This expression is equivalent to

$$\tilde{w}_n^k - w_n^k = \sum_{j,k'} (1 - \beta_n^j) \gamma_n^j \frac{\left( \frac{w_n^{k'}}{\lambda_n^{k'j}(\mathbf{L}_n)} \right)^{1-\epsilon}}{\sum_{k''} \left( \frac{w_n^{k''}}{\lambda_n^{k''j}(\mathbf{L}_n)} \right)^{1-\epsilon}} \frac{1}{\lambda_n^{k'j}(\mathbf{L}_n)} \frac{\partial \lambda_n^{k'j}(\mathbf{L}_n)}{\partial L_n^k} q_n^j(\mathbf{z}) p_n^j(\mathbf{z}) d\Phi(\mathbf{z}).$$

Rearranging and integrating equation (A-52) yields

$$w_n^k L_n^{kj} = (1 - \beta_n^j) \gamma_n^j \frac{\left( \frac{w_n^k}{\lambda_n^{kj}(\mathbf{L}_n)} \right)^{1-\epsilon}}{\sum_{k'} \left( \frac{w_n^{k'}}{\lambda_n^{k'j}(\mathbf{L}_n)} \right)^{1-\epsilon}} \int q_n^j(\mathbf{z}) p_n^j(\mathbf{z}) d\Phi(\mathbf{z}),$$

so that the expression for the deviation of private from social marginal product of labor simplifies further to

$$\tilde{w}_n^k - w_n^k = \sum_{j,k'} w_n^{k'} \frac{L_n^{k'j}}{L_n^k} \frac{\partial \ln \lambda_n^{k'j}(\mathbf{L}_n)}{\partial \ln L_n^k}. \quad (\text{A-61})$$

It is a simple matter to check that if  $\frac{\partial \ln \lambda_n^{k'j}(\mathbf{L}_n)}{\partial \ln L_n^k}$  is the same for all  $j$ , then the expression in Lemma 1 (general version) reduces to the one in Lemma 1 in the text.

## 4.2 Implementation (Proof of Proposition 1)

We now discuss the implementation of the optimal policy. One possible implementation is to combine a direct employment subsidy to firms that is specific to cities and occupations ( $\Delta_n^k$ ), a linear occupation-specific labor income tax ( $t_L^k$ ), combined with occupation-specific transfers ( $R^k$ ).

With externalities in occupations, the social and private marginal products of labor differ. The first step in the implementation of optimal allocations, therefore, is to subsidize firms in different

locations to hire different occupation types. We define  $\tilde{w}_n^k$  to be the after-subsidy wage associated with workers in occupation  $k$  living in city  $n$  such that

$$\tilde{w}_n^k = w_n^k + \Delta_n^k,$$

where  $\Delta_n^k$  is a per-worker subsidy offered to firms in city  $n$  hiring workers of type  $k$ . With these subsidized wages in place, we take advantage of various additional taxes and transfers to implement optimal allocations. In particular, equation (A-2) becomes

$$I_n^k = (1 - t_L^k)\tilde{w}_n^k + \chi^k + R^k, \quad (NK \text{ eqs.}) \quad (\text{A-62})$$

where transfers have to be such that the government budget balances,

$$\sum_{n,k} L_n^k R^k = \sum_{n,k} t_L^k w_n^k L_n^k - \sum_{n,k} (1 - t_L^k) \Delta_n^k L_n^k. \quad (\text{A-63})$$

We also have that labor demand depends only on pre-subsidy wages,  $w_n^k$ ,

$$w_n^k L_n^{kj}(\mathbf{z}) = \frac{\left(\frac{w_n^k}{\lambda_n^{kj}(\mathbf{L}_n)}\right)^{1-\epsilon}}{\sum_{k'} \left(\frac{w_n^k}{\lambda_n^{k'j}(\mathbf{L}_n)}\right)^{1-\epsilon}} \gamma_n^j (1 - \beta_n^j) p_n^j(\mathbf{z}) q_n^j(\mathbf{z}), \quad (NKJ \text{ eqs.}) \quad (\text{A-64})$$

$$x_n^j = B^j \left[ r_n^{\beta_n^j} \left[ \sum_k \left(\frac{w_n^k}{\lambda_n^{kj}(\mathbf{L}_n)}\right)^{1-\epsilon} \right]^{\frac{1-\beta_n^j}{1-\epsilon}} \right]^{\gamma_n^j} \prod_{j'} (P_n^{j'})^{\gamma_n^{j'j}}, \quad (\text{A-65})$$

**Definition 1.** An equilibrium with taxes and transfers is defined as the equilibrium without taxes and transfers but with the additional conditions that i)  $I_n^k$  is given by equation (A-62), ii) the first-order condition describing intermediate goods producers' labor demand is given by (A-64), iii) the cost index  $x_n^j$  is given by (A-65), and iv) the government budget constraint (A-63) is satisfied.

**Proposition.** *Let*

$$t_L^k = \frac{1}{1 + \nu}$$

$$\Delta_n^k = \sum_{k'j} w_n^{k'} \frac{L_n^{k'j}}{L_n^k} \frac{\partial \ln \lambda_n^{k'j}(\mathbf{L}_n)}{\partial \ln L_n^k}$$

and  $R^k$  such that

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

Then, if the planner's problem is globally concave, the equilibrium with taxes and transfers implements the optimal allocation.

*Proof.* 1) The first-order condition for household consumption choice (A-1) is identical to the first order condition for consumption in the planner's problem, (A-46). The modified budget constraint for the household (A-62) implies a relationship between consumption and prices identical to equation (A-57), which is derived from the first order conditions (A-48) and (A-49) in the planner's problem. At the same time, the optimal location decision for the household, (A-4) is identical to the free mobility constraint in the planner's choice (A-44) for a given set of consumption  $C_n^k$ .

2) The first order condition for factor demand for intermediate input producers, (A-5), (A-7) and (A-64) are identical to the first order conditions for the planner's problem (A-52), (A-53) and (A-54) once one uses equation (A-61) to substitute  $\tilde{w}_n^k$  out of (A-52).

3) The condition that a producer in city  $n$  and industry  $j$  imports a variety  $\mathbf{z}$  from city  $n'$  if and only if  $\kappa_{nn'}^j p_n^j(\mathbf{z}) = \min_{n''} \kappa_{nn''}^j p_{n''}^j(\mathbf{z})$  is implied by the first order condition for the planner's problem (A-51), given that  $Q_n^j(\mathbf{z}) = \sum_{n'} Q_{nn'}^j(\mathbf{z})$ ,  $\tilde{p}_n^j(\mathbf{z}) d\Phi(\mathbf{z}) = p_n^j(\mathbf{z})$ .

4) The first order condition associated with the optimal use of different varieties by final goods producers (A-12) is implied by (A-51) given that  $Q_n^j(\mathbf{z}) = \sum_{n'} Q_{nn'}^j(\mathbf{z})$ ,  $\tilde{p}_n^j(\mathbf{z}) d\Phi(\mathbf{z}) = p_n^j(\mathbf{z})$ , and  $P_n^j(\mathbf{z}) = \min_{n'} \kappa_{nn'}^j p_{n'}^j(\mathbf{z})$ .

5) The market clearing conditions for employment (equation A-14), structures (equation A-15), final goods (equation A-16) and intermediate goods (A-17) are identical to the resource constraints faced by the planner, respectively, (A-43) combined with (A-44), (A-45), (A-41) and (A-42).

6) In the planner's solution, equation (A-58) has to hold.

□

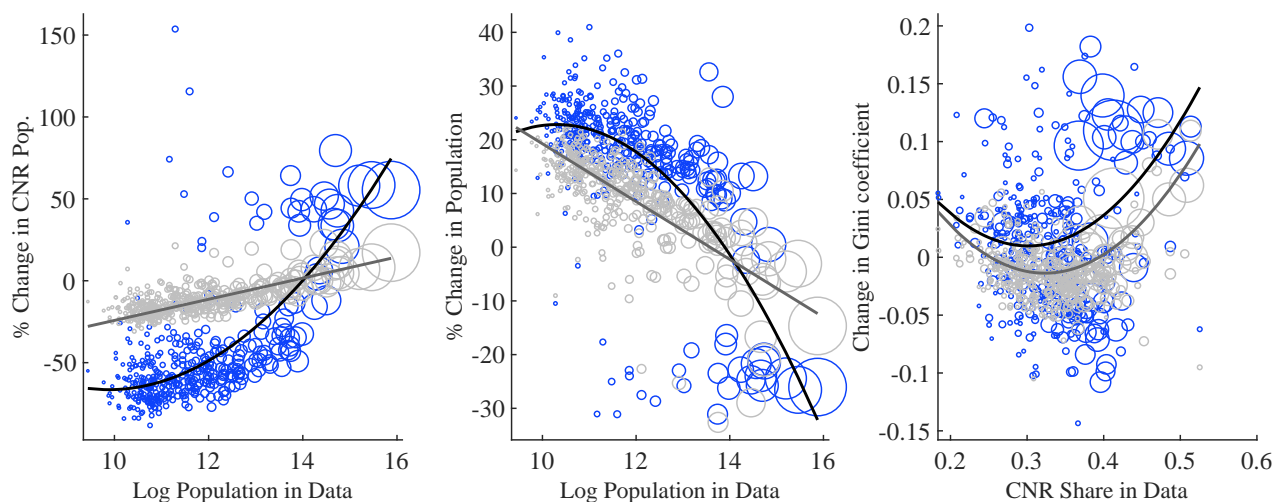
### 4.3 Robustness: Policy Without a Linear Labor Tax

As explained above, implementing the optimal allocation involves imposing a linear tax on labor income. This tax is associated with heterogeneous preferences for location that are complementary to consumption and are drawn from a Fréchet distribution. However, in the paper, we explain the main results with reference to Pigouvian taxes and subsidies captured by  $\Delta_n^k$ , and that allow individuals to internalize the external effects of their location shocks. One may, therefore, reasonably ask about the extent to which the results may depend on the linear labor tax rather than the Pigouvian taxes and subsidies.

We compute counterfactual allocations with Pigouvian taxes reflecting production spillovers (captured by  $\Delta_n^k$  in the text), but not the labor income tax that is implied by  $\nu$ , the curvature parameter of the Fréchet distribution. The results from this exercise are shown in Figure A-3. We find that the policy of concentrating cognitive hubs in large cities is kept in place. In fact, it becomes stronger and convex in city size. The converse result is that smaller cities gain proportionately more in size than larger cities under the policy. Finally, the U-shaped relationship between industry specialization and city-size is maintained.

These results indicate that the optimal policy of incentivizing CNR workers to live in large cities is not an artifact of the particular distribution of idiosyncratic productivity shocks, or how optimal policy reacts to these shocks. On the contrary, our assumptions regarding the distribution of these shocks leads the planner to attenuate the extent to which those incentives are emphasized.

Figure A-3: Allocations without the Labor Income Tax Implied by the Fréchet Parameter



Note: Each observation refers to a CBSA. Marker sizes are proportional to total employment. Blue markers refer to the counterfactual exercise and grey markers to the baseline from the paper. The solid black lines are linear or quadratic fits to the data. The Gini coefficient is constructed using the Lorenz curves depicting within city wage bill and industry rank.

## 5 Quantifying the Model for 1980 and Counterfactual Exercises

### 5.1 Quantifying the Model for 1980

In order to quantify the model for 1980, we follow similar steps as described in Section 2, with modifications to accommodate data constraints.

Regional Price Parities data are not available for 1980. In the baseline model quantification, we used those in order to calculate the productivity of the non-tradable sectors. To obtain the productivity of the real estate sector in 1980, we match instead changes in CoreLogic HPI data, available by county. As for the productivity of the non-tradable sector, we assume that its spatial distribution does not change. In addition, the model inversion exercise carried out for our 2011-15 benchmark does not pin down the national average level of productivity for each industry, only its occupational and spatial variation. In order to obtain the time variation of those levels, we choose average 1980 productivity levels to match national level sectoral price series made available by the BEA.

To obtain wages and the occupational composition of cities and industries, we use the 5% sample of the 1980 Census data which is comparable to the ACS. The 1980 Census has data for 213 MSA's

that account for approximately 85% of U.S. employment in that year. For the remaining MSA’s, we impute wages and employment by occupation and by sector by taking the predicted values of a regression of those variables on 1980 CBP employment by sector and housing prices.

## 5.2 Details of Counterfactual Exercises

In the counterfactual exercises described in Section 7 of the paper, we separate average changes in productivity or amenities from their geographical and occupational dispersion.

The first step is to study the consequences of changing factor shares. We focus on the consequences of those changes to factor demand, while keeping unit costs in individual cities and industries fixed. This exercise implies a set of alternative productivity parameters for 1980, which we then take as our base for comparison with the current period.<sup>12</sup> Productivity changes then refer to changes in  $T_n^{kj} = \left(H_n^j\right)^{\beta_n^j} \left(\lambda_n^{kj}\right)^{\gamma_n^j(1-\beta_n^j)}$ .<sup>13</sup> The average change in productivity between 1980 and 2011-15 for a given industry is a Tornqvist type index: a geometric weighted average of the changes in productivity across cities, with the weights given by the value added by each city/industry as a fraction of total industry value added. Those shares are first calculated separately for the 1980 and 2011-15 periods, and the weights correspond to the arithmetic average of those shares.<sup>14</sup>

The model does not allow us to pin down an aggregate trend in amenities since changing amenities in all cities by a common scaling parameter leaves the equilibrium unchanged. We thus assume that there was no such trend so as to focus on the welfare implications of endogenous changes in equilibrium variables. For the baseline economy, this implies keeping a Tornqvist type index of amenities constant relative to the 2011-15 period: specifically, we keep a weighted geometric average of changes in amenities equal to 1, with the weights given by employment shares by city (again the shares are taken for the baseline and 2011-15 periods separately and the weights are given by an arithmetic average).

## 6 Counterfactuals

### 6.1 Homogeneous Linkages

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

<sup>12</sup>One advantage of this procedure is that, given that changes in factor shares can be city-specific, implied productivity changes may otherwise depend on scaling parameters adopted for the different inputs.

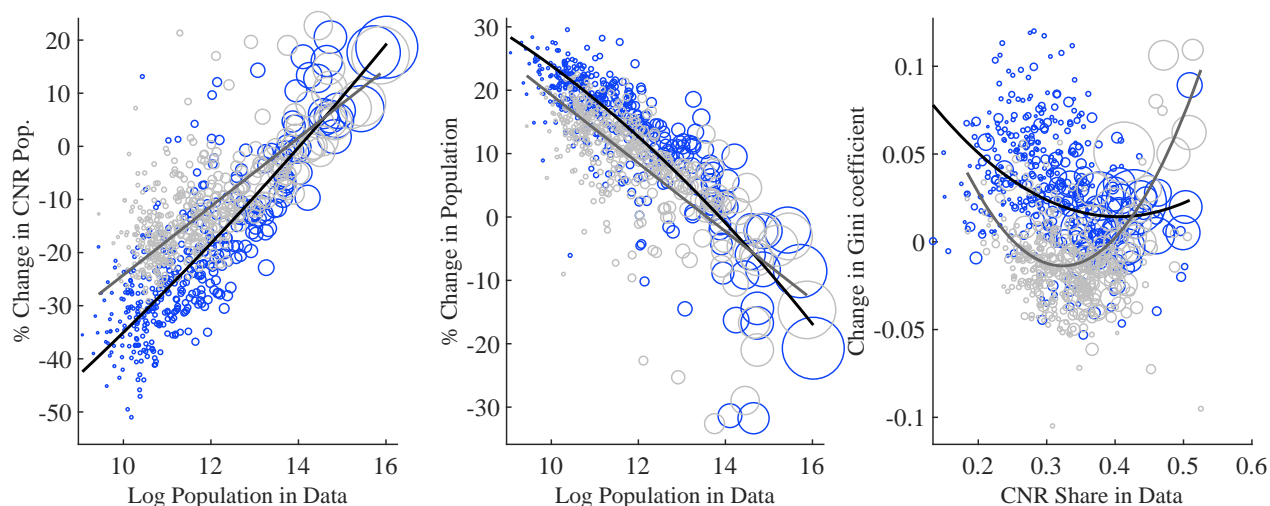
<sup>13</sup>Specifically, when calculating the average change in productivity for a given sector  $j$  and occupation  $k$  we set  $\ln(T_n^{kj, \text{counterfactual}}) = \gamma_n^j \sum_{n'} \omega_n^{kj} \frac{1}{\gamma_n^j} \ln(T_{n'}^{kj})$ , where  $\omega_n^{kj} = \frac{w_n^k L_n^{kj}}{\sum_{n'} w_{n'}^k L_{n'}^{kj}}$ , and analogously for other averages.

<sup>14</sup>We carry out a similar calculation in order to obtain productivity trends by city/industry/occupation

shown in the text, doing away with sectoral heterogeneity reproduces exactly the kind of U-shaped relationship between skill intensity and city size found by Fajgelbaum and Gaubert (2020).

Absent linkages the planner has good reasons to have some small cities specialize to a greater degree in health and education while letting larger cities specialize in communication and financial services. We further underscore this mechanism as a driving force by calculating an additional set of counterfactuals in which (i) we make links homogeneous as explained above and (ii) we remove all cross-city variation in the exogenous productivity component of health and education. In that counterfactual, shown in Figure A-4, the U-shape relationship disappears and the upward sloping line comes back.

Figure A-4: Homogenous Industry Links, Flat Exogenous Productivity in Health and Education



Each observation refers to a CBSA. Marker sizes are proportional to total employment. The Gini is constructed using the Lorenz curves depicting within city wage bill and industry rank.

## 6.2 No Trade Costs

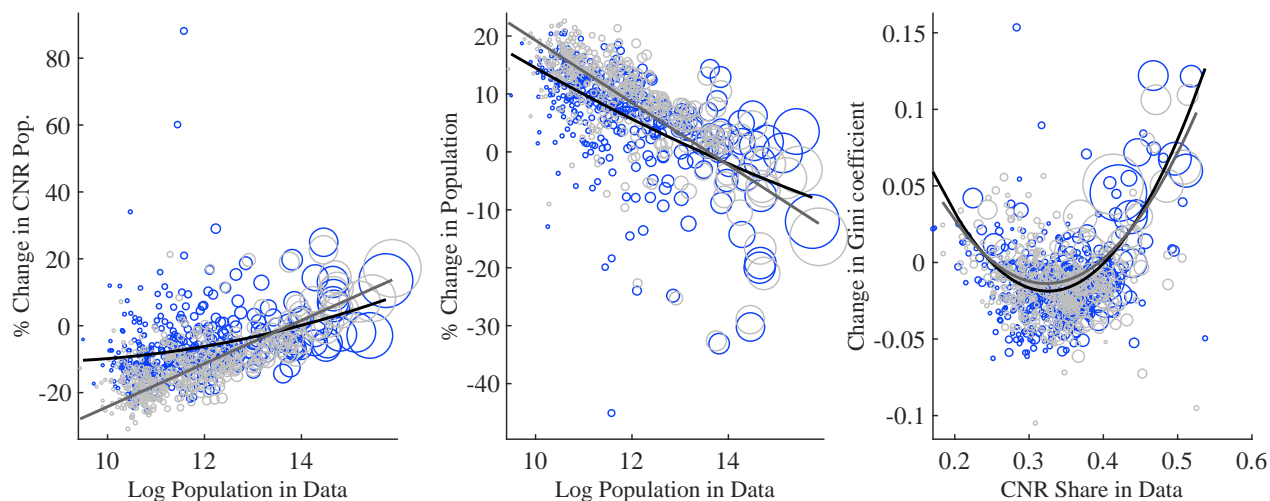
We compute a counterfactual equilibrium without transportation costs. Specifically, we set trade costs in all sectors (other than housing) to zero. The findings from this exercise are illustrated in Figure A-5. Eliminating trade costs does not materially affect our main qualitative findings. However, the reasons behind this result are instructive.

First, absent trade costs, optimal policy allows for a large increase in CNR intensity in some smaller cities that can now easily trade their output from CNR-intensive sectors with other locations. Consequently, the slope tying optimal CNR intensity to city size also becomes less steep. In other words, trade costs play a role in limiting CNR intensity in small cities. This observation aligns with the fact that CNR-intensive sectors in which those cities are most productive tend to be in health and education whose services are relatively difficult to trade (see Section 6.1 in the Paper). Second, and working against this first mechanism, is that 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 through externalities.

The net effect of these two countervailing forces is to leave our benchmark findings relatively unchanged. Said differently, optimal policy still tends to concentrate CNR workers in larger cities even if proximity between CNR-intensive sectors is no longer needed to take advantage of trade linkages.

Figure A-5: No Trade Costs



Comparison between counterfactual equilibrium and optimal allocations. Each observation refers to a CBSA. Marker sizes are proportional to total employment. Blue markers refer to the counterfactual exercise and grey markers to the baseline from the paper. The solid black lines are linear or quadratic fits to the data. The Gini coefficient is constructed using the Lorenz curves depicting within city wage bill and industry rank.

The way in which trade costs affect the optimal policy can be further viewed best in the maps in Figures A-6 and A-7. Unlike in Figure 6 in the revised version of the text, CNR workers end up mostly concentrated mostly in coastal cities such as New York, Washington DC and San Francisco as well as smaller nearby towns, without the formation of inland regional hubs in places such as Chicago, Minneapolis, Dallas or Atlanta.

### 6.3 A Counterfactual Economy Without Endogenous Amenities

We now verify whether the planner solution would be likely to change if one were to adjust local amenities to remove the components that Diamond (2016) argues are likely to be endogenous. For that purpose, we carry out two counterfactual exercises. For both exercises, we first extract the exogenous component of amenities as implied by the mapping of Diamond’s (2016) estimates into amenity spillovers described in Fajgelbaum and Gaubert (2020). Specifically, we calculate a value of  $A_n^{k,exo}$  such that  $A_n^{k,exo} \prod (L_n^{k'})^{\tau_a^{k'k}} C_n^k = 1$ , with  $\tau_a^{CNR,CNR} = 0.77$ ,  $\tau_a^{non-CNR,CNR} =$

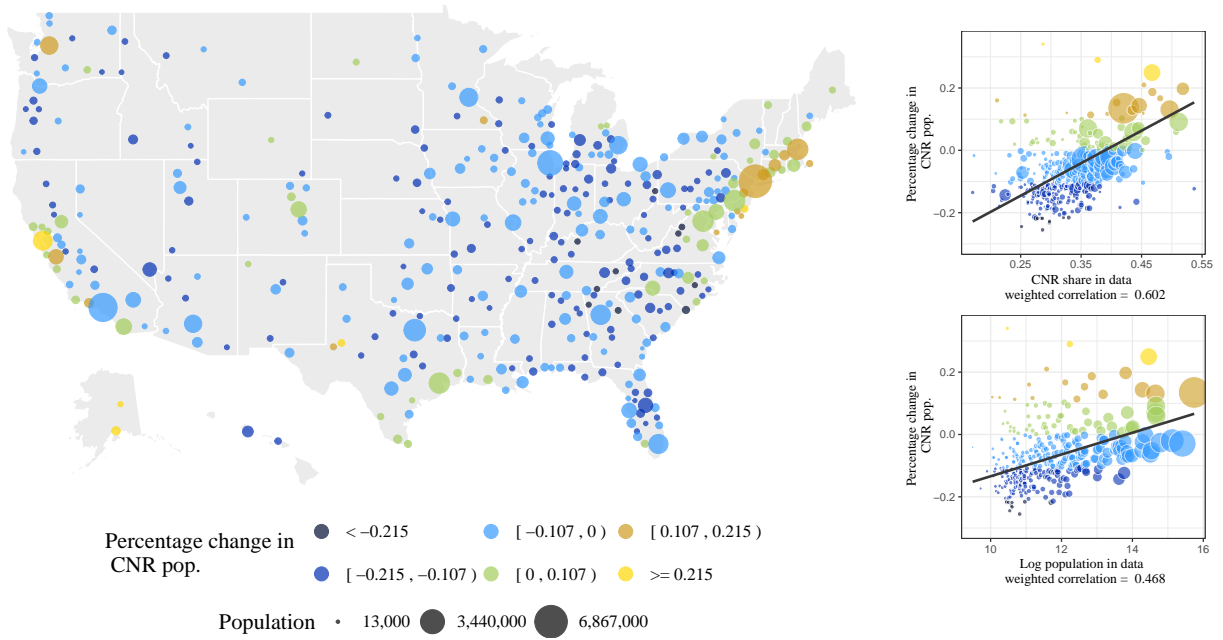


Figure A-6: Change in  $L_n^{\text{CNR}}$  from no trade costs equilibrium to optimal allocation)

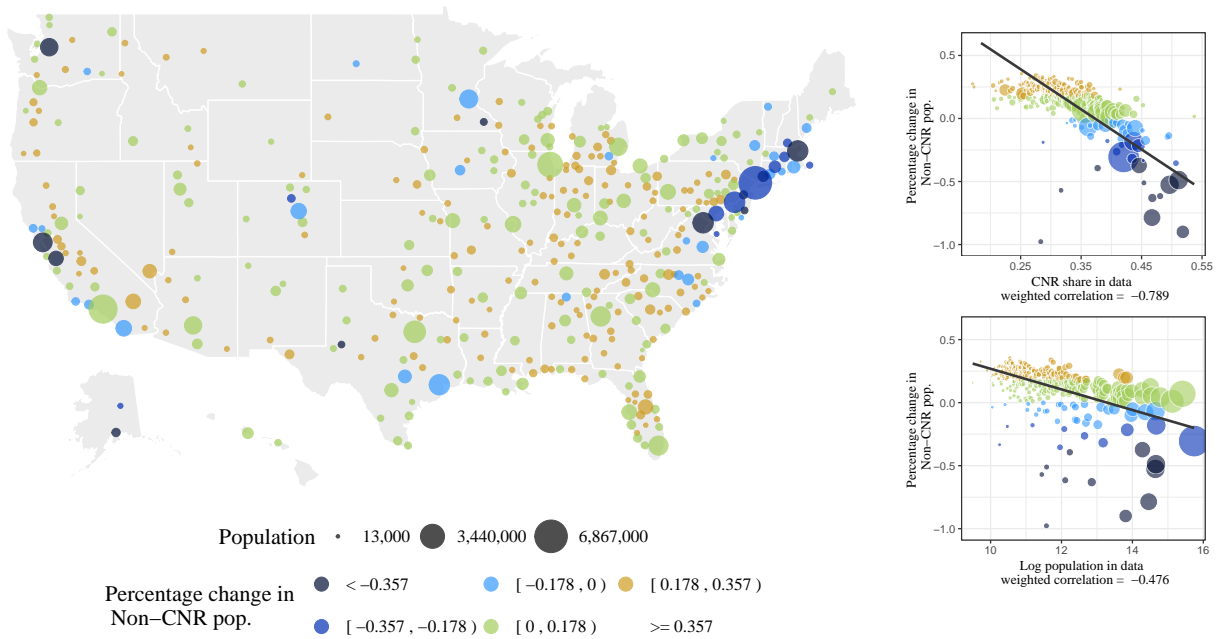


Figure A-7: Change in  $L_n^{\text{non-CNR}}$  from no trade costs equilibrium to optimal allocation

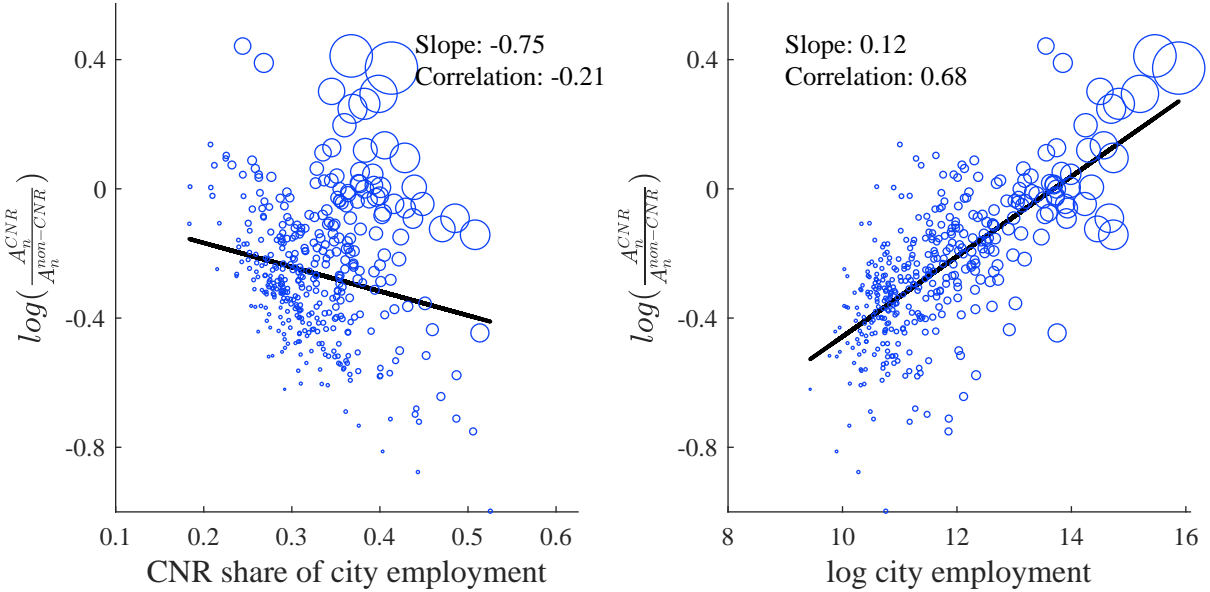


Figure A-8: Relative amenities and city size and composition (exogenous part)

Ratio of occupational-specific amenity parameters for each city obtained after extracting the endogenous part of amenities implied by the parametrization used by [Fajgelbaum and Gaubert \(2020\)](#). Each observation refers to a CBSA. Marker sizes are proportional to total city employment.

$-1.24$ ,  $\tau_a^{\text{CNR,non-CNR}} = 0.18$  and  $\tau_a^{\text{non-CNR,non-CNR}} = -0.43$ . In the first exercise, we calculate a counterfactual equilibrium where the labor supply equations are given by  $L_n^k = \frac{(A_n^{k,exo} C_n^k)^{-\nu^k}}{\sum_{n'} (A_{n'}^{k,exo} C_{n'}^k)^{-\nu^k}}$ . In the second exercise, we calculate the optimal allocation in that counterfactual environment.

Figures [A-8](#) below show the relationship between relative the exogenous part of amenities implied by that exercise and city size and composition, further discussed in the text.

Figure [A-9](#) shows how the distribution of CNR workers in the optimal allocation compares with the counterfactual equilibrium. As in our baseline economy, the planner has an incentive to increase labor market polarization by concentrating proportionately more CNR workers in larger cities. Figure [A-10](#) shows that, as in our baseline analysis, this increased polarization is matched by transfers from the large cities to the small ones.

## 7 Robustness

### 7.1 Stress-Testing Externality Parameters

Because larger cities also happen to have a larger share of CNR workers, it is difficult to disentangle scale from quality effects. Within our empirical approach, this is reflected in the covariance between estimated coefficients  $\tau^{R,k}$  and  $\tau^{L,k}$ .

To explore this question, we carry out a robustness exercise where we increase the external effects of population,  $\tau^{L,CNR}$  and  $\tau^{L,NCNR}$ , to their 84th percentile (i.e., 1 standard error above their point

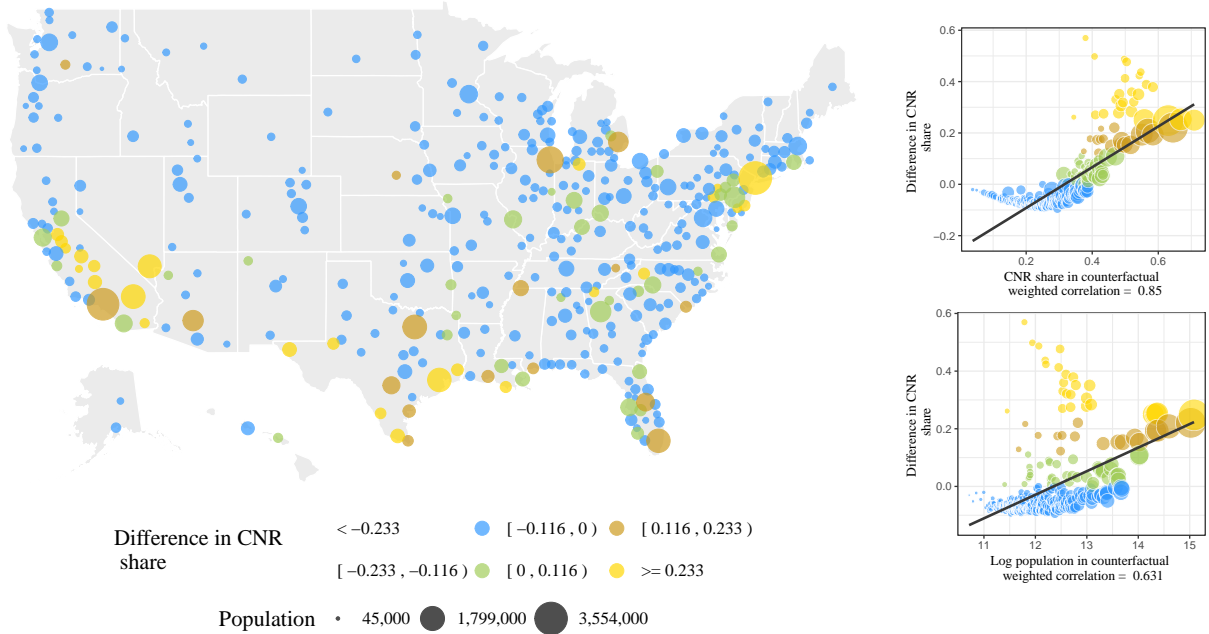


Figure A-9: Optimal  $L_n^{CNR}/L_n$  with counterfactual amenities (change from counterfactual equilibrium)

Each marker in the map refers to a CBSA. Marker sizes are proportional to total equilibrium employment in each city.  $\rho$  and  $\bar{\rho}$  are unweighted and population weighted correlations respectively.

estimates), and where we alter the effects of composition,  $\tau^{R,CNR}$  and  $\tau^{R,NCNR}$ , correspondingly (i.e., for a given choice of  $\tau^{L,k}$ , we set  $\tau^{R,k} = E[\hat{\tau}^{R,k}|\tau^{L,k}]$ ).<sup>15</sup> The resulting coefficients are given in Table A-8.

Table A-8: **Robust Point Estimates for Externalities**

Variables	CNR	non-CNR
$\gamma_n^j \log(\frac{L_n^k}{L_n})$	1.04	1.19
$\gamma_n^j \log(L_n)$	0.41	0.40

This exercise effectively puts the externality from overall population in the upper range of possibilities. Furthermore, given the correlation between population and composition, the externality from CNR workers onto to their own productivity becomes lower, but the externality from non-CNR workers onto their own productivity is now larger. Since the optimal policy of reinforcing cognitive hubs is especially important when CNR productivity is disproportionately affected by the share of CNR workers in a city, this exercise can only attenuate the findings in the paper. That said, results from this exercise, shown in Figure A-11, indicate that our benchmark results are

<sup>15</sup>In particular, using ‘tildes’ to denote robust values and ‘hats’ to denote estimated values, we set for each  $k \in \{CNR, NCNR\}$ ,  $\tilde{\tau}^{R,k} = \hat{\tau}^{R,k} + \frac{cov(\hat{\tau}^{R,k}, \hat{\tau}^{L,k})}{var(\hat{\tau}^{L,k})} (\tilde{\tau}^{L,k} - \hat{\tau}^{L,k})$ , where variances and covariances refer to elements of the covariance matrix of the estimation errors.

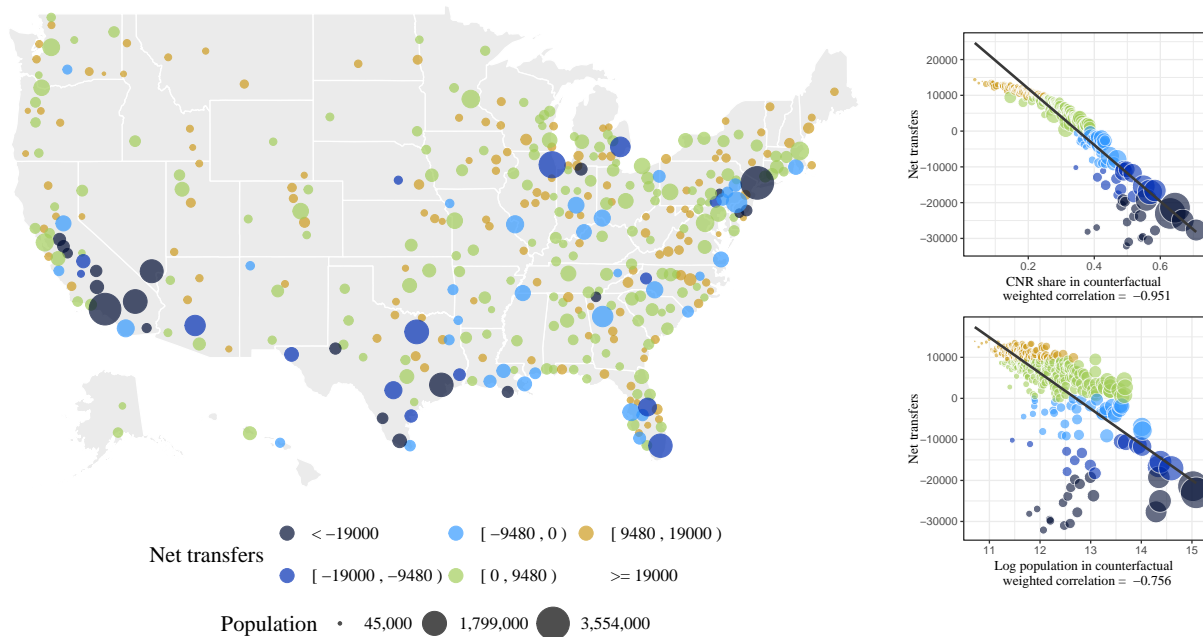


Figure A-10: Optimal transfers with counterfactual amenities

Optimal transfers defined as the difference in the optimal allocation between the value consumed and value added in each city ( $\sum_k P_n C_n^k - \sum_k w_n^k L_n^k - r_n H_n$ ). Each marker in the map refers to a CBSA. Marker sizes are proportional to total equilibrium employment in each city.  $\rho$  and  $\tilde{\rho}$  are unweighted and population weighted correlations respectively.

generally robust to, though mitigated by, this alternative composition of externalities.<sup>16</sup>

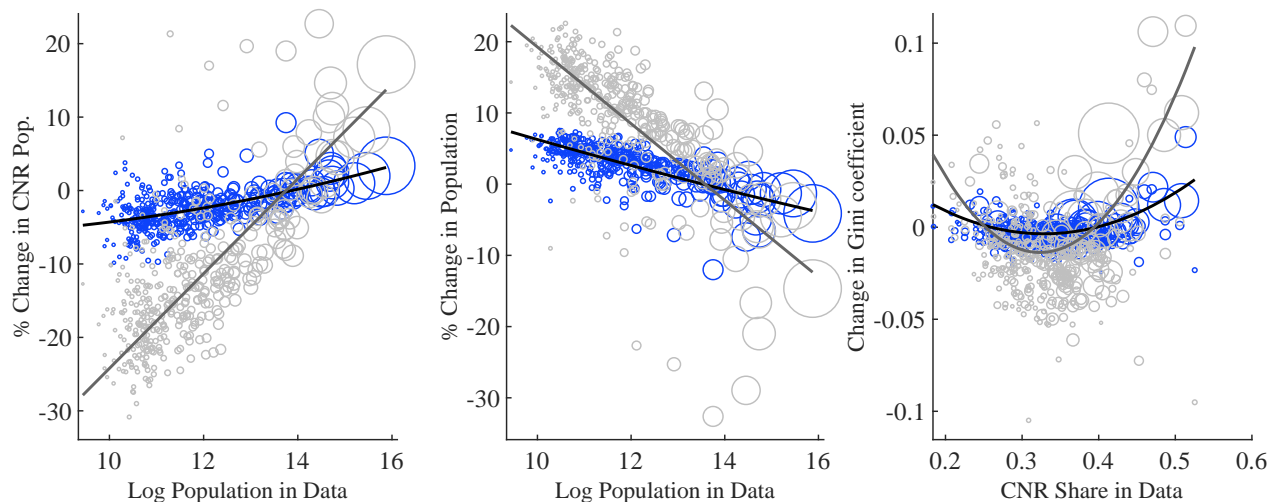
## 7.2 Industry-Specific Externality Parameters

In the paper, we abstract from detailed spillovers at the industry level. However, for reasons explained below, a more detailed parameterization of spillovers does not obviously alter our main policy prescriptions, in particular that which highlights the concentration of CNR workers in large cities.

To address this question quantitatively, we consider a more flexible parameterization of externalities that allow them to be industry-specific, as explored in [Rossi-Hansberg et al. \(2021\)](#). Specifically, the parameterization allows for the possibility that while CNR workers employed in Professional Services may benefit from a high concentration of CNR workers, the same may be somewhat less true, for instance, in Accommodation and Food Services. Conversely, non-CNR workers in Manufacturing may have more to gain from the proximity of other non-CNR workers

<sup>16</sup>If we consider further increasing the effects of externalities from overall population, where  $\tau^{L,CNR}$  and  $\tau^{L,NCNR}$  are now set to their 90th percentile values, the signs and slopes of the relationships remain but their magnitudes are smaller as expected.

Figure A-11: Robust Externalities



Note: Each observation refers to a CBSA. Marker sizes are proportional to total employment. Blue markers refer to the counterfactual exercise and grey markers to the baseline from the paper. The solid black lines are linear or quadratic fits to the data. The Gini coefficient is constructed using the Lorenz curves depicting within city wage bill and industry rank.

than in Professional Services. In order to implement the optimal policy, one needs to rely on the more general expression for  $\Delta_n^k$  derived in Section 4.1 of this Technical Appendix.

The main challenge in implementing this extension is that single industry data is more noisy than the implicit averaging done by including data from all industries in the regression.<sup>17</sup> To mitigate this problem, we combine the industries into 4 groups so that industries within each group share the same externality parameters. Even then, the IV strategy yields noisy estimates. For example, the standard error for  $\tau^{R,CNR}$  and  $\tau^{R,non-CNR}$  in the group including Professional Services, Communication, Finance and Insurance and Other Services are 1.03 and 2.83, respectively, as compared to our benchmark standard errors of 0.38 and 0.51. Those values are of course exacerbated by the overall weakness of the instruments. An alternative is to use model-implied IVs as described in Section 2.7 since this stronger instrument yields more precise estimates.

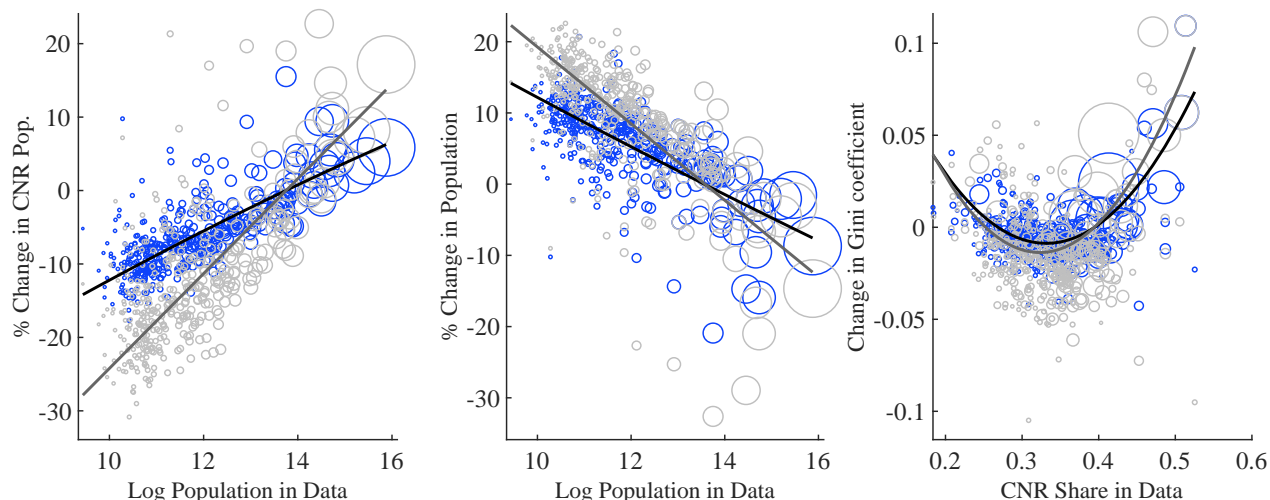
Our approach captures the notion that different industry groups use different worker mixes. In addition, CNR workers confer greater externalities onto CNR intensive sectors and vice-versa. In fact, estimated CNR spillovers are largest among Professional Services and non-CNR spillovers are largest in Manufacturing.<sup>18</sup>

In the end, the results pertaining to the formation of cognitive hubs under this more flexible parameterization of externalities are qualitatively very similar to those obtained in the main text (see Figure A-12 below). However, it is noteworthy that the relationship between overall population and optimal change in CNR employment is somewhat attenuated.

<sup>17</sup>This is true even though we cluster standard errors at the city level, since exogenous productivity variation also has a city/industry component.

<sup>18</sup>Details are given in Rossi-Hansberg et al. (2021)

Figure A-12: Industry-specific externality parameters



Each observation refers to a CBSA. Marker sizes are proportional to total employment. Blue markers refer to the counterfactual exercise and grey markers to the baseline from the paper. The solid black lines are linear or quadratic fits to the data. The Gini coefficient is constructed using the Lorenz curves depicting within city wage bill and industry rank.

Table A-9: Frequency of CNRs and College Graduates

	non-CNR	CNR
Non-Graduate	52.3%	13.5%
College-Graduate	9.5%	24.8%

### 7.3 College Attainment vs. Occupation

We choose to classify workers by occupation rather than academic achievement. The distinction between occupational and academic achievement groups is not merely theoretical. As shown in Table A-9, while it is true that almost 90% of non-CNR workers do not have 4-year college degrees, only about 60% of CNR workers do.<sup>19</sup> The reason that so many CNR workers do not have college degrees is that CNR tasks also include those carried out in occupations such as managers, healthcare practitioners, etc., that do not necessarily require a college degree. In fact, managers alone account for about a third of the CNRs without college. At the same time, non-CNRs with college include a large share of office and administrative assistants and sales-workers. Our occupational classification, therefore, gets at the notion that managers without a college degree more likely generate spillovers to other high earning workers than college educated administrative assistants.

In addition, focusing on occupational rather than educational distinctions also has implications for the spatial distribution of spillovers. Figure A-13 shows that the ratio of CNR workers to college educated workers tends to be larger in larger cities. Therefore, focusing on college educated workers rather than CNRs is likely to overestimate the stock of individuals generating spillovers in large

<sup>19</sup>We use here the same data restrictions as in the paper, focusing on workers who work full-time, for 52 weeks of the year.

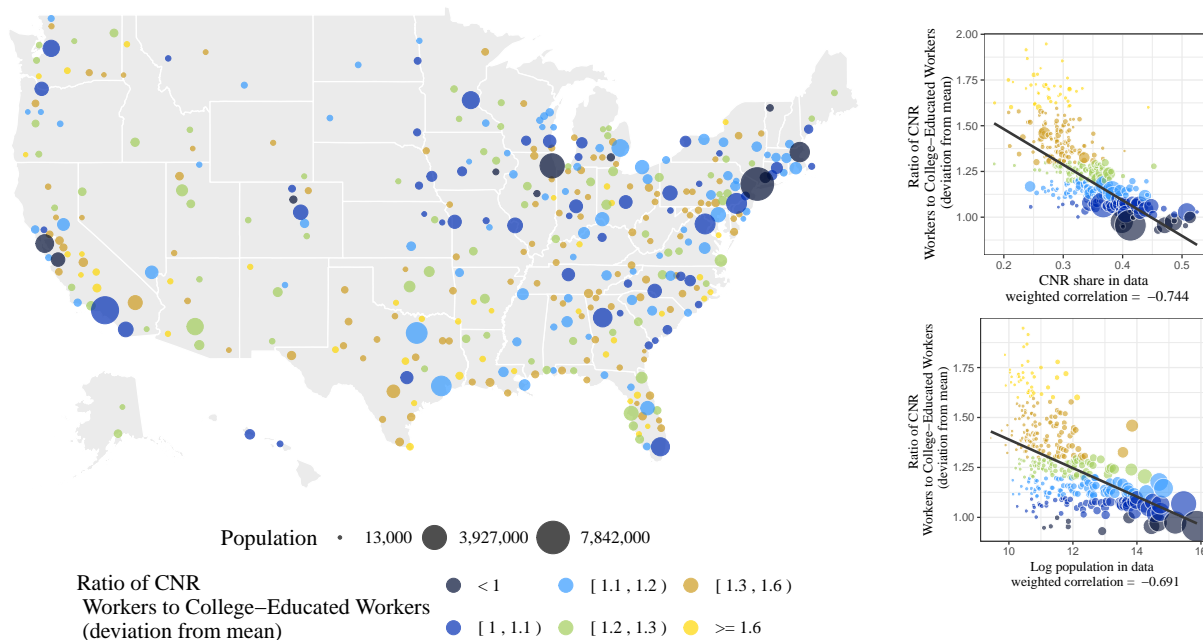


Figure A-13: Ratio of CNR workers to College-Educated Workers

cities.

That said, we nevertheless explore further the implications of classifying workers by occupation rather than educational attainment. Specifically, we carry out an alternative model inversion and re-estimate externalities when workers are classified according to four-year college graduation status. As it turns out, the point estimates for the externalities of high-skilled workers in their own productivity are smaller (see Table A-10). This is consistent with spillovers being more closely tied to occupation rather than college attainment, as one might expect when spillovers operate through learning from others in similar jobs. Accordingly, Figure A-14 shows that the relationships between occupational and industrial makeup across cities, and that between city size and composition, remain though less salient quantitatively.

In conclusion, while CNR workers tend to be college educated, that is not invariably the case. Focusing on college workers would likely exclude from that group many managers and similar workers in smaller towns (where CNR workers are less represented) conferring productive externalities onto others. Furthermore, the choice of worker classification is consequential from a quantitative standpoint even if the main qualitative findings are robust to the distinction between occupational and educational attainment. Bringing these considerations together, we see our classification of workers by occupation rather than college attainment as preferable.

#### 7.4 Restricting CNRs to Graduate Degree Holders

In the paper, we calculate wages for each occupation within each city by first stripping them from the (occupation-specific) effects of years of education using a Mincer regression with separate dummies

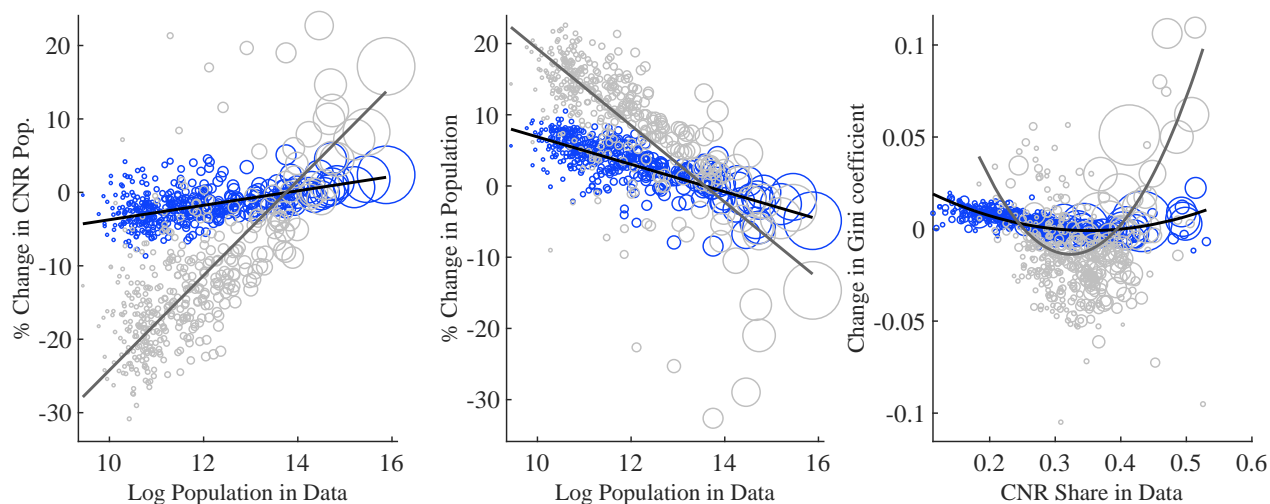
Table A-10: Externality Estimation for Workers Grouped by Education Accomplishment

VARIABLES	(1) <b>OLS</b>		(2) <b>2SLS</b>		(3) <b>CUE</b>	
	CNR	non-CNR	CNR	non-CNR	CNR	non-CNR
$\ln(\frac{L_n^k}{L_n})$	1.08*** (0.11)	1.04*** (0.17)	1.03*** (0.22)	0.71** (0.33)	0.94*** (0.27)	0.75** (0.33)
$\ln(L_n)$	0.36*** (0.04)	0.34*** (0.04)	0.36*** (0.06)	0.29*** (0.04)	0.48*** (0.08)	0.32*** (0.03)
Observations	7,460	7,460	7,460	7,460	7,460	7,460
R-squared	0.47	0.30	0.47	0.30	0.45	0.30
J Test P-Value			0.030	0.38	0.060	0.38
K.P. F			4.55	7.42	4.55	7.42
S.W.F. $L_{nk}$ Share			6.83	11.27	6.83	11.27
S.W.F. $L_n$			7.46	12.9	7.46	12.86

Regressions estimate equation (6) in main text multiplied by  $\gamma_n^j$  to correct for heteroskedasticity in measurement errors (see Footnote 15 in main text). Dependent variable is  $\ln \lambda_n^{kj}$  obtained from model inversion procedure. Individuals are classified by whether they are in CNR occupations and have more years of study than college or not. Standard errors in parentheses, clustered two-ways by city and by industry.

\*\*\*p<0.01, \*\* p<0.05, \* p<0.1

Figure A-14: Workers grouped by educational accomplishment rather than occupation



Each observation refers to a CBSA. Marker sizes are proportional to total employment. Blue markers refer to the counterfactual exercise and grey markers to the baseline from the paper. The solid black lines are linear or quadratic fits to the data. The Gini coefficient is constructed using the Lorenz curves depicting within city wage bill and industry rank.

for each level of education accomplishment in the American Community Survey. This means that our results already allow controls for observable educational outcomes at a fairly fine-grained level.

Concretely, in the data that we use, cities with greater concentration of graduate-degree holders will not feature higher wages simply because those earn more regardless of where they live.

It remains that observed city-specific effects that we attribute to the overall CNR share may instead be driven by the share of post-graduates. If so, one may wonder about the implications for our findings. We assess the role of workers holding post-graduate degrees by redefining our two worker groups so that the upper wage group is now the intersection of CNR workers with those holding post-graduate degrees, and defining the lower wage group as the remainder.

Table A-11 shows the estimated parameters for that case. It is noteworthy that the estimated own-elasticity parameter,  $\tau^R$ , for the upper wage group is now larger. This is expected since this alternative specification downplays the contribution of other CNRs to the productivity of higher income groups while requiring that the smaller newly-defined group account for the same variation in its productivity.

Figure A-15 shows how optimal policy affects real allocations under this parameterization. The creation of Cognitive Hubs becomes if anything more extreme. The reason is intuitive. The new upper wage group accounts for about a third of CNR workers (or about 10% of workers), so that the planner can now create incentives for them to move between cities without generating as much congestion.

Table A-11: Externality Estimation for Graduate CNRs

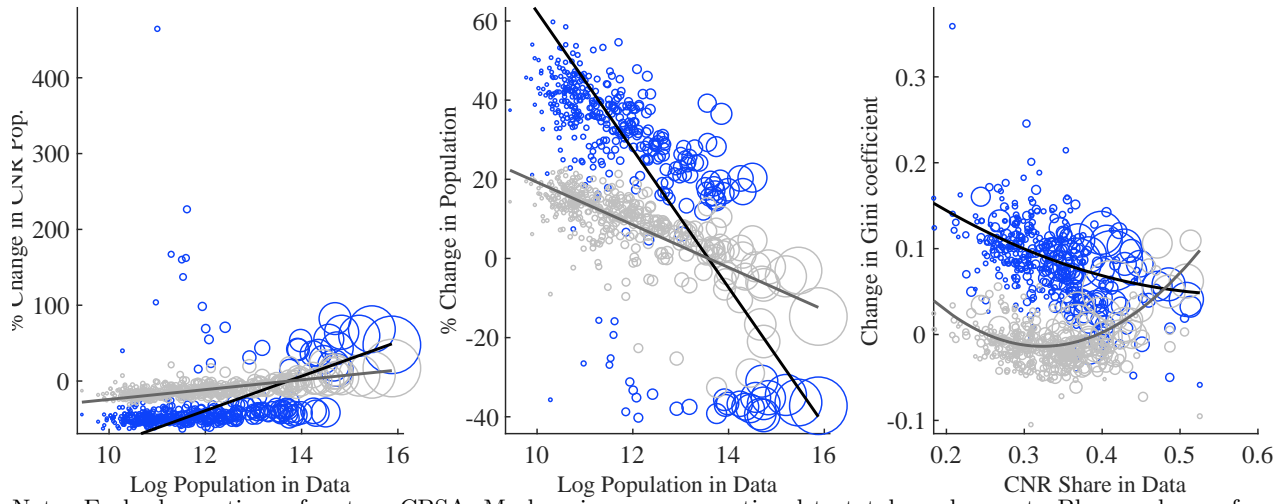
VARIABLES	(1) <b>OLS</b>		(2) <b>2SLS</b>		(3) <b>CUE</b>	
	CNR	non-CNR	CNR	non-CNR	CNR	non-CNR
$\ln(\frac{L_n^k}{L_n})$	0.917*** (0.10)	0.34 (0.41)	1.19*** (0.28)	-0.14 (0.72)	2.08*** (0.49)	-0.06 (0.74)
$\ln(L_n)$	0.42*** (0.06)	0.34*** (0.04)	0.35*** (0.09)	0.31*** (0.04)	0.19* (0.11)	0.34*** (0.03)
Observations	7,460	7,460	7,460	7,460	7,460	7,460
R-squared	0.29	0.41	0.28	0.41	0.21	0.41
J Test P-Value			0.02	0.34	0.05	0.34
K.P. F			5.00	6.65	5.00	6.65
S.W.F. $L_{nk}$ Share			7.52	10.73	7.52	10.73
S.W.F. $L_n$			8.52	13.70	8.52	13.70

Regressions estimate equation (6) in main text multiplied by  $\gamma_n^j$  to correct for heteroskedasticity in measurement errors (see Footnote 15 in main text). Dependent variable is  $\ln \lambda_n^{kj}$  obtained from model inversion procedure. Individuals are classified by whether they are in CNR occupations and have more years of study than college or not. Standard errors in parentheses, clustered two-ways by city and by industry.

\*\*\*p<0.01, \*\* p<0.05, \* p<0.1

In the end, assigning externalities to the full CNR group, as we do in the main text, appears to be conservative for the optimal policy prescription. There are two reasons for this: (i) the full

Figure A-15: Graduate CNRs Externalities



Note: Each observation refers to a CBSA. Marker sizes are proportional to total employment. Blue markers refer to the counterfactual exercise and grey markers to the baseline from the paper. The solid black lines are linear or quadratic fits to the data. The Gini coefficient is constructed using the Lorenz curves depicting within city wage bill and industry rank.

CNR group is larger so that variations in productivity for that group is accounted for by a smaller estimated externality parameter, and (ii) the planner faces more congestion when incentivizing this larger group of workers to move.

## 8 Additional Figures and Tables

Figure A-16 shows the change in optimal industry concentration as a function of city size.

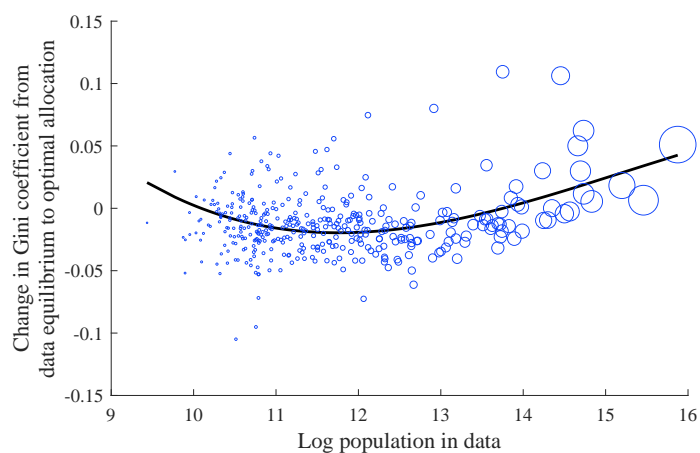


Figure A-16: Changes in the Gini coefficient between the data and optimal allocation.

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

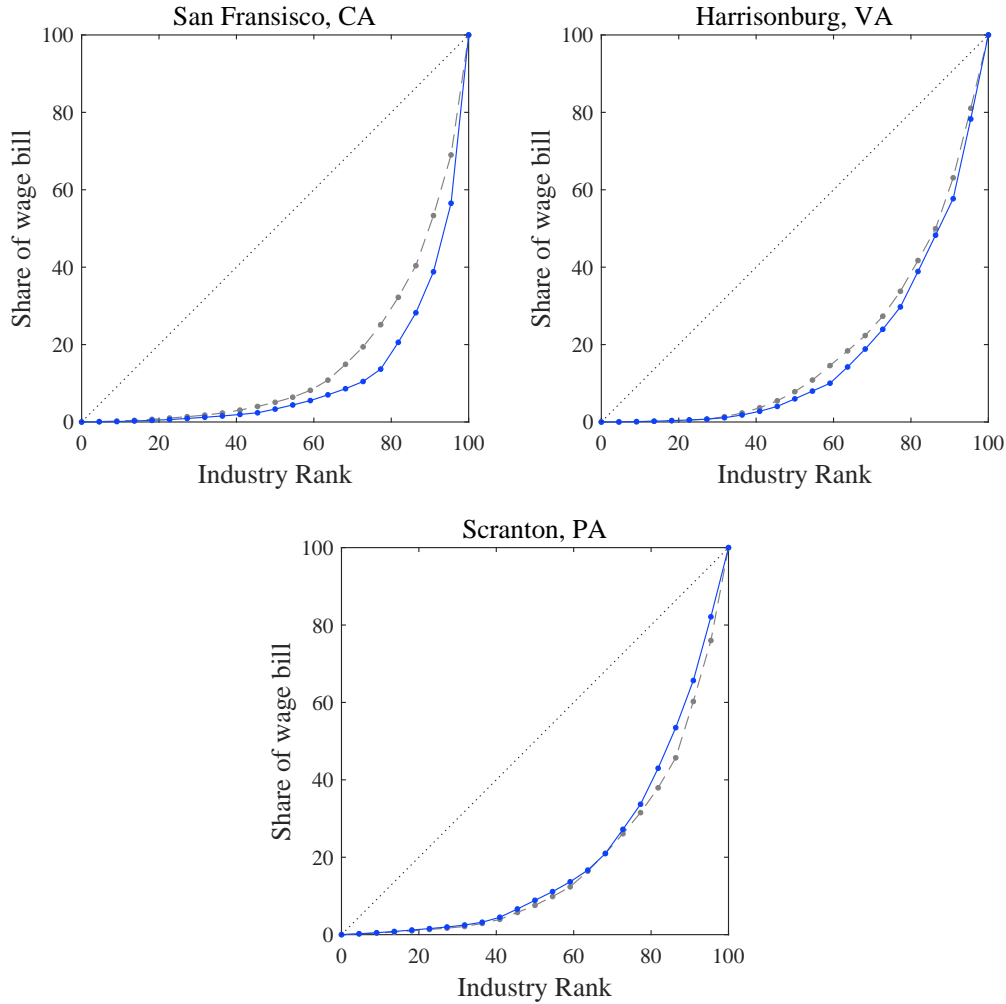


Figure A-17: Shift in the Lorenz curve between the data and optimal allocation.

Note: Each marker refers to an industry, where the dashed gray line shows the Lorenz curve for the data equilibrium and the blue line is for the optimal allocation.

## 9 Data

### 9.1 Definitions and Data Sources

**City:** Our spatial unit, interpreted as cities, is the Metropolitan Statistical Area (MSA) as defined by the Census in 2015. There are 382 such cities in our data.

**Industry:** An industry in this paper is an aggregation of industries as defined in the North American Industry Classification System (NAICS) 1997 vintage. Our 22 such industries are defined according to Table A-12 in this file.

**Occupation:** We define two occupations, Cognitive Non-Routine (CNR) and others (non-CNR). Employed individuals are split into these groups according to their OCCSOC classification,

Table A-12: Industry Definitions

Code	Name	NAICS 1997 Abbreviation
1	Non-Tradables	22 <sub>...</sub> , 23 <sub>...</sub> , 441 <sub>...</sub> , 442 <sub>...</sub> , 443 <sub>...</sub> , 444 <sub>...</sub> , 446 <sub>...</sub> , 447 <sub>...</sub> , 448 <sub>...</sub> , 451 <sub>...</sub> , 453 <sub>...</sub> , 454 <sub>...</sub> , 445 <sub>...</sub> , 452 <sub>...</sub>
2	Food and Beverage	311 <sub>...</sub> , 312 <sub>...</sub>
3	Textiles	313 <sub>...</sub> , 314 <sub>...</sub> , 315 <sub>...</sub> , 316 <sub>...</sub>
4	Wood Product, Paper, and Printing	321 <sub>...</sub> , 322 <sub>...</sub> , 323 <sub>...</sub>
5	Oil, Chemicals, and Nonmetallic Minerals	211 <sub>...</sub> , 2121 <sub>...</sub> , 2122 <sub>...</sub> , 213 <sub>...</sub> , 2123 <sub>...</sub> , 324 <sub>...</sub> , 325 <sub>...</sub> , 326 <sub>...</sub> , 327 <sub>...</sub>
6	Metals	331 <sub>...</sub> , 332 <sub>...</sub>
7	Machinery	333 <sub>...</sub>
8	Computer and Electric	334 <sub>...</sub>
9	Electrical Equipment	335 <sub>...</sub>
10	Motor Vehicles (Air, Cars, and Rail)	3361 <sub>...</sub> , 3362 <sub>...</sub> , 3363 <sub>...</sub> , 3364 <sub>...</sub> , 3365 <sub>...</sub> , 3366 <sub>...</sub> , 3369 <sub>...</sub>
11	Furniture and Fixtures	337 <sub>...</sub>
12	Miscellaneous	3391 <sub>...</sub> , 3399 <sub>...</sub>
13	Wholesale Trade	42 <sub>...</sub>
14	Transportation and Storage	481 <sub>...</sub> , 483 <sub>...</sub> , 484 <sub>...</sub> , 485 <sub>...</sub> , 487 <sub>...</sub> , 488 <sub>...</sub> , 486 <sub>...</sub> , 492 <sub>...</sub> , 493 <sub>...</sub>
15	Professional and Business Services	5111 <sub>...</sub> , 5112 <sub>...</sub> , 5141 <sub>...</sub> , 5142 <sub>...</sub> , 5411 <sub>...</sub> , 5412 <sub>...</sub> , 5413 <sub>...</sub> , 5414 <sub>...</sub> , 5415 <sub>...</sub> , 5416 <sub>...</sub> , 5417 <sub>...</sub> , 5418 <sub>...</sub> , 54190 <sub>...</sub> , 54191 <sub>...</sub> , 54192 <sub>...</sub> , 54193 <sub>...</sub> , 54195 <sub>...</sub> , 54196 <sub>...</sub> , 54197 <sub>...</sub> , 54198 <sub>...</sub> , 54199 <sub>...</sub> , 54194 <sub>...</sub> , 551 <sub>...</sub> , 5611 <sub>...</sub> , 5612 <sub>...</sub> , 5619 <sub>...</sub> , 5613 <sub>...</sub> , 5614 <sub>...</sub> , 5615 <sub>...</sub> , 5616 <sub>...</sub> , 5617 <sub>...</sub> , 562 <sub>...</sub>
16	Other	512 <sub>...</sub> , 5321 <sub>...</sub> , 5322 <sub>...</sub> , 5323 <sub>...</sub> , 5324 <sub>...</sub> , 533 <sub>...</sub> , 711 <sub>...</sub> , 712 <sub>...</sub> , 713 <sub>...</sub> , 8111 <sub>...</sub> , 8112 <sub>...</sub> , 8113 <sub>...</sub> , 8114 <sub>...</sub> , 8121 <sub>...</sub> , 8122 <sub>...</sub> , 8123 <sub>...</sub> , 8129 <sub>...</sub> , 8131 <sub>...</sub>
17	Communication	5131 <sub>...</sub> , 51330 <sub>...</sub> , 51332 <sub>...</sub> , 51333 <sub>...</sub> , 51334 <sub>...</sub> , 51335 <sub>...</sub> , 51336 <sub>...</sub> , 51337 <sub>...</sub> , 51338 <sub>...</sub> , 51339 <sub>...</sub> , 51331 <sub>...</sub> , 513210 <sub>...</sub> , 513220 <sub>...</sub>
18	Finance and Insurance	521 <sub>...</sub> , 5221 <sub>...</sub> , 5222 <sub>...</sub> , 5223 <sub>...</sub> , 523 <sub>...</sub> , 525 <sub>...</sub> , 524 <sub>...</sub>
19	Real Estate	531 <sub>...</sub>
20	Education	6111 <sub>...</sub> , 6112 <sub>...</sub> , 6114 <sub>...</sub> , 6115 <sub>...</sub> , 6116 <sub>...</sub> , 6117 <sub>...</sub> , 611310
21	Health	6211 <sub>...</sub> , 6212 <sub>...</sub> , 6213 <sub>...</sub> , 6214 <sub>...</sub> , 6215 <sub>...</sub> , 6219 <sub>...</sub> , 6216 <sub>...</sub> , 622 <sub>...</sub> , 623 <sub>...</sub> , 6241 <sub>...</sub> , 6242 <sub>...</sub> , 8132 <sub>...</sub> , 8133 <sub>...</sub> , 8134 <sub>...</sub> , 8139 <sub>...</sub> , 6244 <sub>...</sub> , 624310
22	Accommodation	721 <sub>...</sub> , 722 <sub>...</sub>
999	[Omitted]	11 <sub>...</sub> , 482111, 482112, 491110, 814110, 92 <sub>...</sub>

based on the 2010 Standard Occupational Classification system. An individual is considered a CNR worker if the first two digits of the ACS variable "occsoc" is between 11 and 29 inclusively.<sup>20</sup>

**Employment:** Employment counts by industry and city ( $L_n^j$ ) come from 2011 to 2015 tables from the Census Bureau's County Business Pattern (CBP), county level data.<sup>21</sup> It counts "full and part time employees, including salaried officers and executives of corporations, who were on the payroll in the pay period including March 12". For many entries, the CBP data includes ranges rather than values. We use an imputation procedure, described in detail below, that uses available information down to the 6 digit NAICS code, and then aggregated to the industries described in Table A-12. The resulting numbers are split into CNR and non-CNR workers using ratios obtained from the US Census American Community Survey (ACS).<sup>22</sup> See section 9.2 for further details.

**Wages:** Wages in this paper refer to a constructed employee compensation measure using mincerian wage regressions that rely on the ACS variable "incwage" and non-wage compensation imputed using BEA industry-level data.<sup>23</sup> We use the US Census American Community Survey 2011-2015 5-Year Sample, an individual-level 5% weighted sample of the United States. We adjust population weights to match 2013 population and demographics. Consistent with the convention adopted for the 2011-15 ACS data, we keep prices at 2015 dollar. When we invert the model to 1980 data, we use the 1980 5% state sample census.<sup>24</sup> We use of the BEA's NIPA data, (table 6: Income and Employment by Industry), to obtain an estimate of non-wage compensation for workers in different industries. Details on how wages are constructed are provided in Section 9.3.

**Prices** Data on local consumer prices for 2011-15 come from the BEA's 2013 Regional Price Parity (RPP) all items index, which we convert to 2015 values using the Personal Consumption Expenditures (PCE) price index. The same RPP data also includes city-level rents, which we also use for the model inversion. In addition, to invert the model using 1980 we use the PCE price index and changes in industry-level value added prices from the BEA.

**Industry Interactions** Data on industry-to-industry nominal trade flows come from the BEA's "before redefinition, producer value" use table for 2011 through 2015 and the equivalent import tables for the same years. For 1980, since only industry-to-industry nominal trade flows exist, we impute 1980 imports (see section 9.4 for details).

---

<sup>20</sup>IPUMS occsoc description: <https://usa.ipums.org/usa/volii/acsoccsoc.shtml>

<sup>21</sup>CBP Data: <https://www.census.gov/programs-surveys/cbp/data/datasets.html>

<sup>22</sup>For example, if the ACS/Census reports (using weights) that there are 30 CNR and 70 non-CNR workers in Akron, Ohio's Computer and Electric sector, we say that 30% of the employment the CBP reports in this city-industry is CNR workers.

<sup>23</sup>The predicted wages have a 0.88 and 0.83 correlation with the raw wage data for CNR's and non-CNR's respectively.

<sup>24</sup><https://usa.ipums.org/usa/sampdesc.shtml#us1980a>

**Trade** Gravity coefficients for commodity trade are estimated from the US census’s Commodity Flow Survey. Distances between cities are calculated in miles using Gazetteer latitude/longitude coordinates and the Haversine distance formula.<sup>25</sup> Gravity coefficients for services are obtained directly from estimates in Anderson, Milot and Yotov (2014)

## 9.2 Employment

### 9.2.1 MSA-Industry Employment ( $L_n^j$ )

The CBP provides employment by county-industry codes for each year. Different years use different NAICS/SIC vintages for industry classification. When handling the 2011-2015 data we must average over the years. Thus, for 2011-2015 we adjust employment in each year  $t$  by multiplying  $L_n^{j,t}$  by  $\sum_{n,j} L_n^{j,t} / \sum_{n,j} L_n^{j,2013}$  and then average over the 5 years.

For each county/year/industry cell, the data is either reported exactly, or suppressed, with range provided (e.g. 0-19, 20-99, etc.). As one might expect, the data is more often suppressed at higher levels of disaggregation. The 1980 CBP data uses SIC 1972 codes, the 2011 CBP uses NAICS 2007 codes, and the CBP for years 2012-2015 use NAICS 2012 codes, so we need to apply crosswalks (described in section 9.9.2 below) to make them comparable.

We now describe the imputation strategy to obtain employment at the lowest provided industry-county level using language general to both NAICS and SIC codes:

1. For each classification level  $D$ , find the corresponding  $D - 1$  classification level. We use the notation  $j^D \in j^{D-1}$  to denote that  $j^D$  is an industry defined in the  $D$  industry classification and that it belongs to industry  $j^{D-1}$  in the  $D - 1$  classification.

2. Construct for each  $D - 1$  level industries the range of possible values implied by the reported ranges at the  $D$  level, while ignoring the levels actually reported at the  $D - 1$  level. Denote that range by  $[L_n^{j^{D-1},\text{low}}, L_n^{j^{D-1},\text{high}}]$ , where the lower boundary is given by

$$L_n^{j^{D-1},\text{low}} = \sum_{j^D \in j^{D-1}} \min L_n^{j^D}$$

where  $\min L_n^{j^D}$  is the lower end of the range of possible values for  $L_n^{j^D}$ . This lower boundary is equal to the reported value for  $j^D$  when there is no data suppression at the  $D$  level. The upper boundary is defined in analogous manner.

3. For all sectors for which employment is available at the  $D - 1$  level, calculate the ratio:

$$p_n^{j^{D-1}} = \frac{L_n^{j^{D-1}} - \sum_{j^D \in j^{D-1}} L_n^{j^D,\text{known}} - L_n^{j^{D-1},\text{low}}}{L_n^{j^{D-1},\text{high}} - L_n^{j^{D-1},\text{low}}}$$

where  $\sum_{j^D \in j^{D-1}} L_n^{j^D,\text{known}}$  is the total employment in  $j^D$  that is either provided by the CBP (without data suppression) or has already been imputed (this may be true after the initial iteration).

<sup>25</sup>[https://en.wikipedia.org/wiki/Haversine\\_formula](https://en.wikipedia.org/wiki/Haversine_formula)

4. For  $j^D \in j^{D-1}$ , impute

$$L_n^{j^D} = L_n^{j^D, \text{low}} + p_n^{j^{D-1}} (L_n^{j^D, \text{high}} - L_n^{j^D, \text{low}})$$

5. For sectors for which employment is suppressed at the  $D - 1$  level, repeat the procedure at the  $D - 2$  level, and then use imputed  $D - 1$  values to impute  $D$  values.

6. Iterate until all  $D$  digit industries are imputed. By construction, imputed values will aggregate to every reported aggregate, and bounds are respected.<sup>26</sup> When data is suppressed at the county level, we impute employment at the  $j^D$  level by taking the midpoint of the provided range.<sup>27</sup>

### 9.2.2 Occupation Split by MSA-Industry

While we have obtained MSA-industry employment from the CBP ( $L_n^j$ ), we must calculate the occupation split within each MSA-industry ( $L_n^{kj}$ ). In order to do this we use ACS/Census data on occupations to find the share of type  $k$  workers in a given city-industry. We include ACS individuals if (i) the variable “empstat”<sup>28</sup> is “employed”, (ii) they report a positive wage (IPUMS variable “incwage”<sup>29</sup>;0) in the last 12 months, and (iii) report working approximately 50 to 52 weeks of the last year (IPUMS variable “wkswork2” is equal to 6).

**Geography** The ACS provides a geographic identifier for where an individual lives called the Public Use Micro Area, or PUMA, for 2011-2015. PUMAs do not map cleanly into MSA’s, so we map individuals with probability weights to 2010 counties using the PUMA to 2010 county crosswalk described below. We then drop individuals outside of MSA’s. Our procedure results in duplicate observations of a single individual in different locations with an associated probability. These probability weights are multiplied by ACS person weights (perwt<sup>30</sup>) to construct probability-weighted person weights, which are also used to count employment and to run weighted mincerian regressions (discussed further in 9.3).

When handling the 1980 Census, for which PUMA’s are not available, we rely upon state-county identifiers to map individuals into 2015 CBSA’s using the 2010 county to 2015 CBSA crosswalk. Note that when handling the 1980 employment data, we account for minor changes to counties that occurred between 1980 and 2015.<sup>31</sup> Furthermore, since PUMA to county probability weights are not needed for 1980, we simply use the Census provided person weights to count employment.

**Industry** The ACS provides an industry identifier for employed individuals called indnaics, which is based on but not identical to NAICS. We use the system of crosswalks described below to

<sup>26</sup>Using this imputation strategy, bounds are violated only when the reported aggregates and ranges are inconsistent. This happens in about 0.00% and 0.2% of the data coverage for 2011-2015 and 1980 respectively .

<sup>27</sup>This occurs for less than 1% of the sample

<sup>28</sup>IPUMS empstat description: [https://usa.ipums.org/usa-action/variables/EMPSTAT#description\\_section](https://usa.ipums.org/usa-action/variables/EMPSTAT#description_section)

<sup>29</sup>IPUMS incwage description: [https://usa.ipums.org/usa-action/variables/INCWAGE#description\\_section](https://usa.ipums.org/usa-action/variables/INCWAGE#description_section)

<sup>30</sup>IPUMS perwt description: [https://usa.ipums.org/usa-action/variables/PERWT#description\\_section](https://usa.ipums.org/usa-action/variables/PERWT#description_section)

<sup>31</sup>While we found that these changes needed to be implemented in the 1980 CBP data, accounting for county changes was unnecessary in the 1980 Census.

translate those into the industry classification described above. This procedure may result in one individual being in any one of several industries with probability weights. For the 1980 Census, we similarly translate the available ind1990 industry identifier. We then drop individuals who do not work in one of the 22 industries in table A-12. Next, we multiply the probability-weighted person weights (for 2011-2015 ACS) or the person weights (for 1980 Census) by the industry allocation factor from the crosswalk to obtain a final person/industry employment weight. Final employment by industry is then calculated by multiplying the weight by an employment indicator, where the indicator is 1 if a person 1) is currently employed, 2) earns a positive wage, and 3) worked at least 51 weeks in the past year.

**Occupation** In the 2011-2015 ACS, occupations are straightforward to assign using the ACS occsoc variable and the above definition of occupation. Meanwhile, in the 1980 Census we apply a crosswalk as described below. Aggregating the weighted employment counts by MSA, industry, and occupation (i.e. for the ACS we sum over years) thus returns the employment distribution of interest from the ACS/Census.<sup>32</sup> Given this distribution of employment, we calculate the share of CNR workers in each MSA-industry. This occupation split is then multiplied by the MSA-industry employment count obtained from the CBP to achieve the final distribution of employment,  $L_n^{kj}$ , for the 2011-2015 model equilibrium. Since the 1980 Census only has data on a subset of MSA's we must impute the CNR split for part of the 1980 data.

**Imputations for MSA's not observed from 1980 Census data** The 1980 Census only has data on the 213 cities that were classified as an MSA at that time, accounting for approximately 80% of the relevant population.<sup>33</sup> We impute employment for the remaining 169 cities. To do that, we use the fact that the CBP has industry employment for each of the 382 MSA's in 1980, which we can use to our advantage in the imputations. To carry out this imputation, for each industry  $j$  we regress the industrial composition of employment and a 1980 measure of cost of real estate services (see section 9.6) on the logit of the CNR share of employment in industry  $j$ . Specifically, for each industry  $j$  we run

$$\ln \left( \frac{L_n^{CNR,j}}{\sum_k L_n^{k,j}} \right) - \ln \left( 1 - \ln \left( \frac{L_n^{CNR,j}}{\sum_k L_n^{k,j}} \right) \right) = \alpha + \sum_j \beta_j \ln(L_n^j) + \beta_{RE} \ln \left( P_n^{\text{real estate}} \right) + \epsilon_n^j,$$

where  $P_n^{\text{real estate}}$  is the price for real estate services in 1980. After undoing the logistic form on the LHS, we then use the predictions from the regressions as the shares for missing data.

<sup>32</sup>We sum ACSs over the years because we are trying to find occupational shares of industry/city employment. So while ideally we would take weighted averages of ACSs over the years, adding them up amounts to the same.

<sup>33</sup>While we use the 1980 5% state sample that has county identifiers, the 1980 census suppresses the county id in counties with population less than 100,000 people. This suppression eliminates our ability to identify 169 MSA's.

### 9.3 Wages

Wages in the model,  $w_n^k$ , are measured as total adjusted employee compensation after controlling for observable characteristics (see section 9.3.2 for details) and are expressed in 2015 dollars. We primarily use ACS and Population Census data to obtain labor compensation for 2011-15 and 1980, respectively. The sample selection and assignment of individuals to industries, occupations and locations is done as described in Section 9.2.2 above. First, NIPA data is aggregated to our industry classification using the crosswalk, described in section 9.9.2, from BEA 71 industries to our industry classification. Throughout, we express wages in 2015 dollars as implied by the PCE price index.<sup>34</sup>

#### 9.3.1 Accounting for non-wage compensation

To account for nonwage compensation, we construct an employee compensation variable from wages observed in the ACS/Census and non-wage compensation in the NIPA data. To do that, we use the ratio of employee compensation to wages and salaries by industry observed in the BEA's NIPA data to adjust the reported ACS/Census wage for that individual.

For each industry, the BEA provides a measure of total compensation of employees<sup>35</sup> and wages and salaries for each industry.<sup>36</sup> NIPA tables are defined using the BEA's 71 industries, which we map into our  $J$  industries using the two BEA 71 industry crosswalks described in section 9.9.2. We then multiply the compensation to wages ratios by the earnings reported in the ACS to obtain full labor compensation measures for each worker. When doing this, we need to account for the fact that differences in industry classification between the ACS and our reference industry classification may imply that some workers are assigned to different industries with different probabilities (see section 9.2.2).

To aggregate NIPA data between 2011 and 2015, while accounting for differential changes in prices and relative sector size, we adjust industry level values for total compensation and total wage and salaries by the ratio of gross output in 2013 to gross output in each of the years.<sup>37</sup> For years 2011-2015, compensation and wages and salaries are then averaged over the 5 years, after which we can obtain a ratio of compensation to wages and salaries. This ratio is multiplied by the measure of wages provided by the ACS/Census to account for non-wage compensation. Letting the earnings in the ACS be  $e_n^{j,k}$ , the BEA's compensation of employees be  $COE^{t,j}$ , wages and salaries be  $WS^{t,j}$ , and gross output be  $GO^t$  for year  $t$  and industry  $j$ , then our new measure of wages for 2011 to 2015 is

$$\hat{w}_n^{k,j} = e_n^{k,j} \left[ \frac{\sum_t \left( \frac{COE^{t,j}}{GO^t} \right) / 5}{\sum_t \left( \frac{WS^{t,j}}{GO^t} \right) / 5} \right],$$

<sup>34</sup>See BEA Table 2.3.4. Price Indexes for Personal Consumption Expenditures by Major Type of Product.

<sup>35</sup>Found in Table 6.2B (for 1980) and Table 6.2D (for 2011-2015). under the NIPA data

<sup>36</sup>Found in Table 6.3B (for 1980) and Table 6.3D (for 2011-2015) under the NIPA data

<sup>37</sup>Gross output is measured using the BEA's Use tables (after subtracting net imports) as described in section 9.4

where in 1980 it is

$$\hat{w}_n^{k,j} = e_n^{k,j} \times \frac{COE^{1980,j}}{WS^{1980,j}}.$$

### 9.3.2 Mincerian Regression

Take  $\hat{w}_n^k(i)$  to be the employee compensation variable constructed as above for some individual  $i$ . We run the following mincerian regression for CNR and non-CNR workers separately, weighted as described in section 9.2.2:

$$\ln \hat{w}_n^k(i) = c^k + X(i)\beta^k + d_n^k + u_n^k(i)$$

where  $c^k$  is a constant,  $X(i)$  is a set of controls with occupation-specific coefficients  $\beta^k$ ,  $d_n^k(i)$  is an occupation-city dummy, and  $u_n^k(i)$  is an error term. Our controls  $X(i)$  are variables for educational attainment, english ability, marital status, veteran status, race, sex, and whether they've had a child in the last year,<sup>38</sup> and continuous variables for potential experience<sup>39</sup> and number of years in the United States.

### 9.3.3 Adjusted Wages

Given the mincerian regressions above, adjusted wages are defined to be

$$w_n^k = \exp \left( c^k + \left( \frac{1}{L^k} \sum_{i \in k} X(i) \right) \beta^k + d_n^k \right)$$

representing the "average" employee compensation of a worker in occupation  $k$  working in city  $n$  while assuming that workers have otherwise identical characteristics between cities. After calculating adjusted wages, we use the PCE price index to bring 1980 and 2015 dollars to 2013 dollars.

**Imputations for 1980** As explained above we must impute wages for the remaining 169 cities. To carry out this imputation, we follow the same procedure described in section 9.2.2, but with log adjusted wages in the left-hand-side of the imputation regression.

## 9.4 Industry Interactions

To choose production flow parameters of the model, we make use of the BEA Use tables (Before Redefinition) at the 71 industry summary level and also the corresponding import tables, from 2011 to 2015. The use table reports the amount purchased from one US industry by another, as well as value added broken down into total employee compensation and gross operating surplus. The import table reports the amount imported to the US by each US industry from corresponding industries in other countries.

<sup>38</sup>For 1980 we infer this variable by using information on the age of their youngest child.

<sup>39</sup>(potential experience) = (age) - (years in education) - 6, where (years in education) is inferred from reported educational attainment.

We first aggregate values in both the use and import tables to our industry classification using a crosswalk constructed as described in Section 9.9, below.<sup>40</sup> We then subtract the imports from each cell of the use table for each year. We define total output by industry to be the column sum (materials sold plus employee compensation plus operating surplus), and total output in the economy to be the sum of output by industry over the included industries in table A-12. Using the ratio of total output in the economy for each year to 2013, we adjust the values in 2011, 2012, 2014, and 2015 to 2013 terms, then take the average of all 5 years.

In summary, let  $U^{t,jj'}$  be the use materials purchased and  $I^{t,jj'}$  be the materials imported by industry  $j$  from industry  $j'$  in year  $t$ . Furthermore, let  $EC^{t,j}$  be employee compensation and  $OS^{t,j}$  be operating surplus, such that  $EC^{t,j} + OS^{t,j}$  is the total value added by industry  $j$  in year  $t$ . Then we define the industry flows net of imports as  $M^{t,jj'} = U^{t,jj'} - I^{t,jj'}$ , and the 2011-2015 average as

$$M^{jj'} = \left( \sum_t M^{t,jj'} \times \frac{\sum_{jj'} M^{2013,jj'} + EC^{2013,j} + OS^{2013,j}}{\sum_{jj'} M^{t,jj'} + EC^{t,j} + OS^{t,j}} \right) / 5.$$

While the Use table is available for 1980, the import table is not. To obtain trade flows net of imports we assume that the share of imports in gross output is the same in 1980 as in the 2011-2015 averaged period at the industry level, but allow for the national share of imports in gross output to reflect 1980 data. To carry this out, we calculate the share of imports in gross output for the averaged 2011-2015. We then account for national changes to the import share from 1980 to 2011-2015 by multiplying the share of imports in gross output in the averaged 2011-15 period by the ratio of the national import share in 1980 to the national import share in a 2011-2015 average. The 1980 industry flows net of imports, where  $I^{\bar{13},jj'}$  refers to imports in the averaged 2011-2015 period, is thus

$$M^{jj'} = U^{1980,jj'} \left( 1 - \frac{I^{\bar{13},jj'}}{M^{\bar{13},jj'}} \right) \times \frac{I^{1980}/M^{1980}}{I^{\bar{13}}/M^{\bar{13}}},$$

where  $I^{1980}$ ,  $M^{1980}$ ,  $I^{\bar{13}}$ , and  $M^{\bar{13}}$  are national aggregates.<sup>41</sup>

We interpret gross operating surplus reported in this table as compensation for capital investment in equipment and real estate structures. In particular, we attribute equipment income to materials (in that equipment fully depreciates in a static setting where the period lasts indefinitely). In the Greenwood, Hercowitz, and Krussel (1997) measure the share of equipment in value added as 17%. For some industries, 17% of value added is greater than the reported gross industry surplus. So we take capital investment to be the minimum of gross operating surplus or 17% of value added and subtract this from gross operating surplus; hence, capital investment,  $CI^j$ , is defined as

<sup>40</sup>we omit scrap, used goods, noncomprable imports, and government in addition to the omitted industries in Table A-12.

<sup>41</sup>The national ratio can be calculated using import data from BEA International Transactions (Annual) Table 1.1, line 10 and GDP data from BEA Gross Domestic Product (Annual) Table 1.1.5, line 1.

$$CI^j = \min \left( 0.17 \times (EC^j + OS^j), \quad OS^j \right).$$

We then increase purchases of materials from all industries by the amount subtracted (holding the ratios between rows constant) to keep industry output constant. Thus, the input-output matrix becomes

$$M^{jj'} = M^{jj'} \times \frac{\sum_{j'} M^{jj'} + CI^j}{\sum_{j'} M^{jj'}}.$$

The remaining operating surplus is interpreted as compensation for owning structures, and added to purchases from the real estate sector and to real estate's gross operating surplus, such that purchases from real estate becomes

$$M^{j,RE} = M^{j,RE} + (OS^j - CI^j), \quad \forall j \neq \text{real estate}.$$

This results in our input-output table, which has no operating surplus (except in real estate, where it is interpreted as rental income). Operating surplus is then

$$OS^j = \begin{cases} \sum_j (OS^j - CI^j), & j = \text{real estate} \\ 0, & j \neq \text{real estate} \end{cases}.$$

## 9.5 Trade Costs

Trade costs,  $\kappa_{n'n}^j$ , are a function of the distances between MSA's and gravity coefficients. As described above, distances between MSA's are calculated using haversine distances with U.S. Census Gazetteer coordinates. In order to estimate the gravity coefficients, we use the Public Use Microdata (PUM) File for the 2012 Commodity Flow Service. The PUM File includes, among other data, shipment level observations. For each shipment it has information on the total value and on the Great Circle distance covered. It also includes information on CFS area of origin, the CFS area of destination and a 2012 NAICS classification and a weighing variable that we use when consolidating across the individual observations. After using the 2012 NAICS to our  $J$  industries crosswalk, we consolidate the microdata into a data-set where each observation is characterized by an industry, an origin and a destination. This consolidated data-set includes, for each observation, the average Great Circle distance between origin and destination, and the total value shipped. At this point we are interested in estimating the gravity coefficients.

From the model, we have that

$$\pi_{nn'}^j X_n^j = \frac{\left(\kappa_{nn'}^j x_{n'}^j\right)^{-\theta^j}}{\sum_{n'} \left(\kappa_{nn'}^j x_{n'}^j\right)^{-\theta^j}} X_n^j$$

Taking logs and applying  $\kappa_{nn'}^j = \left(d_{nn'}^j\right)^{t_n^j}$ ,

$$\ln\left(\pi_{nn'}^j X_n^j\right) = \ln\left(X_n^j\right) - \theta^j \ln x_{n'}^j - \ln \sum_{n'} \left(\kappa_{nn'}^j x_{n'}^j\right)^{-\theta^j} - \theta^j t_n^j \ln d_{nn'}^j$$

We can therefore estimate  $g^j \equiv \theta^j t_n^j$  from the gravity regressions:

$$\ln\left(\pi_{nn'}^j X_n^j\right) = \mu_n^j + \nu_{n'}^j - g^j \ln d_{nn'}^j$$

where  $\mu_n^j$  is an destination dummy,  $\nu_{n'}^j$  is an origin dummy,  $\ln d_{nn'}^j$  is log distance between origin  $n'$  and destination  $n$  and  $\ln\left(\pi_{nn'}^j X_n^j\right)$  is log bilateral trade-flow from origin  $n'$  to destination  $n$ .<sup>42</sup> We estimate the regression equation by OLS.

We run this regression for each of the  $J$  industries separately, but replace the estimates for the non-tradable industries (i.e. real estate and retail, construction, and utilities) with  $g^j = 9999$ . For service industries we use coefficients estimated by Anderson, Milot and Yotov (2014), Table 1.<sup>43</sup>

## 9.6 Prices

Prices in our model vary across industries and cities. Since data on industry prices alone,  $P^j$ , is only a measure of price level within industry over time, we choose  $P^j$  to be a normalization on prices, which is described further in the appendix in the paper. Below we will discuss how we measure prices across both cities and industries and how we measure price changes within industry over the span of 1980 to 2011 to 2015.

## 9.7 Joint distribution of prices, $P_n^j$

To obtain prices for the non-tradable industries (i.e. real estate and retail, construction, and utilities) we require a spatial distribution of prices on rents, goods, and services. For these prices we use the BEA's Regional Price Parities (RPP) for 2013. In the RPP the BEA explicitly provides measures of goods and services, which they decompose into "rents" and "other", the latter of which we take to be a measure of service prices in our model. Of the industries in the paper, we take the following industries to have prices represented by the RPP's goods measure: Food and beverage, Textiles, Wood, paper, and printing, Oil, chemicals, and nonmetallic minerals, Metals, Machinery, Computer

<sup>42</sup>Note that  $\ln d_{nn'}^j$  here is the distance between the shipment origin and destination provided by the CFS. Only when calculating  $\kappa_{nn'}^j$  do we use distances between MSA's calculated using haversine distances with U.S. Census Gazetteer coordinates.

<sup>43</sup>These industries are wholesale trade, transportation and storage, professional and business services, other, communication, finance an insurance, education, health, and accommodation

and electronics, Electrical equipment, Motor vehicles (air, cars, and rail), Furniture and fixtures, and Miscellaneous manufacturing. We then leave the following service industries to be represented by "other": Non-tradables, Wholesale trade, Transportation and storage, Professional and business services, Other, Communication, Finance and insurance, Education, Health, and Accomodation. Details on how prices in the model are inverted can be found in the appendix in the paper.

**Regional Price Parities for 1980** Since the BEA has only published the RPP as far back as 2008, we assume that the spatial distribution of prices for goods and services has remained constant over time (i.e. we use the same distribution as in 2013). For the distribution of rents in 1980 we use CoreLogic HPI, which tracks changes in the price of rents in each MSA over time. We can then divide the RPP for rents in 2013 by the HPI to obtain the distribution of rents in 1980.<sup>44</sup>

Since HPI data is proprietary, as an alternative to using the HPI data as the default method to calculate rents for the 1980 equilibrium, we also measure rents with the ACS/Census samples by following the rental imputation procedure in Diamond (2016). Following her strategy, imputed rents are the value of (monthly) rent for renters and the home value times 0.0785/12 for home owners.<sup>45</sup> This strategy provides the percent change in rental prices across MSA's for only the 213 MSA's available in the 1980 data. To impute the remaining 169 rental price changes, we run a regression of the industrial composition of employment in city  $n$  on the log of the rental price percent change. Provided this measure of rental changes, we can apply the same procedure as under the HPI data to obtain the distribution of rents in 1980 expressed in 2015 dollars. We find that the correlation is 0.44

To obtain the distribution of prices across cities,  $P_n$ , we have from the model that

$$P_n = \prod_j (P_n^j)^{\alpha_j}$$

As mentioned earlier industry prices are chose to be a normalization for 2011 to 2015. We will now discuss how we make within industry price adjustments for  $P^j$  in 1980.

## 9.8 Industry Level Prices

We adjust  $P^j$  to account for inflation over the span of years from 1980 to 2011-2015 by using the BEA's Chain Type Price Indexes for Value Added by Industry. Provided at the BEA's 71 industry summary level defined with NAICS 2007, we map the provided industries into our own industry definitions using the appropriate crosswalk defined in section 9.9.2. We translate prices from the BEA's classification to ours by taking a value-added weighted geometric means of price indices. Lastly, we multiply the price indexes for 1980 by the ratio of the 2013 PCE to the 1980 PCE to adjust for inflation. When we invert prices from the model for 1980 we can then apply this

<sup>44</sup>As with wages, we use the PCE price index to adjust the 1980 distribution of rents to 2015 dollars.

<sup>45</sup>The specific IPUMS variables used are 'RENT' for monthly rent, 'VALUEH' for home value, and 'OWNERSHP' for the home ownership indicator. As a baseline, we only use the imputed rent for homeowners as our measure.

adjustment to industry level prices from the 2011-2015 inversion in order to obtain a measure of prices resulting from productivity growth over time.

## 9.9 Crosswalks

Our data come from many different sources, many of which use different geographic and industry classifications that need to be made consistent.

### 9.9.1 Geographic Crosswalks

The geographic unit of analysis for the paper is 2015 MSA's; however, the spatial data we use is made available at either the PUMA (ACS/Census) or county (CBP) level. Since MSA's are composed of sets of counties (i.e. no MSA contains a portion of a county), we can cleanly map counties to 2015 MSA's. The mapping from PUMA's to MSA's requires further assumptions. This and all other geographic crosswalks are taken from the Missouri Census Data Center.<sup>46</sup> Below are the geographic crosswalks used throughout the paper.

- **1980/2010 County to 2015 CBSA:** Given that the CBP for 2011 to 2015 is defined with 2010 counties, we download the crosswalk from 2010 counties to 2015 MSA's from the MCDC. When handling the 1980 CBP data we account for any changes to county definitions between 1980 and 2010 by referring to the Census's account of county changes.<sup>47</sup> Otherwise, the crosswalk from 2010 counties to 2015 MSA's can be applied to the 1980 CBP data in order to calculate the 1980 employment level in 2015 defined MSA's.
- **2000/2010 PUMA to 2015 CBSA:** Since the ACS provides 2000 PUMA's for 2011 and 2010 PUMA's for years 2012 through 2015, we download from the MCDC crosswalks from 2000 and 2010 PUMA's to 2010 counties, with an allocation factor constructed from employment proportions. We then merge on the county to MSA crosswalk to obtain the final PUMA to MSA crosswalk. Note that the 1980 Census does not provide PUMA's (only counties), and so the same 2010 county to 2015 MSA crosswalk can be used as described above.

### 9.9.2 Industry Crosswalks

As explained above, the 22 industries in the model ("ModelInds") are based off NAICS 1997 codes. Different industry classifications crosswalks with allocation factors are often not available, but concordances from a given classification to NAICS 1997 usually are or can easily be constructed. We construct crosswalks using these concordances using the following procedure: for each duplicated industry identifier in the given classification, suppose that there is an equal probability the observed value is in each NAICS 1997 code. We then sum these probabilities within our industries, which

---

<sup>46</sup>Missouri Census Data Center: <http://mcdc.missouri.edu/websas/geocorr14.html>

<sup>47</sup><https://www.census.gov/geo/reference/county-changes.html>

produces a crosswalk from the data's original industry classification to our industries.<sup>48</sup> Given the raw data we use throughout the paper, we utilize the following crosswalks.

- **2012, 2007 NAICS to ModelInds** Concordances exist from NAICS 2012 to NAICS 2007 codes, NAICS 2007 to NAICS 2002 codes, NAICS 2002 to NAICS 1997 codes, and the industries defined in the paper (ModelInds) are defined with NAICS 1997 codes. Given this, we merge each concordance and assume that an industry in one classification year has uniform probability of being in any of the matching industry classifications in NAICS 1997. To create the final allocation factor, we sum over the uniform probability for each NAICS year and ModelInd; hence, the result is an allocation factor to bring industries in the NAICS year to ModelInds, where the allocation factor sums to one within the NAICS year industries. See footnote 48 for an example of how this works in practice.
- **NIPA industries to ModelInds** While the BEA's NIPA tables have a separate industry classification a crosswalk exists between them and the 71 BEA industry codes. The mapping to ModelInds follows the process described above.
- **BEA 71 industries to ModelInds** Since we use data from the BEA Use and Import tables at the their 71 industry classification level, we make use of their concordance from their 71 industries to NAICS 2007 codes. The process to obtain a crosswalk from the BEA industries to ModelInds then follows the process described above. Note that the 1980 Use table has already been defined by the BEA with the contemporary 71 industries that map to NAICS 2007.
- **1980 BEA 71 industries to ModelInds** The 1980 NIPA table has industries classified into the 71 BEA industries, but with the industries being defined with 1972 SIC codes. Thus, a concordance is established between the 71 BEA industries in 1980 and 1972 SIC codes. This concordance is then merged with the 1972 SIC codes to ModelInds crosswalk to obtain a final crosswalk from 1980 BEA 71 industries to ModelInds.
- **1972 SIC to ModelInds** Since the 1980 CBP data uses 1972 SIC industry definitions, we use crosswalks going from 1972 SIC to 1977 SIC codes to 1987 SIC codes. These crosswalks come from Fort and Klimek (2016).<sup>49</sup> We then use a crosswalk from 1987 SIC to 1997 NAICS codes provided by the Census.
- **indnaics to ModelInds** The 2011-2015 5-Year ACS provides their own industry classification titled 'INDNAICS', for which they provide a crosswalk to NAICS 2002 codes.<sup>50</sup> Thus, we can construct a crosswalk to ModelInds as described above.

---

<sup>48</sup>For example, say in a concordance from indnaics to naics1997 that the ACS indnaics code 23 is seen matched with 33 distinct naics1997 codes. Each possibility is given probability 1/33. Suppose further that 32 of these naics1997 codes are in our industry "Non-Tradables" and 1 is in our industry "Professional and Business Services". Then indnaics code 23 is given a 32/33 probability of being in the "Non-Tradables" industry and a 1/33 probability of being in the "Professional and Business Services" industry.

<sup>49</sup><http://faculty.tuck.dartmouth.edu/teresa-fort/data>

<sup>50</sup><https://usa.ipums.org/usa/volii/indcross03.shtml>

- **ind1990 to ModelInds** The 1980 Census sample provides Ind1990 industry classifications, for which a Census provided concordance exists to NAICS 1997.<sup>51</sup> <sup>52</sup>

### 9.9.3 Occupation Crosswalks

- **occ2010 to occsoc** Since the 1980 Census sample does not include occsoc, but does include occ2010 (for which IPUMS provides a crosswalk to occsoc), we utilize the provided crosswalk.<sup>53</sup>

## Data References

- [Data1] CoreLogic. “Home Price Index, MSAs”. accessed via the Federal Reserve Bank System. accessed December 2023.
- [Data2] National Institute of Food and Agriculture. “College Partners Directory”. US Department of Agriculture, <https://www.nifa.usda.gov/land-grant-colleges-and-universities-partner-website-directory>. accessed December 2023.
- [Data3] U.S. Bureau of Economic Analysis. “Capital Flows, 1997”. US Department of Commerce. <https://www.bea.gov/industry/capital-flow-data>. accessed December 2023.
- [Data4] U.S. Bureau of Economic Analysis. “Chain-Type Price Indexes for Value Added by Industry”. US Department of Commerce. <https://apps.bea.gov/iTable/?isuri=1&reqid=151&step=1>. accessed December 2023.
- [Data5] U.S. Bureau of Economic Analysis. “GDP by Industry, SIC Classification”. US Department of Commerce. <https://www.bea.gov/itable/gdp-by-industry>. accessed December 2023.
- [Data6] U.S. Bureau of Economic Analysis. “Imports by Industry”. US Department of Commerce <https://www.bea.gov/data/industries/input-output-accounts-data>, note = accessed December 2023.
- [Data7] U.S. Bureau of Economic Analysis. “Input-Output Accounts: Use Tables”. US Department of Commerce. <https://www.bea.gov/data/industries/input-output-accounts-data>. accessed December 2023.
- [Data8] U.S. Bureau of Economic Analysis. “National Income and Product Accounts”. US Department of Commerce, <https://www.bea.gov/itable/national-gdp-and-personal-income>. accessed December 2023.

---

<sup>51</sup><https://www2.census.gov/programs-surveys/demo/guidance/eeo/indcswk2k.pdf>

<sup>52</sup>8 individuals in the 1980 Census have ind1990 codes that are not listed in the provided concordance to NAICS 1997. We manually add these missing codes to the crosswalk by identifying the  $i-1$  industry level the code corresponds to and assuming the code follows the same concordance of the other  $i$  digit level industries within the identified  $i-1$  industry.

<sup>53</sup>[https://usa.ipums.org/usa/volii/acs\\_occtooccsoc.shtml](https://usa.ipums.org/usa/volii/acs_occtooccsoc.shtml)

- [Data9] U.S. Bureau of Economic Analysis. “NIPA Flat Files: Table Register”. US Department of Commerce <https://apps.bea.gov/iTable/?ReqID=19&step=4&isuri=1&1921=flatfiles>. accessed December 2023.
- [Data10] U.S. Bureau of Economic Analysis. “Regional Price Parities by MSA”. US Department of Commerce <https://www.bea.gov/data/income-saving/personal-income-county-metro-and-other-areas>. accessed December 2023.
- [Data11] U.S. Bureau of Economic Analysis. “Value Added by Industry”. US Department of Commerce. <https://apps.bea.gov/iTable/?isuri=1&reqid=151&step=1>. accessed December 2023.
- [Data12] U.S. Census Bureau. “1980 US census”. Retrieved from IPUMS <https://usa.ipums.org/usa/>. accessed December 2023.
- [Data13] U.S. Census Bureau. “2012 Commodity Flow Survey”. US Department of Commerce, <https://www.census.gov/data/datasets/2012/econ/cfs/historical-datasets.html>. accessed December 2023.
- [Data14] U.S. Census Bureau. “American Community Survey, 5 year sample 2011-2015”. Retrieved from IPUMS <https://usa.ipums.org/usa/>. accessed December 2023.
- [Data15] U.S. Census Bureau. “CBSA region and division reference files”. US Department of Commerce, <https://www.census.gov/geographies/reference-files.html>. accessed December 2023.
- [Data16] U.S. Census Bureau. “Census Gazetteer Reference Files”. US Department of Commerce <https://www.census.gov/geographies/reference-files/time-series/geo/gazetteer-files.html>. accessed December 2023.
- [Data17] U.S. Census Bureau. “County Business Patterns”. US Department of Commerce, <https://www.census.gov/programs-surveys/cbp.html>. accessed December 2023.
- [Data18] U.S. Census Bureau. “Geography Crosswalks”. Retrieved from the Missouri Census Data Center <https://mcdc.missouri.edu/applications/geocorr.html>. accessed December 2023.
- [Data19] U.S. Census Bureau. “INDNAICS to NAICS Crosswalk”. US Department of Commerce via IPUMS [https://usa.ipums.org/usa-action/variables/indnaics#description\\_section](https://usa.ipums.org/usa-action/variables/indnaics#description_section). accessed December 2023.
- [Data20] U.S. Census Bureau. “North American Industry Classification System Concordances”. US Department of Commerce, <https://www.census.gov/naics/>. accessed December 2023.
- [Data21] U.S. Economic Research Service. “Geographic Amenities”. US Department of Agriculture, <https://www.ers.usda.gov/data-products/natural-amenities-scale/>. accessed December 2023.

## References

- Altonji, J. G., T. E. Elder, and C. R. Taber (2005). Selection on observed and unobserved variables: Assessing the effectiveness of catholic schools. *Journal of political economy* 113(1), 151–184.
- 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.
- Baum-Snow, N., M. Freedman, and R. Pavan (2018). Why has urban inequality increased? *American Economic Journal: Applied Economics* 10(4), 1–42.
- Baum-Snow, N. and R. Pavan (2013). Inequality and city size. *Review of Economics and Statistics* 95(5), 1535–1548.
- 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.
- Card, D. (2001). Immigrant inflows, native outflows, and the local labor market impacts of higher immigration. *Journal of Labor Economics* 19(1), 22–64.
- Ciccone, A. and R. E. Hall (1996). Productivity and the density of economic activity. *The American Economic Review* 86(1), 54.
- de la Roca, J. and D. Puga (2017). Learning by working in big cities. *The Review of Economic Studies* 84(1), 106–142.
- 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.
- Eaton, J. and S. Kortum (2002). Technology, geography, and trade. *Econometrica* 70(5), 1741–1779.
- Fajgelbaum, P. and C. Gaubert (2020). Optimal spatial policies, geography, and sorting. *The Quarterly Journal of Economics* 135(2), 959–1036.
- 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.
- 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.
- Moretti, E. (2004b). Workers’ education, spillovers, and productivity: evidence from plant-level production functions. *American Economic Review* 94(3), 656–690.
- Neal, D. (1997). The effects of catholic secondary schooling on educational achievement. *Journal of Labor Economics* 15(1, Part 1), 98–123.

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.