DISCUSSION PAPER SERIES

DP14352

IMPUTING MISSING VALUES IN THE US CENSUS BUREAU'S COUNTY BUSINESS PATTERNS

Peter Schott, Natalie Yang, Fabian Eckert and Teresa Fort

INTERNATIONAL TRADE AND REGIONAL ECONOMICS

LABOUR ECONOMICS

MACROECONOMICS AND GROWTH

CE PR

IMPUTING MISSING VALUES IN THE US CENSUS BUREAU'S COUNTY BUSINESS PATTERNS

Peter Schott, Natalie Yang, Fabian Eckert and Teresa Fort

This Discussion Paper is issued under the auspices of the Centre's research programmes:

- International Trade and Regional Economics
- Labour Economics
- Macroeconomics and Growth

IMPUTING MISSING VALUES IN THE US CENSUS BUREAU'S COUNTY BUSINESS PATTERNS

Abstract

The County Business Patterns data published by the US Census Bureau track employment by county and industry from 1946 to the present. Two features of the data limit their usefulness to researchers in practice: (1) employment for the majority of county-industry cells is suppressed to protect confidentiality, and (2) industry classifications change over time. We address both issues. First, we develop a linear programming method that exploits the large set of adding-up constraints implicit in the hierarchical arrangement of the data to impute missing employment. Second, we provide concordances to map all data to a consistent set of industry codes.

Peter Schott - peter.schott@yale.edu
Yale School of Management and CEPR

Natalie Yang - natalie.yang@chicagobooth.edu
University of Chicago, Booth School of Business

Fabian Eckert - fabian.eckert@princeton.edu
Princeton University

Teresa Fort - teresa.fort@dartmouth.edu
Tuck School at Dartmouth

Imputing Missing Values in the US Census Bureau's County Business Patterns*

Fabian Eckert[†]   Teresa C. Fort[‡]   Peter K. Schott[§]   Natalie J. Yang[¶]

First Draft: August 2018
This Draft: January 2020

Abstract

The County Business Patterns data published by the US Census Bureau track employment by county and industry from 1946 to the present. Two features of the data limit their usefulness to researchers in practice: (1) employment for the majority of county-industry cells is suppressed to protect confidentiality, and (2) industry classifications change over time. We address both issues. First, we develop a linear programming method that exploits the large set of adding-up constraints implicit in the hierarchical arrangement of the data to impute missing employment. Second, we provide concordances to map all data to a consistent set of industry codes.

[†]Princeton University; fabian.eckert@princeton.edu
[‡]Tuck School at Dartmouth, NBER & CEPR; teresa.fort@dartmouth.edu
[§]Yale School of Management & NBER; peter.schott@yale.edu.
[¶]University of Chicago, Booth School of Business; natalie.yang@chicagobooth.edu

1. INTRODUCTION

The US Census Bureau's County Business Patterns (CBP) files offer the most detailed view of the United States' industrial structure available to the public. They contain administrative data on employment, payroll, and establishment counts for approximately 1,000 industries by county and year annually since 1964, and sporadically between 1946 and 1962. Many economic studies have used the CBP data to examine a wide range of topics, and recent interest in the spatial distribution of economic activity within countries promises to increase their popularity among researchers in international trade, urban, and macroeconomics.[1]

Despite their benefits, the CBP files have two important limitations. First, employment for many county-industry cells is suppressed to preserve confidentiality, substantially reducing the number of counties and industries that can be examined either in the cross section or over time. Second, industry classifications change periodically, hampering researchers' ability to construct panels. This paper addresses both shortcomings. We develop and implement a new method to impute information for suppressed cells and offer industry concordances to map the resulting employment counts to a consistent set of industry codes. We currently provide these improvements for the 1977 to 2016 files, and plan to extend them back in time to 1946 in future drafts.

Each edition of the CBP contains three files which record national, state, and county employment by industry. We refer to the geography-by-industry bins as "cells." Both dimensions of the data are hierarchical. Counties are small geographic units whose boundaries are always contained within a single US state, and the union of states makes up the country as a whole.[2] For each geographic unit, employment counts are given for industries *and* aggregations of industries, which we refer to as "roots." For example, if an industry code has four digits, e.g., 3571, then that industry has three roots at the $k \in \{1, 2, 3\}$ digit level, i.e., 3xxx, 35xx, and 357x.

Census suppresses the employment counts of a large fraction of cells in each year to prevent users from inferring information about any particular firm. Intuitively, this suppression is more likely the finer the geography or industry of a cell. For example, suppression is more frequent among counties than states (more than half of county-industry cells are suppressed), and among industries than roots. Importantly, when Census does suppress the employment of a cell, it provides a lower and upper bound on its actual employment.

---

[1] For example, Glaeser et al. (1992) test for regional convergence, Autor et al. (2013) study the impact of increased Chinese import penetration on US manufacturing employment, Holmes and Stevens (2004) examine economic specialization, and Hershbein and Kahn (2018) investigate whether recessions accelerate skill-biased technical change. For recent studies using regional data to study the industrial composition of regions see Caliendo et al. (2015), Caliendo et al. (2014), Diamond (2013), Hornbeck and Moretti (2018), Bernard et al. (2013), Eckert (2019), Eckert et al. (2019), Fort et al. (2018), and Ding et al. (2019).

[2] After 1994, the CBP files contain tabulations at the zip code level. We plan to apply our imputation method to this geographic unit in a future draft.

In the first part of the paper, we develop a method for imputing suppressed employment that uses the geographic and industrial hierarchies as constraints in a linear program. The intuition for our method is straightforward. For each year, we pool the data from the national, state, and county files and assign a lower and upper bound to the employment of each cell. For unsuppressed cells, lower and upper bounds coincide. For suppressed cells, the lower and upper bounds differ, and are provided by Census. The key insight of our approach is that a given cell's employment is restricted not just by its own lower and upper bounds, but by the adding-up constraints implied by the employment of all cells along the geographic and industrial hierarchies implicit in the union of the county, state, and national files.[3]

The algorithm chooses employment counts for each cell subject to two sets of constraints. First, n-digit industry counts add up to those of $n-1$ level roots, $n-1$ level roots add up to those of $n-2$ level roots, and so on. Second, within each industry or root, county employment adds to state employment, and state employment adds to national employment. For a given suppressed cell, our baseline linear program identifies the employment count that is closest to the midpoint of its lower and upper bounds, while satisfying these industrial and geographical adding-up constraints.[4] While previous attempts to impute missing information in the CBP have made use of some of these adding-up constraints, ours is the first to provide imputed values that obey *all* these restrictions.[5] Moreover, we show that our method can be used to find and correct inconsistencies in the bounds reported in the raw CBP files, and to identify employment at implied "partial" codes, defined in Section 4.3, during the SIC era.

The CBP use two different industry classification systems. Prior to 1998, employment is classified using the Standard Industry Classification (SIC) system, while employment thereafter is expressed according to the North American Industry Classification System (NAICS). In the second part of the paper, we supplement industry concordances provided by Fort and Klimek (2018) to reclassify the employment totals in the 1977 to 2016 editions of the CBP

---

[3]As an example of restrictions imposed by geographical hierarchy, consider bakery employment in Mercer County, NJ. Together with employment in bakeries in all other NJ counties (from the county file), Mercer's count has to add up to bakery employment for New Jersey overall (from the state file) which in turn, along with bakery employment in all other states, has to yield the national bakery employment count (from the national file). Industry hierarchies impose constraints *within* each geographic units. Bakery employment within Mercer County has to add up to Food Manufacturing employment along with all other food production employment in Mercer, which in turn adds to the total Manufacturing employment count in Mercer (all from the county file).

[4]We use the midpoint in our baseline estimates given its popularity as an imputation method in many prior studies (e.g., Glaeser et al. (1992); Holmes and Stevens (2004)). We plan to discuss the sensitivity of our imputed estimates to this objective function as well as methods for bootstrapping confidence intervals for our imputes in a future draft.

[5]Autor et al. (2013), for example, use industry hierarchies to tighten missing cells' lower and upper bounds and then select point estimates within those bounds subject to industry adding-up constraints. Our method is most closely related to the procedure proposed by Isserman and Westervelt (2006), who use both industry and geography hierarchies to tighten the lower and upper bounds of suppressed cells iteratively, and employ simulated annealing to choose point estimates within those bounds.

to the NAICS 2012 vintage of industry codes. We use NAICS for three reasons. First, this reclassification renders the data compatible with future releases of the CBP. Second, NAICS is comparable with the Canadian and Mexican classification systems at the four-digit level, facilitating analysis of North America as a whole. Third, NAICS assigns all establishments to industries based on what the establishment does, whereas SIC assignments are based on a number of different concepts that vary across sectors and are therefore less transparent.[6] Lastly, we discuss changes in geographic units over time. While state boundaries are constant and their number unchanged, the number of counties and their geographic delineations change over time.

The remainder of this paper is structured as follows. Section 2 provides a detailed description of the CBP data. Section 3 describes our method for imputing missing values and summarizes its results. Sections 5 and 6 discuss industry and county concordances. Section 7 concludes. Raw data sets, imputed employment, industry concordances, and all code used to generate our results are in the data appendix located at www.fpeckert.me/cbp.

2. The County Business Patterns Files

This section describes the County Business Patterns (CBP) files in detail and highlights key features related to our imputation procedure and industry concordances.

CBP files are available for 1946 to 1951, 1953, 1956, 1959, 1962 and annually from 1964 to 2016.[7] Depending on the year, these files record employment during the week of March 12, first quarter and annual payroll by county and industry, and establishment counts.[8]

Beginning in 1975, state and national files with the same structure as the county files are also available. The data in the CBP files are extracted from the US Census Bureau's Business Register (BR), a database constructed from administrative tax records of all private, non-farm employer establishments in the United States. The BR provides the underlying frame for all of the economic data collected by the Census Bureau. Census supplements the BR data with additional information from various other sources, including its Economic Censuses, Annual Surveys, Current Business Surveys, and Company Organization Surveys, as well as data from other agencies including the Bureau of Labor Statistics and the Social

---

[6]Employment at even the most broadly defined sectoral aggregates differs substantially between SIC and NAICS, so it is critical to address the change in systems for any time-series analysis. For example, 9 percent of manufacturing employment under the SIC system was reclassified outside of manufacturing under NAICS (Fort and Klimek (2018)).

[7]The sources of the files we use are detailed in Appendix B. The US Census Bureau's CBP website posts CBP files from 1986 to the 2016 (the most recent file available). CBP files for earlier years are available at the National Archives.

[8]Excluded sectors are: crop and animal production; rail transportation; the National Post Service; pension, health, welfare and vacation funds; trusts, estates and agency accounts; private households; and public administration. Table 2 shows the total number of industries contained in the data for each year.

Security Administration.[9]

2.1 Unit of Analysis

Since 1974, establishments are the fundamental reporting unit underlying the CBP tabulations. Employment is assigned to industries and locations by aggregating the employment reported at establishments with the same industry and location code. Before 1974, multi-location employers outside manufacturing were permitted to aggregate their employment to a single location, while multi-location employers within manufacturing reported employment according to individual establishments' locations (see United States Census Bureau (1986)). As a result of this change, there is an unbridgeable break in the reporting of non-manufacturing employment in 1974.

The CBP files also report the distribution of establishments across the 12 establishment size bins noted in Table 1 within each geographical unit and industry.[10] Summing establishments across size bins within a location yields the total number of establishments separately reported for each location-industry cell. Unlike employment counts, establishment counts are never suppressed.

2.2 Industry Classification

The CBP employs two different industry classification systems: the Standard Industrial Classification (SIC) codes through 1997 inclusive, and the North American Industry Classification System (NAICS) codes after 1997. The most detailed industry codes available under SIC are four-digit codes. The most detailed industry codes under NAICS are six-digit codes. Table 2 shows that the number of four-digit SIC codes varies between 900 and 1,000 over time. NAICS offers more detail with around 1,200 industries, with some fluctuation in total codes also evident in Table 2. The two systems are structured similarly in that for each system, there is a set of "economic divisions," such as manufacturing, that comprise all activity under broad categories. Under SIC and NAICS, there are 10 and 20 divisions, respectively.

In the CBP data, economic divisions are denoted by two-digit numbers followed by a

---

[9]Census applies various automated and analytical edits to the raw data to remove anomalies and to validate geographic coding, addresses, and industry classifications. For further information on these edits, see Census' technical documentation of the CBP program at https://www.census.gov/programs-surveys/cbp/technical-documentation.html, copies of which are available in our data appendix at www.fpeckert.me/cbp.

[10]Prior to 1983, establishment counts are based on whether establishments are active in the fourth quarter of the year. Starting in that year, establishments are counted if they are active at any point in the year. Before 1974 the largest establishment size bracket is 500 employees or more. Starting in that year, it is 5,000 employees or more.

TABLE 1: ESTABLISHMENT SIZE BINS IN THE CBP FILES

| | Employment | |
Flag	Min	Max
1	1	4
2	5	9
3	10	19
4	20	49
5	50	99
6	100	249
7	250	499
8	500	999
9	1,000	1,499
10	1,500	2,499
11	2,500	4,999
12	5,000	or More

Source: Supplementary Materials to the CBP files froms the US Census Bureau County Business Patterns website.

string of hyphens.[11] For example, the division code for the manufacturing sector is 20-- under the 1987 SIC and 31---- under the 1997 NAICS. Subtotals for aggregates between these economic division totals and the most detailed industries available are also reported using codes that are a combination of the sector roots, non-zero digits, and either trailing zeros (SIC) or forward slashes (NAICS). For example, in 1987 SIC, under the manufacturing economic division code 20--, the code 2000 represents the subtotal for "food and kindred products," and under that, the code 2010 represents the subsubtotal for "meat products," and under that, the code 2011 covers the "meat packing plants" industry.

Similarly, in 1997 NAICS, under the manufacturing division code 31----, the code 311/// is the subtotal for "food manufacturing," the code 3112// is the sub-subtotal for "grain and oilseed milling," the code 31121/ is the sub-sub-subtotal for "flour milling and malt manufacturing," and, at the lowest aggregation level, the code 311211 is for "flour milling." Note that in the CBP data, industry codes do not always have the same first two digits as

---

[11]Some of the raw SIC CBP files contain a division code '19--' in lieu of '20--', e.g., for 1986. Communication with employees at the County Business Pattern unit at the Census Bureau confirmed this as a mistake. In our imputed data files and concordances we replaced all occurrences of '19--' with '20--.'

Industry Classification System	Years Active	Digits of Most Detailed Industry	Number of Industries
1957 SIC	1962 − 1967	2-digit	926
1967 SIC	1968 − 1973	4-digit	921
1972 SIC	1974 − 1976	4-digit	1,003
1977 SIC	1977 − 1987	4-digit	1,004
1987 SIC	1988 − 1997	4-digit	1,006
1997 NAICS	1998 − 2002	6-digit	1,169
2002 NAICS	2003 − 2007	6-digit	1,179
2007 NAICS	2008 − 2011	6-digit	1,175
2012 NAICS	2012 − 2016	6-digit	1,065

Source: Constructed by the authors by drawing on the County Business Patterns datasets provided by the National Archives prior to 1986 and the US Census Bureau County Business Patterns website thereafter.

their economic division. For example, in 1987 SIC, the economic division total code 20-- covers SIC codes with first two digits ranging from 20 to 39. Similarly, the NAICS division code 31---- in the CBP data contains the total employment for all industry codes that start with 31, 32, or 33.

As indicated in Table 2, SIC was updated every five to ten years, while NAICS is updated every five years.[12] As a result, there are several vintages within each system. Both systems are hierarchical in the sense that their $k \in \{1 : n − 1\}$ digit roots (or sectors) encompass all n-digit industries sharing those roots.

We introduce the terms "parent" and "child" to clarify our discussion of CBP industrial and geographical hierarchies. The parent of a given industry code (or root) contains one less digit than the industry code (or root) that it nests. For example, NAICS 115--- is the parent of all industry codes and four-digit roots that begin with 115, which, collectively are the children of 115---. This terminology also extends to the geographical hierarchy. For example, for a given geography, a state is the parent of the counties it contains, i.e., its children counties.

---

[12]Technically, SIC was also updated every five years. However, in some years, changes were negligible enough that the updated codes were considered a "supplement" to the previous vinage and not a separate new vintage.

TABLE 3: DATA SUPPRESSION FLAGS

	Employment	
Flag	Min	Max
A	0	19
B	20	99
C	100	249
E	250	499
F	500	999
G	1,000	2,499
H	2,500	4,999
I	5,000	9,999
J	10,000	24,999
K	25,000	49,999
L	50,000	99,999
M	100,000	or More

Source: Supplementary Materials to the CBP files froms the US Census Bureau County Business Patterns website.

A typical geographic unit has many cells, some at the industry level, and more at the root level. Since the root-level employment is also captured by the industry-level employment, summing employment across all cells in the raw data for a given geographic unit generally produces an employment total that is higher than the true employment for that cell. Aggregation without double counting under NAICS is straightforward since *all* employment in a given root is accounted for by its industrial children, i.e., the roots or industries nested directly below it. Thus, to obtain total employment within a given geographical unit, one can sum employment across all roots of a given order, or across all industries.

Aggregation is more complicated for SIC years because a county may have reported employment for a root but not have cells for all/any industries within that root. In other words, the employment count of a given root is weakly larger than the employment count obtained from summing employment across its industrial children. We introduce "partial" codes—discussed further in Section 4.3—to capture the difference in employment between a root and the reported industrial children under that root. Without partial codes, the only way to produce an accurate national total from the state or county files under SIC is to sum across division codes or 2-digit roots. Census always reports all employment for these codes, even if it does not report how this employment is distributed over more detailed roots or

TABLE 4: NOISE INFUSION FLAGS

Flag	Meaning
G	0 to $< 2\%$ noise (low noise)
H	2 to $< 5\%$ noise (medium noise)
D	Set to 0 to avoid disclosing data for individual companies
S	Set to 0 because estimate did not meet publication standards

Source: Supplementary Materials to the 2007-2017 CBP files from the US Census Bureau County Business Patterns website.

industries.

As related to the authors by CBP employees, this feature of the CBP files in the SIC years was driven by insufficient information. When CBP data were tabulated under the SIC classification system, the programming allowed data to be published at the k-root level. As a result, there are some establishments that are known to be in root k, but whose $k+1$ root or industry cannot be determined by Census. For example, it may be that employment counts are stated for SIC 0760 (a 3-digit root), but that not all of it can be assigned to the more detailed corresponding industry codes 0761 and 0762. As a result, the employment tabulated for 0761 and 0762 would not add up to employment tabulated for the root 0760.

2.3 Suppression

By law (US Code, Title 13, Section 9), the US Census Bureau cannot publicly release any data that would disclose the operations of an individual firm. Prior to 2007, Census satisfied this restriction by reporting 0 employment along with an "employment suppression flag" for a subset of cells. These flags, listed in Table 3, indicate which of 12 mutually exclusive ranges contain a cell's true level of employment.[13] In 2007, Census additionally introduced "noise infusion" for unsuppressed cells. This technique uses a random noise multiplier to perturb the true employment of cells that might otherwise be suppressed. In those years, Census reports the perturbed employment and one of two noise infusion flags, listed in Table 4,

---

[13]As there is no upper bound for the final code, "M", we use a conservative upper bound of 100,000,000 in the few instances in which this code appears in 2011, 2013 and 2015. In principle, information on the establishment size distribution within each geography-industry cell (discussed in Section 2.1) can be used to construct an alternative set of upper and lower employment bounds for each suppressed cell. In practice, we find that these imputed bounds are often inconsistent with the lower and upper bounds of employment implied by the employment suppression flags, a feature of the CBP also noted in Autor et al. (2013). As a result, our baseline estimates are based solely on the employment bounds, i.e., we ignore the establishment size distribution.

which indicate the range of the random noise multiplier used to produce it.[14] Suppressed cells also are given a noise suppression flag, but these flags merely indicate why employment for the cell is suppressed, i.e., either to avoid disclosing data for an individual establishment, or because the employment count does not meet (unspecified) publication standards. From 2007 to 2016, the four noise infusion flags G, H, D, and S apply to roughly 60, 10, 12, and 8 percent of observations in the raw county data respectively.

Figure 1 reports the total number of county-industry observations in the 1977 to 2016 county CBP files as well as the number of such cells which are suppressed, i.e., are associated with a flag from Table 3. As indicated in the figure, 55 to 70 percent of cells are suppressed, depending upon the year. Figure 1 also highlights the jump in the number of county-industry cells from roughly 0.9 to 2.1 million that occurs in the transition from SIC to NAICS. This increase has two sources. First, the NAICS classification scheme is more detailed than the SIC, as illustrated in Table 2. Second, Census reports all industry codes under roots during the NAICS years but not during the SIC years, as discussed in Section 2.2. Prior to 2015, cell values with a "high noise" flag (J) were suppressed from publication by replacing the cell value and associated noise flag with an "S." Starting in 2015, employment for these cells starts to be reported, but with a "high noise" flag (J). This change explains the drop in suppressed cells for the years after 2014 seen in Figure 1.

Figure 2 reports the share of suppressed cells by division codes for SIC (left panel) and NAICS (right panel), respectively.

Figure 3 plots the total employment implied by the lower and upper bounds provided in the raw CBP county files against aggregate US employment as recorded in the national files. In computing these lower and upper bounds, we ignore noise infusion after 2007 to focus solely on the gap associated with the suppression flags listed in Table 3. Furthermore, for the SIC years, we ignore the existence of partial codes discussed in Section 2.2 and include only the finest industry or root observation available in each county-industry hierarchy.[15] As illustrated in the figure, the gap between upper and lower bounds ranges between 20 and 40 million workers in most years. The gaps are particularly large in 2011, 2013 and 2015 due to presence of "M" suppression flags in those years. As noted above, we conservatively set the upper bound for those cells to 100 million.

2.4 Exploiting Hierarchies

Table 5 exhibits excerpts from the 2010 CBP county, state, and national files. The top panel displays a subset of the observations for Autauga County, Alabama, while the middle

---

[14]For a more detailed discussion of noise infusion, see https://www.census.gov/programs-surveys/susb/technical-documentation/methodology.html. A copy of this documentation is provided in our data appendix.

[15]That is, if a given county has an industry reported, we use the employment reported for that industry, but if employment is reported only for roots, we use the most disaggregated roots available for each industry hierarchy.

TABLE 5: EXCERPT FROM COUNTY BUSINESS PATTERNS, 2010

Index	File	State	County	NAICS	Employment	Employment Flag	Noise Flag
C1	County	Alabama	Autauga	//////	10,167		H
C2	County	Alabama	Autauga	11----	33		G
C3	County	Alabama	Autauga	113///	27		G
C4	County	Alabama	Autauga	1133//	27		G
C5	County	Alabama	Autauga	11331/	27		G
C6	County	Alabama	Autauga	113310	27		G
C7	County	Alabama	Autauga	115///	0	A	D
C8	County	Alabama	Autauga	1151//	0	A	D
C9	County	Alabama	Autauga	11511/	0	A	D
C10	County	Alabama	Autauga	115112	0	A	D
C11	County	Alabama	Autauga	21////	34		H
S1	State	Alabama		//////	1,568,111		G
S2	State	Alabama		11----	5,984		G
S3	State	Alabama		113///	4,364		G
S4	State	Alabama		1131//	225		G
S5	State	Alabama		11311/	225		G
S6	State	Alabama		113110	225		G
N1	National			//////	111,970,095		G
N2	National			11----	156,055		G
N3	National			113///	53,525		G
N4	National			1131//	2,059		G
N5	National			11311/	2,059		G
N6	National			113110	2,059		G

Source: Extracted by the authors from the 2010 CBP county-, state, and national-level files from the US Census Bureau County Business Patterns website.
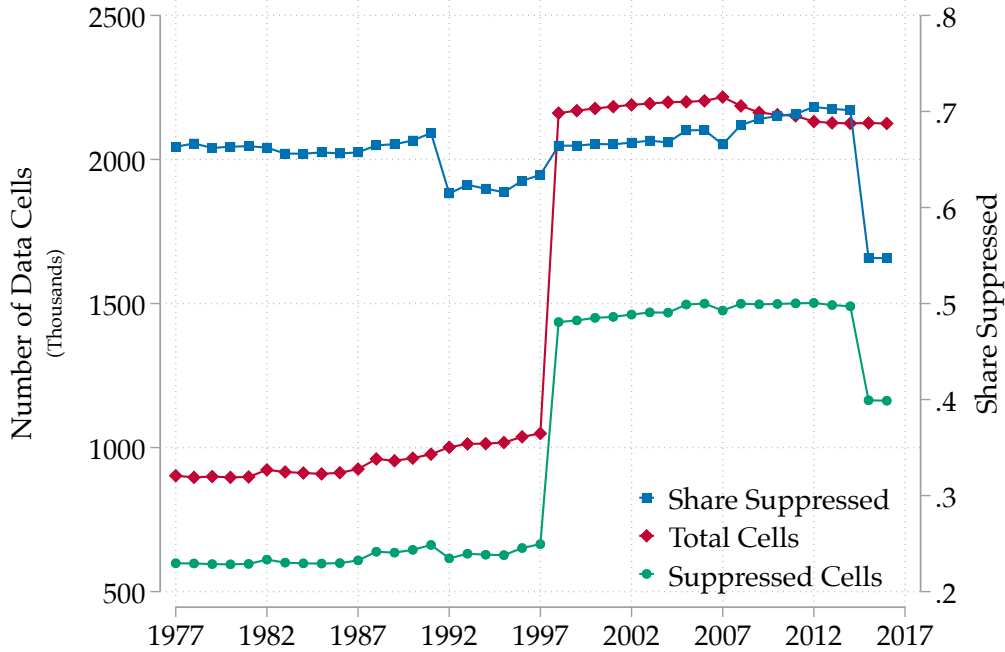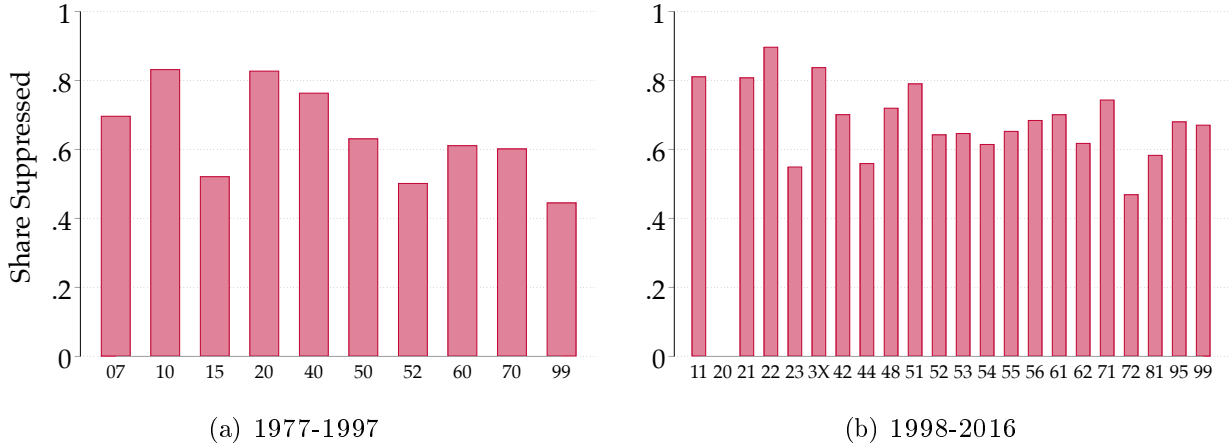
Source: 1977 to 2017 CBP files and authors' calculations. Figure displays the number of cells in the county files in each year, the number of those cells that are suppressed, and the share of cells that are suppressed. Suppressed cell counts do not include the addition of noise infusion to all cells starting in 2007. Industry classification switches from SIC to NAICS in 1998.

and lower panels report information for Alabama as a whole and the United States as a whole, respectively. The first column in each panel indexes the cells. Columns 2 through 5 identify the geographic and industry unit of observation. Column 6 reports the employment total and, if relevant, Column 7 reports the employment suppression flag. The final column reports the noise suppression flag.

In cell C6 of the top panel, the six-digit NAICS industry 113310 is the only "child" of the NAICS 5-digit root 11331/ in cell C5. As a result, its employment must add to that of cell C5. Likewise for cells C4 and C3. Employment in cell C2, however, is larger than that for cell C3 because Autauga County has employment in root 115/// as well as 113///. Cell C1 gives the total employment for Autauga County. This employment must be the sum of employment across all two-digit roots, i.e, C2, C11, and a number of other cells which are not included in this excerpt. Likewise, US employment in two-digit root 11----, in cell N2 of the bottom panel, must be the sum of employment in that root across states, including cell S2 for Alabama. The number of children a parent has depends on the breadth of the geographic unit's underlying industrial structure. For example, counties with a wider range of activities have subtotals for employment reported for a longer list of two-digit roots.

FIGURE 2: SUPPRESSION BY TWO-DIGIT ROOTS

(a) 1977-1997   (b) 1998-2016

Source: 1977 to 2017 CBP files and authors' calculations. Figure displays the share of the cells that are suppressed, by SIC division code for 1997 to 1997 (right panel) and two-digit NAICS sector (right panel). Suppressed cell counts do not include those subject to noise infusion in 2007. NAICS sector 3X contains 31, 32 and 33; NAICS sector 44 cotains 44 and 45; and NAICS sector 48 contains 48 and 49. Industry classification switches from SIC to NAICS in 1998.

Table 5 provides intuition for how industry and geography hierarchies can be exploited to impute employment for suppressed cells. Cell C7 in the top panel of the table is suppressed, indicated by its employment of 0 and the existence of a suppression flag, "A." As shown in Table 3, the "A" flag signifies that Autauga employs between 1 and 19 workers in three-digit NAICS root 115///. Unsurprisingly, the industry children of cell C7 (cells C8, C9, and C10), including six-digit industry 115112, are also suppressed, since their employment must be weakly less than that of cell C7. The middle panel of Table 5 shows that while Alabama's employment in the roots of 115212 is not suppressed, its employment in that industry is also suppressed, with the same suppression code. National employment in both these roots and the industry are not suppressed.

Ignoring noise infusion, employment counts for cells C3 and C7 must add up to cell C2, their common industrial parent: i.e., employment in cell C7 must be 6 (33 less 27), which does indeed fall within the lower and upper bounds of the 1 to 19 suppression flag for this cell. Accounting for noise infusion complicates this example, as the totals in cells C2 and C3 have to be replaced with lower and upper bounds (of 32.3 to 33.7 and 26.4 to 27.6) that incorporate the noise.[16] Even in that case, however, it is clear that information in cells C2

---

[16]The noise suppression flag for these observations (G) indicates noise of up to 2 percent (see Table 4). The upper and lower bounds implied by this noise are computed by multiplying the reported employments by 1/1.02 and 1/0.98. As discussed in Section 3 below, we ignore noise infusion and just use the reported bounds in our imputation procedure.

FIGURE 3: EMPLOYMENT IN COUNTY VS NATIONAL FILES
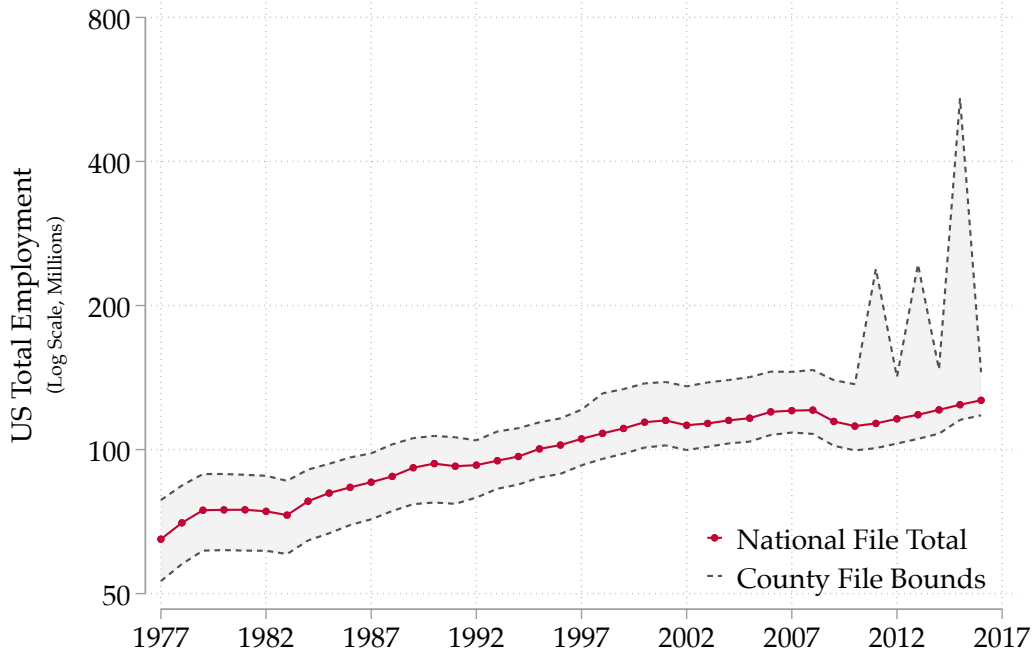
Source: 1977 to 2017 CBP files and authors' calculations. Figure displays the sum of employment across suppressed and unsuppressed lower and upper bounds in the county file against the aggregate US employment contained in the national files. SIC years include only the most disaggregate observation available in each county-industry hierarchy. Industry classification switches from SIC to NAICS in 1998.

and C3 can be used to help pin down employment in cell C7.[17] That is, the industry and geography hierarchical constraints can narrow bounds even in instances where we cannot infer the exact employment count.

Table 6 provides an excerpt from the SIC era, 1990. The top panel reveals that Autauga has employment in roots 1510 and 1530, the latter of which is suppressed with flag "B."[18] Together, the employment for these cells must sum to the employment noted for 1500. The employment for 15-- is substantially greater than that for 1500, however, because the former

---

[17]Autor et al. (2013) use industry hierarchies to tighten missing cells' lower and upper bounds before selecting point estimates within the tightened bounds. Here, for example, the 0 to 19 bounds implied by the employment suppression flag can be narrowed to 4.7 (=32.3-27.6) to 7.3 (=33.7-26.4). To our knowledge, all previous attempts to use industry and geography adding-up constraints employ such bound tightening iteratively. Isserman and Westervelt (2006), for example, propose cycling between industry and geography bound tightening before using simulated annealing to pick point estimates between any remaining non-convergent lower and upper bounds. We do not follow this approach because it is not clear one could achieve convergence even if sufficient computing power were not an issue (for us, it is). Thus, even with bound tightening, the solution to a linear program such as the one proposed here would be necessary.

[18]As noted in Section 2.2, SIC codes ending in one or more zeros are roots.

TABLE 6: EXCERPT FROM COUNTY BUSINESS PATTERNS, 1990

Index	File	State	County	SIC	Employment	Flag
C1	County	Alabama	Autauga	----	6,639	
C2	County	Alabama	Autauga	15--	381	
C3	County	Alabama	Autauga	1500	123	
C4	County	Alabama	Autauga	1510	68	
C5	County	Alabama	Autauga	1530	0	B
S1	State	Alabama		----	1,342,993	
S2	State	Alabama		15--	100,301	
S3	State	Alabama		1500	36,279	
S4	State	Alabama		1510	29,074	
S5	State	Alabama		1530	1,620	
N1	National			----	93,476,087	
N2	National			15--	5,239,067	
N3	National			1500	1,352,043	
N4	National			1510	899,616	
N5	National			1530	120,522	

Source: Extracted by the authors from the 2010 CBP county-, state-, and national-level files from the US Census Bureau County Business Patterns website.

also includes employment in industries with two-digit roots 16 and 17 (not shown).

2.5 Inconsistent Codes

For each year, our imputation procedure uses the county, state, and national CBP files simultaneously. Comparing these files reveals two types of inconsistencies in industry codes. The first type of inconsistency arises from industry codes that do not appear in all files for a given year. For example, SIC code "8631" appears in the county file for 1980, but is not present in either the state or national files for that year. The second inconsistency stems from codes that appear in the data files but that do not appear in the reference list of SIC or NAICS codes included in the CBP documentation for that year.

In implementing our imputation procedure, we treat these inconsistencies as follows. First, we drop industry codes that do not appear in all three files in a given year. Second, we drop all codes that do not appear in the industry reference file of a given year, but only

for codes that have no further industrial children or siblings. Table A.3 in Section C.1 of the Appendix lists codes dropped in each year, noting which type of inconsistency they represent. Because the employment in these "inconsistent codes" is accounted for in their root codes, dropping them does not result in lost employment.

3. Imputing Suppressed Cells as Linear Program

The hierarchical nature of the CBP data, combined with the suppression flags provided by Census, offer three sets of linear constraints that can be used to impute missing values. The precise formulation of these constraints differs under the NAICS and SIC classification systems. We discuss each in turn.

3.1 Linear Program for the NAICS Era

Let $\mathcal{I}$ be the set of all NAICS industry codes or roots appearing in the CBP files in a given year, and index individual industries by i, i'. Let $\mathcal{G}$ be the set of all geography codes appearing in the current CBP data set and index individual geographies by g, g'. Let $\{x_{i,g}\}_{i \in \mathcal{I}, g \in \mathcal{G}}$ be the employment in industry i and geography g, the free variable. Also denote by $i \prec i'$ the set of industries that are industrial children of industry i'. For example, drawing on Table 5, if $i' = 113///$ then $i \prec i' = \{1131//, 1132//\}$. Similarly, denote by $g \prec g'$ the set of geographies that are geographical children of geography g'. For example, in Table 5, if $g' = \{Alabama\}$ then $g \prec g' = \{Autauga, \dots\}$, where "$\dots$" is a placeholder for all the other counties in the state of Alabama besides Autauga. With this notation in hand we can now formalize the problem of inferring suppressed employment as a linear program.

The first constraint on each cell is the lower and upper bounds provided by the suppression flags. The employment count for industry i and geography g, $x_{i,g}$, has to satisfy:

$$(1) \qquad\qquad lb_{i,g} \leq x_{i,g} \leq ub_{i,g}.$$

For years before 2007, $lb_{i,g} = ub_{i,g}$ for unsuppressed cells. Starting in 2007, when noise infusion is introduced, even cells with unsuppressed reported employment have modified bounds, $\bar{lb}_{i,g}$ and $\bar{ub}_{i,g}$, such that $\bar{lb}_{i,g} \equiv lb_{i,g}/(1 + \rho) < \bar{ub}_{i,g} \equiv ub_{i,g}/(1 - \rho)$, where ρ is given by the noise suppression flag, e.g., 2 percent. As Figure 1 shows for the majority of county-industry cells, $lb_{i,g} < ub_{i,g}$ holds even before noise infusion. In practice, our baseline results ignore noise infusion, i.e., we use $lb_{i,g}$ and $ub_{i,g}$ rather than $\bar{lb}_{i,g}$ and $\bar{ub}_{i,g}$.

Second, within each geography, i.e., within each county, within each state, and for the United States as a whole, employment counts of industrial children have to add up to the

employment counts of their industrial parents:

$$(2) \qquad x_{i,g} = \sum_{\underline{i} \prec i} x_{\underline{i},g} \qquad \forall i,g \in \mathcal{I},\mathcal{G}.$$

Third, within each industry or root, employment counts across counties within a state have to add up to the respective state totals, and employment counts across all US states have to add up to the national total within that industry:

$$(3) \qquad x_{i,g} = \sum_{\underline{g} \prec g} x_{i,\underline{g}} \qquad \forall i,g \in \mathcal{I},\mathcal{G}.$$

The constraints in equations 1 to 3 define a *feasible set* of employment count vectors, $\{x_{i,g}\}_{i \in \mathcal{I}, g \in \mathcal{G}}$. Members of this set are consistent with all the information and restrictions implicit in the national, state, and county CBP files taken together.

To select an individual vector $\{x_{i,g}\}_{i \in \mathcal{I}, g \in \mathcal{G}}$ from the feasible set, we need to choose an objective function to minimize over this set. For our baseline estimates, we choose $\{x_{i,g}\}_{i \in \mathcal{I}, g \in \mathcal{G}}$ to be as close as possible to the midpoint between the upper and lower bounds of cell (i, g), conditional to $\{x_{i,g}\}_{i \in \mathcal{I}, g \in \mathcal{G}}$ being in the feasible set defined by equations 1 to 3. We choose this objective function given the large number of studies that uses the midpoint of suppressed cells as the imputed employment value (e.g., Glaeser et al. (1992); Holmes and Stevens (2004)).[19] The resulting linear program can be written as follows:

$$(4) \qquad \min_{\{x_{i,g}\}_{i \in \mathcal{I}, g \in \mathcal{G}}} \sum_{i,g} \mid x_{i,g} - \frac{ub_{i,g} + lb_{i,g}}{2} \mid$$

$$\text{s.t. } lb_{i,g} \leq x_{i,g} \leq ub_{i,g},$$

$$x_{i,g} = \sum_{\underline{i} \prec i} x_{\underline{i},g} \; \forall i,g \in \mathcal{I},\mathcal{G},$$

$$x_{i,g} = \sum_{\underline{g} \prec g} x_{i,\underline{g}} \; \forall i,g \in \mathcal{I},\mathcal{G}.$$

Despite the absolute value operators, it is easy to reformulate the objective function in problem 4 as linear. As a result, it can be implemented as a classic linear programming problem, albeit a very large one as the vector $\{x_{i,g}\}_{i \in \mathcal{I}, g \in \mathcal{G}}$ contains all possible geography-industry combinations, the size of which varies between 5 and 7 million depending upon the year.[20] The scale of the optimization problem necessitates the use of industrial grade optimization software. We use Gurobi on Yale's High-Powered Computing (YHPC) network,

---

[19]In future drafts, we hope to consider three alternative objective functions: (1) distance from lower bound; (2) distance from upper bounds; and (3) distance from a random point within each set of bounds.

[20]Rewriting this objective function as a linear objective doubles this number.

where identifying the minimizing vector for a particular year takes about 20 minutes.

3.2   Modified Linear Program for the SIC Era

As discussed in the previous section, only the first two industry child-parent relationships in equations 1 to 3 hold with equality under the SIC classification. Thus, when formulating the problem for SIC codes, we re-write constraint 2 above as two equations:

$$(5) \qquad x_{i,g} = \sum_{\underline{\mathrm{i}} \prec \mathrm{i}} x_{\underline{\mathrm{i}},g} \ \forall i, g \in \mathcal{I}, \mathcal{G} \text{ s.t. } i \text{ is 1-,2- digit,}$$

$$x_{i,g} \geq \sum_{\underline{\mathrm{i}} \prec \mathrm{i}} x_{\underline{\mathrm{i}},g} \ \forall i, g \in \mathcal{I}, \mathcal{G} \text{ s.t. } i \text{ is 3-,4- digit.}$$

The linear programming problem we solve for the SIC years is then identical to the one in equation 4 above, with constraint 2 replaced by constraint 5.

3.3   Closest Feasible Model Procedure

As noted in the next section, we find that for many of the years before 2001, there is no vector $\{x_{i,g}\}_{i \in \mathcal{I}, g \in \mathcal{G}}$ for which the constraints in 1-3 hold simultaneously. We interpret this lack of a solution as being due to errors in the employment of unsuppressed cells or the employment bounds of the suppressed cells. That is, if $lb_{i,g}$ and $ub_{i,g}$ are the lower and upper bounds of a given cell, then the "true bounds" on this cell can be expressed as

$$lb_{i,g} - a_{i,g} \leq x_{i,g} \leq ub_{i,g} + b_{i,g},$$

where $a_{i,g}$ and $b_{i,g}$ are non-negative constants.[21] $a_{i,g} \geq 0$ and $b_{i,g} \geq 0$ are adjustments to the bounds stated in the CBP files so that the new bounds are the minimal bounds that contain the true $x_{i,g}$. We call this solution, i.e., the data set with bounds adjusted in this way, the *closest feasible model.*

---

[21]We ignore noise infusion in this exposition since it was introduced in 2007, whereas the feasible set is empty only in data prior to 2001.

This extended problem also can be stated as a linear program:

(6)

$$\min_{\{a_{i,g}\}_{i\in\mathcal{I},g\in\mathcal{G}}\geq 0,\{b_{i,g}\}_{i\in\mathcal{I},g\in\mathcal{G}}\geq 0} \sum_{i\in\mathcal{I}}\sum_{g\in\mathcal{G}}(a_{i,g}+b_{i,g}) \text{ and then } \min_{\{x_{i,g}\}_{i\in\mathcal{I},g\in\mathcal{G}}} \mid x_{i,g} - \frac{ub_{i,g}+lb_{i,g}}{2} \mid$$

$$\text{s.t. } lb_{i,g} - a_{i,g} \leq x_{i,g} \leq ub_{i,g} + b_{i,g},$$

$$x_{i,g} = \sum_{\underline{i}\prec i} x_{\underline{i},g} \ \forall i,g \in \mathcal{I},\mathcal{G},$$

$$x_{i,g} = \sum_{\underline{g}\prec g} x_{i,\underline{g}} \ \forall i,g \in \mathcal{I},\mathcal{G}.$$

Equation 6 states the problem for the NAICS years; the formulation for the SIC years is analogous with some of the equality constraints turned into inequality constraints as in equation 5 above.

The closest feasible model always has a feasible solution for $\{x_{i,g}\}_{i\in\mathcal{I},g\in\mathcal{G}}$ that satisfies all constraints.[22] In cases where the original constraint set is non-empty, the problem in equation 6 always results in $a_{i,g} = b_{i,g} = 0 \, \forall i \in \mathcal{I}, \forall g \in \mathcal{G}$, and hence delivers the same solution as the problem in equation 4 above. In cases where the original data are internally inconsistent and do imply an empty constrained set, the program in equation 6 finds the minimal adjustments to the data necessary to permit a solution.

4. Discussion of Imputation Results

The implementation of the above linear programming procedure on CBP files from 1977 to 2016 yields imputed employment for all suppressed cells in all files in all years. In this section, we discuss the resulting estimates. The imputed data are available at www.fpeckert.me/cbp.

4.1 Inconsistent Bounds

Internally inconsistent observations (i.e., non-zero estimates for $a_{i,g}$ and $b_{i,g}$) are found in 14 out of 40 years of data. Appendix Table A.4 provides a set of summary statistics on the adjustments we make in these cases. For the years in which they are necessary, the median adjustment in terms of employment is 10, and the median total adjustment per year (i.e., the sum of adjustments within a year) is 624 to the lower bounds and 279 to the upper bounds. Table 7 lists the adjustments made in the 1990 county-level file. Of the nine necessary changes, one is made to the national total for root 3990 and another eight are to the state-level totals for that and various other roots. There are no adjustments after 2001,

---

[22]In fact, this is easy to see: $a_{i,g} = lb_{i,g} \, \forall i \in \mathcal{I}, \, \forall g \in \mathcal{G}$ and $x_{i,g} = 0 \, \forall i \in \mathcal{I}, \, \forall g \in \mathcal{G}$ is always a trivial solution to the problem in equation 6.

SIC	Geography	LB	UB
3990	National	0	+107
8900	MA	-17	0
8900	MN	-8	0
2800	OH	-28	0
3300	OH	-20	0
3500	OH	-6	0
3900	OH	-209	0
3990	OH	-339	0
8900	PA	-52	0

Notes: Adjustments as put out by least feasible model.

which may indicate that Census changed its data handling protocol in a way that prevented these minor errors.
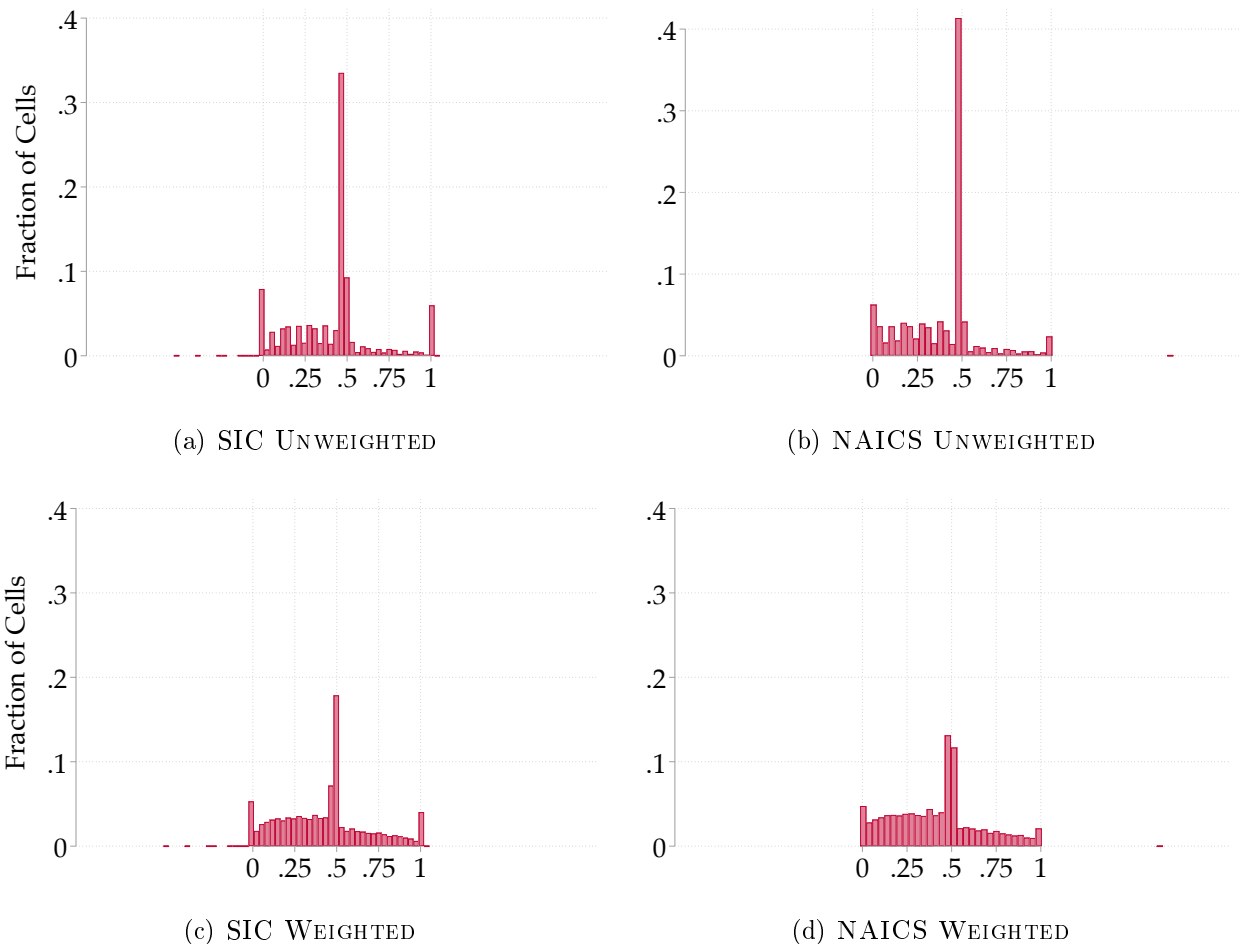
4.2 Baseline Estimates

Our baseline estimates are derived from a linear program which minimizes deviations from cells' midpoints. Figure 4 plots the distribution of the location of our imputed estimates within the bounds provided by Census. This figure focuses solely on cells with employment suppression flags, and separate distributions are provided for the SIC (left panels) and NAICS (right panels) years. In both cases, as in Figure 3, we include suppressed cells only if they are the finest observations available in each county-industry hierarchy. For the SIC years, this observation might be a four-digit SIC code or a root. For the NAICS years, this observation is a six-digit NAICS code. Distributions in the upper panels are unweighted, while the distributions in the lower panels are weighted by cells' imputed employment.

As indicated in the figure, most imputes in both the SIC and NAICS years lie at cells' midpoints, though sizable shares—roughly 10 percent—lie at cells' lower and upper bounds. The distributions in the lower panel, however, reveal that imputes at cells' midpoints are less likely for larger cells, especially during the NAICS era.

One way to compare our baseline estimates to "naïve" estimates which merely choose the midpoint of each cell is to plot their implied aggregate employment. This comparison is provided in Figure 5, which restricts attention to the most disaggregated observation in

(a) SIC UNWEIGHTED

(b) NAICS UNWEIGHTED

(c) SIC WEIGHTED

(d) NAICS WEIGHTED

Source: 1977 to 2017 CBP files and authors' calculations. Figure reports the distribution of imputed employment within the bounds of suppressed cells. Lower and upper bounds are 0 and 1. Position can be below zero (i.e., less than the lower bound) or above 1 in cases where the bounds are found to be inconsistent, as discussed in Section 3. Observations are weighted by employment. Left panel focuses on the SIC era, while right panel summarizes NAICS years. Upper panel distributions are unweighted. Lower panel distributions are weighted by imputed employment. SIC years include only the most disaggregate observation available in each county-industry hierarchy. NAICS years include only industries.

each industry hierarchy. As indicated in the figure, and consistent with Figure 3, the total employment for our imputed observations is smaller than that of imputations based solely on the midpoints, but the two series are highly correlated. The average difference in employment between these two series across years is 13 million, or 6 million if 2011, 2013, and 2015 are excluded. The correlation between the two series is 0.69 across all years, and 0.99 if 2011, 2013, and 2015 are excluded.
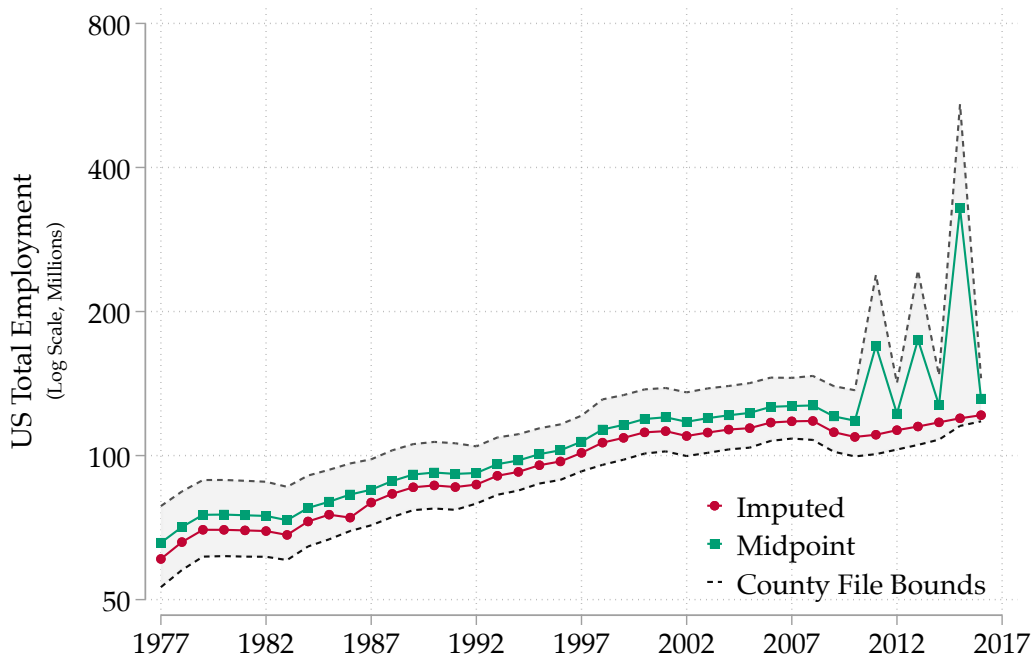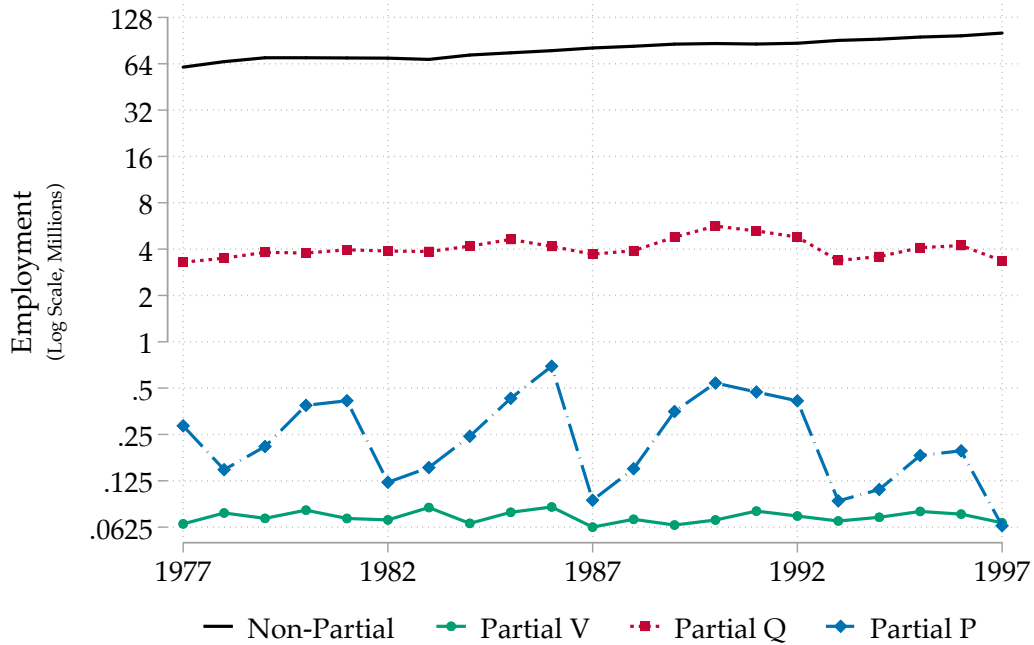
Source: 1977 to 2017 CBP files and authors' calculations. Figure compares the sum of imputed values for suppressed cells to the sum of the midpoints of the Census-provided lower and upper bounds of those cells, across counties. Figure ignores noise suppression. Upper bounds and midpoints for 2011, 2013 and 2015 contain "M" codes from Table 3. Industry classification switches from SIC to NAICS in 1998. SIC years include only the most disaggregate observation available in each county-industry hierarchy. NAICS years include only 6-digit NAICS industries (i.e., roots are excluded).

4.3 SIC Partial Codes

During the SIC era, "partial" codes capture the difference in employment between a root and the finer roots or industries that lie under that root. For example, the CBP data may report employment of 100 for root 1510, and list one industry under that root with employment of 80. The remaining employment of 20 is known to be in 1510 but cannot be further attributed, so it does not receive an entry in the CBP files.

We assign employment to these implied "partial" codes at three different levels of aggregation by comparing employment totals along industry hierarchies. These codes take the form ddxxxP for employment that can be associated with a three-digit root (but nothing below that), ddxxQQ for employment that can be associated with a two-digit root (but nothing below that), and ddVVVV for employment that can be associated with a division code (but nothing below that). In these synthetic partial codes, dd corresponds to a division code, xxx and xx correspond to 3- and 2-digit SIC roots, and P, Q and V are filler letters used to

FIGURE 6: PARTIAL CODE EMPLOYMENT

signify the level of the partial code. The filler letters ensure that the overall partial code is six characters long, to match the way we report SIC codes in the imputed data provided by our data appendix. Figure 6 plots the employment associated with non-partial and partial codes by year over the SIC era. The latter is broken down into the three constituent parts, i.e., the employment accounted for by codes ddxxxP, ddxxQQ and ddVVVV. As noted in the figure, partial-code employment is about 4 million per year. It tends to drop in years in which the Census conducts its Economic Censuses (years ending in 2 and 7), presumably because information on establishments' industries is most likely to be known in those years.

5. DESCRIPTION OF INDUSTRY CONCORDANCES

In this section, we describe the various industry classification systems in the CBP, the concordances we use to map employment across these systems, and the treatment of auxiliary establishments.

23

5.1 Industry Classification Changes

Between 1962 and 2016, the CBP industry classification system changed eight times, in 1967, 1972, 1977, 1987, 1997, 2002, 2007, and 2012.[23] These updates were issued by the Office of Management and Budget, which oversees industrial classification. The most significant change occurred when the Census Bureau moved from using SIC to NAICS. The first vintage of NAICS was introduced in 1997, though NAICS was not incorporated in the CBP data until 1998. Fort and Klimek (2018) describe the rationale for the switch from SIC to NAICS, and the resulting changes that occurred in how establishments, and thus workers, are classified. Table 2 summarizes the changes to the classification system and what years of the CBP are covered by each system.

5.2 Concordances

We collect official published concordances for the eight industry classification changes listed above and reported in Table 2. Details on the sources of these concordances are in Section D of the appendix. Each concordance contains a mapping from the most detailed industry of the outgoing classification to the most detailed industry of the incoming classification. These mappings between each vintage are of four types: (1) one-to-one, where a single outgoing code corresponds to a single incoming code; (2) many-to-one, where more than one outgoing code is combined into a single new code; (3) one-to-many, where a single outgoing code is broken into more than one new code; and (4) many-to-many, where several outgoing codes map to several incoming codes. We use these concordances to assign employment in all years to a NAICS 2012 basis.

Assigning employment to a corresponding NAICS industry is trivial when the mappings between different systems are one-to-one or many-to-one. In cases in which a single industry under the system used in year $t-1$ maps to multiple industries under the system used in year t, we must allocate the employment in $t-1$ across those industries in t. This is particularly important for the SIC to NAICS transition, where 321 out of a total of 891 industries map to multiple NAICS industries. Some of these changes were dramatic; 96 SIC industries have some (or all) employment reassigned to entirely different aggregate divisions.

In order to deal with these transitions as cleanly as possible, we use the 1997 Economic Census (EC) data provided by Fort and Klimek (2018) to construct an employment-weighted concordance between SIC to NAICS. The 1997 EC collected establishment-level information on both a SIC and a NAICS basis. For a particular SIC industry, we can thus calculate the

---

[23]There typically is a one- or two-year delay between the publication of a new classification system and its use in the CBP. The system (adoption) years are as follows: 1967 (1968), 1972 (1974), 1977 (1977), 1987 (1988), 1997 (1998), 2002 (2003), 2007 (2008), and 2012 (2012). It is not uncommon for older vintage codes to appear occasionally in the data (e.g., a SIC 1972 code may be present in the 1989 data). The 1977 version was considered a supplement to the 1972 version and contains only a few minor changes.

share of employment that maps to each of its corresponding NAICS industries. Because the CBP contains data on some industries that are out of scope of the Economic Census, we must supplement this concordance with mappings for those industries from the official published concordance.[24] Unfortunately, we do not have weights for those mappings and thus simply share SIC employment equally across all the NAICS industries to which that particular SIC code maps.[25] In future drafts, we plan to construct similar concordances using the published 2002, 2007, and 2012 EC data for transitions across the various NAICS vintages. In the current version, we share employment in a NAICS year t vintage code equally across the NAICS $t+1$ vintage codes to which they map in the official NAICS concordances. For the earlier transitions between different SIC vintages, there does not seem to be similarly published data, so we rely on the official, published concordances and simply share employment in the SIC t year vintage industry equally across the multiple SIC $t+1$ year vintage industries.

The concordances we collect supply mappings for the most detailed industry codes, i.e. 4-digit SIC and 6-digit NAICS. From these, we construct concordances between economic division SIC and NAICS codes as well as between 2- and 3-digit SIC roots and 3- and 4-digit NAICS roots, respectively, by replacing digits with zeros and slashes as appropriate and collapsing the results.[26]

5.3 Auxiliary Establishments

An important distinction between SIC and NAICS is the treatment of "auxiliary" establishments. These are establishments that provide support functions for other establishments of the firm, e.g., a headquarter plant or a research and development (R&D) lab.[27] Under SIC, the employment of auxiliaries is classified according to the primary sector of the establishments they serve. For example, if a motor vehicle manufacturing firm has a R&D establishment that develops new products for its automotive manufacturing plants, it is an auxiliary and its SIC industry code reflects the two-digit root of its motor vehicle manufacturing plants. In contrast, if that R&D establishment primarily serves outside customers, it

---

[24]We also identify 251 mappings from SIC to NAICS that are not in the EC data, but that are in the official concordance and appear to be relevant in the CBP data. In particular, for a given county and SIC code in the 1997 CBP data, we consider a mapping "possible" if a corresponding NAICS code appears in the 1998 data in that county. If a particular mapping is "possible" in at least 50 percent of the counties in which that SIC code appears, then we supplement the EC concordance with that mapping.

[25]For example, SIC code 0133 ("sugar cane and sugar beets") maps to NAICS 111930 ("sugar beet farming") and NAICS 111991("sugarcane farming"). We thus share SIC code 0133 as 50 percent to NAICS 111930 and 50 percent to NAICS 111991.

[26]For example, from the mapping from the 4-digit SIC code 0133 to 6-digit NAICS codes 111930 and 111991, we can obtain a mapping from 3-digit SIC 0130 to 4-digit NAICS 1119//.

[27]The Census defines auxiliaries as "establishments that are primarily engaged in providing supporting services for other establishments of the same company rather than for the general public or for other business firms" (Office of Statistical Standards, 1972). Another example, from the 1972 SIC manual, is a warehouse storing a firm's own goods.

is not an auxiliary and is given an SIC industry code reflecting its R&D activity.

Under NAICS, auxiliaries are classified according to their primary activity irrespective of whether they primarily serve their own firm. Note that under NAICS, a headquarters establishment (NAICS 551114) is a plant that performs two or more functions generally classified under NAICS 54, "professional, scientific and technical services."[28]

In the CBP, auxiliary employment under SIC is aggregated to the SIC division level using a code that is the combination of a two-digit root followed by an "8" or a "9" followed by a "\."[29] To concord auxiliary employment between SIC and NAICS, we use the mapping between SIC auxiliary codes and NAICS codes provided in the 1997 Economic Census. Unfortunately, this concordance only provides mappings to relatively coarse NAICS codes at the 2-, 3-, and 4-digit levels.[30]

5.4 Assigning Employment to a NAICS Basis

We use the concordances described above to assign employment to a NAICS basis, in particular, NAICS 2012. There are at least three limitations to this approach. First, the Economic Census concordances that we build are based on data in a particular Census year. We must therefore assume that the relevant industries to which a particular SIC industry maps are the same across time. This is a particularly strong assumption for the SIC to NAICS transition, in which we use the 1997 data to concord all historical SIC data to a NAICS basis. Second, we use national-level information to share employment in a particular SIC industry across the multiple corresponding NAICS industries to which it maps. It is possible that these transitions, which Fort and Klimek (2018) show to vary by firm size, vary by geography, too. The implied NAICS employment in the SIC years may thus appear less spatially concentrated than it is in reality.[31]

---

[28]For a more detailed discussion of how SIC auxiliary establishments are classified under NAICS, see https://www.census.gov/eos/www/naics/history/docs/cm_3.pdf.

[29]The full set of auxiliary codes is: 098\ (for 07--), 149\ (for 10--), 179\ (for 15--), 399\ (for 19--), 497\ (for 40--), 519\ (for 50--), 599\ (for 52--), 679\ for (60), and 899\ for (70--). Thus, division employment for a geographic area is the sum of all employment at lower levels plus employment at auxiliaries. For example, the manufacturing (20--) employment for a county would be the sum of all two digit roots in that county (2000 to 3900) plus the employment in 399\.

[30]There are four SIC codes used after 1974 that "are not comparable to [those] published in prior years": 5063 (electrical apparatus and equipment), 1730 (electrical work), "auxiliary" code 3600, which captures administrative and auxiliary employment related to electric and electronic equipment, and auxiliary code 4800, which captures administrative and auxiliary employment related to communication industries.

[31]We are not able to map establishment counts from a SIC to a NAICS basis in a convincing way. It does not seem sensible to share establishment counts given the fact that in reality, an establishment can only belong to a single industry. We encourage researchers to bear these limitations in mind when using the concorded data.
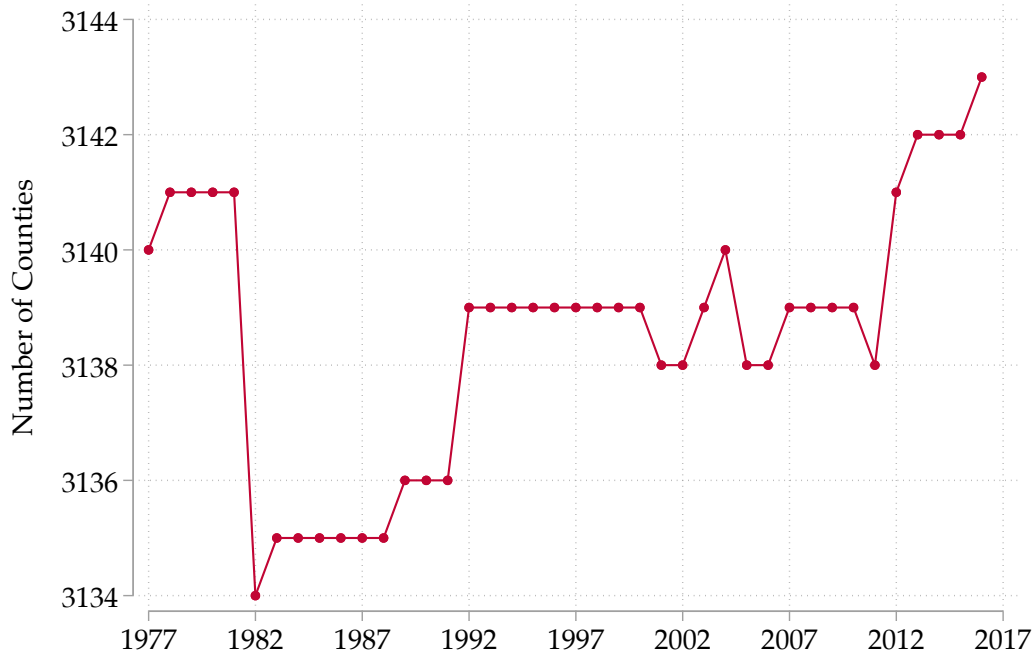
Source: 1977 to 2017 CBP files and authors' calculations. Figure reports the number of counties in each year's raw CBP county file.

6. COUNTY CONCORDANCES

In addition to time-varying industry classifications, the number and boundaries of spatial units are not constant over time. In particular, Figure 7 shows the number of counties for which data is reported in the CBP Files in each year. The number fluctuates between 3,134 and 3,143. A file documenting which counties are present in each year is available in our data appendix available at www.fpeckert.me/cbp.

Even during periods for which the number of counties is stable, there may be changes in the geographic delineations of counties. Eckert et al. (2018) overlay GIS county maps for all decades of US history to create a crosswalk connecting counties in each year to the 1990 county boundaries.[32] This crosswalk allows researchers to work with a fixed geographic delineation. A copy of this crosswalk is also available in our data appendix.

---

[32]See Eckert and Peters (2018) for an application of the crosswalk to US Census data from 1880 to 2000.

7. CONCLUSION

The County Business Pattern files present a unique public data source for the study of the spatial economy of the United States and its evolution over time. In this paper, we outline procedures to overcome the key limitations of the raw data sources: suppressed employment counts and different industry classification systems used throughout the decades. By providing imputed data for all data cells and a consistent classification system over time, we hope to encourage more researchers to study the spatial dimension of the economy of the United States.

References

Autor, D. H., D. Dorn, and G. H. Hanson (2013): "The China Syndrome: Local Labor Market Effects of Import Competition in the United States," *American Economic Review*, 103, 2121–68.

Bernard, A. B., S. J. Redding, and P. K. Schott (2013): "Testing for Factor Price Equality with Unobserved Differences in Factor Quality or Productivity," *American Economic Journal: Microeconomics*, 5, 135–63.

Caliendo, L., M. Dvorkin, and F. Parro (2015): "Trade and Labor Market Dynamics," NBER Working Paper 21149.

Caliendo, L., F. Parro, E. Rossi-Hansberg, and P.-D. Sarte (2014): "The Impact of Regional and Sectoral Productivity Changes on the U.S. Economy," NBER Working Paper.

Diamond, R. (2013): "The Determinants and Welfare Implications of US Workers' Diverging Location Choices by Skill: 1980-2000," Stanford Graduate School of Business, mimeo.

Ding, X., T. Fort, Redding, and P. K. Schott (2019): "Structural Change Within Versus Across Firms: Evidence from the United States," Working Paper.

Eckert, F. (2019): "Growing apart: Tradable services and the fragmentation of the US economy," Working Paper.

Eckert, F., S. Ganapati, and C. Walsh (2019): "Skilled tradable services: The transformation of US high-skill labor markets," *Available at SSRN 3439118*.

Eckert, F., A. Gvirtz, J. Liang, and M. Peters (2018): "A Consistent County-Level Crosswalk for US Spatial Data since 1790," Working Paper.

Eckert, F. and M. Peters (2018): "Spatial Structural Change," Working Paper.

Fort, T. C. and S. K. D. Klimek (2018): "The Effects of Industry Classification Changes on US Employment Composition," Working Paper 18-28, CES.

Fort, T. C., J. R. Pierce, and P. K. Schott (2018): "New perspectives on the decline of US manufacturing employment," *Journal of Economic Perspectives*, 32, 47–72.

Glaeser, E., H. D. Kallal, J. Scheinkman, and A. Shleifer (1992): "Growth in Cities," *Journal of Political Economy*, 100, 1126–52.

HERSHBEIN, B. AND L. B. KAHN (2018): "Do recessions accelerate routine-biased technological change? Evidence from vacancy postings," *American Economic Review*, 108, 1737–72.

HOLMES, T. J. AND J. J. STEVENS (2004): "Geographic concentration and establishment size: analysis in an alternative economic geography model," *Journal of Economic Geography*, 4, 227–250.

HORNBECK, R. AND E. MORETTI (2018): "Who benefits from productivity growth? Direct and indirect effects of local TFP growth on wages, rents, and inequality," Tech. rep., National Bureau of Economic Research.

ISSERMAN, A. M. AND J. WESTERVELT (2006): "1.5 Million Missing Numbers: Overcoming Employment Suppression in County Business Patterns Data," *International Regional Science Review*, 29, 311–335.

OFFICE OF STATISTICAL STANDARDS (1957): *Standard Industrial Classification Manual*, Office of Statistical Standards, Bureau of the Budget.

——— (1967): *Standard Industrial Classification Manual*, Office of Statistical Standards, Bureau of the Budget, https://archive.org/details/standardindustri00offi_2.

——— (1972): *Standard Industrial Classification Manual*, Office of Statistical Standards, Bureau of the Budget.

UNITED STATES CENSUS BUREAU (1986): *County Business Patterns, 1974-84, Technical Documentation*, U.S. Department of Commerce, Bureau of the Census.

——— (2019a): "County Business Patterns (CBP)," https://www.census.gov/programs-surveys/cbp/data/datasets.html.

——— (2019b): "North American Industry Classification System (NAICS), Concordances," https://www.census.gov/eos/www/naics/concordances/concordances.html.

UNITED STATES NATIONAL ARCHIVES (2019): "County Business Patterns (CBP)," https://research.archives.gov/id/613576.

Online Appendix (Not for Publication)

This online appendix contains detailed additional empirical results as well as more detailed explanations of data used in the main text.

A. SIC and NAICS Hierarchies

Table 2 lists the industry classification system in use during the years over which the CBP is available. Tables A.1 and A.2 provide examples of the hierarchical nature of the root codes under both the SIC and NAICS systems.

B. Data Sources

We use two primary sources for the raw CBP data files. The National Archives (United States National Archives (2019)) website provides the files for 1967 and 1968 to 1985. Beginning with 1986, our files are downloaded directly from the CBP webpage of the Census Bureau website (United States Census Bureau (2019a)). The data for each year consist of comma-separated text files, with one each for US-, state-, and county-level data.[33] The Census Bureau website also provides industry and geography reference files. Copies of all of our input files are available on our data appendix website www.fpeckert.me/cbp

C. Data Adjustments and Cleaning

C.1 Inconsistent Codes Dropped Before Imputation

Before implementing our imputation procedure, we clean the raw data by dropping certain codes and their associated employment in all files. We remove two types of codes. First, we remove codes that do not appear in all three data files —national, state, and county—in a given year. Column 2 of Table A.3 lists the codes dropped for this reason. Second, we remove all codes that do appear in the data files but not in the official industry reference file for that year. These codes are listed in column 3 of Table A.3. Note that Table A.3 is restricted to years in which at least one code is dropped.

---

[33]Beginning in 1993, data at the metropolitan area level were published. Beginning in 1994, data at the ZIP code level were published. Data for Puerto Rico started being published in 1998. We plan to include these additional geographies and the additional adding up constraints they imply in future releases.

C.2   Identifying Bound Adjustments

The first step of the imputation procedure tests whether there exists at least one vector of employment counts that satisfies all constraints implicit in the union of the three data sets (national, states, and county). If not, we implement the procedure outlined in Section 3.3 above to find the minimal adjustments to any cell upper and lower bounds in the data set so as to allow for at least one feasible solution. Summary statistics for these adjustments are noted in Table A.4.

D.   CONCORDANCE CONSTRUCTION

In order to construct the concordance file from each SIC and NAICS vintage to the NAICS 2012 set of codes, we iteratively walk each data file from vintage to vintage until 2012 is reached. Table A.5 list the sources we use for the concordances for each vintage-to-vintage pair.

Each raw concordance contains a list of mappings from an industry code in the outgoing vintage to an industry code in the incoming vintage. In order to create a complete concordance, we use these industry mappings to construct mappings for the root codes. For example, in the 1987 SIC-to-1997 NAICS concordance, the SIC code 1521 maps to NAICS code 233210 and 1522 maps to 23220 and 233320. From this, we know that SIC root code 1520 maps to NAICS root codes 2332// and 2333//. We do this iteratively for each code and root to get mappings for all possible root levels. When mapping from SIC to NAICS, we map 1-digit SIC to 2-digit NAICS, 2-digit SIC to 3-digit NAICS, and 3-digit SIC to 4-digit NAICS.

In order to match the codes that appear in the CBP data, we make two adjustments to the SIC codes in the raw concordances. First, from the industry descriptions, we determined that the code 1510 ("General Building Contractors, Residential Buildings") that appears in the CBP industry reference file is equivalent to 1520 in the official SIC. Second, the code 8310 ("Social Services, n.e.c."), which appears in the CBP industry reference file, does not appear in the official SIC, so we group this code with its root code, 8300.

We also add two mappings to the SIC concordances for codes that appear in the CBP but not in the official SIC concordances or in the concordance constructed from the EC. First, the agricultural auxiliary code 098\, which appears in the CBP, does not appear in the EC. This discrepancy arises because the EC does not include agricultural industries. (They are tabulated in the separate Census of Agriculture.) Based on the NAICS description for the code 115/// ("Support Activities for Agricultural and Forestry"), we assign all employment from auxiliary SIC code 098\ to NAICS code 115///. Second, between 1998 and 2002, the CBP include an auxiliary total code 95----, which appears neither in the subsequent CBP years nor in the official NAICS code list. In order to remain consistent with codes that

appear in the EC, we assign this code to the NAICS code 94999, which Fort and Klimek (2018) describe as "Unclassified Auxiliary Establishments."

For all concordances except the concordance for 1987 SIC to 1997 NAICS, we compute weights for a particular outgoing code by sharing employment equally across all the outgoing industries to which that particular outgoing code maps. For example, SIC code 0133 maps to NAICS 111930 and NAICS 111991. We thus share SIC code 0133 as 50 percent to NAICS 111930 and 50 percent to NAICS 111991. For the 1987 SIC-to-1997 NAICS concordance, we compute weights using employment given in the EC.

E. Online Data Repository Overview

Our online data appendix at www.fpeckert.me/cbp offers an extensive repository with all codes, raw data, and imputed data. In this section, we provide an overview of the sets of files found in the online data repository.

E.1   Raw Data Files

We provide the raw national, state, and county CBP files for each year. The sources for these files are discussed in Section B above.

E.2   Imputed Data

For each year we provide the imputed data at the finest industry level. The imputed data we present are based on the raw data bounds given by Census to which we apply the two step procedure described in Section 3.

E.3   Concordances

For each vintage of SIC and NAICS codes we provide a concordance file with weights to translate industry employment to NAICS 2012 industry codes.

E.4   Codes

We provide the Python files we used to clean the raw data and to run the linear program on the resulting cleaned data.

TABLE A.1: NAICS INDUSTRY OVERVIEW, 2010

Level	NAICS	Industry	Employment
1	------	All Sectors	111,970,095
2	31----	Manufacturing	10,862,838
3	311///	Food manufacturing	1,432,843
4	3111//	Animal food manufacturing	50,442
5	31111/	Animal food manufacturing	50,442
6	311111	Dog and cat food manufacturing	22,163
6	311119	Other animal food manufacturing	28,279
4	3112//	Grain and oilseed milling	54,926
5	31121/	Flour milling and malt manufacturing	15,543
6	311211	Flour milling	11,027
6	311212	Rice milling	3,686
6	311213	Malt manufacturing	830
5	31122/	Starch and vegetable fats and oils manufacturing	24,532
	. . .	. . .	. . .

Source: Excerpt from the 2010 CBP files which employ the 2007 vintage of the NAICS codes. The files are available on the Census website: https://www.census.gov/programs-surveys/cbp.html

TABLE A.2: SIC INDUSTRY OVERVIEW, 1990

TABLE A.2: SIC INDUSTRY OVERVIEW, 1990

Level	SIC	Industry	Employment
1	----	All Sectors	93,476,087
2	20--	Manufacturing	19,173,382
3	2000	Food and kindred products	14,52,803
4	2010	Meat products	371,386
5	2011	Meat packing plants	119,172
5	2013	Sausages and other prepared meats	78,799
5	2015	Poultry slaughtering and processing	170,850
4	2020	Dairy products	140,154
5	2021	Creamery butter	1,903
5	2022	Cheese, natural and processed	34,570
5	2023	Dry, condensed, evaporated products	13,178
5	2024	Ice cream and frozen desserts	20,737
5	2026	Fluid milk	69,308
4	2030	Preserved fruits and vegetables	187,979
	. . .	. . .	. . .

Source: Excerpt from the 1990 CBP files which employ the 1987 vintage of the SIC codes. The files are available on the Census website: https://www.census.gov/programs-surveys/cbp.html

TABLE A.3: INCONSISTENT INDUSTRY CODES

Year	Codes in CBP Files Dropped due to Inconsistency...	
	...across CBP Files	...with Reference File
1977	785 2031 2433 2442 3611 3791 3803 3821 5122	3716 5129 6798 7012 7013 8310 8361
1978	2015 2433 2442 2661 3611 3791 3803 3821 4582 8062 8084 8411 8800 8810	3716 6070 6798 7012 7013 8310 8361
1979	759 785 079/ 1625 2036 2940 2942 3073 3239 3481 3716 5192 5212 5513 5780 5781 5820 5821 5991 6122 6406 7012 7060 7065 7328 7380 7388 7626 7638 7835 7912 7994 8120 8126 8500 8560 8562 8680 8800 8810 8811	5129 6590 6599 6798 7013 8310 8361
1980	1629 3716 64– 8631	6798 8310 8361
1981	1540 1542 8051	6798 8310 8361
1982	1321 2771 8800 8810 8811	3716 6590 6599 6798 8310 8361
1983	1711 6793	3716 4229 6590 6599 6798 8310 8361
1984	5380 5580	3716 4229 6590 6599 6798 8310 8361
1985	5380 5580	3716 6798 8310 8361
1986	5380 5580 7840	1111 1481 1531 1611 4121 4131 4151 4231 4411 4431 4441 4712 4811 4821 4899 4911 4941 4961 4971 5380 6410 6610 8110 8361
1987	1540 4214 5399 6410 8320 8321 8330 8331 8350 8351 8390 8399	
1988	5399	
1990	8990	
1992	5810	
1993	5810	

Source: The codes in the second column of this table table are identified as inconsistent by comparing the national, state, and county files with one another for each CBP year from 1977 to 2016. The codes in the third column of this table table are identified as inconsistent by comparing the national, state, and county files with the appropriate industry reference files for each CBP year from 1977 to 2016. Years in which no codes are dropped are ommitted from the table.

TABLE A.4: SUMMARY STATISTICS ON ADJUSTMENTS TO RAW DATA

Year	Bounds Adjusted	Total LB Adjustment	Total UB Adjustment	Avg LB Adjustment	Avg UB Adjustment
1977	8	4,732	22	591.5	2.75
1978	135	3,863	859	28.40	6.32
1979	55	2,095	6,140	37.41	109.64
1980	30	3,990	238	128.71	7.677
1981	7	123	70	15.38	8.75
1982	10	568	320	51.64	29.09
1984	1	0	54	0	27
1985	1	1	0	.5	0
1986	2	463	0	154.33	0
1987	530	18,424	13,523	34.7	25.47
1989	15	347	1,347	21.69	84.19
1990	9	679	107	67.9	10.7
1999	9	989	1,837	98.9	183.7
2001	14	51	1,221	3.4	81.4

Source: Calculations by authors based on output from Closest Feasible Model Procedure applied to raw bounds in CBP files for 1977 to 2016. Years where adjustments were not necessary are not listed.

Table A.5: Sources for Concordances

Crosswalking		
From	**To**	**Source**
SIC '57	SIC '67	Constructed by authors using Office of Statistical Standards (1957) and Office of Statistical Standards (1967)
SIC '67	SIC '72	Constructed by authors using Office of Statistical Standards (1967) and Office of Statistical Standards (1972)
SIC '72	SIC '77	Using Crosswalk from Fort and Klimek (2018)
SIC '77	SIC '87	Using Crosswalk from Fort and Klimek (2018)
SIC '87	NAICS '97	Using Crosswalk from Fort and Klimek (2018)
NAICS '97	NAICS '02	Using Crosswalk from United States Census Bureau (2019b)
NAICS '02	NAICS '07	Using Crosswalk from United States Census Bureau (2019b)
NAICS '07	NAICS '12	Using Crosswalk from United States Census Bureau (2019b)

Source: The table provides the sources used to construct the crosswalk from any SIC vintage to the 2012 NAICS industry classification.