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 2021

Preliminary Draft!

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: (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. Finally, we construct a user-friendly, 1975 to 2016 county-level panel that classifies industries according to a consistent set of 2012 NAICS codes in all years.

*The imputed data sets and the industry concordances are available in the Data Appendix to this paper located at www.fpeckert.me/cbp. We thank Anubhav Agarwal, Soumya Gottipati, Udit Jain, and Yunus Tuncbilek for outstanding research assistance, and the Yale Economics Department and the Princeton IES Section for financial support. We also thank Chris Harshaw and Dan Spielman at the Yale Network Institute for help with the Gurobi software and the Yale High-Powered Computer Cluster staff for their assistance. We thank Don Davis, Thomas Holmes, and Stephen Redding for helpful comments. We note that all imputations in this paper are to assist researchers in conducting statistical analyses and impart no information about the underlying firms in counties or industries with suppressed data.

†University of California, San Diego; fpe@ucsd.edu
‡Tuck School at Dartmouth, CEPR & NBER; teresa.fort@dartmouth.edu
§Yale School of Management & CEPR & NBER; peter.schott@yale.edu
¶Columbia University; natalie.yang@columbia.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 1975 to 2016 files, and plan to extend them back in time to 1946, and forward as new versions are released, 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

[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.

provides a lower and upper bound on its actual employment.

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.[3] 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.[4] We stress that while our imputations are internally consistent, there is no guarantee that they represent the true geography-industry employment that suppression was designed to obscure, nor do they reveal any information about firm identities.

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.[5]

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 discuss these classification systems and the challenges they present to constructing a long CBP panel.[6] In the final section of the paper, we describe how we

[3]The CBP files also contain information on the establishment size distribution within each geography-industry cell. While this information could in principle be use to help impute employment, in practice these imputed bounds are often inconsistent. As a result, we do not use this information, at least for now.

[4]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).

[5]We use the midpoint in our baseline estimates since many prior studies (e.g., Glaeser et al., 1992; Holmes and Stevens, 2004) using the CBP simply chose the midpoint as imputed employment for suppressed cells. We plan to discuss the sensitivity of our imputed estimates to this objective function and methods for bootstrapping confidence intervals for our imputes in a future draft.

[6]We also discuss changes in geographic units. While state boundaries are constant during our sample period, the number of counties and the geographic borders of some of them change over time.

augment SIC-SIC and SIC-NAICS concordances provided by Fort and Klimek (2018), and use them, along with official NAICS-NAICS concordances, to reclassify employment in the 1975 to 2016 editions of the CBP to the NAICS 2012 vintage of industry codes. We use NAICS 2012 in constructing this panel for three reasons. First, this reclassification most accurately captures current economic activity, and renders the data compatible with future releases of the CBP.[7] 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.[8] We provide the input datasets and code to assign the CBP data to a NAICS 2012 basis, as well as a dataset from 1975 to 2016 with county-level employment on a NAICS 2012 basis. While we cannot always provide a full six-digit NAICS codes for every observation, this dataset is relatively user-friendly since researchers can simply aggregate all the observations to arrive at correct industry, county, or national employment totals. By contrast, the raw CBP data during the SIC years cannot be aggregated to obtain employment totals since they include aggregated industry codes that sometimes are the sum of the more detailed codes, but other times contain employment for which their disaggregated-industry-code employment is incomplete.

Our paper contributes to previous efforts to impute missing values in the CBP data and make them available to other researchers. Our method is most closely related to the procedure proposed by Isserman and Westervelt (2006), who also use the industry and geography hierarchies as constraints.[9] Our method differs in how we choose a vector of employment estimates that simultaneously satisfies all constraints. Isserman and Westervelt (2006) use simulated annealing to find a solution vector satisfying all bounds.[10] Instead, we define a linear objective function, transforming the problem into a simple linear program.[11] Given that linear programming is a mature technology, this transformation allows us to exploit well-established and

[7] Consider Computer Systems Design as an example. It is an important and growing sector that did not exist under SIC.

[8] 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)).

[9] Autor et al. (2013) exploit industry hierarchies to tighten suppressed cells' lower and upper bounds, and then select point estimates within those bounds subject to industry adding-up constraints.

[10] Isserman and Westervelt (2006) also attempt to incorporate information implied by the establishment size distribution provided in the raw Census files. In fact, this information can be inconsistent with the employment bounds we focus on in our method, and frequently the respective employment bounds do not overlap. As a result, for our baseline estimates, we use only the employment bounds directly stated in the raw CBP files. For more information about this issue, please contact the authors.

[11] We note that while the problem is *conceptually* simple, solving a linear program with millions of choice variables is computationally challenging. In practice, our imputed employments for all years can be estimated in a single day on Yale's High Powered Computing Cluster.

understood algorithms that guarantee exact, globally optimal solutions in non-prohibitive time (see Bertsimas and Tsitsiklis, 1997). As such, our method has three advantages over Isserman and Westervelt (2006). First, because we can solve the linear program problem exactly, all of our estimates satisfy all constraints. Simulated annealing, by contrast, yields estimates that satisfy constraints only approximately. Second, simulated annealing as used in Isserman and Westervelt (2006) is a heuristic without proof of convergence, while the theory behind linear programs is extensively studied and discussed.[12] Third, our approach can be used to find and correct inconsistencies in the bounds reported in the raw CBP files by using data correction methods from the linear programming literature. Such corrections are especially useful for data prior to 2000, which contain typos and errors implying there exists no vector of employment consistent with all bounds as stated.[13]

We also are the first to provide CBP data on a consistent NAICS basis for the entire 1975 to 2016 period. We provide all the input files and computer code needed to concord the raw imputed data to NAICS 2012, noting our concordance assumptions so that they can be adapted by other researchers as needed. As part of this effort, we construct *synthetic* "partial" SIC codes that reconcile inconsistencies in reported employment at different levels of aggregation, which we add to the user-friendly dataset. We provide concordances for these synthetic codes to NAICS, and show how they can be used to develop a long NAICS CBP panel containing county-industry observations at the finest level of aggregation. This panel can then be then be collapsed to provide employment at any coarser level of aggregation, e.g., national employment totals.

Finally, our method to impute missing data using a linear programming approach contributes to a robust literature on data correction. The idea to exploit hierarchical constraints to impute missing data using linear programming techniques and to correct data inconsistencies extends beyond the context of the CBP files. Our method can be applied to any data collected at different levels of aggregation, both to infer missing information and to resolve internal inconsistencies. For example, it could be used to infer sub-national outcomes in situations where complete data on the aggregate economic performance of a country are available, but data across regions is sparse. Or, it could be useful in identifying inconsistencies in data reported at the two levels.

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. Section 4 describes and analyzes the employment estimates resulting from applying our method to the

[12] Our implementation of the approach adopted in Isserman and Westervelt (2006) using the C coding language did not converge to an exact solution in reasonable time for any of the years we study.

[13] Estimates from Isserman and Westervelt (2006) are not public and therefore not available for comparison. Bartik et al. (2018) use the same approach to impute employment during the NAICS years, i.e., starting in 1998. We compare our imputations to theirs over this period in Section 7.2.

CBP. Sections 5 and 6 discuss industry and county concordances. Section 7 describes how we create a user-friendly, county-level CBP panel that classifies industries according to a consistent set of 2012 NAICS codes in all years. This panel can be collapsed along both industries and geographies to yield correct national totals. Section 8 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.[14] 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.[15]

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 the 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 Security Administration.[16]

2.1 Unit of Analysis

Since 1974, establishments have been 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

[14]The sources of the files we use are detailed in Appendix A.2. The US Census Bureau's CBP website posts CBP files from 1986 to the 2018 (the most recent file available). CBP files for earlier years are available at the National Archives.

[15]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.

[16]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's 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 located at www.fpeckert.me/cbp.

Table 1: Establishment Size Bins in the CBP Files

	Employment	
Flag	**Minimum**	**Maximum**
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.

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 spatial break in the reporting of non-manufacturing employment in 1974.[17]

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.[18] 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 before 2017.[19]

[17]We discuss a second spacial break, related to the classification of professional employer organizations (PEOs), is Section 2.6.

[18]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.

[19]Beginning with reference year 2017, a cell is only published if it contains three or more establishments. In all other cases, the cell is not included in the release (i.e., it is dropped from publication). In prior years, payroll values for cells with fewer than three establishments would have been suppressed and an employment size range (EMPFLAG) would have been provided; however, the number of establishments would have been published.

Table 2: CBP Industry Codes

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: CBP and authors' calculations. CBP datasets provided by the National Archives prior to 1986 and the US Census Bureau County Business Patterns website thereafter.

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 1,150 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 string of hyphens.[20] 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

[20]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 discrepancy as a mistake. In our imputed data files and concordances we replaced all occurrences of '19--' with '20--.'

covers the "meat packing plants" industry. Note that Census does not always provide a detailed breakout of the industry employment. As a result, summing all industry employment does not always yield total national employment. Instead, it is necessary to sum all employment for the detailed industry codes, as well as parts of of the employment under the corresponding root codes, in order to aggregate the raw SIC data to obtain national totals. We discuss this in more detail in Section 4.4 below.

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."

It is important to note that in the CBP data, industry codes do not always have the same first two digits as their economic division. For example, in 1987 SIC, the economic division code 20-- covers SIC codes with first two digits ranging from 20 to 39. Similarly, the NAICS division code 31---- contains the total employment for all industry codes that start with 31, 32, or 33. We therefore construct "six-digit SIC codes" in the data files we provide that contain the division codes followed by the SIC code reported by the CBP to facilitate aggregation and imputation.

As indicated in Table 2, SIC was updated every five to ten years, while NAICS is updated every five years.[21] 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.

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 under NAICS, one can sum employment across all roots of a given order, or across all industries.

[21] 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.

Aggregation is more complicated for SIC years because a county may have reported employment for a root but not have cells for some or all of the industries within that root. This feature of the SIC years complicates collapsing of the data to yield national totals, as the employment count of a given root can be weakly larger than the employment count obtained from summing employment across its industrial children. As a result, we create supplementary, synthetic "partial" codes – discussed further in Section 4.4 below – that can be inserted into the raw or imputed data to capture the difference in employment between a root and the reported industrial children under that root. In Section 7, we create a user-friendly, county-level 1975 to 2016 CBP panel with consistent 2012 NAICS codes which contains these "partial" SIC entries, concorded to NAICS. Without these 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 industries.

This feature of the CBP files during the SIC years is driven by missing information. Via email correspondence with Census, we have been informed that when CBP data were tabulated during the SIC years, employment at the $k-1$ level could be larger than the sum of employment at the k level as there may have been establishments that were known to be in root $k-1$, but whose k root or industry employment was not known. For example, according to an April 9, 2019 email from the Census help desk, EWD County Business Patterns (CENSUS/EWD):

> "When CBP was tabulated under the SIC classification system, the programming allowed data to be published at the "xxx0" level where "xxx" equals the first three digits of the industry group code. Based on available industry data, there were some establishments that were known to be in one of these categories, but it was not clear which category was correct, so they were included in the 0760 total, but not included in either the 0761 or 0762 tabulations."

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.[22] In 2007, Census additionally introduced "noise

[22] As there is no upper bound for the final code, "M", we use a conservative value of 100,000,000 in the few instances in which this code appears, in 2011, 2013 and 2015. We note that flag "A" includes zero employment. All cells included in the CBP have at least one establishment. While it is possible that an establishment has zero employment, actual employment for these cells may not be zero. In principle, information on the

Table 3: Data Suppression Flags

	Employment	
Flag	**Minimum**	**Maximum**
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.

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, which indicate the range of the random noise multiplier used to produce it.[23] 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 1975 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. To avoid double counting, this figure is restricted to the most detailed industry codes available for each county.[24] As indicated in the figure, 55 to 70 percent of cells

establishment size distribution within each geography-industry cell (see Section 2.1) could 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, in the baseline estimates included in this paper and our data appendix, we ignore information implied by the establishment size distribution. For more information about this issue, please contact the authors.

[23] 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 located at www.fpeckert.me/cbp.

[24] For example, if a county reports employment for root 2050 as well as employment for industries 2051, 2052 and 2053 under that root, the observation for 2050 is not included. We refer to these codes as *level*4 or *level*6

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 2007-2016 CBP files from the US Census Bureau County Business Patterns website.

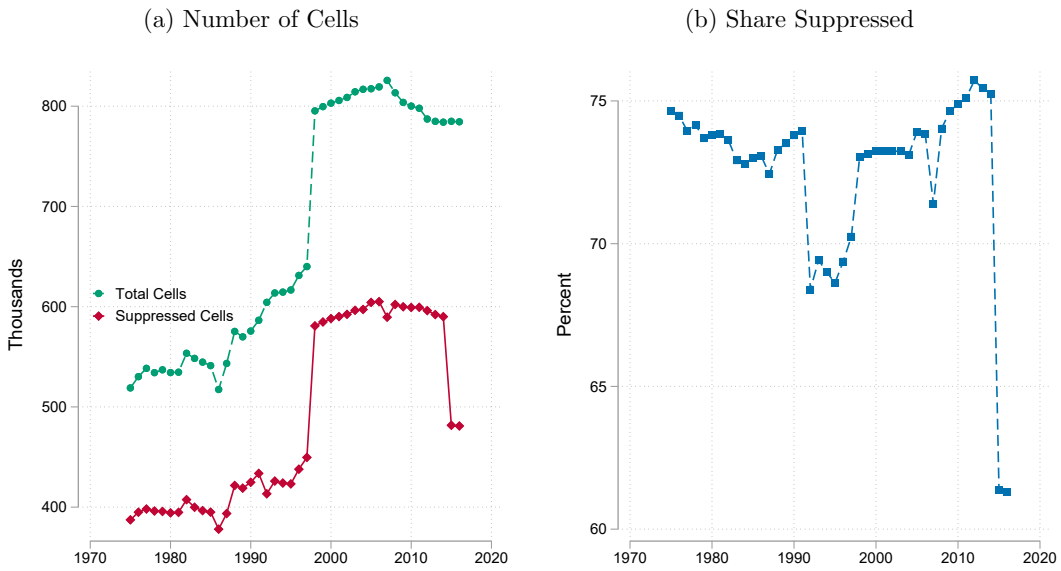
are suppressed, depending upon the year. Figure 1 also highlights the jump in the number of county-industry cells from roughly 0.9 to 2.1 million that occurs in the transition from SIC to NAICS. This increase has two sources. First, the NAICS classification scheme is more detailed than the SIC, as illustrated 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 in the raw CBP files by division code for SIC (left panel) and NAICS (right panel) years, 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 from the national files (which is always unsuppressed). Here, too, both figures only include cells with the most disaggregate industry observation available for each county. 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.[25] 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.

in Section 7.

[25] 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.

Figure 1: Suppressed County-Industry Cells



Source: 1975 to 2017 CBP files and authors' calculations. Figure displays the number of cells in the raw CBP county files in each year, the number of those cells that are suppressed, and the share of cells that are suppressed. Cell counts include only the most disaggregate industry observation available for each county. The number of cells affected solely by noise infusion starting in 2007 are not included in these counts. Industry classification switches from SIC to NAICS in 1998.

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 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.

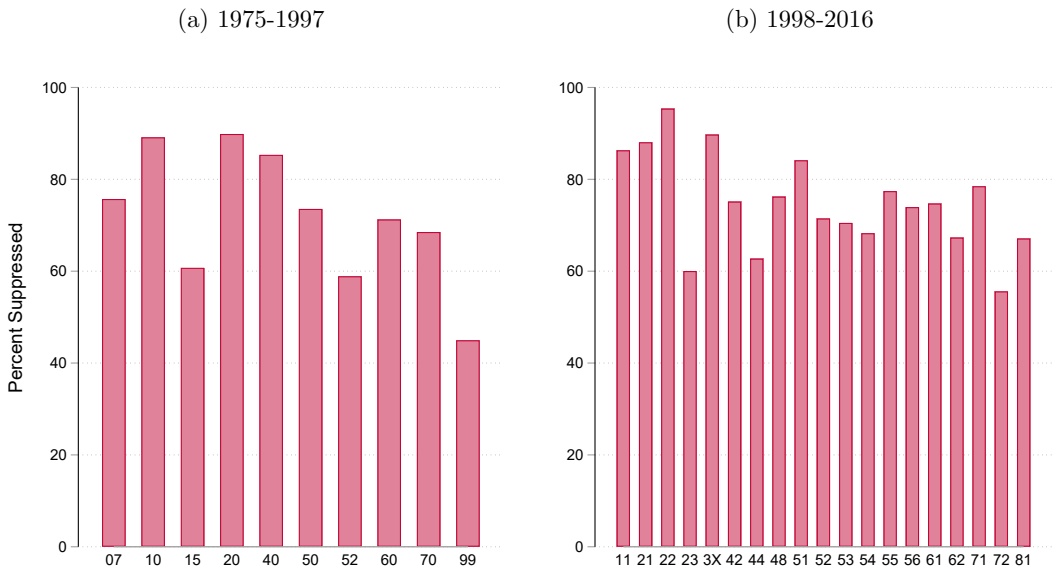
In cell C6 of the top panel, the six-digit NAICS industry 113310 is the only "child" of the NAICS 5-digit root 11331/ in cell C5. As a result, its employment must add to that of cell C5. Likewise for cells C4 and C3. Employment in cell C2, however, is larger than that for cell C3 because Autauga County 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

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: CBP and authors' calculations.

Figure 2: Suppression by Two-Digit Roots



Source: 1975 to 2017 CBP files and authors' calculations. Figure displays the share of cells in the raw CPB files that are suppressed, by SIC division code for 1997 to 1997 (right panel) and two-digit NAICS sector (right panel). Cell counts include only the most disaggregate industry observation available for each county. 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.

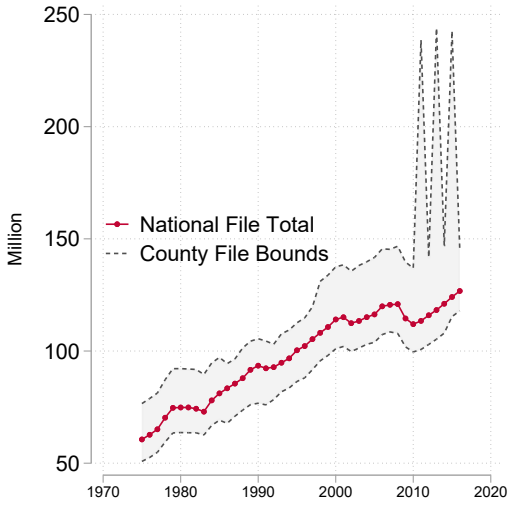
employment reported for a longer list of two-digit roots.

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.[26] Even in that case, however, it is clear that information in cells C2 and C3 can be used

[26]The noise suppression flag for these observations (G) indicates noise of up to 2 percent (see Table 4). The

Figure 3: Employment in County vs National Files



Source: 1975 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. Only cells containing the most disaggregate industry observation available for each county are included. Industry classification switches from SIC to NAICS in 1998.

to help pin down employment in cell C7.[27] 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."[28] 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 also includes employment in industries with two-digit roots 16 and 17 (not shown).

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.

[27] 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.

[28] 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: CBP and authors' calculations.

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 for the SIC years. 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 codes included in the CBP documentation for that year.

In implementing our imputation procedure, we treat these inconsistencies as follows. First, we drop all observations with industry codes that do not appear in all three files in a given year. The second inconsistency does not impact our imputation procedure, which only requires that the three input data sets are consistent with one another. As a result, we do not drop codes that do not appear in the official reference file. Users of our imputation are free to do so. Table A.3 in Section A.3.1 of the Appendix lists all codes associated with dropped observations. We note that because the employment associated with any of these dropped observations appears in more aggregate roots, dropping it does not affect our adding up constraints. Moreover, in Section 4.4 below, we explain how any employment associated with these codes is captured by

the creation of synthetic "partial" codes.

2.6 County-less Employment

The CBP county data files contain a row for each county and industry combination as discussed above. Additionally, they contain an extra "artificial" county code, "999," for each state. These 999 cells record industry employment that can be attributed to a state but not to an individual county within that state. For example, since 2007, professional employer organizations (PEOs) have become the employer of record for an increasing number of workers.[29] The Census Bureau has taken the position that assigning these workers to establishments is too difficult and instead allocates them to county 999.[30]

As indicated in Figure 4, employment assigned to county 999 codes is less than 1 percent until 2000, at which point it begins to rise steady, hitting 4 percent in 2016.

2.7 Change in Census Suppression Protocol After 2016

The Census Bureau changed its suppression protocol with the 2017 release of the County Business Patterns data. From the County Business Patterns website:

> "Prior to reference year 2017, the number of establishments in a particular tabulation cell was not considered sensitive; therefore, counts of establishments were released without any disclosure avoidance methods applied. Beginning with reference year 2017, cells with fewer than 3 establishments have been omitted from the release."

As a result, it is impossible to know whether industry-county combinations missing from the raw data reflect a suppression by the Census or an actual absence of establishments, rendering direct application of our method infeasible. We plan to update our approach to handle the new disclosure regime in a future draft.

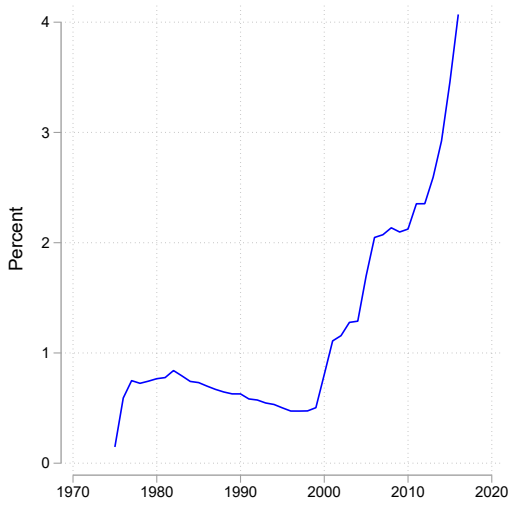
3 Imputing Suppressed Cells as Linear Program

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

[29] Communication with the CBP office at the Census Bureau has revealed that "the decision to start assigning PEOs as statewide was a more casual internal decision. Currently there is no documentation stating this policy and we only move them when we come across them in our review so not all PEOs will be statewide. We are discussing updating the documentation to reflect the change, but unfortunately I don't have anything official to provide at this time."

[30] We thank Teresa M. Lynch, Founding Principal at Mass Economics, for alerting us to this issue.

Figure 4: Share of Employment in 999 Counties



Source: CBP and authors' calculations. Figure reports the share of overall employment in each year accounted for by "artificial" county code "999", which captures employment that can be attributed to a state but not to an individual county within that state. Cells are included in the total only if they represent the most disaggregate industry observation available for the regular or 999 county code (i.e., the long NAICS panel described in Section 7).

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 "..." 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:

$$lb_{i,g} \leq x_{i,g} \leq ub_{i,g}. \tag{1}$$

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:

$$x_{i,g} = \sum_{\mathrm{i} \prec \mathrm{i}} x_{\mathrm{i},g} \qquad \forall i, g \in \mathcal{I}, \mathcal{G}. \tag{2}$$

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:

$$x_{i,g} = \sum_{\mathrm{g} \prec \mathrm{g}} x_{i,\mathrm{g}} \qquad \forall i, g \in \mathcal{I}, \mathcal{G}. \tag{3}$$

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)).[31] The resulting linear program can be written as follows:

[31] 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.

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

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

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

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

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.[32] 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, 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:

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

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

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

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

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.[33] $a_{i,g} \geq 0$ 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*.

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

$$\min_{\{a_{i,g}\}_{i\in\mathcal{I},g\in\mathcal{G}}\geq 0,\{b_{i,g}\}_{i\in\mathcal{I},g\in\mathcal{G}}\geq 0} \sum_{i\in\mathcal{I}}\sum_{g\in\mathcal{G}}(a_{i,g}+b_{i,g}) \text{ and then } \min_{\{x_{i,g}\}_{i\in\mathcal{I},g\in\mathcal{G}}} \mid x_{i,g} - \frac{ub_{i,g}+lb_{i,g}}{2} \mid \quad (6)$$

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

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

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

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.[34] 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 1975 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.

[33] 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.

[34] 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.

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

Table 7: Adjustments to Internally Inconsistent Bounds, 1990

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

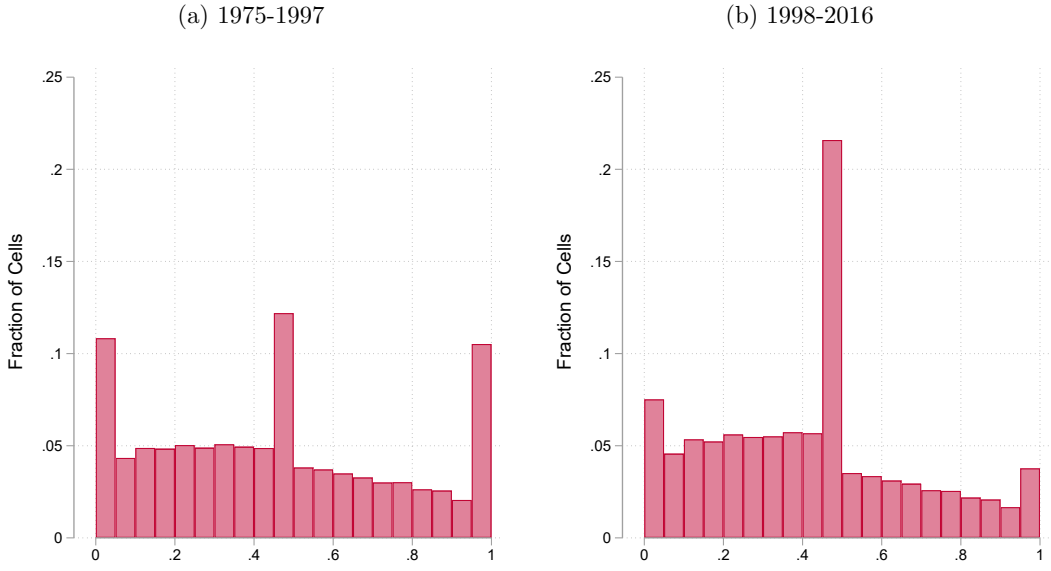
Source: CBP and authors' calculations. Table displays adjustments required by the least feasible model in the 1990 imputation.

4.2 Baseline Estimates

Our baseline estimates are derived from a linear program which minimizes deviations from cells' midpoints. Figure 5 plots the imputed-employment-weighted 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 avoid double counting by including suppressed cells only if they are the finest industry observations available in each county-industry hierarchy.

As indicated in the figure, most imputes in both the SIC and NAICS years lie at cells' midpoints, though 10 to 20 percent lie at cells' lower and upper bounds.

Figure 5: Position of Imputed Employment in Suppressed Cell-Bounds



Source: CBP 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. Left panel focuses on the SIC era, while right panel summarizes NAICS years. SIC years include only the most disaggregate observation available in each county-industry hierarchy. NAICS years include only industries. Distributions are weighted by imputed employment.

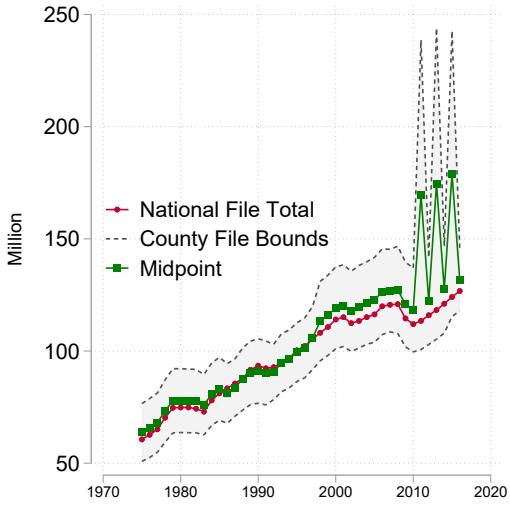
4.3 Imputed County Total vs Raw County Totals

We compare the total imputed employment to the total employment of the lower and upper bounds in the raw CPB files as well as the total employment of the midpoint of these bounds in Figure 6. As above, this figure is restricted to the observations for each county that are at the finest level of aggregation. As indicated in the figure, and consistent with Figure 3, the total imputed employment is smaller than midpoint employment in most years between 1975 and 2016. The average difference in employment between these two series across years is 13 million, or 6 million if 2011, 2013, and 2015 (i.e., the years with "M" suppression codes) are excluded. The correlation between the two series is 0.69 across all years, and 0.99 if 2011, 2013, and 2015 are excluded.

4.4 SIC Partial Codes

As noted above, during the SIC era the employment for a root can exceed that of its underlying children because employment at some of the children is not known, or may be suppressed. We create what we refer to as *synthetic* "partial" codes to capture these differences. For example, the CBP data may report employment of 100 for root 2050, and list one industry under that

Figure 6: Imputed vs Midpoint Employment for Disaggregate Observations



Source: CBP and authors' calculations. Figure compares total imputed employment in each year to three raw totals implied by the lower and upper bounds in the raw data and their midpoints. For cells not subject to suppression, the latter three quantities are the same. Upper bounds and midpoints spike in 2011, 2013 and 2015 due to the presence of "M" suppression codes (see Table 3). Industry classification switches from SIC to NAICS in 1998. Cells are included in the figure only if they contain the most disaggregate industry observations.

root (e.g., 2051) with employment of 80. The remaining employment of 20 is known to be in 2050, but its allocation across the remaining children within 2050 (i.e., 2052 and 2053) is not known. Moreover, Census does not include any observation at a finer level of detail than 2050 to mark this unknown employment of 20. As a result, if one were to sum just the most detailed four-digit codes during the SIC era, they would not sum to the national total. We address this issue by adding a synthetic partial code with SIC code 2050 (the most detailed information available for the missing employment) and employment of 20 to the data so that it plus the employment of 80 for 2051 adds to the employment of 100 for the root code 2050.

The synthetic partial codes we create are six characters long to match the way we report SIC codes in the imputed data.[35] They can appear at three different levels of aggregation: they take the form $ddxxxP$ for employment that can be associated with a three-digit root (but nothing below that) as in the example just given, $ddxxQQ$ for employment that can be

[35]We add two-digit division code prefixes to the standard four-digit SIC codes in our imputation algorithm for technical reasons related to ensuring that imputations obey the industry adding-up constraints. We therefore use the same 6-digit convention in assigning the synthetic partial codes discussed in this section. These codes can be converted back to standard four-digit SIC codes by dropping the prefixes and adding zeros in place of P, QQ and $VVVV$.

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 or xx correspond to 3- and 2-digit SIC roots, and P, Q and V are filler letters used to signify the level of the partial code.

For the example of SIC 2050 given above, the included observation of 80 workers for 2051 would have the six-digit code 202051. We then add to the data an observation with synthetic partial code 20205P and employment of 20, so that it, plus the employment of 80 for 202051 yields the proper total for root 205, which is 100. Identifying these synthetic partials is a necessary step for creating long panels using a consistent industry definition, such as those we describe in Section 7. For example, to create a long NAICS panel, we want to map SIC to NAICS as precisely as possible. In our running example, it is preferable to separately map 2051 with employment of 80 and 205P with employment of 20 to NAICS than to map 2050 with employment of 100 to NAICS, since the former exploit the highest level of detail available.[36]

The raw CBP data also contain what we refer to as *native* "partial" codes, i.e., two- or three-digit SIC codes that end in 00 or 0, respectively, under which no detailed four-digit codes are reported. In the example provided above, the original 2050 observation with employment of 100 is not a native partial code because employment for at least one more detailed code under the 205 root is reported (for 2051). However, if employment of 100 for SIC 2050 were listed for a county, but no employment for any codes under that root (i.e., for 2051, 2052, or 2053) were reported, that instance of 2050 would constitute a native partial code. In creating the long, NAICS-level panel discussed in Section 7, employment in these native partials is allocated across all NAICS codes to which the underlying four-digit SIC 205x codes map.

Our procedure for identifying synthetic and native partial codes is contained in the program *p1_partial_code_employment_20201231.do*, available in our Data Appendix. This program creates the output file *efsy_partial_19771997_01_posted.csv*, which contains two industry fields: *sic6* and *sic4*. The former contains the six-digit codes used in our imputation, discussed above. Synthetic partial codes have the patterns noted above, i.e., $ddxxxP$, $ddxxQQ$ or $ddVVVV$. Native partial codes have the pattern $ddxxx0$ or $ddxx00$. The field *sic_old* contains the original codes from the raw CBP data. For the synthetic codes, which represent new observations that we add to the raw data, we convert the six-digit *sic* codes back to four-digit *sic4* codes by dropping the *dd* prefix and substituting "0" or "00" as needed for P and QQ and $VVVV$.

An accurate "bottom-up" total of US employment during the SIC years for a particular

[36]We note that our approach to concording synthetic partial codes from SIC to NAICS assumes that some of the employment of 20 at 2050 will be attributed to 2051 as well as 2052 and 2053. This assumption is correct if the missing industry detail is due to incomplete industry information for certain establishments. It would be incorrect if the missing industry detail is due to suppression of employment for the remaining detailed industries (e.g., 2052 and 2053 in our example). Since we cannot distinguish these two possibilities, we follow the simpler approach of mapping employment at the synthetic partials using an aggregation of all possible detailed SIC to NAICS mappings.

geography, e.g., the national total, can be computed as the sum of three types of codes: (i) four-digit SIC codes that do not end in zero, i.e., the most detailed codes available; (ii) native partial codes; and (iii) synthetic partial codes.

In the user-friendly, county-level 1975 to 2016 panel of imputed employment discussed in Section 7, we retain only the most detailed codes needed to create a "bottom-up" total for any geography, e.g., the national total. In these panels, in the case of our running example, SIC 202050 with employment of 100 would be dropped, the SIC 202051 observation with employment of 80 would be retained, and the synthetic partial SIC code 20205P with employment of 20 would be added.[37]

Figure 7 plots the employment of partial and non-partial codes of this panel on a log scale by year over the SIC era. The sum of all lines shown in the figure is total US employment. Partial employment is broken down into its synthetic and native constituent parts, where the legend in the figure refers to these codes' suffix. As noted in the figure, native partial code "0" employment is about as large as non-partial employment, i.e., employment associated with the most detailed four-digit SIC codes. The next largest level of employment is for synthetic partial P and QQ codes and native "00" employment. Employment attributed to P and QQ codes cycles around the years in which the Census conducts its Economic Censuses (years ending in 2 and 7), since more complete and accurate information on establishments' industries is most likely to be known in those years due to extensive survey efforts. Industry information in non-EC years, and for establishments not surveyed by the EC, is collected from administrative data and thus less complete. Figure 7 also depicts a rise in the partial employment in native partial codes with "00" in 1988, perhaps due to the transition to the SIC 1987 vintage.

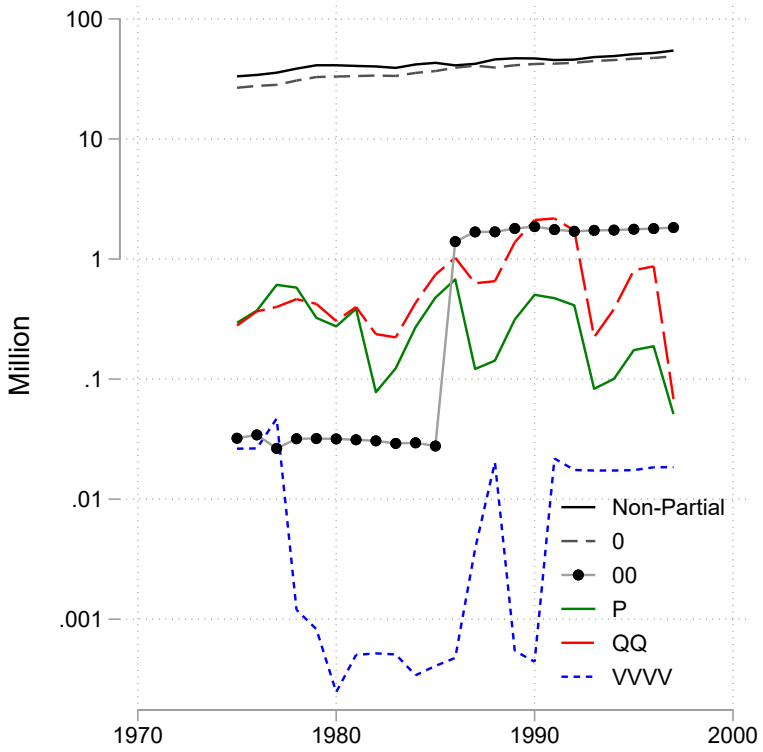
Finally, Figure 8 plots overall imputed county employment with and without synthetic partials and imputed national employment as shares of national employment reported in the raw national files. As indicated in the figure, we match the raw totals exactly until 1997, after which we are shy by about 1 percent until 2002 and then less then half a percent thereafter. Our exact match before 1998 is driven by our addition of synthetic partial codes. We believe that gaps after 2002 are driven by rounding issues in our algorithm, and suspect that the larger mismatch from 1998 to 2002 is related to concordances bridging the transition from SIC to NAICS. We plan to address these gaps further in a future draft.

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

[37] We note that partial codes in a very small number of cells -1 due to rounding. These can be reset to zero as needed. For transparency, we leave this task to consumers of our imputation.

Figure 7: Synthetic and Native Partial Code Employment



Source: CBP and authors' calculations. Figure reports the employment accounted for by SIC "partial codes" in the long NAICS panel described in Section 7. Cells are included in these totals only if they represent the most disaggregate industry observations for the county. Codes P, QQ, and VVVV correspond to 3, 2, and 1 digit synthetic partial codes, respectively. Codes 0 and 00 represent native partial codes.

establishments.

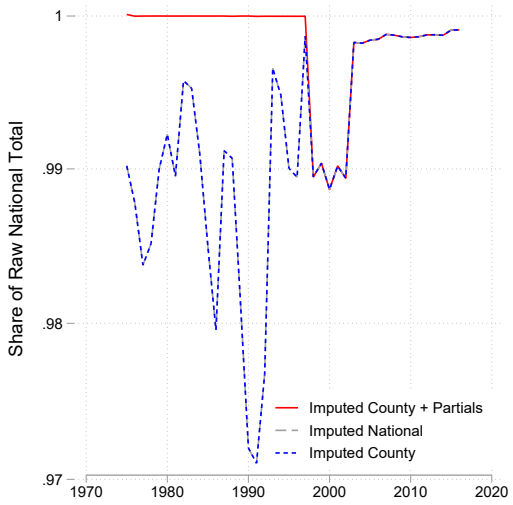
5.1 Industry Classification Changes

Between 1962 and 2016, the CBP industry classification system changed eight times, in 1967, 1972, 1977, 1987, 1997, 2002, 2007, and 2012.[38] 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 incor-

[38]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.

Figure 8: Comparison of National Totals



Source: CBP and authors' calculations. Figure displays three ratios. The first is the imputed national employment divided by the national employment in the raw, national files. The second and third are the sum of imputed county employment, with and without the addition of the partial codes discussed in Section 4.4. Sum of imputed county totals come from the "long panel" described in Section 7.

porated in the CBP data until 1998. Fort and Klimek (2018) describe the rationale for the transition 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 A.4 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.[39]

Assigning employment to a corresponding NAICS industry is trivial when the mappings between different systems are one-to-one or many-to-one. In one-to-many or many-to-many cases where 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 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.[40] 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.[41] 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.

[39] In Appendix Section A.4, we discuss two adjustments to the concordances necessary to match the SIC codes that appear in the CBP data to NAICS. 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.

[40] We 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. To determine which mappings from the official concordance need to be added to supplement the EC concordance, we consider the frequency with which codes appear in the 1997 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.

[41] For example, SIC code 0133 ("sugar cane and sugar beets") maps to NAICS 111930 ("sugar beet farming") and NAICS 111991("sugarcane farming"). These mappings do not appear in the EC concordance, so we cannot compute employment-based weights for them. We thus share SIC code 0133 as 50 percent to NAICS 111930 and 50 percent to NAICS 111991.

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.[42] Please note that these aggregated mappings should only be used for partial code observations, discussed above in Section 4.4. It is never sensible to try to map, for example, the entire manufacturing sector using our mappings from 2-digit SIC to 3-digit NAICS. Instead, users need to use the most detailed information available for every observation and then collapse the data to their aggregation level of interest. We provide do-files for this approach on the website.[43]

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.[44] 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 an 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 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, the headquarters industry (NAICS 551114) includes establishments that perform two or more functions generally classified under NAICS 54, "professional, scientific and technical services," primarily for their own firm.[45]

In the CBP, auxiliary employment under SIC is aggregated to the SIC division level using

[42]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//.

[43]For users of pre-1974 data, there are four SIC codes used after 1974 that "are not comparable to [those] published in prior years" according to the CBP data manuals: 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.

[44]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.

[45]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.

a code that is the combination of a two-digit root followed by an "8" or a "9" followed by a "\."[46] 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. Users interested in time series analyses of these sectors therefore need to address these limitations by aggregating the data in the SIC years in those sectors that contain auxiliary employment. For example, users who are interested in employment in "research and development in Biotechnology" (NAICS code 541711) should note that in the SIC years, all auxiliary employment in this area will be assigned 541700, and is thus not distinguishable from employment in 541712 and 541720.

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 9 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.[47] This crosswalk allows researchers to work with a fixed geographic delineation. A copy of this crosswalk is also available in our Data Appendix.

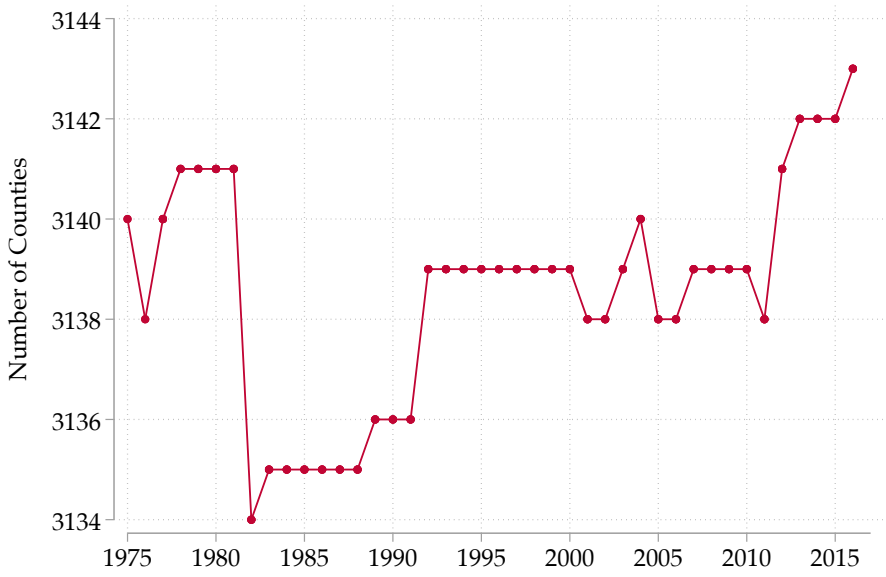
7 Creating a Long Panel

In this section we describe how we use our imputed data to construct a user-friendly, county- and NAICS-level panel that can be collapsed along any dimension to yield the proper national totals. The panel we currently construct runs from 1975 to 2016. Both this panel and its precursor, which retains each year's native industry codes, are described below and are available in our Data Appendix located at www.fpeckert.me/cbp, along with the programs used to create them, which also are listed below.

[46] 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\.

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

Figure 9: Number of Counties in the CBP Files



Source: CBP and authors' calculations. Figure reports the number of counties in each year's raw CBP county file.

For the NAICS panel, we use the concordances described in Section 5.2 to assign employment in each county-year pair to a 2012 vintage NAICS code. We note that there are at least three limitations to this approach. First, the Economic Census concordance that we build is based on data from the 1997 Census. 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 information for the US as a whole to assign employment in a particular SIC industry to one or more NAICS industries. It is possible that these transitions, which Fort and Klimek (2018) show vary by firm size, also vary by geography. As a result, *concorded* NAICS employment during the SIC years may appear less spatially concentrated than *actual* employment.[48] Third, the actual and synthetic partial SIC codes in the data can only reasonably be mapped to partial NAICS codes.[49] Auxiliary employment in the SIC years also can only be linked to partial NAICS codes. We encourage researchers to bear these limitations in mind when using the NAICS long panel.

[48] 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 use establishment counts given that Census assigns each establishment to a single industry.

[49] In particular, we caution that due to the presence of "partial" codes during the SIC years (see Section 4.4), not all codes in either file are at the most disaggregate 4- or 6-digit levels. As a result, it may not be possible to analyze the employment of a county along the narrowest industry codes over time. In those cases, however, it will be possible to analyze employment along a consistent higher aggregate, e.g., 3341 or 33411 *in lieu* of 334112.

Our user-friendly 1975 to 2016 long panels are created via the following programs:

– *p0_create_long_panel_*20201231.*do*: This Stata program sets directories and calls the remaining programs in this list.

– *p1_partial_code_employment_*20201231.*do*: This Stata program reads in the raw imputed CBP files during the SIC years (e.g., $efsy\_cbp\_YYYY.csv$), creates the *synthetic* partial codes and identifies the *native* partial codes described in Section 4.4, and adds the former to the raw imputed SIC files. The output dataset, $efsy\_partial\_19751997\_01.csv$ contains the following additional variables compared to those in the raw imputed files: *partial*, *level*0, *level*1, *level*2, *level*3 and *level*4. These variables indicate whether the entry is a *synthetic* "partial code" and whether, for summing purposes, it is: (i) a national total SIC code; (ii) an SIC division code; (iii) a 2-digit SIC; (iv) a 3-digit SIC; or (v) one of either a proper 4-digit SIC, a *native* partial code, or a *synthetic* partial code.[50] The sum of *level*4 entries produces the national total. In creating the panel in the next step, we first drop non-*level*4 codes during the SIC era, and analogous non-*level*6 codes in the NAICS era.

  - Input files: $efsy\_cbp\_YYYY.csv$
  - Output file: $efsy\_partial\_19751997\_01.csv$

– *p2_build_concordances_*20201231.*do*: This programs concords industry codes across each SIC and NAICS revision, using raw concordances from the US Census Bureau as inputs. These concordances are adapted to include the partial codes identified in the previous step. The raw census concordances called in this step are also available in our Data Appendix. The output files of this program are the "final" concordances used in the next step.

  - Input files
    - $bridge\_s77\_s87.csv$
    - $SIC\_NAIC\_concordance.csv$
    - $1987\_SIC\_to\_1997\_NAICS.xls$
    - $1997\_NAICS\_to\_2002\_NAICS.xls$
    - $2002\_to\_2007\_NAICS.xls$
    - $2007\_to\_2012\_NAICS.xls$
  - Output files

[50] If a *level*4 code ends in zero and *partial* = 0, it is a *native* partial code.

- $full\_sic77\_sic87.csv$
- $full\_sic87\_naics97.csv$
- $full\_naics97\_naics02.csv$
- $full\_naics02\_naics07.csv$
- $full\_naics07\_naics12.csv$

– $p3\_concord\_and\_append\_long\_panel\_20201231.do$: This Stata program uses the output files of the above programs as well as the raw imputed files for the NAICS years (e.g., $efsy\_cbp\_YYYY.csv$) to create the two long panels mentioned in the main text above:

- Input files
  - $efsy\_cbp\_YYYY.csv$
  - $efsy\_partial\_19751997\_01.csv$
  - $full\_sic77\_sic87.csv$
  - $full\_sic77\_sic87.csv$
  - $full\_naics97\_naics02.csv$
  - $full\_naics02\_naics07.csv$
  - $full\_naics07\_naics12.csv$
- Output files
  - $efsy\_partial\_19752016\_01\_NAICS\_20201231.csv$
  - $efsy\_partial\_19752016\_01\_NATIVE\_20201231.csv$

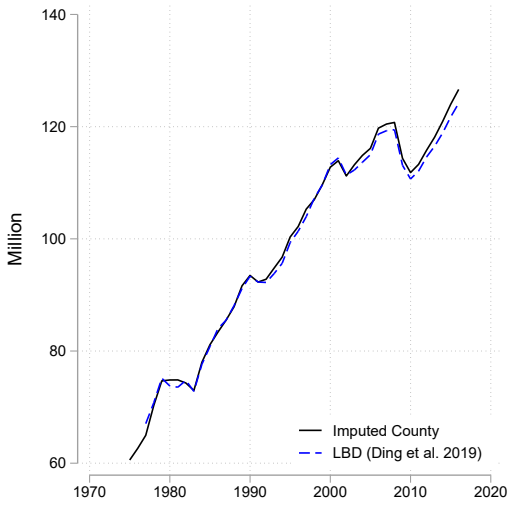
– $p4\_figures\_for\_paper\_20201231.do$: This Stata program uses the output files of the above programs as well as the raw data to create the figures in this paper.

7.1 Comparison of our Long Panel to the LBD

Figure 10 compares total non-agricultural employment in the long NAICS panel to that reported for the Census Bureau's Longitudinal Business Database (LBD) in Ding et al. (2019).[51] As indicated in the figure, imputed CBP employment is highly correlated with LBD employment, but is slightly above it at both the beginning and end of the sample period. Small discrepancies between the series likely are driven by differences in CBP and LBD processing rules as well as the particular sectors deemed to be "in scope."

[51]For this comparison, we drop agricultural employment,i.e., SIC $0xxx$ and NAICS $11xxxx$, as it is excluded in the LBD data we cite.

Figure 10: US Employment CBP vs LBD



Source: LBD, CBP, and authors' calculations. Figure reports total US employment in the Census Bureau's Longitudinal Business Database (LBD), as reported in Ding et al. (2019), and authors' imputed NAICS panel.

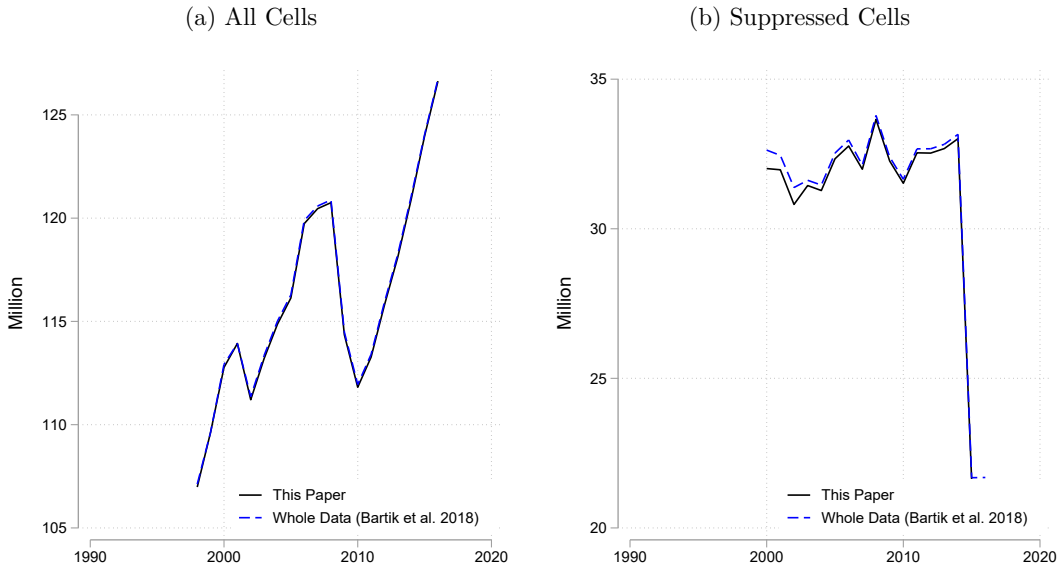
7.2 Comparison of our Long Panel to *Whole Data*

Bartik et al. (2018) implement the alogorithm described in Isserman and Westervelt (2006) to impute suppressed employment starting in 1998. Their estimates, known as *WholeData*, are available upon request from the W.E. Upjohn Institute for Employment Research. In this section, we compare the long panel described above to their imputations for the overlapping years.[52] Figure 11 compares total employment across counties in our long panel to the analogous total in *Whole Data*. As indicated in the left panel, they are quite close overall, with employment in our data less than one-tenth of a percent lower in most years. In the right panel of the figure, we compare total employment across suppressed cells only, using the *Whole Data* indicator for suppression. This figure reveals that we are lower in aggregate until the final two years, where we match exactly.

At the county level, the two series are highly correlated. For all observations, the correlations are 100 percent for 1998 to 1999, about 96 percent for 1999 to 2002, and then above 99.5

[52] Bartik et al. (2018), like us, provide a version of the data in which all NAICS codes are concorded to the 2012 vintage. We drop all non-county and non-6-digit NAICS observations from their dataset and then merge it into ours. The merge is perfect starting in 2008. For years 1998 to 2007, there are roughly 200 observations per year in our data but not theirs. For 1998 to 2002, there are 26 thousand obersvations in their dataset but not ours. For 2003 to 2007, that drops to about 1500 per year. Virtually all of the observations in our dataset but not theirs are for NAICS 525990. The others are 238220, 239990, 425110, 425120, and 454111. Observations in their dataset but not ours span a much wider range of NAICS codes, most heavily concentrated in manufacturing, especially 326122.

Figure 11: US Employment CBP vs *Whole Data*



Source: Whole Data, CBP, and authors' calculations. Figure compares total imputed US employment in the W.E. Upjohn Institute for Employment Research's Whole Data database, which implements approach described in Isserman and Westervelt (2006), to authors' imputed NAICS panel.

percent thereafter. We plan to discuss these discrepancies further in a future draft.

8 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.

Bartik, T. J., S. C. Biddle, B. J. Hershbein, and N. D. Sotherland (2018): " WholeData: Unsuppressed County Business Patterns Data: Version 1.0 [dataset]," Tech. rep., W. E. Upjohn Institute for Employment Research.

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.

Bertsimas, D. and J. N. Tsitsiklis (1997): *Introduction to linear optimization*, vol. 6, Athena Scientific Belmont, MA.

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. C. Fort, S. J. Redding, and P. K. Schott (2019): "Structural change within versus across firms: Evidence from the United States," Tech. rep., Technical Report, Discussion Paper). Dartmouth College.

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 additional empirical results and a more detailed description of the data used in the main text.

A.1 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.

A.2 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.[53] 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

A.3 Data Adjustments and Cleaning

A.3.1 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.

[53] 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.

A.3.2 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.

A.4 Concordance Construction

In order to implement the concordance files from each SIC and NAICS vintage to the NAICS 2012 set of codes in the long NAICS panel, we iteratively walk each data file from vintage to vintage until 2012 is reached. Table A.6 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.

A.5 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.

A.5.1 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 A.2 above.

A.5.2 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.

A.5.3 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.

A.5.4 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

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	Inconsistent Codes across CBP Files
1975	['1311', '3777', '2666', '3077', '2241', '3007']
1976	['3211', '4300', '2031', '2742', '3821', '3323', '4310', '2442', '2432', '8062', '4311', '3611', '2433', '3352', '3481', '3831', '3548']
1977	['4311', '4310', '3611', '3803', '2442', '4300', '2031', '5122', '2433', '3821', '3791', '0785']
1978	['8411', '8062', '2661', '3611', '3803', '2015', '4582', '8800', '2442', '8810', '2433', '3791', '3821', '8084', '0785']
1979	['5991', '5513', '5821', '5781', '7835', '7388', '7012', '2942', '7912', '7065', '8680', '7060', '7626', '8800', '8562', '3073', '7380', '2940', '7638', '6122', '3716', '5192', '5212', '1625', '6406', '8126', '8500', '8560', '3239', '8810', '5820', '3481', '2036', '8811', '6113', '0759', '1092', '7328', '7994', '8120', '5780', '0785']
1980	['1629', '6113', '1092', '8631']
1981	[6113', '1542', '1540', '8051']
1982	[8811', '6113', '8800', '1092', '2771', '1321', '8810']
1983	['1711', '6113', '1092', '3572', '6793']
1984	['5380', '3572', '3673', '5580']
1985	[3572', '5580', '5380']
1986	['4231', '4712', '1531', '4411', '6410', '4941', '4441', '4821', '1111', '4971', '5970', '5380', '1611', '4961', '7840', '4911', '4899', '1481', '4131', '8110', '4811', '8361', '6610', '4431', '4151']
1987	['8351', '4214', '8330', '8390', '8331', '1540', '8320', '8321', '8399', '8350', '5399', '6410']
1988	['5399']
1991	['5810']
1992	['5810']
1993	['5810']
1994	['5810']
1995	['5810']
1996	['5810']
1997	['2067', '5810']

Source: This table lists the industry codes that do not appear in all three CBP files for a given year, i.e., the national and state and county files of a given year. 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
1975	870	62554	84098	71.90	96.66
1976	945	40160	80391	42.50	85.07
1977	1159	5456802	378238	4708.20	326.35
1978	242	7925	1529646	32.75	6320.85
1979	88	3823	7585	43.44	86.19
1980	75	6921	1318	92.28	17.57
1981	8	772	64	96.5	8
1982	16	1017	34	63.56	2.13
1983	5	77	13	15.4	2.6
1984	6	96	54	16	9
1985	5	491	50	98.2	10
1986	1	463	0	463	0
1987	591	21996	11434	37.22	19.35
1989	38	1673	692	44.03	18.21
1990	16	656	124	41	7.75
1999	30	3481	2450	116.03	81.67
2001	10	70	1202	7	120.2

Source: Calculations by authors based on output from Closest Feasible Model Procedure applied to raw bounds in CBP files for 1975 to 2016. Years where adjustments were not necessary are not listed.

Table A.5: Mappings for NAICS codes with Auxiliary Component

NAICS codes with auxiliary component		**Mappings from SIC:**		**Share of 1997 SIC employment**	
Code	**Description**	**AUX SIC codes**	**Non-AUX SIC codes**	AUX SIC codes	Non-AUX SIC codes
115///	Support activities for agriculture and forestry	098\	0700, 0800	.007	.993
484///	Truck transportation	149\, 179\, 399\, 497\, 519\, 599\, 899\	4200	.053	.947
4931//	Warehousing and storage	149\, 179\, 399\, 497\, 519\, 599\, 679\, 899\	4220	.756	.244
514210	Data processing services	149\, 179\, 399\, 497\, 519\, 599\, 679\, 899\	7374, 7379	.069	.931
5411//	Legal services	149\, 399\, 497\, 519\, 599\, 679\, 899\	6540, 7380, 8110	.081	.919
5412//	Offices of notaries	149\, 179\, 399\, 497\, 519\, 499\, 679\, 899\	7290, 7810, 8720	.083	.917
5417//	Scientific research and development services	149\, 179\, 399\, 497\, 519\, 599\, 679\, 899\	3720, 3760, 8730	.161	.839
5418//	Advertising and related services	149\, 399\, 497\, 519\, 599\, 679\, 899\	5190, 7310, 7330, 7380, 8740	.091	.909
551114	Corporate, subsidiary, and regional managing offices	149\, 179\, 399\, 497\, 519\, 679\, 899\		1	0
5613//	Employment services	149\, 179\, 399\, 497\, 519\, 599\, 679\, 899\	7290, 7360, 7810, 7920	.004	.996
56161/	Investigation, guard, and armored car services	149\, 399\, 497\, 519\, 599\, 899\		1	0
5617//	Services to buildings and dwellings	149\, 399\, 519\, 599\, 679\, 899\	0780, 4580, 4950, 7210, 7340, 7380, 7690	.003	.997
811///	Repair and maintenance	149\, 179\, 399\, 497\, 519\, 679\, 899\	3700, 7200, 7300, 7500, 7600	.042	.958
949999	Unclassified auxiliaries	149\, 179\, 399\, 497\, 519\, 599\, 679\, 899\		1	0

Source: These mappings come from the authors' constructed 1987 SIC to 1997 NAICS concordance.

Table A.6: 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.