Data Appendix

1 Overview

This file describes in detail how we cleaned and transformed data from original sources to obtain the values used in the model inversion. Throughout, we use 2015 US dollar as our monetary unit. We use the US Census's TIGER/Line Shapefiles for mapping data and results.

2 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 1 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, based on the 2010 Standard Occupational Classfication system. An individual is considered a CNR worker if the first two digits of the ACS variable "occsoc" is between 11 and 29 inclusively.[1]

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.[2] 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 1. The resulting numbers are split into CNR and non-CNR workers using ratios obtained from the US Census American Community Survey (ACS).[3] See section 3 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.[4] 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.[5] 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 4.

[1] IPUMS occsoc description: `https://usa.ipums.org/usa/volii/acsoccsoc.shtml`

[2] CBP Data: `https://www.census.gov/programs-surveys/cbp/data/datasets.html`

[3] 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.

[4] The predicted wages have a 0.88 and 0.83 correlation with the raw wage data for CNR's and non-CNR's respectively.

[5] `https://usa.ipums.org/usa/sampdesc.shtml#us1980a`

Table 1: Industry Definitions

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

Prices Data on local consumer prices for 2011-15 come from the BEA's 2013 Regional Price Parity all items index, which we convert to 2015 values using the Personal Consumption Expenditures (PCE) price index. 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 5 for details).

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.[6] Gravity coefficients for services are obtained directly from estimates in Anderson, Milot and Yotov (2014)

3 Employment

3.1 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 8.2 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 $\left[L^{j^{D-1},\text{low}}, L^{j^{D-1},\text{high}}\right]$, 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:

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

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).

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

[6] https://en.wikipedia.org/wiki/Haversine_formula

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

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.[7] When data is suppressed at the county level, we impute employment at the j^D level by taking the midpoint of the provided range.[8]

3.2 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"[9] is "employed", (ii) they report a positive wage (IPUMS variable “incwage"[10]>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).

3.2.1 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[11]) to construct probability-weighted person weights, which are also used to count employment and to run weighted mincerian regressions (discussed further in 4).

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.[12] Furthermore, since PUMA to county probability weights are not needed for 1980, we simply use the Census provided person weights to count employment.

3.2.2 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 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 1. 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.

[7] Using this imputation strategy, bounds are violated only when the reported aggregates and ranges are inconsistent. This happens in about 0.002% and 0.2% of the data coverage for 2011-2015 and 1980 respectively .

[8] This occurs for less than 1% of the sample

[9] IPUMS empstat description: `https://usa.ipums.org/usa-action/variables/EMPSTAT#description_section`

[10] IPUMS incwage description: `https://usa.ipums.org/usa-action/variables/INCWAGE#description_section`

[11] IPUMS perwt description: `https://usa.ipums.org/usa-action/variables/PERWT#description_section`

[12] 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.

3.2.3 Occupation

In the 2011-2015 ACS, occupations are straightforward to assign using the ACS occsoc variable and the above defintion 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.[13] 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.

3.2.4 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.[14] 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 7) on the logit of the CNR share of employment in industry j. Specifically, for each industry j we run

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

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.

4 Wages

Wages in the model, w_n^k, are measured as total adjusted employee compensation after controlling for observable characteristics (see section 4.2 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 3.2 above. First, NIPA data is aggregated to our industry classification using the crosswalk, described in section 8.2, from BEA 71 industries to our industry classification. Throughout, we express wages in 2015 dollars as implied by the PCE price index.[15]

4.1 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[16] and wages and salaries for each industry.[17] NIPA tables are defined using the BEA's 71 industries, which we map into our J industries using the two BEA 71 industry crosswalks descibed in section 8.2. We then multiply

[13] 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.

[14] 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.

[15] See BEA Table 2.3.4. Price Indexes for Personal Consumption Expenditures by Major Type of Product.

[16] Found in Table 6.2B (for 1980) and Table 6.2D (for 2011-2015). under the NIPA data

[17] Found in Table 6.3B (for 1980) and Table 6.3D (for 2011-2015) under the NIPA data

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 3.2.2).

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.[18] 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

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

where in 1980 it is

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

4.2 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 3.2.1:

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

where c^k is a constant, $X(i)$ is a set of controls with occupation-specific coefficients β^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,[19] and continuous variables for potential experience[20] and number of years in the United States.

4.3 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.

4.3.1 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 3.2.4, but with log adjusted wages in the left-hand-side of the imputation regression.

[18] Gross output is measured using the BEA's Use tables (after subtracting net imports) as described in section 5

[19] For 1980 we infer this variable by using information on the age of their youngest child.

[20] (potential experience) = (age) - (years in education) - 6, where (years in education) is inferred from reported educational attainment.

5 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.

We first aggregate values in both the use and import tables to our industry classification using a crosswalk constructed as described in Section 8, below.[21] 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 1. 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

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

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 $I^{1\bar{3},jj'}$ refers to imports in the averaged 2011-2015 period, is thus

$$M^{jj'} = U^{1980,jj'} \Big(1 - \frac{I^{1\bar{3},jj'}}{M^{1\bar{3},jj'}} \Big) \times \frac{I^{1980}/M^{1980}}{I^{1\bar{3}}/M^{1\bar{3}}},$$

where $I^{1980}, M^{1980}, I^{1\bar{3}}, and M^{1\bar{3}}$ are national aggregates.[22]

We interpret gross output reported in this table as compensation for capital investment in equipment and real estate structures. To make the data consistent with modeling assumptions of zero profits, we adjust the table in the following way. 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

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

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

[21] we omit scrap, used goods, noncomprable imports, and government in addition to the omitted industries in Table 1.

[22] 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.

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

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

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

6 Trade costs

Trade costs, $\kappa^j_{n'n}$, 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

$$\pi^j_{nn'} X^j_n = \frac{\left(\kappa^j_{nn'} x^j_{n'}\right)^{-\theta^j}}{\sum_{n'} \left(\kappa^j_{nn'} x^j_{n'}\right)^{-\theta^j}} X^j_n$$

Taking logs and applying $\kappa^j_{nn'} = \left(d^j_{nn'}\right)^{t^j_n}$,

$$\ln\left(\pi^j_{nn'} X^j_n\right) = \ln\left(X^j_n\right) - \theta^j \ln x^j_{n'} - \ln \sum_{n'} \left(\kappa^j_{nn'} x^j_{n'}\right)^{-\theta^j} - \theta^j t^j_n \ln d^j_{nn'}$$

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

$$\ln\left(\pi^j_{nn'} X^j_n\right) = \mu^j_n + \nu^j_{n'} - g^j \ln d^j_{nn'}$$

where μ^j_n is an destination dummy, $\nu^j_{n'}$ is an origin dummy, $\ln d^j_{nn'}$ is log distance between origin n' and destination n and $\ln\left(\pi^j_{nn'} X^j_n\right)$ is log bilateral trade-flow from origin n' to destination n. [23] We estimate the regression equation by OLS.

[23] Note that $\ln d^j_{nn'}$ here is the distance between the shipment origin and destination provided by the CFS. Only when calculating $\kappa^j_{nn'}$ do we use distances between MSA's calculated using haversine distances with U.S. Census Gazetteer coordinates.

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.[24]

7 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.

7.1 Joint distribution of prices, P_n^j

To obtain prices for the non-tradable industries (i.e. real estate and retail, construction, and utilites) 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 nonmetalic minerals, Metals, Machinery, Computer 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.

7.1.1 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.[25]

Since HPI data is proprietary, as an alternative to using he 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.[26] 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

$$P_n = \prod_j \left(P_n^j\right)^{\alpha^j}$$

[24] These industries are Wholesale trade, Transportation and storage, Professional and business services, Other, Communication, Finance an insurance, Education, Health, and Accommodation

[25] As with wages, we use the PCE price index to adjust the 1980 distribution of rents to 2015 dollars.

[26] 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.

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.

7.2 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 8.2. 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 adjustment to industry level prices from the 2011-2015 inversion in order to obtain a measure of prices resulting from productivity growth over time.

8 Crosswalks

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

8.1 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.[27] 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.[28] 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.

8.2 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 produces a crosswalk from the

[27] Missouri Census Data Center: `http://mcdc.missouri.edu/websas/geocorr14.html`

[28] `https://www.census.gov/geo/reference/county-changes.html`

data's original industry classification to our industries.[29] 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 29 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 inustries 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).[30] 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.[31] Thus, we can construct a crosswalk to ModelInds as described above.

- **ind1990 to ModelInds** The 1980 Census sample provides Ind1990 industry classifications, for which a Census provided concordance exists to NAICS 1997.[32] [33]

8.3 Occupation crosswalks

- **occ2010 to occsoc** Since the 1980 Census sample does not include occosc, but does include occ2010 (for which IPUMS provides a crosswalk to occsoc), we utilize the provided crosswalk.[34]

[29] 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 "Profesisonal and Business Services" industry.

[30] `http://faculty.tuck.dartmouth.edu/teresa-fort/data`

[31] `https://usa.ipums.org/usa/volii/indcross03.shtml`

[32] `https://www2.census.gov/programs-surveys/demo/guidance/eeo/indcswk2k.pdf`

[33] 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 identied $i-1$ industry.

[34] `https://usa.ipums.org/usa/volii/acs_occtooccsoc.shtml`