6-25-2018

Estimating state-industry employment, with an application to industrial localization

Andrew J. Cassey
Washington State University

Ben O. Smith
University of Nebraska at Omaha, bosmith@unomaha.edu

Estimating State-Industry Employment, With an Application to Industrial Localisation

Andrew J Cassey[a] and Ben O Smith[b]

[a]School of Economic Sciences, Washington State University; [b]College of Business Administration, University of Nebraska at Omaha

ARTICLE HISTORY
Compiled April 20, 2017

ABSTRACT

We describe a method to construct an industry-by-state repeated cross-section of employment at the most disaggregated level publicly available, covering 1963–2012. Nondisclosed data are estimated with a procedure using the hierarchical information structure. To illustrate the usefulness of the procedure, the resulting estimated data are tested to determine if industrial localisation of the processed food sector has changed over the last fifty years in the United States. Our findings suggest it has not changed systemically despite variation in levels of localisation within industries.

KEYWORDS
Nondisclosed data; processed food; localisation; filling method

JEL CLASSIFICATION
L11; Q19; R14

1. Introduction

The purpose of this paper is to describe a method for assembling a complete repeated cross-section of employment at the finest level of detail publicly provided, the state-industry level. We compile this repeated cross-section from public releases of the *Census of Manufactures* every five years from 1963 to 2012. Because of the fineness of the data at the state-industry level, many of the observations are either not disclosed or obfuscated by the Census Bureau. Thus we estimate those entries. This paper describes our estimation algorithm.

We expect that our estimation procedure will prove useful to the many scientists and researchers confronted with datasets containing nondisclosed entries. As an illustration of the usefulness of our method, we apply the resulting estimated data to the question of whether industrial localisation is changing over time in the United States. Industrial localisation occurs when industrial employment is geographically concentrated beyond the level observed in general economic activity, or in our case, overall manufacturing employment. Thus to study changes in industrial localisation, we need a repeated cross-section of employment at the state-industry level because that level is detailed enough for us to assess 1) if the level of localisation is changing within industries and 2) if the distribution of the level of localisation across industries is changing.

CONTACT Ben O Smith: bosmith@unomaha.edu

We measure the localisation of each industry in each time period using the Ellison and Glaeser (1997, *EG*) index. Compared to alternative measures such as the location quotient, the Gini index, or the Hoover index (1936), the *EG* measure is the appropriate measure to use when the data are highly disaggregated and some industries have few plants. Compared to measures such as that in Duranton and Overman (2005), the *EG* measure is preferred because it does not require the address of each plant or the distance between plants. That amount of information is too burdensome to acquire on a large scale.

We examine the processed food and kindred products industries to see if industrial localisation is changing over time individually and in aggregate. Food and kindred products is an ideal sector to study because it is a large and important sector in U.S. manufacturing, accounting for almost 10% of manufacturing employment. Therefore it is big enough to yield results that are not random and small enough that it does not define the manufacturing employment distribution. Furthermore, because one of the main inputs is agricultural output that is largely fixed in place by the soil and land, any changes to the overall localisation of the sector is unlikely to be due to changes in the location of its inputs.

Though localisation is known to occur widely in the United States (Holmes and Stevens 2004), it is not currently known if industrial localisation is changing over time. Krenz (2012) finds evidence of increased localisation in the European Union from 1970 to 2005 and Brakman, Garretsen, and Zhao (Forthcoming) find evidence of increasing industrial localisation in China from 2002 to 2008. Evidence from Kim (1995), however, suggests localisation may be decreasing in the United States from the 1940s through the 1980s. In particular, Kim shows the locational Hoover index of the U.S. processed food industry decreased from 0.196 in 1967 to 0.153 in 1987. This evidence, however, is not definitive: Kim uses aggregate sectoral data that could mask the trend at the industry level as well as regional geographic data. (Kim and Margo (2004) survey the literature on changes in economic geography in the United States over time, but do not discuss changes in localisation.)

To determine if localisation is changing at the level of each individual industry within the processed food sector, for each industry-year observation of localisation value, we use the Cassey and Smith (2014) procedure to create a 95% confidence interval. We then examine if there are statistically significant changes within an industry over time. To determine if localisation is changing at the level of the processed food sector, we construct the distribution of *EG* statistics for each time period and compare.

We find there are statistically significant changes in the levels of localisation within industries over time, but that the overall distribution of industrial localisation do not differ statistically from one another. Thus there is no systemic pattern of change to the distribution of localisation levels of processed food and kindred product industries in the United States since the 1960s. That result contrasts with the claims of Kim (1995) that localisation is decreasing in the U.S. processed food sector recently.

2. Measuring and Testing for Localisation

The Ellison and Glaeser (1997) measure for localisation is a ratio of the share of industry employment in a U.S state to the share of overall manufacturing employment in that state adjusted to account for the employment distribution of the plants in that industry. The advantage of the *EG* statistic over measures such as the locational Gini or Hoover coefficient is that it controls for the industrial organisation of each industry.

That is, the *EG* measure accounts for when there are a small number of plants or there is a small number of large plants in an industry. Thus the *EG* measure is best used on highly disaggregated industry-level data.[2] Unlike continuous-distance measures of localisation such as Duranton and Overman (2005) and Billings and Johnson (2016), the data requirements for the *EG* measure are met in principle by the published information available in the *Census of Manufactures*. In particular, the public data does not include the address of each plant, which is needed to calculate the distance between plants in the Duranton and Overman or Billings and Johnson localisation measures.[3] The *EG* statistic is a particularly useful measure of localisation because its values can be compared across industries, time, and levels of geographic aggregation.

The Ellison and Glaeser (1997) index for localisation requires three data inputs:

(1) for each U.S. state i, the state share of total manufacturing employment in year t, x_{it},

(2) for each U.S. state i, the state share of industry k employment in year t, s_{ikt}, and

(3) the plant-Herfindahl for industry k in year t, $H_{kt} = \sum_{j=1}^{N_{kt}} z_{jt}^2$, where N_{kt} is the number of plants in industry k in year t, and z_{jt} is plant j's share of industry k employment in year t.

The *EG* statistic for industry k in year t:

$$EG_{kt} = \frac{\overbrace{\sum_{i=1}^{51}(s_{ikt} - x_{it})^2}^{G_{kt}} - (1 - \sum_i x_{it}^2)H_{kt}}{(1 - \sum_i x_{it}^2)(1 - H_{kt})}.$$

The summation in the numerator goes to 51 since we consider the 50 U.S states and the District of Columbia. $G_{kt} = \sum_{i=1}^{51}(s_{ikt} - x_{it})^2$ is a raw geographic concentration measure unadjusted by distribution of plant employment in the industry.

Ellison and Glaeser prove that zero is the expected value of the *EG* statistic if there is no natural, economic, or political force for localisation regardless of industry parameters. The larger the statistic, the greater the indication of industrial localisation. Though the mean of the *EG* statistic does not depend on industry parameters, the distribution of the *EG* statistic does. Therefore, Cassey and Smith (2014) develop a procedure to test if an industry with a positive *EG* value is localised statistically. That test uses information about the industry to simulate 100,000 random *EG* data points and then compares the actual industry *EG* to a critical value associated with a significance level of 5%. The data required to perform this test are state employment shares, the number of plants in the industry, and the industry plant-Herfindahl. This method requires an implicit assumption that every industry has a log-normal plant

[2] Kim's (1995) findings are obtained using the Hoover index (a measure similar to the Gini coefficient except using absolute difference instead of squared difference) on sector level data at a regional, rather than state, geography. Thus he avoids the issue of small plant counts that the *EG* measure controls for.

[3] The improvement in measuring localisation with continuous distance measures over the *EG* statistic is that continuous distance measures avoid the modifiable areal unit problem (MAUP) of which the *EG* statistic is subject. That is, though the *EG* measure is robust to the fineness of the partition of geographic space, it is not robust to moving or modifying the borders of that partition. The cost in terms of avoiding the MAUP is the data requirement that the distance between each establishment must be known. Because we compare the *EG* statistic using the same geographic partition (U.S. States) over time, the MAUP is not an important problem in our context in that we care about changes in the localisation measure rather than the measurement itself and U.S. state boundaries have not changed since the 1960s.

employment distribution, though the parameters of that distribution are allowed to differ.

For interpretation, the larger the EG measure is, the more the industry is localised relative to the geographic distribution of manufacturing employment. Large but negative values indicate an industry is diffused in comparison to manufacturing employment. It is, however, not straightforward to compare the EG values across industries. That is because though the expected value of the EG measure is zero in the absence of agglomeration forces, the distribution of the EG stat for a given industry depends on industry parameters. Thus an EG value of 0.005 may be "large" for one industry but "small" for another. Hence we apply the Cassey and Smith (2014) procedure on each EG value. The Cassey and Smith procedure calculates the likelihood that the observed EG value could have arisen purely from chance. In this way, we asses the strength of the EG measure across industries.

3. Estimating Nondisclosed Observations

We study the highly disaggregated industries falling under Food and Kindred Products, Standard Industrial Classification (SIC) 20 from 1963–1992, and Food Manufacturing, North American Industrial Classification System (NAICS) 311 from 1997–2012. The industries considered include meat packing; poultry slaughtering and processing; dairy products; canned or frozen fruits and vegetables; beverages, liquors, and sodas; and processed seafood.

All of the information required to calculate the EG index are available in the U.S. Census Bureau's *Census of Manufactures*, which is released every five years. We use every release from 1963 to 2012. We begin our sample with 1963 as that release was the first to use administrative records to assist with identifying very small firms. The SIC and NAICS are hierarchical industrial categorisations in which broad sector labels are given in the higher, super-set level and then narrows progressively to the sub-set industrial level. We consider SIC 20 and NAICS 311 to be the comparable sectoral data. However, due to the difficulty in mapping SIC codes to their NAICS counterparts at our level of disaggregation, we perform our analysis between all SIC and NAICS years but not to each other.

We consider the industrial level to be the most disaggregated publicly available: 4-digit SIC and 6-digit NAICS. (For convenience, we will refer to the SIC levels of 2-digit as sector, 3-digit as subsector, and 4-digit as industry when discussing both data given by SIC and the corresponding NAICS 3-digit, 4-digit, and 6-digit levels.) The Census Bureau modified the names and inclusions of each category over time. Of course the major revision comes in 1997 with the end of SIC and the beginning of NAICS. Other major revisions occurred in 1967, 1972, and 1987. Minor revisions occur each release. We use the Census Bureau definitions given in each year, which is why we consider our data a repeated cross-section rather than a panel.

The "Geographic Area Statistics" of the *Census of Manufactures* gives the total manufacturing employment of each state, allowing us to calculate x_{it}. It also contains data on the national and state employment of each industry at the industry level. This allows us to calculate s_{ikt}. However, due to redactions to prevent the identification of individual plant operations, the Census Bureau does not report the employment total in all state-industry cases. There are three restrictions on the reported data. First, there are no observations for a state-industry in which employment does not reach a threshold. That threshold changes from release to release. For those state-

Table 1. Information Availability By Type and Year

SIC 20 4-digit	1963	1967	1972	1977	1982	1987	1992
			(percent of observations)				
Provided Data	68	97	95	94	95	97	95
– Numeric value	44	36	28	32	30	25	32
– Bin and range	24	61	67	62	65	73	62
Single missing value fill	1	1	1	1	1	1	1
Parent code weight fill	31	2	4	5	4	2	4

NAICS 311 6-digit	1997	2002	2007	2012
		(percent of observations)		
Provided Data	84	86	86	85
– Numeric value	38	38	38	37
– Bin and range	46	48	48	48
Single missing value fill	15	14	14	13
Parent code weight fill	1	0	0	2

The percent of cells filled by procedure. For the cases when a numeric value was not provided, an employment bin with a range of values was usually given or could be applied. Other observations were calculated by virtue of being a single missing value with a populated parent observation. In cases where no range was provided, weights were used from the parent code.

industries whose employment level exceeds the threshold and thus are reported, the employment datum is either reported as the numeric value rounded to the nearest hundred or assigned to a bin with a range of employments. Table 1 shows the percent of observations provided by the Census Bureau by each type and year. The numerical value is directly reported for about 30% of observations for the SIC years and 38% for the NAICS years. Data provided in the bin and range format account for the vast majority of the remaining observations, or about 66% in the SIC years and just under 50% in the NAICS years. No information on the remaining observations is disclosed. We categorise those observations by whether we calculate the numeric value by virtue of it being a single missing entry with a populated parent observation or by applying the share from a parent observation as a weight, as described below.

For the state-industry-employment entries that are either provided as a bin and range or are not disclosed at all, we create a procedure to estimate their precise numeric value. The estimation procedure works from the bottom up first and then back down. From the bottom up, if there is an industry in which all state-employment observations are reported except for one, then we take the sum of industry employments from the other 51 "states" and subtract from the reported total industry employment in the parent level. Likewise if there is a state in which all state-employment observations are reported except for one industry, then we take the sum of industry employments from the other industries and subtract from the reported total state employment in the parent level. These are the two adding up constraints: one for employment across states within an industry and the other for employment across industries within a state.

Figure 1 shows the first steps in the "bottom-up" procedure. In the figure, the dashed areas were initially nondisclosed. The entry for State 4 Industry A can be filled in using the Industry adding up constraint: $600 - (100 + 200 + 200) = 100$. The entry for State 2 Industry C can be filled in using the State adding up constraint: $1810 - (200 - 1500 - 100) = 10$. Once those entries are complete, then the entry for State 4 Industry C can be filled in even though it could not be filled in originally. Though there are not many observations we can complete in the SIC years using this "single missing value fill" method, as table 1 shows, we can fill in as many as 15%

	Industry A	*Industry B*	*Industry C*	*Industry D*	*Total State*
State 1	100	1000	10	200	**1310**
State 2	200	1500	10	100	**1810**
State 3	200	1700	1000	500	**3400**
State 4	100	1000	20	100	**1220**
Total Ind.	**600**	**5200**	**1040**	**900**	

Figure 1. The estimation "bottom up" procedure. Areas enclosed with a dashed line are initially nondisclosed. The State 2 Industry C observation and State 4 Industry A observation are filled in using one of the adding up constraints. That allows State 4 Industry C to be filled in next.

	Industry Employment	*Constraint*	*Parent Employment*	*% of Unallocated*
State 1	300		20000	**0.00**
State 2	451*.66=298 ⟶ 249	100-249	40000	**0.00|0.66|0.44**
State 3	500*.33=165 ⟶ 49	0-49	30000	**0.00|0.00|0.33**
State 4	202*1.0=202 ⟶ 202	None	20000	**1.00|0.33|0.22**
Total	**800**		**110,000**	

Figure 2. The estimation "top down" procedure. Dashed areas are initially nondisclosed. The state share of parent employment for nondisclosed entries is recalculated each step.

percent of observations in the NAICS years.

Once all nondisclosed entries that can be filled in using one of the two adding up constraints have been entered, we then move up and look if there is a nondisclosed entry at the parent level. This would have been the case in figure 1 if State 2 Industry A had also been nondisclosed. If there is a single nondisclosed entry at the state-subsector level we can fill in the observation value using the state or industry total at the state-sector level and the adding up constraint. Finally, we move up to the top at the state-sector level—where there are no nondisclosed entries.

With all state-sector entries filled, the estimation procedure works its way down to the state-subsector level. It is sometimes possible to tighten the Census Bureau bins by seeing how the state total compares to the industry sum when either the minimum amount of employment in each bin is assigned or the maximum amount is assigned. Even if it is not possible to tighten the employment bins, we can estimate the nondisclosed observations recursively. Beginning with the smallest employment bin, we use the relative employment ratio from the parent category *of the states with nondisclosed observations only* to fill in the blanks. We then update the relative employment ratio from the parent category because one of the nondisclosed observations has been filled in and move on to the next largest bin size. We start with the smallest bin because there is the least amount of uncertainty and thus by assigning unallocated employment we reduce the uncertainty of the larger bins. All remaining employment is assigned to the top-coded entries. At the end of the process we recursively check if our assignment can be improved. Once all the sub-sector level entries are filled in, we move down to the industry level entries, using the sub-sector data as the parent value adding up constraints. This procedure continues until no other cells can be assigned.

Figure 2 illustrates how the "top down" procedure works. From the difference in the sum of industry employment and the U.S. total, there are 500 unallocated employees across States 2, 3, and 4 within the industry. We know there are 40,000 employees in State 2 in the parent subsector level that includes this industry and many others. Of the three states with unallocated industry employment, the parent share of State 2 is $40,000/(40,000 + 30,000 + 20,000) = 44\%$. We start with the state that has the smallest bin range. In figure 2 that is State 3, whose bin only has a range of 50 employees compared to 150 for State 2 and infinite for State 4. We apply State 3's 33% share of parent employment to the 500 unallocated employees from the industry. As $500 \times 0.33 = 165$, the share runs up against the top of State 3's bin constraint. Hence we enter the maximum value in the bin, 49 employees. We then update the share of unallocated workers among the two remaining states so that State 2 has a share of $40,000/(40,000 + 20,000) = 66\%$. That share for State 2 is applied to the remaining 451 unallocated employees, $451 \times 0.66 = 297.66$, which is above the bin constraint. Hence 249 is recorded as the entry and the remaining 202 employees are assigned to State 4.

For the state-industry observations that we cannot estimate using our procedure, we take the difference between national employment and the sum of the states including our estimated values and assign a value by using the overall state relative employment. Other than 1963, our procedure gives usable data or estimates for at least 96% of observations. Thus our application of the parent code weight fill is less than 4% of observations.

For some years, the Census Bureau releases the firm-Herfindhal for each industry, which includes only up to the top 50 largest firms. However, the firm-Herfindahl is conceptually different from the plant-Herfindahl when there are multi-plant firms. Thus we do not use the Herfindahl reported by the Census Bureau. Instead we calculate the Herfindahl in (2) from the data in the "Statistics for Industry Groups and Industries" of the *Census of Manufactures*. Based on its employment, the Census Bureau assigns each establishment into one of ten bins. The bins, which do not change over time, are 1–4, 5–9, 10–19, 20–49, 50–99, 100–249, 250–499, 500–999, 1000–2499, and 2500+ employees. For each industry, the Census Bureau reports the total employment of each bin, which is made up of all the plants assigned to that bin.

From this, we estimate the plant-Herfindahl using the Schmalensee (1977) method. This method takes the employment total for each bin and assigns that number to each value in the bin range so that employment shares are equal and backs out the plant count to allow for the adding up constraint (p.187). For example, if there are 120 employees for the 1–4 bin, the method says to assign 30 employees to plants with 1 worker and thus implying 30 such plants, 30 employees to plants with 2 workers implying 15 such plants, and so on. Then move on to the 5–9 bin and continue up. The linear distribution within the 1000–2499 bin is extended to cover the open-ended top-code bin. This gives us an estimate for the number of plants in each industry nationwide in each year N_{kt} as well as the employment in each of those plants. From that we can calculate the plant-Herfindahl, H_{kt}.

As part of this paper, we provide the code for our estimation procedure and the data for our example sector of processed foods. The code and data are obtainable in the accompanying online materials as well as https://goo.gl/nK1fqs. It is hoped that the program and data will be used by other scientists and researchers. The appendix contains further detail on how our program includes checks and balances to ensure accuracy and prevent mistakes in the estimation, in particular in preventing errors in estimates based on other estimated data.

4. Assessing the Quality of the Estimation Procedure

We compare our filing procedure to one common filling procedure used: the midpoint procedure. With the midpoint procedure, the practitioner uses the median of the range of values for the bin indicated in the entry cell provided by the Census Bureau as the estimated value.

To compare our procedure to the midpoint procedure, we simulated the structure of each of the forty-three 2012 NAICS-311 industries 250 times using the number of plants and mean plant size as the parameters for the underlying characteristics of each industry. For each industry k, we generate N_k random plants from a lognormal distribution with parameters (μ_k, σ_k). We parameterise μ_k so that the expected median of the distribution is equal to the mean employment of the Census Bureau provided data for industry k. Given the underlying σ_k is unknown in the data, we conduct our simulation using multiple plausible values of σ_k: 1.00, 0.50, and 0.25. Each of the N_k plants randomly drawn from the industry-specific distribution is then probabilistically assigned a geographic location based on the share of state employment. Once these N_k plants of different sizes are located, we sum the employment of the plants to generate state and national level employment tables similar to those provided by the Census Bureau. With these “complete” simulated data, we calculate the “true” EG index value for each of the simulated industries. We then apply a censoring procedure to the simulated data to obfuscate the data in a way that we believe represents the Census Bureau’s own nondisclosure procedure.

The Census Bureau’s nondisclosure procedure is not publicly known. Clearly, however, one of the goals of the Census Bureau nondisclosure procedure is to not reveal the specific employment of any individual plant. Plant-specific employment could be deduced if either there are a few plants in a given employment state-industry cell or a large share of employment from a single employer in a given cell. Thus, our censoring method calculates a plant Herfindahl index H_{ik} for each state-industry level from the employment of plants in that state-industry pair. (We similarly calculate a plant Herfindahl index for each industry employment cell by establishment size in the national data). If the calculated H_{ik} is above a threshold, then we do not disclose that simulated data entry, but rather report a bin entry.[4]

To establish the H cut-off for nondisclosure, we search across all possible thresholds and choose the one that best matches the actual nondisclosed entries for the 2012 data. We find an H threshold of 0.25 best matches the set of nondisclosed entries in the actual 2012 data. At an H cut-off of 0.25 and $\sigma = 1.00$, 44% of our simulated data are obfuscated as a bin-and-range entry, whereas it is 48% in the actual data. Furthermore, not only were we able to closely recreate the number of nondisclosed entries with our simulated data, we were also able to recreate that all of the possible bin ranges were used in the obfuscation, just as in the actual data. The H cut-off at 0.25 also works best when $\sigma = 0.50$ and $\sigma = 0.25$.

Once we have our simulated data with nondisclosed entries, we run our estimation procedure. After the bottom-up and top-down algorithms estimate the nondisclosed entries, we calculate the EG index for each of the industries. Similarly, we estimate the nondisclosed entries using the midpoint procedure and calculate a corresponding EG index. Thus we have three EG values for each industry: one from the complete simulated data with no nondisclosed entries, one from the simulated data whose nondis-

[4] All of our nondisclosure edits are recorded as a bin entry as the midpoint procedure requires a range. Our general estimation procedure, however, works with missing entries as well as bins.

closed entries were estimated with our procedure, and one from the simulated data whose nondisclosed entries were estimated with the midpoint procedure.

We assess if our estimation procedure outperforms the midpoint procedure by calculating the percent difference between the *EG* value from each procedure and the "true" simulated *EG* value. We find that our estimate procedure outperforms the midpoint procedure. When $\sigma = 1.00$, our procedure is 40% closer to the "true" simulated *EG* value at the median. (Our procedure is 35% closer on average.) When $\sigma = 0.50$, our procedure is 30% more accurate at the median, and when $\sigma = 0.25$, our procedure is about 40% more accurate at the median than the midpoint procedure. Furthermore, for $\sigma = 1.00$, our procedure results in 20% fewer *EG* values that are more than 10% away from the "true" simulated EG value than the midpoint procedure. Thus we believe our estimation procedure results in a strong improvement in accuracy over the midpoint estimation procedure.

5. Results for the Processed Food Sector

As an illustration of why having a complete repeated cross-section of state-industry employment data is useful, we calculate the *EG* statistic for each of the disaggregated industries in processed food and kindred products sector from 1963 through 2012 using the industry description and code at the time the data was released. These results may be seen in tables 2 and 3. A * indicates when the level of localisation is statistically significant at the 5% level and a $^\diamond$ indicates when the level of localisation is significantly different from the previous period at the 5% level.

One issue with the Cassey and Smith (2014) test is that when the distribution of plant employment sizes is not known, there is a range of critical values that depends on the parameters of the unknown distribution. A conservative or strict way to apply the test is to require that the *EG* measure for the industry is statistically significant for all plausible parameters of the plant employment distribution. We apply this conservative approach in our results.

5.1. *Within Industry Results*

Consider the share of processed food industries that are statistically localised over time in tables 2 and 3. Despite our strict application of the Cassey and Smith (2014) test, tables 2 and 3 show localisation is common in the industries making up the processed food sector. In each time period cross-section, at least 60% of industries have levels of localisation that are statistically different from zero with 95% confidence. The percent of localised industries fluctuates over time, but there is no trend.

Nonetheless, we see some individual industries experience dramatic changes in localisation. There are 32 SIC-4 industries in the processed food and kindred products group whose data are present all eight years. Of those 32, 17 (44%) change from levels of localisation that are statistically significant to not statistically significant (or vice versa) at least once in the eight time periods. Eleven (34%) change more than once. Similarly, there are 37 NAICS-6 industries in the processed food and kindred products group available each applicable year. Of those, 16 (48%) have levels of localisation that have changed from statistically significant to not statistically significant (or vice versa). Four have changed more than once.

More formally, we compare the change in *EG* values year to year to the most conservative Cassey and Smith critical value given the the number of plants in the

Table 2. *EG* Values for Food Processing Industries from 1963–1992

Description	SIC	1963	1967	1972	1977	1982	1987	1992
Meat Packing Plants†	2011	.033*	.033*	.037*	.031$^{*\diamond}$	.037$^{*\diamond}$	.039*	.050$^{*\diamond}$
Meat Processing Plants*†	2013	.011*	.006$^{*\diamond}$	.007*	.004$^{*\diamond}$	.004*	.007*	.009$^{*\diamond}$
Poultry Dressing Plants	2015	.039*	.043$^{*\diamond}$				.057*	.056*
Poultry and Egg Processing	2016			.052*	.048$^{*\diamond}$	.046*		
Creamery Butter	2021	.130*	.127*	.081$^{*\diamond}$	.078*	.071*	.048*	.154$^{*\diamond}$
Natural, Processed Cheese†	2022	.158*	.150$^{*\diamond}$	.132$^{*\diamond}$	.145$^{*\diamond}$	.139$^{*\diamond}$	.136*	.118$^{*\diamond}$
Condensed and Evaporated Milk†	2023	.053*	.038$^{*\diamond}$	.033*	.044$^{*\diamond}$	.035*	.045*	.018$^{*\diamond}$
Ice Cream and Frozen Desserts†	2024	.000	−.000	.002	.001	−.001	−.003	−.007*
Fluid Milk	2026	.003*	.003*	.004*	.001	.006$^{*\diamond}$	.007*	−.001 $^{\diamond}$
Canned and Cured Seafoods	2031	.083*	.077*					
Canned Specialties	2032	.009	−.029$^{*\diamond}$	−.005	−.017*	−.015*	−.012	−.012
Canned Fruits and Vegetables	2033	.037*	.037*	.046$^{*\diamond}$	.058$^{*\diamond}$		.040*	.040*
Dehydrated Food Products*	2034	.393*	.216$^{*\diamond}$	.152$^{*\diamond}$	.174$^{*\diamond}$	.270$^{*\diamond}$	.269$\ast$	.230$^{*\diamond}$
Pickles, Sauces, Salad Dressing†	2035	.008*	.008	.003	.010*	.011*	.003	.003
Fresh or Frozen Packaged Fish	2036	.064*	.051$^{*\diamond}$					
Frozen Fruits and Vegetables†	2037	.028*	.016$^{*\diamond}$	.039$^{*\diamond}$	.078$^{*\diamond}$	.069*	.068*	.058*
Frozen Specialties	2038			−.002	−.001	−.001	.005	.003
Flour Mills*†	2041	.021*	.016*	.019*	.019*	.017*	.111$^{*\diamond}$	.013$^{*\diamond}$
Prepared Feeds for Animals and Fowls	2042	.014$\ast$	.017$^{*\diamond}$					
Cereal Preparations*	2043	.212*	−.055$^{*\diamond}$	−.055*	.018$^{*\diamond}$	−.022	.006	.013
Rice Milling	2044	.161*	.160*	.146*	.163*	.158*	.159*	.155*
Blended and Prepared Flour*	2045	−.012	−.003	.005	.047$^{*\diamond}$	−.008 $^{\diamond}$	.005	.002
Wet Corn Milling	2046	−.212*	−.002 $^{\diamond}$	−.003	.085$^{*\diamond}$	.093*	.132$^{*\diamond}$	.146*
Dog, cat, and Other pet Food*	2047			.010	.000	.004	.006	.005
Prepared Feed, nec	2048			.020*	.021*	.021*	.020*	.017*
Bread, and Related Products†	2051	−.000	.000	−.001	−.001	−.001	.003$^{*\diamond}$	−.000$^{\diamond}$
Biscuit, Cookies and Crackers*†	2052	.006	.010	.018*	.017*	.014*	.011*	.012*
Raw Cane Sugar	2061	.407*	.388$^{*\diamond}$	.267$^{*\diamond}$	.263*	.246*	.411$^{*\diamond}$	.156$^{*\diamond}$
Cane Sugar Refining	2062	.193*	.000 $^{\diamond}$	.001	.001	.035$^{*\diamond}$	−.048$^{*\diamond}$	−.060*
Beet Sugar	2063	.040*	.033*	.025*	.037$^{*\diamond}$	.046$^{*\diamond}$		.017*
Confectionery Products	2065			.035*	.042*	.030$^{*\diamond}$		
Chocolate and Cocoa Products	2066			.285*	.244*	.635$^{*\diamond}$	.092$^{*\diamond}$	.171$^{*\diamond}$
Chewing Gum	2067			−.164*	−.051$^{*\diamond}$	−.034	−.023*	
Confectionery Products	2071	.027*	.030*					
Chocolate and Cocoa Products	2072	.213*	.217$^{*\diamond}$					
Chewing Gum	2073	.161*	.415*					
Cottonseed Oil Mills	2074			.096*	.112$^{*\diamond}$	.113*	.009$^{\diamond}$	.122$^{*\diamond}$
Soybean Oil Mills	2075			.019*	.079$^{*\diamond}$	.086*	.051$^{*\diamond}$	.089$^{*\diamond}$
Vegetable Oil Mills, nec	2076			−.024	−.017	−.038	.043$^{*\diamond}$	−.026 $^{\diamond}$
Animal and Marine Fats and Oils	2077			.003	.005*	.007*	.012*	.009*
Shortening and Cooking Oils	2079			.010	.006	.015*	.017	.005
Malt Beverages†	2082	.010	.011	−.011 $^{\diamond}$	.006	−.011	−.003	.012
Malt	2083	.150*	.143*	.054$^{*\diamond}$	.103$^{*\diamond}$	.175$^{*\diamond}$	.227$^{*\diamond}$	.136$^{*\diamond}$
Wines and Brandy*	2084	.341*	.278$^{*\diamond}$	.341$^{*\diamond}$	.447$^{*\diamond}$	.491$^{*\diamond}$	.524$^{*\diamond}$	.572$^{*\diamond}$
Distilled Liquor, Except Brandy	2085	.119*	.121*	.014 $^{\diamond}$	.064$^{*\diamond}$	.089$^{*\diamond}$	.075*	.131$^{*\diamond}$
Bottled and Canned Soft Drinks†	2086	.006*	.005*	.004*	.004*	.004*	.006$^{*\diamond}$	−.000 $^{\diamond}$
Flavorings*†	2087	.020*	.019*	.033$^{*\diamond}$	.037*	.013$^{*\diamond}$	−.003 $^{\diamond}$	.017$^{*\diamond}$
Cottonseed Oil Mills*	2091	.109*	.130$^{*\diamond}$	.087$^{*\diamond}$	.100*	.131$^{*\diamond}$	.053$^{*\diamond}$	.072*
Soybean Oil Mills*	2092	.106*	.089*	.046$^{*\diamond}$	.042$^{*\diamond}$	.046*	.034$^{*\diamond}$	.082$^{*\diamond}$
Vegetable Oil Mills, nec	2093	−.008	−.025					
Animal and Marine Fats and Oils	2094	−.000	.003					
Shortening and Cooking Oils*	2095	.024*	.030*	.059$^{*\diamond}$	.047*	.038*	.018$^{*\diamond}$	−.001 $^{\diamond}$
Roasted Coffee*	2096	.022*	.026*				.011*	.006
Manufactured Ice†	2097	.020*	.016$^{*\diamond}$	.016*	.017*	.018*	.013$^{*\diamond}$	.002 $^{\diamond}$
Macaroni and Spaghetti	2098	.002	−.011	−.010	−.001	.001	−.018$^{*\diamond}$	−.004
Food Preparations, nec†	2099	.001	.003$^{*\diamond}$	.001	.004$^{*\diamond}$	.005*	.010$^{*\diamond}$	.011*
Share localised		.773	.750	.660	.787	.783	.750	.667
Sectoral *EG*		.023*	.003*	.003*	.003*	.003*	.004*	.005*
Sectoral *G*		.022	.003	.003	.003	.004	.004	.005

Notes: A "*" indicates that the industry is localised beyond randomness with 95% confidence using the most conservative critical value. A "$^{\diamond}$" indicates that the difference in *EG* from the previous period is greater than the 95% critical value. Blanks indicate the code does not exist for the given year. SIC code descriptions are based on the first year the data are available. Descriptions with a "$\star$" indicate a definition change within reported years. The details may be found in table 6. The 15 industries with the fewest estimated cells are indicated with a "†." Raw Geographic Concentration: $G = \sum_{i=1}^{51} (s_{it} - x_{it})^2$.

Table 3. *EG* Values for Food Processing Industries from 1997–2012

Description	NAICS	1997	2002	2007	2012
Dog and cat Food Manufacturing†	311111	$-.006$	$.012^{\diamond}$	$.003$	$.024^{*\diamond}$
Other Animal Food Manufacturing†	311119	$.013^{*}$	$.010^{*\diamond}$	$.012^{*}$	$.012^{*}$
Flour Milling	311211	$.010^{*}$	$.009^{*}$	$.010^{*}$	$.003$
Rice Milling	311212	$.148^{*}$	$.183^{*\diamond}$	$.051^{*\diamond}$	$.235^{*\diamond}$
Malt Manufacturing	311213	$.126^{*}$	$.120^{*}$	$.011^{\diamond}$	$.057^{*}$
Wet Corn Milling	311221	$.121^{*}$	$.150^{*}$	$.165^{*}$	$.123^{*\diamond}$
Soybean Processing	311222	$.077^{*}$	$.090^{*}$	$.086^{*}$	
Other Oilseed Processing	311223	$-.005$	$-.010$	$-.030$	
Soybean and Other Oilseed Processing	311224				$.044^{*}$
Fats and Oils Refining and Blending	311225	$.009$	$.015$	$.019^{*}$	$.020^{*}$
Breakfast Cereal Manufacturing	311230	$.007$	$.021$	$.022$	$.050^{*}$
Sugarcane Mills	311311	$.089^{*}$	$.102^{*}$	$.044$	
Cane Sugar Refining	311312	$-.069^{*}$	$-.117^{*}$	$-.002^{\diamond}$	
Cane Sugar Manufacturing	311314				$.073^{*}$
Beet Sugar Manufacturing	311313	$.019$	$.038$	$.051^{*}$	$.048$
Chocolate and Confectionery Manufacturing from Cacao Beans	311320	$.161^{*}$	$.044^{*\diamond}$	$.068^{*}$	
Confectionery Manufacturing from Purchased Chocolate	311330	$.030^{*}$	$.023^{*}$	$.024^{*}$	
Non-chocolate Confectionery Manufacturing†	311340	$.027^{*}$	$.033^{*}$	$.020^{*}$	$.041^{*\diamond}$
Chocolate and confectionery manufacturing from cacao beans	311351				$.059^{*}$
Confectionery manufacturing from purchased chocolate	311352				$.015^{*}$
Frozen Fruit, Juice, and Vegetable Manufacturing	311411	$.046^{*}$	$.069^{*\diamond}$	$.076^{*}$	$.086^{*}$
Frozen Specialty Food Manufacturing†	311412	$.008^{*}$	$.011^{*}$	$.012^{*}$	$.009$
Fruit and Vegetable Canning†	311421	$.032^{*}$	$.032^{*}$	$.026^{*\diamond}$	$.036^{*\diamond}$
Specialty Canning	311422	$-.006$	$-.029^{*}$	$-.027$	$-.029$
Dried and Dehydrated Food Manufacturing	311423	$.186^{*}$	$.135^{*\diamond}$	$.074^{*\diamond}$	$.075^{*}$
Fluid Milk Manufacturing†	311511	$.017^{*}$	$.001^{\diamond}$	$.000$	$-.001$
Creamery Butter Manufacturing	311512	$.190^{*}$	$.106^{*\diamond}$	$.032$	$.027$
Cheese Manufacturing†	311513	$.118^{*}$	$.119^{*}$	$.121^{*}$	$.124^{*}$
Dry, Condensed, and Evaporated Dairy Product Manufacturing	311514	$.009$	$.012$	$.017^{*}$	$.027^{*}$
Ice Cream and Frozen Dessert Manufacturing†	311520	$-.005$	$-.006$	$-.012$	$-.012$
Animal (Except Poultry) Slaughtering†	311611	$.046^{*}$	$.046^{*}$	$.045^{*}$	$.041^{*}$
Meat Processed from Carcasses†	311612	$.016^{*}$	$.013^{*}$	$.019^{*\diamond}$	$.022^{*\diamond}$
Rendering and Meat Byproduct Processing†	311613	$.005$	$.002$	$.012^{*\diamond}$	$.008^{*\diamond}$
Poultry Processing†	311615	$.059^{*}$	$.058^{*}$	$.064^{*\diamond}$	$.053^{*\diamond}$
Seafood Canning	311711	$.082^{*}$	$.067^{*}$	$.097^{*}$	
Fresh and Frozen Seafood Processing	311712	$.075^{*}$	$.077^{*}$	$.117^{*\diamond}$	
Seafood Product Preparation and Packaging	311710				$.122^{*}$
Retail Bakeries†	311811	$.011^{*}$	$.009^{*}$	$.011^{*}$	$.010^{*}$
Commercial Bakeries†	311812	$.001$	$.001$	$.002$	$.002^{*}$
Frozen Cakes, Pies, and Other Pastries Manufacturing†	311813	$.013$	$.004$	$.010$	$.003$
Cookie and Cracker Manufacturing†	311821	$.018^{*}$	$.027^{*}$	$.015^{*}$	$.017^{*}$
Flour Mixes and Dough Manufacturing from Purchased Flour	311822	$.008$	$.009$	$.006$	
Dry Pasta Manufacturing	311823	$-.000$	$.011$	$.005$	
Dry Pasta, Dough, and Flour Mixes Manufacturing from Purchased Flour	311824				$.011^{*}$
Tortilla Manufacturing	311830	$.077^{*}$	$.079^{*}$	$.064^{*\diamond}$	$.046^{*\diamond}$
Roasted Nuts and Peanut Butter Manufacturing	311911	$.076^{*}$	$.076^{*}$	$.071^{*}$	$.071^{*}$
Other Snack Food Manufacturing	311919	$.009^{*}$	$.015^{*}$	$.017^{*}$	$.022^{*}$
Coffee and tea Manufacturing	311920	$.423^{*}$	$-.009^{\diamond}$	$-.008$	$-.018^{*}$
Flavoring Syrup and Concentrate Manufacturing	311930	$.009$	$-.006$	$.035^{*\diamond}$	$.030^{*}$
Mayonnaise, Dressing, and Other Prepared Sauce Manufacturing†	311941	$-.003$	$-.006$	$-.001$	$-.003$
Spice and Extract Manufacturing†	311942	$.004$	$.004$	$.005$	$.012^{*}$
Perishable Prepared Food Manufacturing†	311991	$.020^{*}$	$.035^{*\diamond}$	$.031^{*}$	$.039^{*}$
All Other Miscellaneous Food Manufacturing†	311999	$.009^{*}$	$.011^{*}$	$.007^{*}$	$.014^{*}$
Share localised		$.660$	$.638$	$.638$	$.791$
Sectoral *EG*		$.005^{*}$	$.006^{*}$	$.006^{*}$	$.006^{*}$
Sectoral *G*		$.005$	$.006$	$.006$	$.006$

Notes: A "*" indicates that the industry is localised beyond randomness with 95% confidence using the most conservative critical value. A "$^{\diamond}$" indicates that the difference in *EG* from the previous period is greater than the 95% critical value. Blanks indicate the code does not exist for the given year. NAICS code descriptions are based on the first year the data are available. The 20 industries with the fewest estimated cells are indicated with a "†." Raw Geographic Concentration: $G = \sum_{i=1}^{51} (s_{it} - x_{it})^2$.

industry and plant-Herfindahl. Throughout tables 2 and 3 we indicate with a $^\diamond$ the EG values where the difference from the previous year is greater than the largest 95% critical value. Even using this conservative approach, statistically significant changes are common in tables 2 and 3, occurring about 45% of the time in SIC years and about 20% of the time for the NAICS years.

Thus we see a lot of churn and changes within industry, but in general, there does not appear to be a pattern of increasing or decreasing levels of localisation. Rather some industries increase their levels of localisation, others decrease their levels of localisation, some stay the same, and many fluctuate between levels of localisation. One concern is that the reason for the churn in EG values is because of our estimation procedure rather than the data itself. For evidence this is not the case, see the robustness section below as well as the appendix, which details the checks in our program to ensure quality.

5.2. *Sectoral Results*

We now consider the value of the EG measure for the processed food sector as a whole. The results may be found in the bottom three rows of tables 2 and 3. We list the percent of industries that are localised at a statistically significant level. We also list the sectoral EG value (calculated using sector shares rather than industry shares) and the raw geographic concentration, G. Absent the results for 1963, which differ quite strongly from every other period, these measures show a bit of fluctuation, but no trend. Thus using relatively simple measures of sectoral localisation, it appears there has been no systemic change over time. Contrast this with findings of Kim (1995) using only sectoral data. Kim reports the Hoover index decreased from 0.196 in 1967 to 0.153 in 1987.[5]

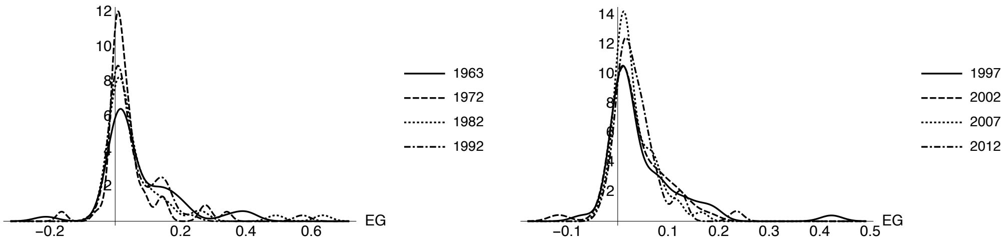
Now consider the distribution of the EG measure at the industry level over time. Figures 3(a) and 3(b) show the histogramme of industry EG values for the processed foods and kindred products sector for each time period.[6] For the SIC years, we include data from each decade only in order to have a clearer image. If localisation were changing systemically, we would expect to see either the peak moving upwards or the tail stretching and widening period-after-period. But instead we do not find a pattern in either the SIC or NAICS distributions. All four distributions cross each other repeatedly in both left and right panels. There is no trend in successive periods for either the peak increasing or the tail getting fatter.

Another way to see this is in figures 4(a) and 4(b). The EG value for each industry and year is on the x-axis. For the y-axis, we ordered industries by rank of their localisation. Thus the industry with the highest level of localisation in each period is ranked 1. We then convert those ranks into percentiles so that the figures are conceptually similar to a cumulative distribution function.

If localisation were increasing over time, then the distribution of each new decade would be shifted further out in the figures. We neither see this nor do we see that the distribution from newer periods shifted in as would be the case if localisation were decreasing over time. In the left panel, the curves shift back and forth and cross each other, sometimes multiple times, thus indicating there is no first-order dominance

[5] Kim reports data every twenty years from 1860 to 1987, thus our time overlaps only twice in 1967 and 1987. Kim reports the following Hoover index values for the processed food sector: 1860-0.322, 1880-0.311, 1900-0.215, 1914-0.231, 1927-0.249, 1947-0.260, 1967-0.196, 1987-0.153. Unlike our procedure using Census Bureau data from each state, Kim grouped U.S. states into nine geographic regions before calculating index values.

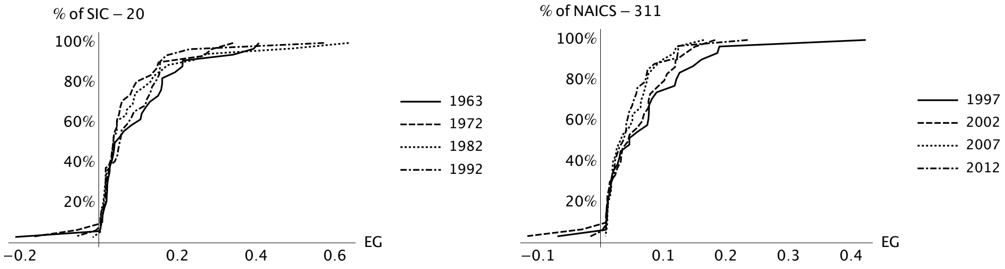
[6] For smoothed plots in the manuscript, we use Silverman's method (1986, p.48) to select the bandwidth.



(a) 1963–1992

(b) 1997–2012

Figure 3. Smooth histogramme of the *EG* values: 1963–2012. Frequency is on the vertical axis.



(a) 1963–1992

(b) 1997–2012

Figure 4. *EG* values against percentile of localised industries

of one distribution of localisation over another. In the right panel, the lines cross, but there does appear to be a shifting-in that would be consistent with decreasing localisation.

To see if there is in fact decreasing localisation, or if it is a trick of the eye in figure 4(b), we formally test if the *EG* index distributions for processed foods and kindred products from each time period are the same. Table 4 contains the first four moments from each distribution by period and SIC and NAICS classification system. First we test if the first four moments match pairwise. Though the means of the distribution change from year to year, we cannot reject the null hypothesis that the distribution centres are the same across the distributions. This result is based on a Mann-Whitney median test which does not rely on the normality assumption. (We support that finding with a t-test.) Using the Fisher Ratio and Conover tests, we also cannot reject the null hypothesis that the variances are the same across the distributions pairwise. For the third and fourth moments, the standard tests rely on a normality assumption that do not apply to our data. We therefore test using a bootstrapping technique. We cannot reject that these moments are the same pairwise.

In addition to testing if the moments match, we use a pairwise Kolmogorov-Smirnov (KS) test to see if the entire distributions are the same. Those results, available in table 5, indicate that we cannot reject the null hypothesis that the distributions are the same. The table lists the p-values for the two-sample KS test. Each result has a p-value much greater than the 0.05 threshold. Thus, despite the differences in appearance among the four distributions in figure 4(b), the distributions are not statistically different from one another.

Table 4. Moments from the Industry Distribution of Localisation Values by Period

Moment	SIC	1963	1967	1972	1977	1982	1987	1992
Mean		0.073	0.065	0.042	0.056	0.068	0.058	0.057
Standard Deviation		0.112	0.103	0.084	0.087	0.127	0.106	0.098
Skewness		1.269	1.895	1.662	2.515	2.934	2.811	3.218
Kurtosis		5.220	6.264	7.251	10.673	12.220	11.346	16.750
Moment	NAICS	1997		2002		2007		2012
Mean		0.049		0.036		0.032		0.038
Standard Deviation		0.079		0.053		0.040		0.047
Skewness		2.601		0.476		1.249		2.006
Kurtosis		12.156		4.195		4.523		8.575

Notes: The first four moments. In all cases, we cannot reject the null hypothesis that the pairwise difference in values are zero.

Table 5. Kolmogorov-Smirnov P-Values

Year	SIC	1963	1967	1972	1977	1982	1987
1967		.814					
1972		.324	.816				
1977		.462	.788	.508			
1982		.791	.997	.886	.900		
1987		.557	.915	.571	.841	.960	
1992		.183	.923	.656	.371	.777	.960
Year	NAICS	1997		2002		2007	
2002		.996					
2007		.508		.844			
2012		.453		.683		.559	

Notes: The p-values from a Kolmogorov-Smirnov test on each pair of *EG* year-distributions.

5.3. *Robustness of Results*

To ensure that our results are robust, we consider five issues. The first issue is that compared to the other years, 1963 has many observations for which we use the parent code fill. But, as table 4 shows, the skewness and the kurtosis of the 1967–1992 distributions are not statistically different from the 1963 distribution. Nor does the KS test reject that the 1963 distribution is statistically different from any other year. Further, our analysis of the NAICS years comes to the same conclusion that localisation is not changing systemically. Thus, our results hold when 1963 is removed from consideration.

The second issue is that we used the industry classification from each year. The industry classification scheme, however, changes over time. Sometimes the changes in classification are minor. SIC 2034, 2043, 2045, and 2047 are examples where the definition undergoes a minor revision to be more precise. It is unlikely that there was a radical change in plant counts or structure that would alter the *EG* stat for this industry because of the classification revision. Sometimes the changes in classification are more substantial. For example, roasted coffee was SIC 2096 in 1963 but 2095 in 1967–1992. Worse, roasted coffee replaced Shortening and cooking oil as SIC 2095. Thus to see changes in localisation within this industry one needs to use 2096 in 1963 but then 2095 for all other years. Table 6 shows the minor and major changes in industry classification over all available periods. (While some NAICS codes are created, and others are joined together, for example 311222 and 311223 merged to

Table 6. Four Digit SIC-20 Definition Changes 1963–1992

Code	1963	1967	1972–1982	1987	1992
2013	Meat processing plants	Sausages and Other Prepared Meats	Sausages and Other Prepared Meats	Sausages and Other Prepared Meats	Sausages and Other Prepared Meats
2034	Dehydrated food products	Dehydrated food products	Dehydrated fruits, vegetables, and soups	Dehydrated fruits, vegetables, and soups	Dehydrated fruits, vegetables, and soups
2041	Flour mills	Flour and other grain mill products	Flour and other grain mill products	Flour and other grain mill products	Flour and other grain mill products
2043	Cereal preparations	Cereal preparations	Cereal breakfast foods	Cereal breakfast foods	Cereal breakfast foods
2045	Blended and prepared flour	Blended and prepared flour	Blended and prepared flour	Prepared flour mixes and doughs	Prepared flour mixes and doughs
2047			Dog, cat, and other pet food	Dog, cat, and other pet food	Dog and cat food
2052	Biscuit, cookies and crackers	Cookies and crackers	Cookies and crackers	Cookies and crackers	Cookies and crackers
2084	Wines and brandy	Wines, brandy, and brandy spirits	Wines, brandy, and brandy spirits	Wines, brandy, and brandy spirits	Wines, brandy, and brandy spirits
2087	Flavorings	Flavoring extracts, syrups nec	Flavoring extracts, syrups nec	Flavoring extracts, syrups nec	Flavoring extracts, syrups nec
2091	Cottonseed oil mills	Cottonseed oil mills	Canned and cured seafood	Canned and cured seafood	Canned and cured seafood
2092	Soybean oil mills	Soybean oil mills	Fresh or frozen packaged fish	Fresh or frozen packaged fish	Fresh or frozen prepared fish
2095	Shortening and cooking oils	Roasted coffee	Roasted coffee	Roasted coffee	Roasted coffee
2096	Roasted coffee	Shortening and cooking oils		Potato chips and similar snacks	Potato chips and similar snacks

become 311224, no code changed definition.)

Despite the different classification definitions, it does not seem likely that the classification changes are driving our result. Nonetheless, we applied the KS test to the subset of industries that have remained constant over time. Thus these industries constitute a balanced panel. The results may be found in table 7. As can be seen, the *smallest* p-value among the pairwise tests (amongst both SIC and NAICS years) is 0.388, indicating the distributions using just the subset of industries that were not reclassified are not statistically significantly different from each other.

Table 7. Kolmogorov-Smirnov P-Values Same-Definition *EG* Values

Year	SIC	1963	1967	1972	1977	1982	1987
1967		.773					
1972		.773	.944				
1977		.773	.773	.944			
1982		.998	.773	.773	.998		
1987		.651	.587	.388	.651	.858	
1992		.773	.998	.944	.944	.944	.982
Year	NAICS	1997		2002		2007	
2002		.894					
2007		.528		.723			
2012		.528		.723		.528	

Notes: The p-values from a Kolmogorov-Smirnov test on each pair of *EG* year-distributions for codes that did not change definitions.

Table 8. Kolmogorov-Smirnov P-Values for Industries with Fewest Estimated Observations

Year	SIC	1963	1967	1972	1977	1982	1987
1967		.678					
1972		.938	.678				
1977		.938	.678	.938			
1982		.938	.938	.938	.938		
1987		.938	.678	.678	.938	.938	
1992		.678	.938	.678	.938	.999	.999
Year	NAICS				1997	2002	2007
2002		.983					
2007		.983		.983			
2012		.983		.983		.983	

Notes: The p-values from a Kolmogorov-Smirnov test on each pair distributions for the 15 SIC and 20 NAICS industries with the fewest estimated observations.

Table 9. Kolmogorov-Smirnov P-Values for Distribution of Plant-Herfindahls by Period

Year	SIC	1963	1967	1972	1977	1982	1987
1967		.999					
1972		.563	.909				
1977		.553	.961	.999			
1982		.436	.668	.746	.973		
1987		.429	.601	.897	.818	.986	
1992		.248	.615	.897	.818	.963	.853
Year	NAICS	1997		2002		2007	
2002		.956					
2007		.956		.996			
2012		.473		.333		.870	

Notes: The p-values from a Kolmogorov-Smirnov test on each pair of HHI year-distributions.

The third issue we consider is to assess if the churn in the EG values seen in tables 2 and 3 are due to our estimation procedure rather than the true underlying data. To see this is not the case, we repeat our analysis using only the 15 SIC and 20 NAICS industries that needed the fewest estimates. The rows in tables 2 and 3 marked by a "†" indicate those industries. As can be seen, the results still show churn in levels of localisation, but no systemic pattern. We also compare the sectoral results from the industry distributions. The KS test p-values in table 8 may be compared to those in table 5. As can be seen, our results from this subsample of industries using the fewest estimated observations is qualitatively the same as for the full sample. Thus it does not appear that our results are being driven by the industries whose data relied most heavily on our estimation procedure. See the appendix for details of the checks and balances in our algorithm to ensure the accuracy of our data estimates.

The fourth issue is to assess if our results are being driven by changes in industry structure as measured by the distribution of plant-Herfindahls. Because of the westward migration of labour over time, we know the result of localisation stability in the processed food industries cannot be due to the population distribution across U.S. states. However, no pairwise combination of Herfindahl distributions resulted in a KS test p-value below the 0.05 threshold. Thus there is constancy in industrial organisation as well as industrial localisation over time.

The fifth issue is about which benchmark we use to compare industrial employment. The most commonly used measures of localisation such as the location quotient, the locational Gini coefficient, and the *EG* index are relative measures. That is, industrial localisation is assessed by comparing the geographic distribution of industry employment to the geographic distribution of manufacturing employment. Since we are interested in the localisation of industries within the processed food sector, we repeat our test using employment in the processed food sector as our benchmark. We have performed all tests described in this paper using processed food manufacturing employment as the baseline geography rather than total manufacturing employment. All pairwise tests are statistically insignificant for the NAICS years. All pairwise tests for the SIC years are statistically insignificant, unless they involve the year 1963. Given that 1963 is our lowest quality dataset with many missing values, we suspect data quality is likely the underlying reason for this result. The smallest p-value of any Kolmorogov-Smirnov test, outside of those involving 1963, is 0.358.

6. Conclusion

We describe a procedure to estimate the nondisclosed and obfuscated observations of state-industry employment provided by the public releases of the Census Bureau's *Census of Manufactures.* In an online appendix, we provide the code as well as the data for the repeated cross-section of state-industry employment from 1963–2012. We hope the data or this procedure can be used by other researchers to estimate state-industry employment for other data sources and for their own applications.

To show the usefulness and accuracy of our estimation procedure, we analyse our estimated data to consider if industrial localisation for the processed food and kindred products industry has changed in the last fifty years in the United States. Though it is well-known that industrial localisation is not rare, it remains an open question if localisation is changing over time at the industrial and sectoral level. We focus on the processed food industries because it is one of the largest manufacturing sectors by employment and it pulls one of its major inputs from the land, which is fixed geographically.

Though we find that levels of localisation are changing significantly within industries, we do not find a pattern overall in that some industries have levels of localisation that are increasing, some have levels of localisation that are decreasing, some have levels of localisation that remain fairly constant, and others have levels of localisation that bounce around. Additionally, we do not find that the distribution of levels of localisation are changing within the sector. Our finding of little to no change in the distribution of localisation over time contrasts with evidence from the European Union and China that indicates industrial localisation is increasing and evidence from Kim (1995) that levels of localisation are decreasing in the United States.

References

Billings, Stephen B., and Erik B. Johnson. 2016. "Agglomeration within an Urban Area." *Journal of Urban Economics* 91 (1): 13–25.

Brakman, Steven, Harry Garretsen, and Zhao Zhao. Forthcoming. "Spatial Concentration of Manufacturing Firms in China." *Papers in Regional Science* DOI:10.1111/pirs.12195.

Cassey, Andrew J., and Ben O. Smith. 2014. "Simulating Confidence for the Ellison-Glaeser Index." *Journal of Urban Economics* 81 (1): 85–193.

Census. Various years. *Census of Manufactures.* Washington, DC: Bureau of the Census.

Duranton, Gilles, and Henry G. Overman. 2005. “Testing for Localisation Using Micro-Geographic Data.” *Review of Economic Studies* 72 (4): 1077–1106.

Ellison, Glenn, and Edward L. Glaeser. 1997. “Geographic Concentration in U.S. Manufacturing Industries: A Dartboard Approach.” *Journal of Political Economy* 105 (5): 889–927.

Holmes, Thomas J., and John J. Stevens. 2004. “Spatial Distribution of Economic Activities in North America.” In *Handbook of Urban and Regional Economics*, 2797–2843. Amsterdam, Netherlands.

Hoover, Edgar M. 1936. “The Measurement of Industrial Organization.” *The Review of Economics and Statistics* 18 (4): 162–171.

Kim, Sukkoo. 1995. “Expansion of Markets and the Geographic Distribution of Economic Activities: The Trends in U.S. Regional Manufacturing Structure, 1860–1987.” *Quarterly Journal of Economics* 110 (4): 881–908.

Kim, Sukkoo, and Robert A. Margo. 2004. “Historical Perspectives on U.S. Economic Geography.” In *Handbook of Urban and Regional Economics*, edited by J. Vernon Henderson and Jacque-François Thisse, Vol. 4, Chap. 66, 2981–3019. Amsterdam: North Holland.

Krenz, Astrid. 2012. “Industrial Localization and Countries’ Specialization in the European Union.” SSRN `http://dx.doi.org/10.2139/ssrn.1645897`.

Schmalensee, Richard. 1977. “Using the H-index of Concentration With Published Data.” *Review of Economics and Statistics* 59 (2): 186–193.

Silverman, Bernard W. 1986. *Density Estimation for Statistics and Data Analysis*. London.

Appendix A. Robustness of Estimation Procedure

There are two kinds of potential errors in the estimates from our method: miskeyed entries and incorrect estimates due to other entries almost entirely derived from other estimated data. To prevent these errors, we have programmed checks and balances that when in violation, output an error to the program log. For instance, if a four digit code exists, then the three digit parent must also exist. Similarly, if national or state employment is non-numerical, or differs by more than 10%, the program outputs to a log. If national employment is distributed in a way that is counter to the assumptions in Schmalensee (1977), for example that the slope of plant distribution changes direction, then this occurrence is output to the log. We also have programmed checks for if there is a state with zero reported employment and if the algorithm has to climb more than one level to find weights. Additionally, we use the repeated cross-section aspect of the data and our estimates to detect errors. We can compare our results across time for a given code five years before and ahead. Our procedure outputs an error to the log if there is a period-to-period change of an order of magnitude for any of the following: EG, geographic raw concentration G, number of plants, plant Herfindahl H, total state employment, state with the largest employment, total national employment, and national bin with the largest employment.

Every entry from the structural and time-series logs is further investigated manually. We examine the published *Census of Manufacturing* data to confirm that our keyed data are accurate. Additionally, we check each parent code of the logged error to confirm that a error is not cascading down the tree structure of the data. In the case of a time-series log entry, we check both years in question. Once an entry error is located, the procedure is run again, regenerating the entire dataset. In addition, we check to make sure outlier industries are not unduly affecting our state-industry estimates.

We would note that, if anything, our procedure would tend to decrease the industry level volatility. The Census Bureau groups industries together that are similar in production methods. For instance, all SIC-4 codes under a 3-digit parent are similar. As noted in the main body of the paper, our procedure uses these aggregate values as weights when no other information are available. As aggregate localisation is constant over time (as seen at the bottoms of tables 2 and 3), aggregate weights for each state will trend towards the average, which is nearly constant over time. For example, if there were to be an industry with no data at the 4-digit level, our procedure would estimate the 4-digit data by proportionally matching the parent code's employment at the state and national level. Therefore, given no change in localisation at the aggregate level, this industry would show no change in localisation at the industry level.

Despite this tendency, while we show no overall change in localisation, our estimated data are volatile at the industry level. A little less than half of the tested industries have changed more than what would be expected by chance from one time period to the next. Our results suggest that localisation is constant at the aggregate level due to changes at the industry level that offset each other – not constant levels of localisation at the industry level.