NBER WORKING PAPER SERIES

THE EARLY COUNTY BUSINESS PATTERN FILES: 1946-1974

Fabian Eckert
Ka-leung Lam
Atif R. Mian
Karsten Müller
Rafael Schwalb
Amir Sufi

Working Paper 30578
http://www.nber.org/papers/w30578

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
October 2022

We would like to thank the Center for Equitable Growth and the National Science Foundation (NSF Award 1949504) for funding. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2022 by Fabian Eckert, Ka-leung Lam, Atif R. Mian, Karsten Müller, Rafael Schwalb, and Amir Sufi. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

The Early County Business Pattern Files: 1946-1974
Fabian Eckert, Ka-leung Lam, Atif R. Mian, Karsten Müller, Rafael Schwalb, and Amir Sufi
NBER Working Paper No. 30578
October 2022
JEL No. E0

ABSTRACT

The County Business Pattern ("CBP") files contain employment and establishment counts for detailed industry codes covering all counties in the United States. The contribution of this project is to digitize, clean, and prepare the CBP files from 1946-1974. We also apply the methods developed in Eckert, Fort, Schott, and Yang (2020a) to impute missing employment observations in the raw data. We provide three digital data products for public use: (1) the cleaned CBP files for each year, (2) a consolidated panel data set of employment and establishment counts for about 20 industries and all US counties,(3) estimates for suppressed employment counts for each year.

Fabian Eckert
Department of Economics
UC San Diego
9500 Gilman Dr
La Jolla, CA 92093
f@fpeckert.me

Karsten Müller
National University of Singapore
15 Kent Ridge Dr
Singapore, Sing 119245
Singapore
kmueller@nus.edu.sg

Ka-leung Lam
Princeton University
jl88@princeton.edu

Rafael Schwalb
Department of Economics
Princeton University
schwalb@princeton.edu

Atif R. Mian
Princeton University
Bendheim Center For Finance
26 Prospect Avenue
Princeton, NJ 08540
and NBER
atif@princeton.edu

Amir Sufi
University of Chicago
Booth School of Business
5807 South Woodlawn Avenue
Chicago, IL 60637
and NBER
amir.sufi@chicagobooth.edu

Data for this paper is available at
https://www.dropbox.com/sh/xwrf9t3ci5bf5n6/AAAUrMfjSbNJ1u2ma_O-oNpva?dl=0

INTRODUCTION

The County Business Patterns ("CBP") is a data product by the U.S. Census Bureau that reports employment and establishment counts by industry and county. The CBP has been published for most years since 1946 and annually since 1964. The CBP has several advantages relative to other publicly available data sets: (1) it has rich industry-level information with close to 1,000 industries in many years, (2) it has granular spatial information with data reported on the county level, (3) it covers a long period of time at annual frequency, (4) the information is of administrative quality.[2]

However, despite these important advantages, the CBP data before 1975 had not been systematically digitized and made available for public use. This paper describes a multi-year effort to systematically digitize and clean the employment and establishment data in the "early" CBP files from 1946 to 1974. We provide three data products for public use. First, we make the raw CBP data available since 1946 in digitized format. Second, we present a harmonized panel data set featuring employment and establishment counts for 20 industries and about 3,000 counties from 1946 to 1974. Third, we provide "revised" data files in which we replaced suppressed employment counts in the original data files with imputed values.

Up to now, the CBP data files were available in digitized format only from 1969 onward.[3] We obtained physical copies of the CBP publications from 1946 to 1969 and digitized the data tables they contain using manual data transcription services. In the digitization process, we sought to minimize transcription errors by applying large-scale random error-checking, investigating outliers, and testing for data validity using the hierarchical nature of the data. We also cleaned the digitized files available from the National Archives for 1969 to 1974, since these data are not readily-usable, and added them to our dataset; Eckert et al. (2020a) provide the cleaned data from 1975 onward.

The cleaned and digitized files contain suppressed cells, especially for more detailed industry codes. The US Census does not disclose the employment counts for these cells due to privacy laws. Recently, Eckert et al. (2020a) (EFSY henceforth) developed a technique to impute values for such missing cells in the CBP exploiting adding-up constraints implicit in the data's hierarchical structure; they then applied this technique to the previously available CBP data for 1975-2016. We employ the EFSY imputation algorithm to fill in suppressed employment counts in our earlier data. Together with the data in EFSY, our paper provides a county-industry panel that describes the

[2]The CBP data is a tabulated version of the tax data in the Census' Business register.

[3]The files are available from the websites of the National Archive for the years 1969 to 1986 and from the CBP website of the Census for the years after 1985. To the best of our knowledge, the pre-1975 CBP data have not been widely used in economic research.

changing industrial structure of US countires from 1946 to 2016. The printed CBP publications also contain information about quarterly payroll on the county level. We did not digitize the payroll data because our focus was on creating a consistent employment panel.

We benchmark the early CBP data to existing aggregate data series. In particular, we use our panel data to compute statistics on aggregate employment by industry. When aggregated across counties, our panel's national industry employment counts track those reported by the Bureau of Economic Analysis (BEA) and the Bureau of Labor Statistics (BLS) closely. The discrepancies that do exist are largely due to differences in the sample frames of the data sets.

In summary, we present the first county-level panel data set on the spatial-industrial structure of the US economy between 1946 and 1974. The raw annual data files, the imputed files, and the processed panel data are available on the authors' websites for public use.

I. Digitizing the Early CBP Data

In this section, we provide a brief general overview of the structure of the CBP data. We then describe the digitization, cleaning, and error-checking process we applied to the digitized files. Finally, we benchmark the employment counts in the CBP data to those from other trusted data sources. The technical appendix contains additional details on data construction and classification changes.

I.1 The Early County Business Patterns Files: An Overview

The CBP files are a data product published by the U.S. Census Bureau since 1946. The files are a collection of tabulated data from the administrative records of all private, non-farm employer establishments in the United States. Before 1962, the Census Bureau obtained these records from the old-age and survivors insurance program of the US Treasury. Thereafter they come from Form 941 of the US Treasury and the data is supplemented with a Census-run survey for multi-establishment firms.

The CBP files are available annually between 1946 and 1951, for the years 1953, 1956, 1959, and 1962, and again annually from 1964 onward. The CBP files record total employment during the week of March 12 and first quarter total payroll for each county and industry in the United States. In addition, the files contain establishment counts by industry, county, and establishment size class. The U.S. Census Bureau also published separate files containing employment and establishment counts on the state and

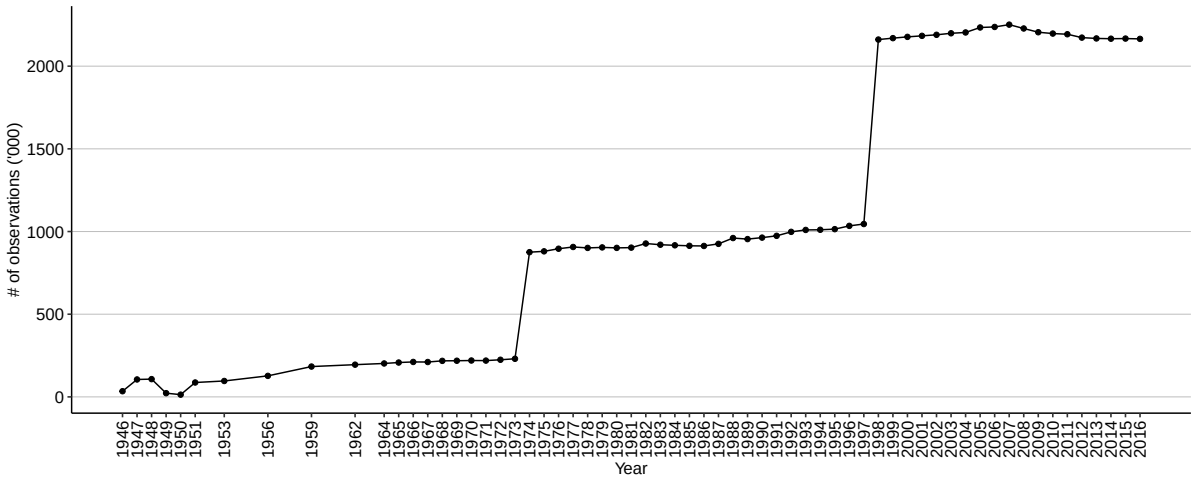


FIGURE 1: Number of observations

Notes: This figure plots the total number of observations in the CBP dataset for 1946-1974 described in this paper combined with the data for 1975-2016 described in Eckert et al. (2020a). The jump in 1974 is due to Census reporting previously omitted small county-industry combinations starting in 1974, while the jump in 1997 is due to the switch to the NAICS classification system.

national level, again by industry, which we provide for almost all years [4]. Our digitization effort focused on employment and establishment counts, which we provide in all our files. [5]

Figure 1 plots the number of observations in the CBP over time. The increase in the number of observations reflects several factors. First, the CBP added more detailed industry identifiers over time. Second, initially the CBP lumped together small counties into "county groups" and reported employment for them but then moved to reporting for individual counties over time. Third, as the US economy grew over time, the number of industries with non-zero employment in each county expanded.

Changing industry identifiers are an important part of the CBP data product. Figure 2 shows the industry classification system used in each year, and the number of industries in each data file. The number of industries reported changes discretely whenever industry classification systems change. The most fundamental change was the move from SIC to NAICS reporting between 1998 and 1999 (see Fort, Klimek et al. (2019)).

In the following, we restrict ourselves to a description of the "early CBP files" between 1946 and 1974. EFSY present a discussion of the CBP files after 1974. Table 1 provides a more detailed overview of the codes used in the early CBP files which are the focus of this paper. The early CBP files use different vintages of the Standard Industrial Classification (SIC) codes to index industries; between 1946 and 1948, non-manufacturing in-

[4] The exceptions are 1946 and 1970-1974.

[5] We also provide payroll information for a small subset of counties before 1970 and for all counties past 1970, as well as for all state and national files.

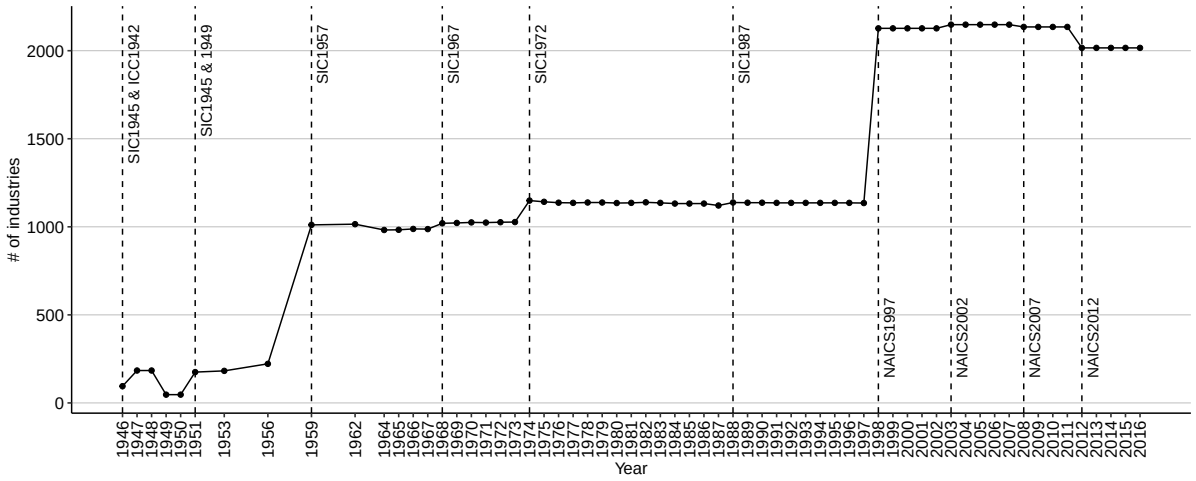


FIGURE 2: Number of industries

Notes: This figure plots the number of industries in the CBP files for 1946-1974 described in this paper and the 1975-2016 data described in Eckert et al. (2020a). Vertical dashed lines indicate the timing of major industry classification changes.

dustries were classified using the Social Security Industrial Classification Code (ICC), which are similar to the SIC classification.

We first describe features that all SIC vintages have in common. First, all CBP files (county, state, or total US) contain total employment for each spatial unit they cover. These totals are not assigned an industry code, since they represent aggregates, and we assign them the code "- - - -." Second, all CBP files list employment for each spatial unit in the following "industries:" Agriculture, Forestry & Fishing; Mining; Construction; Manufacturing; Transportation & Public Utilities; Wholesale Trade; Retail Trade; Finance, Insurance & Real Estate; Services; Public Administration; Non-classifiable Establishments. These industries are not assigned industry codes in the early files. We hence assign the codes used in the post-1985 CBP files provided on the Census Bureau's CBP website. These codes are, in the same order: 07--, 10--, 15--, 20--, 40--, 50--, 52--, 60--, 70--, 99--. Third, for each industry there are "subindustries" that provide more detail. Subindustries are indicated in the data with two digits and there are several subindustry codes for each "parent" industry. The two dashes in the end of the industry codes help distinguish them from the 2-digit subindustry codes which do not end in dashes. For example, "15--" indicates the "Construction" industry of which "General Contractors" (code "15") and "Special Contractors" (code "17") are subindustries. Fourth, some of these 2-digit subindustries further have 3-digit subindustries which provide additional detail; 3-digit subindustries share the first two digits with the 2 digit industries they provide additional detail for.[6] Fifth, while 2-digit subindustry employment counts always add up to industry totals, the employment

[6]The same is not necessarily true for industries and their corresponding 2-digit subindustries.

counts of 3-digit subindustries do not necessarily add up to those of 2-digit subindustries. This is because the Census Bureau cannot always allocate all 2-digit employment to more 3-digit detailed subindustries.

In addition to these common features, there are some important differences between the SIC vintages used in the early CBP files. First, beginning in 1956, some 3-digit subindustries also feature 4-digit subindustries; 4-digit subindustries share the first three digits with the 3-digit industries they provide additional detail for. The employment counts of 4-digit subindustries do not necessarily add up to those of 3-digit subindustries. These 4-digit subindustry codes start appearing in the 1956 files, but only in wholesale and retail, and even there only selectively. After 1956, 4-digit subindustry codes become a prominent feature of all CBP files. Table 1 shows the total number of different industry codes in the data set. Years with earlier SIC vintages have fewer industry codes.

Second, before 1951, 3-digit subindustries contain a residual category that contains all employment that cannot be assigned to a precise 3-digit subindustry. The residual category code shares the first 2-digits of the parent 2-digit subindustry industry and ends in a "0." Summing across 3-digit subindustries and the residual category yields total employment of the parent 2-digit subindustry. From 1951 onwards, such residual categories no longer exist.

Third, before 1951, there are no employment counts provided for 2-digit industries that have 3-digit subindustries. However, since the 3-digit subindustries include a residual category (ending in 0), they can be added up to retrieve total employment of the 2-digit parent industry. When the residual categories disappear in 1951, employment counts for industries start to be listed.

Fourth, most industries have a 2-digit subindustry called "Auxiliary and Administrative" which has the code "- -" or "...", and contains employment in auxiliary establishments of that industry not directly involved in the industry, e.g., it contains employment in headquarter establishments of manufacturing firms. Occasionally auxiliary codes inherit the code of the corresponding industry but instead of ending in "- -" they end in "\\". We use the following codes for the auxiliary establishments for each industry: agricultural (098/), mining (149/), construction (179/), manufacturing (399/), transportation and utilities (499/), wholesale trade (519/), retail trade (599/), finance and banking (679/), services (899/).

Lastly, the CBP files for 1948 and 1949 only contain manufacturing employment. The CBP program suspended collection of employment in other industries in those years due to a funding shortfall.

Figure 3 plots the number of counties covered in the CBP data and the number of

counties reported by the US Census for each year of existence of the CBP. The top panel shows the total number of counties for all available states. The bottom panel shows the number for states on the US Mainland (48 States and District of Columbia) excluding Alaska, Hawaii and overseas territories. Unidentified counties (i.e., those assigned FIPS999) are excluded from this calculation.

Table 1 shows the precise number of counties in each year of the early CBP files. The CBP files prior to 1964 did not contain data for all counties. Instead, they aggregated smaller adjacent counties into so-called county-groups until they reached a certain threshold of total employment. After 1964, the CBP files contain data on *all* U.S. counties. The number of counties in the United States has increased over time (see Eckert, Gvirtz, Liang, and Peters (2020b)).[7] The increasing number of counties in Table 1 reflects both of these developments.

The early CBP files have the following structure. In most years, the CBP release contains separate files for national, state-level, and county-level data. Each file reports separate employment counts for each SIC industry and its 2-, 3-, and 4-digit subindustries for its respective spatial unit. We refer to the combination of a spatial unit (e.g., a county) and an industry or subindustry code (e.g., Construction ("15--")) as "cells." An important difference to the employment listed for a given cell in the early CBP compared to later files concern multi-plant manufacturing firms. Prior to 1974, employment of multi-plant manufacturing firms was reported at the location of the largest establishment of that firm. As a result, county-level employment numbers may not always be representative of the actual employment within that county and year.

Table 2 provides an excerpt from the 1956 county file for illustration. The table shows employment and establishment counts for Arkansas County in the state of Arkansas. The county had a total of 3,514 employees in 1956, of which 79 worked in Agriculture. For the construction industry, the data contains more detail: it shows that 41 of the 545 construction workers were general contractors. Importantly, summing employment counts for more detailed industries does not always yield the employment total of less detailed industries. For example, summing counts for Plumbing and Concrete Contractors does not yield the total number of Special Contractors. As discussed above, this is because the Census omits workers that it cannot allocate to more detailed industry counts for 3- and 4-digit subindustry codes. In other words, at each level of aggregation there is a "shadow" residual category which contains the workers that cannot be allocated at that level of aggregation. This number of "residual" workers grows larger the finer the level of industry aggregation considered. Summing all em-

[7] New counties result from splitting up earlier counties, which means that the geographic boundaries of the some of the spatial units in the data change. Eckert et al. (2020b) provide a crosswalk to create a panel of counties with constant boundaries. Also see Eckert, Lam, Mian, Müller, Schwalb, and Sufi (2022b) for a discussion of boundary changes in the CBP.

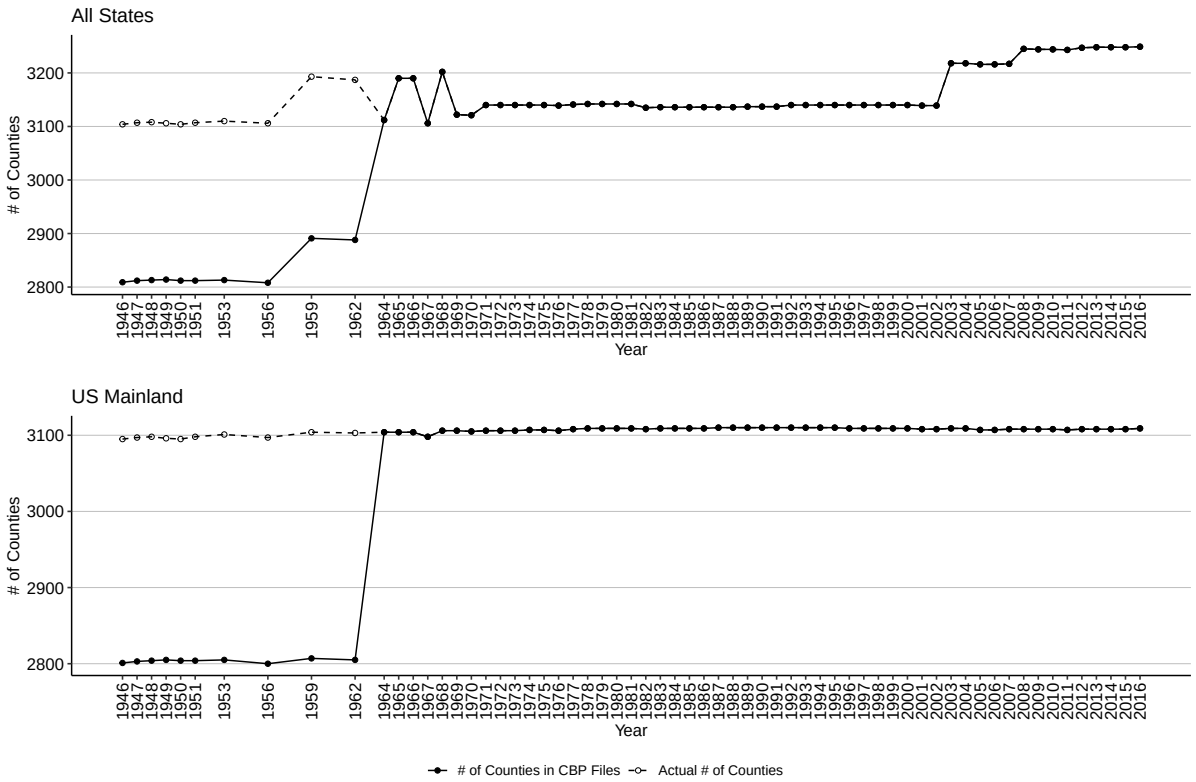


FIGURE 3: Number of county observations

Notes: This figure plots the count of counties appearing in the CBP files and the actual number of counties (according to the US Census) for each year. These counts are based on the CBP data described in this paper for 1946-1974 and the data in Eckert et al. (2020a) for 1975-2016. The upper panel shows the total number of counties for all states. The lower panel only reports the number for the mainland US (48 states and the District of Columbia), excluding Alaska, Hawaii and overseas territories. Unidentified counties (i.e., those assigned FIPS999) are excluded from this calculation.

TABLE 1: FILE OVERVIEW

Year	Industries	Counties	SIC Vintage	Observations	Industry Detail
1946	95	2,809	1945†	34,256	2-digit SIC
1947	184	2,862	1945†	105,352	3-digit SIC
1948	184	2,850	1945†	107,295	3-digit SIC
1949	47*	2,864	1947	22,455	3-digit SIC
1950	47*	2,826	1947	13,556	3-digit SIC
1951	175	2,861	1947	86,997	3-digit SIC
1953	182	2,863	1947	96,213	3-digit SIC
1956	223	2,822	1947	126,957	4-digit SIC
1959	1009	2,941	1957	181,986	4-digit SIC
1962	1016	2,938	1957	194,835	4-digit SIC
1964	982	3,162	1957	199,586	4-digit SIC
1965	983	3,240	1957	205,222	4-digit SIC
1966	989	3,223	1957	208,161	4-digit SIC
1967	990	3,151	1957	207,487	4-digit SIC
1968	1,020	3,244	1967	217,809	4-digit SIC
1969	1,029	3,156	1967	217,838	4-digit SIC
1970	1,025	3,160	1967	219,672	4-digit SIC
1971	1,024	3,160	1967	219,444	4-digit SIC
1972	1,026	3,190	1967	224,316	4-digit SIC
1973	1027	3,190	1967	230,401	4-digit SIC
1974	1147	3,190	1972	875,094	4-digit SIC

Notes: Cells are not suppressed in 1946. † indicates that, between 1946 and 1948, non-manufacturing industries are classified using the 1942 Social Security Industrial Classification Code (ICC). * indicates that, in 1949 and 1950, the CBP only report data on manufacturing industries. Before 1964, counties refers to county groups. Between 1964 and 1967, the CBP also uses the 1963 SIC, which supplements the 1957 SIC. See B and the text for more details.

ployment in Arkansas County for 2-digit subindustry codes yields the corresponding industry total stated under the associated industry code ending in "--," since the Census can always allocate workers to these broader, 2-digit industry codes.

Table 2 also shows the establishment counts contained in the data set. The CBP files provide the total number of establishments in each cell, overall and by establishment size class. However, similar to the employment counts, the Census cannot always match establishments to finer industries: there are 46 construction establishments in Arkansas county, but only 42 of them can be assigned to the general or special contractor category.

Table 2 also highlights that employment counts below a certain threshold are "suppressed," i.e., not reported, which we indicate by "NaN." Suppression occurs by a law that prohibits the U.S. Census Bureau from publicly releasing data that might disclose the operations of an individual firm (US Code, Title 13, Section 9). So cells with too few establishments in them have their associated employment counts suppressed, since else employment could be attributed to individual establishments. However, establishment counts themselves are never suppressed. The establishment counts by establishment size class can be used to construct an upper and lower bound on the employment for each suppressed cell.[8] An upper bound is given by multiplying the upper bound of each establishment size bracket by the number of establishments in that bracket, and vice versa for the lower bound. Table 3 shows the employment bounds implied for each suppressed cell; for non-suppressed cells the bounds coincide and are equal to the reported employment count.

In addition to the cell-specific bounds, the hierarchical structure of the files implies restrictions on the employment counts in suppressed cells: within a given county, summing employment counts across 4-digit industry cells has to yield an employment count weakly lower than that of the corresponding 3-digit parent industry, which is reported separately. For instance, in Table 2, summing the employment counts for Concrete and Plumbing Contractors has to yield a number smaller than 448, which means there cannot be more than $438 - 30 = 408$ Concrete Contractors in Arkansas. Likewise, summing employment in an industry across counties within the same state has to yield the employment reported for that state and industry cell in the state files. EFSY show how to exploit these bounds and hierarchical restrictions to impute suppressed employment numbers. Below, we discuss how we apply their technique to the early CBP data.

[8] An important difference between the CBP files before and after 1974 is that, in the early files, the Census did not directly report employment bounds for suppressed cells. This means we need to use the establishment size distribution to identify them.

TABLE 2: EXCERPT FROM THE 1956 CBP FILE

					Establishment Size Class (#Emp)							
State	County	SIC	Emp	Est	1-3	4-7	8-19	20-49	50-99	100-249	250-499	500+
AR	Arkansas	- - - -	3541	465	276	86	70	22	7	4	0	0
AR	Arkansas	07- -	79	9	6	0	1	2	0	0	0	0
AR	Arkansas	10- -	NaN	1	0	0	0	1	0	0	0	0
AR	Arkansas	15- -	545	46	29	9	4	2	0	2	0	0
AR	Arkansas	15	41	10	4	6	0	0	0	0	0	0
AR	Arkansas	17	448	32	23	3	3	1	0	1	0	0
AR	Arkansas	171	30	10	8	1	1	0	0	0	0	0
AR	Arkansas	177	NaN	2	0	0	0	1	0	1	0	0
AR	Arkansas	20- -	940	32	11	7	2	6	4	2	0	0

Notes: This table shows an excerpt of the first ten rows from the County Business Patterns county file from 1956.

I.2 Digitization Process

The CBP was originally a physical publication similar to the Statistical Abstract of the United States. The publication contained the tables with the employment count data described above. Today, the CBP website of the US Census Bureau only offers digitized version of the CBP files since 1986.[9] The study by Eckert et al. (2020a) used earlier digitized data for the period 1975-1986 from the National Archives. For the period before 1969, neither the National Archive nor the CBP website provides digitized data at the time of writing.

We obtained copies of the physical publications of the CBP for the years 1947-1970 from the Princeton Library, its system of partner libraries, and the Hathi Trust website.[10] The data from 1971 onward were obtained from the National Archives; the 1946 file comes from the website of the Inter-university Consortium for Political and Social Research (ICPSR).[11] While the original CBP publications contain additional information on quarterly payroll by industry, we only digitized employment and establish-

[9] See https://www.census.gov/programs-surveys/cbp.html.

[10] The 1946 and 1947 publications were titled "Business Establishments, Employment and Taxable Pay Rolls Under Old-Age and Survivors Insurance Program," instead of County Business Patterns, but contained the same type of tabulate data.

[11] The National Archives data does not contain the state and national files, which limits our ability to impute employment counts for missing cells. We plan on digitizing the 1971-1974 national and state files in future data releases.

TABLE 3: ESTABLISHMENT COUNT-IMPLIED EMPLOYMENT BOUNDS

State	County	SIC	Industry	Emp (LB)	Emp (UB)
AR	Arkansas	- - - -	Total	3514	3514
AR	Arkansas	07- -	Agriculture	79	79
AR	Arkansas	10- -	Mining	20	49
AR	Arkansas	15- -	Construction	545	545
AR	Arkansas	15	General Contractors	41	41
AR	Arkansas	17	Special Contractors	448	448
AR	Arkansas	171	Plumbing	30	30
AR	Arkansas	177	Concrete	120	298
AR	Arkansas	20- -	Manufacturing	940	940

Notes: This table shows the employment bounds derived from the establishment size distributions in Table 2. When employment is not suppressed, the bounds coincide, otherwise they are the sum of lower and upper bounds of the establishment sizes, respectively.

ments counts.

To digitize the data, we first scanned the pages of the physical CBP books for every year between 1947 and 1970. We then employed a data entry firm to manually transcribe the data from the scanned files into Excel files. In transcribing, we used the double key method, meaning that each cell in our data set was independently transcribed by at least two different data entry workers who identified the same value for it.

Once the data were transcribed into Excel files in this way, we applied a battery of error-detection checks to check for errors induced in the transcription process.[12] Since the publication of the early CBP data involved manual transcription and typesetting of the data, the resulting publications likely contain typos. We did not change numbers if they were flagged by our error detection methods but appeared as such in the original CBP publications. In the next section, we discuss an algorithm developed by EFSY that can help detect and correct typos in the original CBP publications.

We use the raw data to make two data products available to researchers. First, the raw CBP data for 1946-1974 in digitized format. Second, a harmonized panel data set that features employment and establishment counts for 2-digit SIC industries and about 3,000 counties from 1946 to 1974 [13]. We take no responsibility for any remaining tran-

[12] A technical Appendix describing our error-detection methods is available on request from the authors.

[13] Before 1964, when small counties were sometimes grouped into county groups, we allocate employment from these county groups to the individual counties based on their share of the total employment

scription mistakes in these data. However, we would be grateful to users for help in identifying potential discrepancies or errors so we can address these in future releases.

I.3 Benchmarking the Data

We benchmark our county-level employment panel data set in three ways. First, in our panel data, we sum employment across counties to produce US-wide employment totals and industry-specific employment totals. We compare these series to the corresponding data published by the Bureau of Economic Analysis (BEA) and Bureau of Labor Studies (BLS). Second, we sum employment across industries and counties to produce employment series for US States and compare them to corresponding from the BLS. Third, we use our panel to study the industrial structure of three major American cities over time.

Industry Employment We use data on total employment by industry from the BEA and Current Employment Statistics on employees by industry published by the BLS.[14] Figure 4 shows the total employment series from all three sources, and Figure 5 shows the three series for each two-digit industry (data for Agriculture is only available from the BLS from 1976 onward).

For most industries, the employment series generally align well, and are highly correlated. In three industries substantial discrepancies occur: (1) Construction, (2) Transportation, and (3) Other Services. As a result of these industry-specific discrepancies, the aggregate totals in the three data sets differ, too.

To understand the discrepancies in these industries, note the following differences in these data sets' sample frames. First, the CBP excludes crop and animal production; rail transportation; Postal Service; pension, health, welfare, and vacation funds; trusts, estates, and agency accounts; office of notaries; private households; and public administration. The CBP also excludes most establishments reporting government employees. Second, the BEA data is collected in May, while the CBP data is reported for March of each year. The BLS data is collected monthly, and we choose the March series for comparability.[15] The differences in transportation and other services, which contains government workers, are due to the CBP not including these industries. The differ-

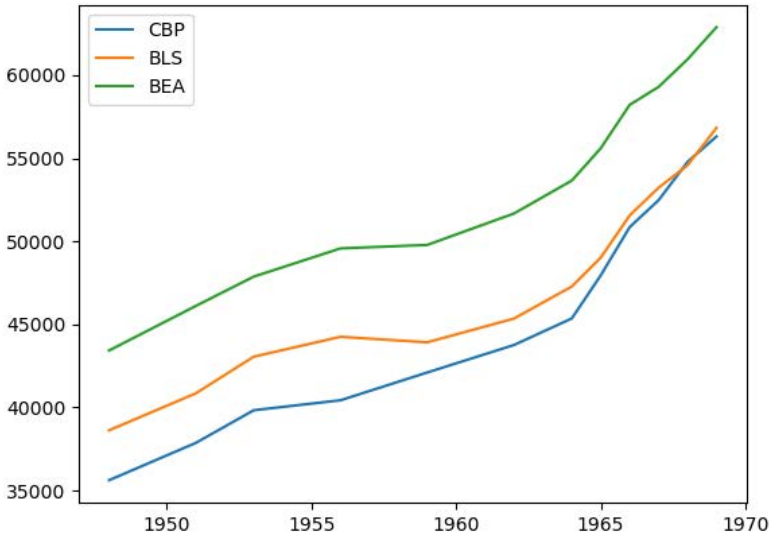
in the county group in 1964.

[14]The BEA table is available at https://www.bea.gov/data/employment/employment-by-industry. Specifically, we use Table 6.4 in the NIPA section on full-time and part-time employees by industry. The BLS data can be found at https://www.bls.gov/ces/data/, specifically the series EEU00500001: All total private employees, not seasonally adjusted.

[15]The pre-1959 CBP data also contains employment on ocean-borne vessels; the BEA and the BLS do not. However, this does not seem to be large enough to meaningfully affect the series, because the CBP data series show no structural break in 1959.

FIGURE 4: Total Employment Comparison to BLS and BEA data. Correlations reported in parenthesis

(A) Total (BLS: 0.996, BEA: 0.997)



Notes: This Figure shows the total private employment (in thousands) from the CBP and the two external data sources (sources listed in footnote 14).

ences in the construction sector between BLS and CBP relative to the BEA appear to be explained by seasonality in construction activity combined with the different sample times of the data sets.

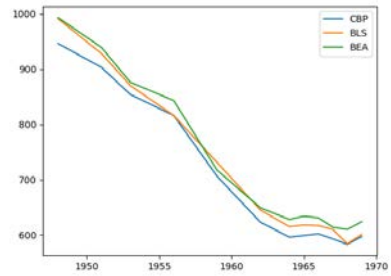
State Employment The most spatially disaggregated data made available by the BLS for this period is for US States.[16] We hence aggregate our county panel to the level of states, and compare the resulting employment counts across data sets. In particular, we compute the percentage deviation from the state employment count provided by the BLS. To avoid showing numbers for each state individually, we compute the mean deviation across states within each Census Region. In Figure 6, we plot these mean deviation for each Census Region. We also show the nationwide mean deviation.

Overall employment counts are routinely higher in the BLS data. The most likely reason for this discrepancy is that the CBP does not include government employment, whereas the BLS data does. To confirm, we compare the deviation for the District of Columbia (DC) with its high proportion of government workers to that of all the other Census Regions. Figure 7 plots the deviation between the BLS and the CBP series for DC as well as the average across all other states. DC exhibits a much higher deviation

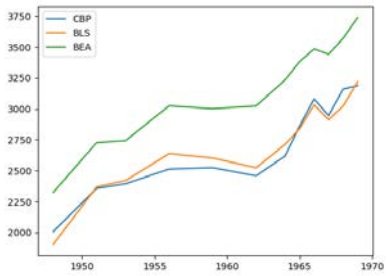
[16]Source: https://www.bls.gov/sae/data/. Neither the BLS nor the BEA publish county-level employment numbers for the period of the early CBP files.

FIGURE 5: Industry Comparison to BLS and BEA

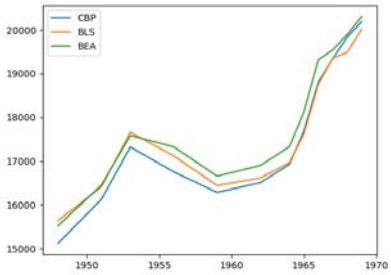
(A) Mining (BLS: 0.997, BEA: 0.998)



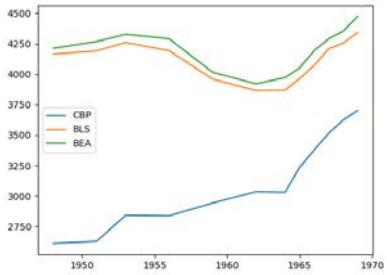
(B) Construction (BLS: 0.977, BEA: 0.977)



(C) Manufacturing (BLS: 0.995, BEA: 0.997)

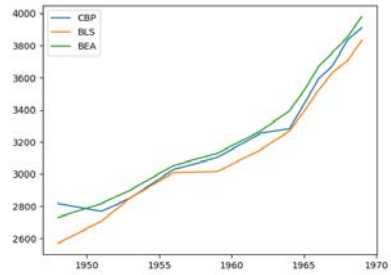


(D) Transportation (BLS: 0.234, BEA: 0.325)

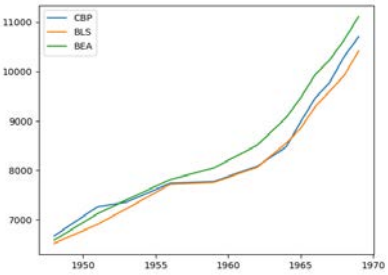


(E) Wholesale (BLS: 0.986, BEA: 0.994)

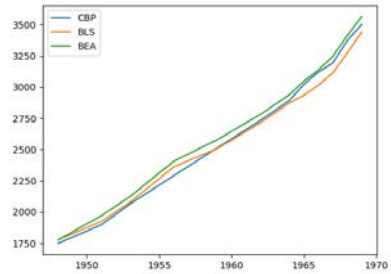
(F) Retail (BLS: 0.995, BEA: 0.993)

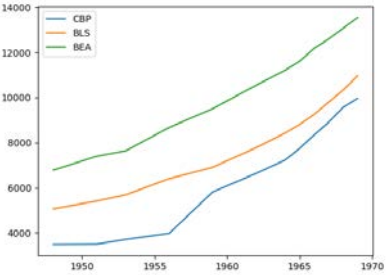


(G) Finance (BLS: 0.998, BEA: 0.999)



(H) Service (BLS: 0.991, BEA: 0.990)





Notes: The figures shows employment (in thousands) from the CBP and the two external data sources for different industries (sources listed in footnote 14. Correlation between the BLS and BEA data series with the CBP reported in parenthesis, respectively.)

of the BLS data compared to the CBP, which provides some suggestive evidence that the differences are in fact driven by government employees.

The Changing Industrial Structure of San Francisco, Houston, and Detroit, 1946–1974 Figure 8 shows time series of employment and employment shares for three well-known counties in the United States between 1946 and 1974: (1) Harris county in Texas, which includes Houston, known for oil production, (2) Wayne County in Michigan, which includes Detroit, representative of mid-western manufacturing counties, and (3) San Francisco County, which includes San Francisco, as a representative of a large city economy with skilled service employment.

The figures show smoothly evolving employment counts over time. Further, the industrial composition of these counties accords with intuition. San Francisco, a large city, has more service than manufacturing employment. The local manufacturing industry consists mainly of the food-producing sector, as expected from an urban economy. Harris County was growing fast following World War II and saw an expansion in all sectors of employment. It has more manufacturing than San Francisco and, within manufacturing, petroleum and chemicals are disproportionately important. Finally, Wayne County – driven by Detroit – has relatively stable overall employment with a majority of workers in manufacturing for most of the years in the sample. Reassuringly, we see the a much larger share of electric machinery, which contains cars, in local manufacturing relative to San Francisco.

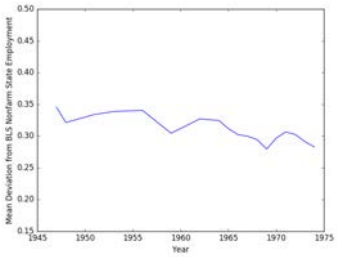
II. Imputing Missing Values

We now turn to addressing the data suppression in the early CBP data. Figure 9 shows the extent of suppression for each year of the newly digitized data. The left panel shows the absolute number of cells with suppressed employment counts, the right panel shows the share out of the total number of cells in the data that year. In the early CBP data, an average of 25% of cells are suppressed each year.

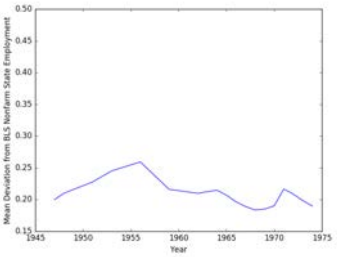
Table 4 further shows that most of the suppression occurs in the county-level files and at the most detailed industry level. The Table shows the average fraction of suppressed cells for benchmark years for the national, state, and county files, separately. For the county-level data files, we also show the fraction of suppressed cells for industries and 2- , 3- , and 4-digit subindustries, separately. In 1967, the suppression rate for counties is 0.27, for states 0.15, and for the national files zero. Table 1 above showed an increase in the number of industries and the number of spatial units reported over time. Since these additions added more disaggregated cells with higher suppression rates, the overall suppression rate shows a slight upward trend between 1947 and 1974.

FIGURE 6: State Total Deviations from BLS, All States and by Census Division

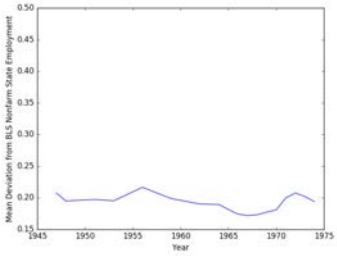
(A) All States



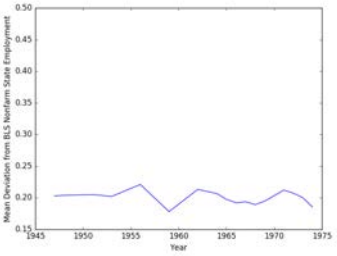
(B) New England



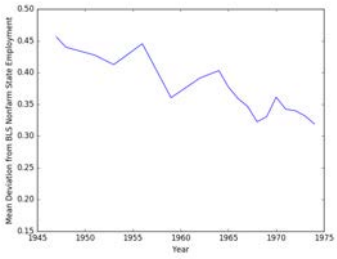
(C) Middle Atlantic



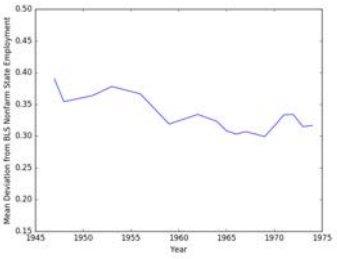
(D) East North Central



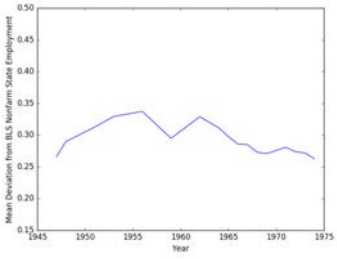
(E) West North Central



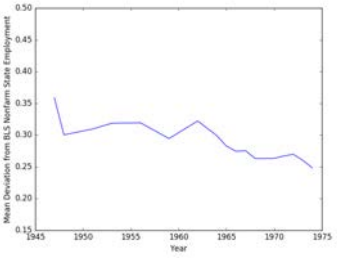
(F) South Atlantic



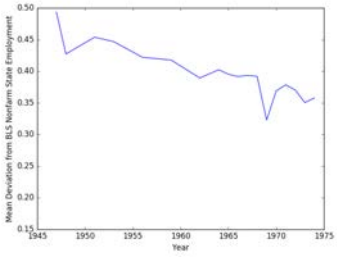
(G) East South Central



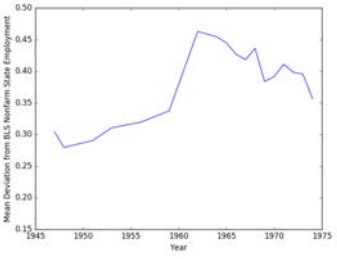
(H) West South Central

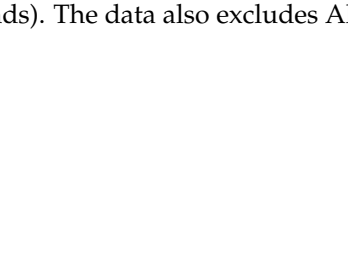


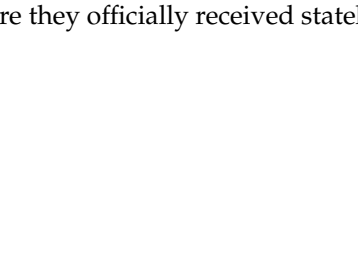
(I) Mountain



(J) Pacific

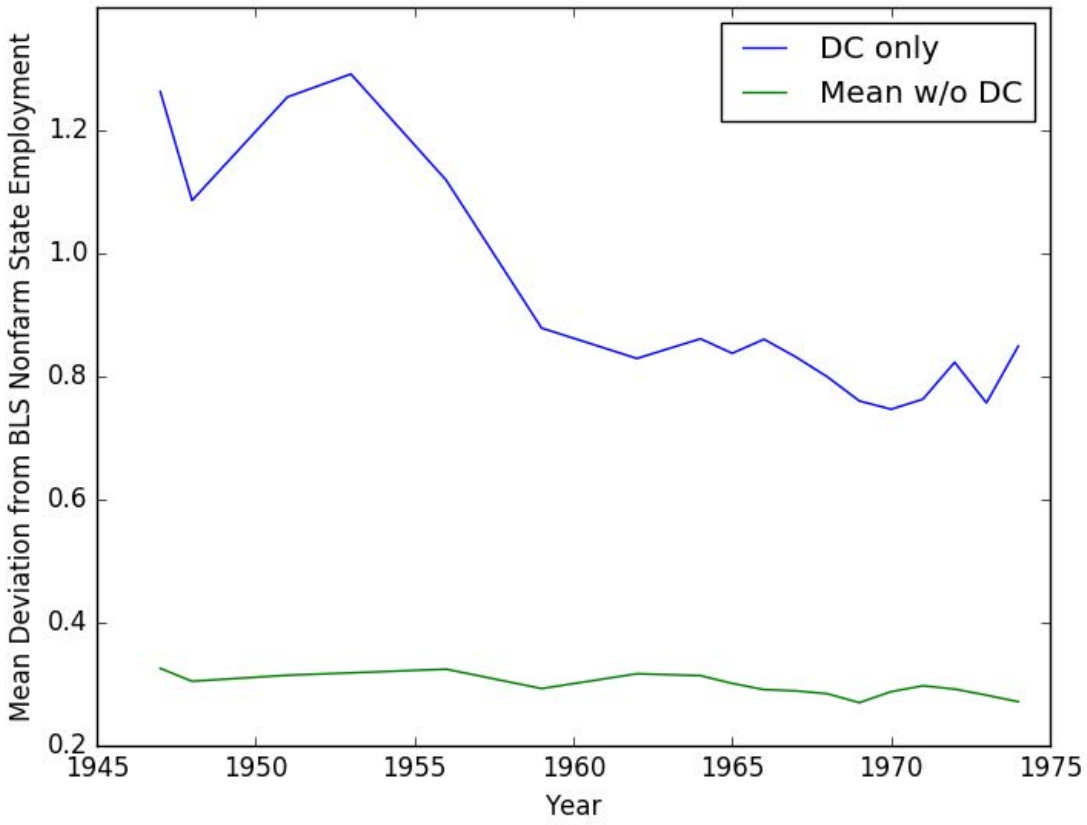






Notes: Mean Deviation from BLS Non-farm State Employment. The BLS series is systematically higher as it includes government employees (BLS state data on private employees only begins after our sample ends). The data also excludes Alaska and Hawaii before they officially received statehood

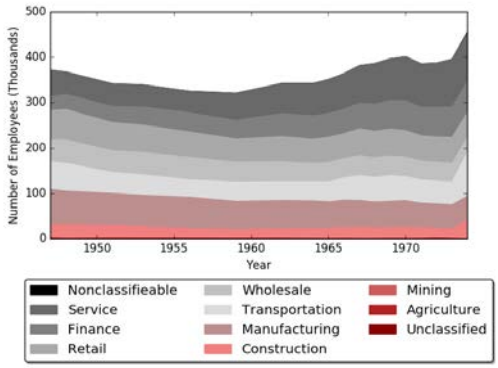
FIGURE 7: State Total Deviations from BLS, DC vs. Mean of All Other States



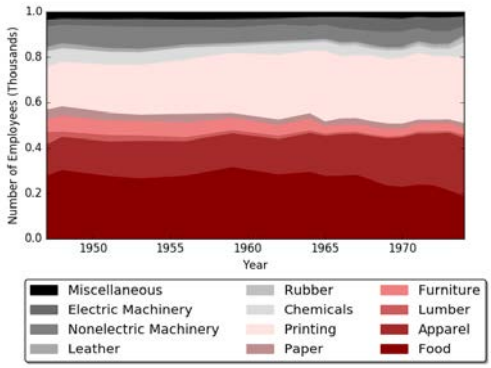
Notes: Mean Deviation from BLS Nonfarm State Employment, separately for the District of Columbia and the mean of all other states.

FIGURE 8: Breakdown of Total Employment and Manufacturing Employment in three sample counties

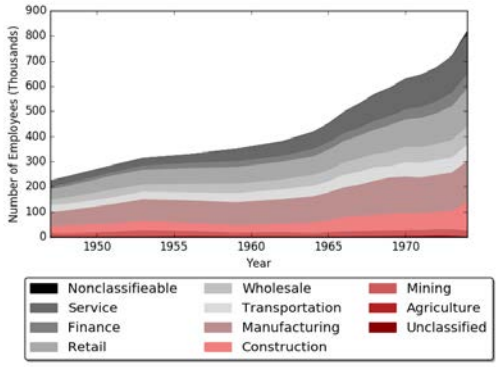
(A) San Francisco County (CA), Total Employment



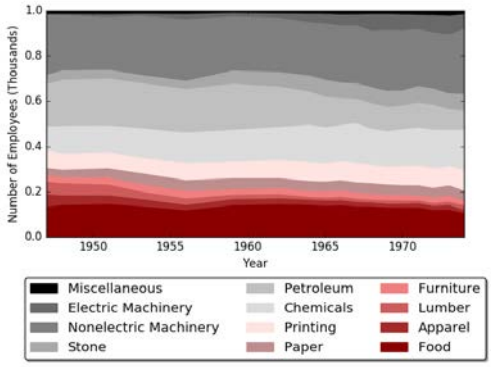
(B) San Francisco County (CA), Manufacturing Shares



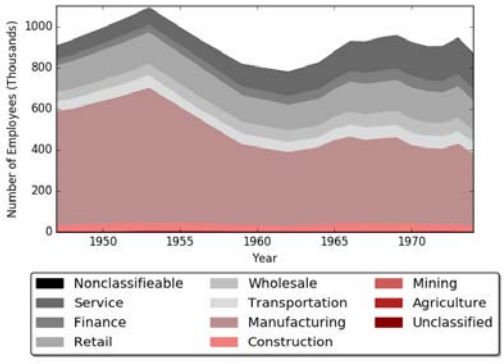
(C) Harris County (TX), Total Employment



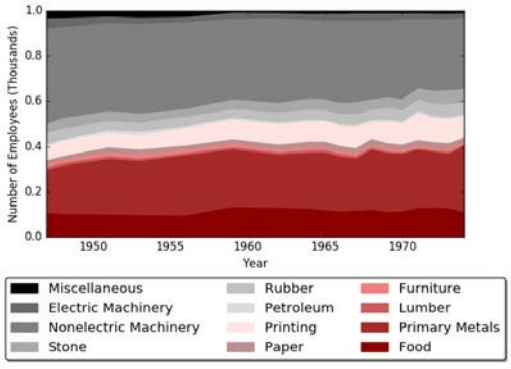
(D) Harris County (TX), Manufacturing Shares



(E) Wayne County (MI), Total Employment



(F) Wayne County (MI), Manufacturing Shares



Notes: Raw data (pre-imputation). The right graph excludes industries auxiliary to Manufacturing, as well as sub-industries with suppressed cells, so total slightly lower than Manufacturing employment total.

TABLE 4: FRACTION OF SUPPRESSED CELLS IN VARIOUS CBP FILES

	1947	1959	1967	1974
By Geographical Aggregation				
County	0.18	0.20	0.27	0.66
State	0.25	0.10	0.15	N/A
National	0.01	0.00	0.00	N/A
By SIC Aggregation (County-File)				
Industry	0.13	0.16	0.19	0.22
2-digit	0.30	0.09	0.15	0.46
3-digit	0.14	0.22	0.29	0.67
4-digit	N/A	0.37	0.47	0.81

Note: This table shows the fraction of cells suppressed in various aggregations of the County Business Patterns data county files for 1947, 1959, 1967, and 1974. In 1947, the suppression share is lower on the county than on the state level because most counties only report numbers for the coarsest industry classification, which are rarely suppressed. Likewise, 3-digit codes exhibit a lower suppression rate than 2-digit codes since they are only reported for large counties. The drop between 2- and 3-digit codes in 1959 and 1967 is due to similar selection effects: only larger county-industry pairs provide the dis-aggregation into 2-digit codes and beyond. *NA* means not available. We do not have state and national aggregates for 1974 or 4-digit SIC detail for 1947.

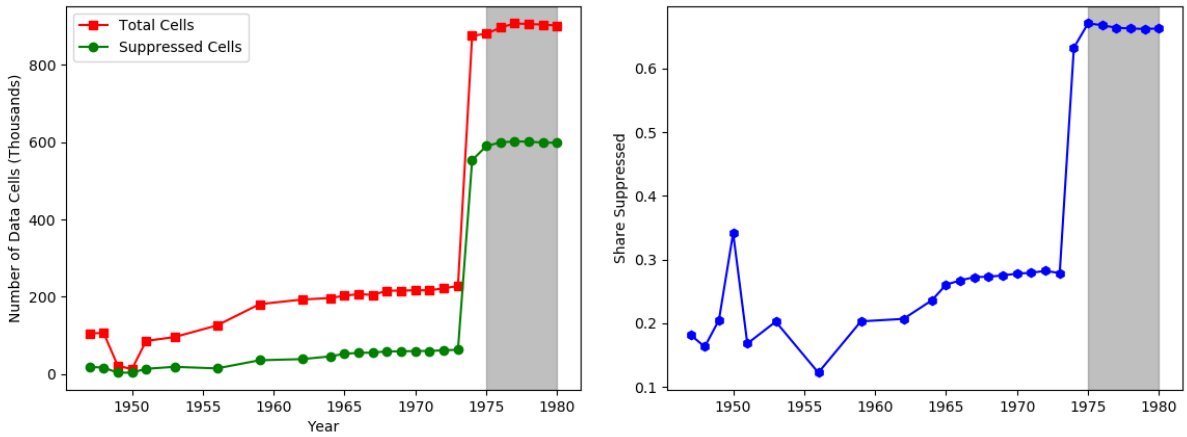
To understand Table 4, it is important to understand two competing effects when adding more industrial or geographical detail. First, mechanically, when adding more detail, more cells are added. On average, these additional cells contain less employment and are hence more likely to be suppressed. Second, small industries may appear more rarely in more disaggregated files. For example, suppose there is a single county with employment in industry X in the US. No other county has employment in industry X. In the national file, industry X appears and is suppressed. Since there are few cells overall in the national file, one additional missing cell raises the percentage of suppressed cells substantially. However, the corresponding county file contains many more cells overall. So the one additional suppressed cells has a smaller impact on the overall fraction of suppressed cells. As a result, the fraction of suppressed cells decreases as we move from industry to 2-digit subindustry codes.

Figure 11 graphs the share of cells suppressed in each county against a county's total employment. Suppression rates are highest in very small counties. Larger counties rarely have more than 20% of their cells suppressed.

After 1973, the Census started reporting data for almost four times as many cells as in the years prior. Since the additional cells are more disaggregated by definition, the suppression rate increased notably after 1973. Note that the suppression rate for

FIGURE 9: Suppressed County-Industry Cells

(A) Number of Cells (B) Share Suppressed



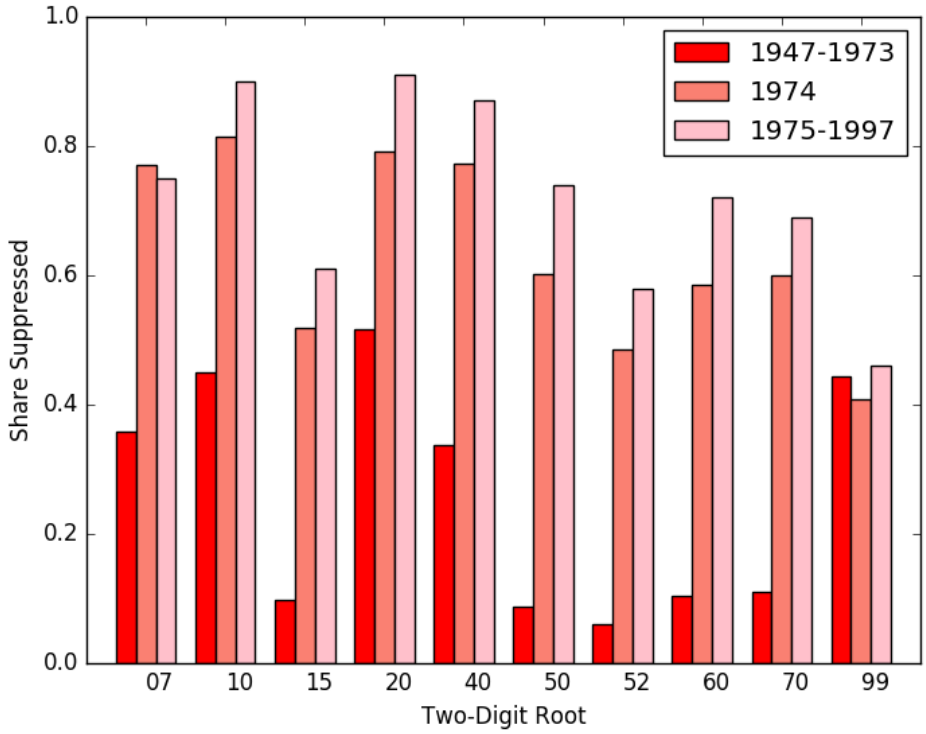
Notes: Data for years from 1975 on from Eckert et al. (2020a) is included with grey background for comparison. The jump in 1974 is due to Census reporting previously omitted small county-industry combinations starting in 1974.

1974 is in line with the rate reported in EFSY for 1975 onward, which we also plot for comparison.

We also look at suppression rates by industry. Figure 10 shows considerable differences across industries. Over the period from 1947 to 1973, the highest level of suppression occurs in manufacturing (SIC 20), mining (SIC 10), and other (SIC 99). The lowest level of suppression occurs in retail (SIC 52), construction (SIC 15), and Wholesale Trade (SIC 50). In general, non-tradable industries with small average establishment sizes have relatively low suppression rates because there are usually multiple such establishments in each county, which helps in clearing the suppression threshold. However, any given county tends to have fewer of the typically large establishments in industries that allow for economies of scale and trade such as manufacturing. As a result, the suppression rate in the data is high for such industries.

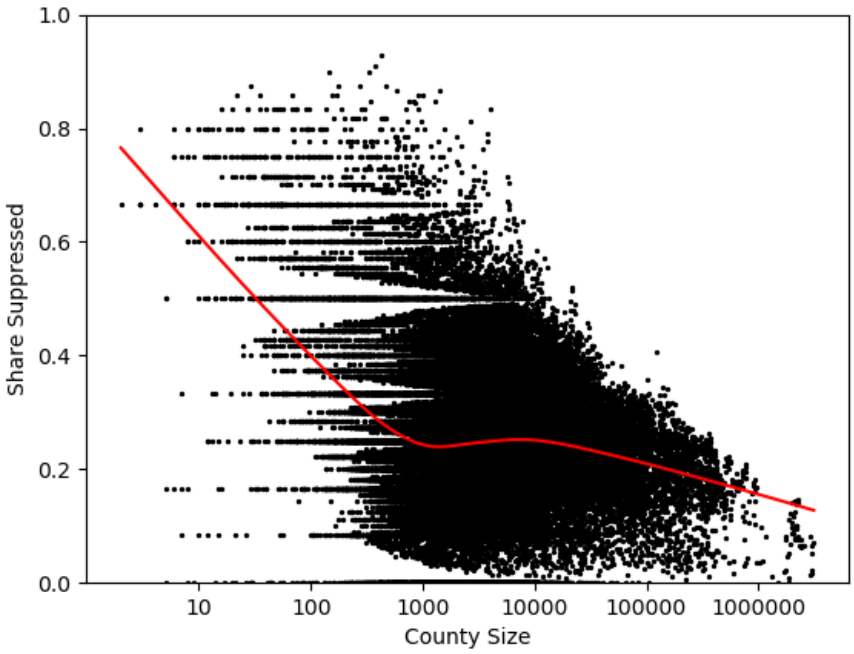
Overcoming the Suppression Problem The hierarchical structure of the CBP data makes it possible to infer values for the employment counts in the suppressed cells. There are three types of constraints on the employment count in a suppressed cell. First, the establishment size distribution in the CBP files allows us to derive a set of employment bounds for each suppressed cell. Second, for a given industry, county totals have to add up to state totals, which in turn have to add up to totals in the national file. Third, for a given spatial unit, e.g., a county, the most disaggregated industry employment counts have to add up to a number weakly smaller than the employment counts for the next more aggregated level of reporting, and so on.

FIGURE 10: Suppression by Two-Digit Roots



Notes: Suppression rates by two-digit industries for the years 1947-1973, 1974 and 1975-1997. 1974 is separate since its imputation rates are more similar to the years 1975-1997 covered in Eckert et al. (2020a). Qualitatively, the differences between high- and low-imputation industries prevail across the different periods, although in the later years imputation rates are more similar due to the higher granularity of the reported data in all county-industries. The only major qualitative difference is the root '99' covering non-classifiable industries, which declined in importance over the years as data quality improved.

FIGURE 11: County Size vs. Suppression Rate



Notes: The plot shows the suppression share across industries in each county, plotted against the county size (measured by total employment). Data from 1974 is omitted due to the level shift in suppression share shown in Figure 9, and 1949 and 1950 are omitted since they only cover manufacturing. The red line shows a LOWESS fit of the data, indicating that larger counties tend to have lower suppression shares, and that very high suppression shares exclusively occur in the smallest counties. The horizontal lines are due to the fact that for small counties, especially in the early years, there are only very few numbers of industries covered, and thus only a few discrete fractions of cells can be suppressed.

EFSY propose an imputation method that exploits these three sets of constraints systematically. They define an objective function that is the sum of the absolute distance of the value of each cell from the midpoint of its bounds. The imputation problem can then be cast as a linear program that minimizes this objective subject to three sets of constraints. We use the code provided by EFSY to implement their imputation technique in our data. We refer to their paper for a detailed discussion of the technique and its advantages and disadvantages.[17]

Relative to the later CBP data in EFSY, the early CBP data poses two problems when applying the imputation method. First, for the years 1971-1974, our data files lack the state and national files.[18] In such cases, the imputation method still works, but there are fewer implicit constraints on each cell, which makes the estimates less reliable.

Second, the early CBP files contain two types of errors: potential transcription errors from the CBP publications that occurred in our digitization process and transcription errors in the original CBP publications due to Census workers transcribing from the underlying data sources. Both types of errors imply that the adding up constraints the EFSY algorithm seeks to exploit may sometimes not hold. For example, in a given year, the total number of butchers across counties in New York State in the county file may be larger than the number reported by the corresponding state file. For such cases, EFSY propose an algorithm that makes the smallest possible adjustments to bounds in the data set so that all all adding up constrains hold. For example, lowering the lower bound on the number of butchers in New York County by 10 may bring the sum across counties to match the state total from the state files. Note that bound adjustments do not affect overall county employment since this number is never suppressed and as a result all county employment estimates add up to the correct county total, and hence also state and national totals.

The bound-adjustment algorithm can be viewed as an error-detection method: it finds inconsistencies and corrects them by assuming the transcribers made the smallest possible mistake. We investigated large adjustments (above 1 million) that were usually caused by data transcription errors, corrected the underlying errors, and re-ran the imputation. The remaining adjustments are likely due to inconsistencies in the CBP publications. In our imputed data set, we flag all cells for which bounds had to be

[17]The method proposed in EFSY leverages a simple insight: minimizing the sum of deviations from the mean of the bounds subject to a set of adding up and inequality constraints constitutes a large linear program. Industrial-scale linear programming solvers can solve such large problems effectively. We use the Gurobi linear program solver with a license provided by the Princeton University Computing Center.

[18]The state and national files are available in the printed version of the CBP. However, the files we obtained from the National Archive contain only the county files, not the state and national files. We plan to digitize the state and national files for these missing years and include them in future releases of our data products.

FIGURE 12: Adjustments over Time

Notes: The plot shows the time series of the fraction of adjusted county-industry cells and the share of total employment that the adjustments represent.

adjusted.

Figure 12 depicts both the frequency and magnitude of adjustments over time. Both the share and the magnitude were low in the initial few years. Since there were fewer industry codes, there were fewer adding up constraints that could be violated and hence adjusted. This is particularly true for the years 1949 and 1950, for which only manufacturing data is reported. As granularity increases with the inclusion of 4-digit SIC codes starting in 1956, the increased number of constraints is reflected in a higher frequency at which they have to be adjusted. Lastly, the adjustment rates and sizes are very low in the 1970s, when no independent state and national data is used. Thus, the only constraints are industry constraints within counties, which are mostly inequality constraints, and necessitate few adjustments.

This concludes the description of the third data product we provide on our websites: the imputed data files for each year, i.e., the output of applying the EFSY algorithm to the raw data files.

III. CONCLUSION

This paper presents newly digitized data from the US Census' County Business Patterns for 1946-1974. Together with the data in Eckert et al. (2020a), our data allows researchers to construct a consistent county-level panel of employment by industry

for 1946 until today. To the best of our knowledge, our dataset is unique in providing information on employment by industry and county for the period 1946-1974. We believe these data have many potential applications. For example, it could be used to deepen our understanding of the changes in the spatial industrial structure of the United States over the last 80 years. It also opens up new possibilities in analyzing major policy reforms with varying impact across industries and localities in the pre-1980 period. We hope the data we provide will be useful for exploring these and other topics going forward.

REFERENCES

ECKERT, F., T. C. FORT, P. K. SCHOTT, AND N. J. YANG (2020a): "Imputing Missing Values in the US Census Bureau's County Business Patterns," Working Paper 26632, National Bureau of Economic Research.

ECKERT, F., A. GVIRTZ, J. LIANG, AND M. PETERS (2020b): "A Method to Construct Geographical Crosswalks with an Application to US Counties since 1790," Tech. rep., National Bureau of Economic Research.

ECKERT, F., K.-L. LAM, A. MIAN, K. MÜLLER, R. SCHWALB, AND A. SUFI (2022a): "The Early County Business Pattern Files: 1946-1974," Tech. rep.

——— (2022b): "On County Boundary Changes in the Early County Business Pattern Files, 1946-1974," Tech. rep.

FORT, T. C., S. D. KLIMEK, ET AL. (2019): "The effects of industry classification changes on US employment composition," .

TECHNICAL APPENDIX
(NOT FOR PUBLICATION)

This technical appendix provides a detailed description of the structure of the database introduced in Eckert, Lam, Mian, Müller, Schwalb, and Sufi (2022a). Section A introduces the file system of the database and the definitions of main variables in the datasets. Section B introduces the county and industry classifications used in the CBP and their evolution over time. A separate technical note (Eckert et al., 2022b) details changes in county boundaries that can affect the time series of individual counties between 1946 and 1974.

A. Detailed Data and Variable Description

In this section, we introduce all available data files in the historical CBP database, including their structure and relationship to each other, as well as the definitions of the main variables in each dataset. Table A.1 presents the files in the database. The dataset consists of four parts: (1) a collection of raw CBP archive files, (2) ready-to-use files with estimated values based on the imputation algorithm we use, (3) a panel of 2-digit industries for all counties and years, and (4) county/industry classification files.

Section A.1 describes the raw and imputed county/state files and the variables they contain. Section A.2 describes the county/industry reference files accompanying each dataset, and how concordances are recorded in these files. Section A.3 introduces the industry classification files including various editions of official US Census industry classification standards and their concordance tables.

TABLE A.1: CBP Database File System

Archive	Dataset	Num	Description
CBP[year].zip	CBP[year].csv	21	County file.
	CBP[year]_Agg.csv	15	State/National file.
	CBP[year]_CountyList.csv	21	County reference file.
	CBP[year]_IndustryList.csv	21	Industry reference file.
	CBP[year]_Imp.csv	20	Imputed file.
Classification.zip	FIPS2018.csv	1	Official FIPS codes 2018.
	FIPS_Changes.csv	1	Historical changes of county boundaries, names and FIPS codes.
	ICC1942.csv	1	Industrial Classification Codes 1942.
	SIC[year].csv	5	Standard Industrial Classification codes.
	SIC1963_Supplement.csv	1	Standard Industrial Classification supplement file.
	Concordance_[IC1]_vs_[IC2].csv	2	Concordance table between Industrial Classification 1 and 2.
	Panel.csv	1	Panel of 2-digit industries for all years and counties

A.1 County and State Files

The main body of the database consists of 21 archive files, one for each year the CBP was published between 1946 and 1974. Each of these archive files in turn contains at least three data files in CSV format: a county file, a county reference file, and an industry reference file. In most years, there is also a state/national summary file (with the exception of 1946 and 1970-1974) and an imputed file (with the exception of 1946, when there was no suppression).

The county files and state/national summary files contain information on employment, payroll, and establishment counts by county/state and industry. In the following, we explain the main variables in these datasets.

- The **primary identifier** is a combination of county classification code and industry classification code (see details in Section B). County codes are a combination of the variables `fipstate` and `fipscty`, which are a 2-digit FIPS code designating the state and a 3-digit FIPS code designating the county, respectively. Industry codes are represented by the variable `sic`.

- **Employment count** (variable `emp`) is the number of employees during the pay period that includes March 12 of the year.

- **First-quarter payroll** (variable `qp1`) is the combined amount of wages paid, tips reported, and other compensation paid to employees at any time during the first quarter of the year before deductions for social security, income tax, insurance, union dues, etc. **Total annual payroll** (variable `ap`, only available in 1974) is the combined amount of payroll as defined above but covering the entire year.

- **Establishment count** Before 1974, the CBP tabulated **reporting units** (variable `rpunit`). In manufacturing industries, each manufacturing location of a company was counted as a separate unit, and hence "reporting units" are conceptually the same as "establishments." In non-manufacturing industries, however, employers (separate legal entities) are counted once in each county for each industry in which they operate, regardless of the number of establishments. Beginning in 1974, (variable `est`) is the number of establishments active in the fourth quarter of the year. An establishment is a single physical location where business is conducted.

- There are also several variables describing the **establishment size** distribution. For example, variable `n0_4` represents the number of establishments/reporting units with 0–4 employees during the mid-March pay period.

A.2 County and Industry Reference Files

The county and industry codes in the CBP do not always correspond perfectly to the official US Census county FIPS or SIC codes. Where they do not, we use a "FPN" system to record many-to-one matches. In particular, each record in the reference files is assigned a `type` variable, which contains the letters "F", "P" or "N". "F" stands for "full", "P" for "partial", and "N" for "none". The first letter of the `type` variable represents the proportion of one county/industry, while the second represents that of the other county/industry. For example, a "FP" concordance between county X and county Y means the entire county X is equivalent to part of county Y.

Examples. We show some examples from the county reference files in Table A.2. Variables `fipscty` and `ctyname` are the county code and county name reported in the CBP. `fipscty_ref` and `ctyname_ref` are the official 3-digit county FIPS code and county name from the US Census. `type` indicates the nature of each record, i.e., what fraction of CBP county observation corresponds to what fraction of the official county entity. The example records in Table A.2 can be interpreted as follows:

- In 1946, the CBP county observation Banks, Georgia (FIPS13011) corresponds exactly to the official US Census code for Banks County, Georgia (FIPS13011).
- In 1946, the CBP county observation Atkinson & Clinch, Georgia (FIPS13003,065) corresponds to two counties: Atkinson County (FIPS13003) and Clinch County (FIPS13065). See details in Section B.1.
- In 1947, the CBP defines a county Statewide, Georgia (FIPS13999) that does not correspond to any official FIPS code. See details in Section B.1.

Examples from the industry reference files are shown in Table A.3. `code_cbp` and `name` are the industry code and name reported in the CBP. `code_ref` is the official US Census industry classification code. `type` indicates the nature of each concordance. The example records in Table A.3 can be interpreted as follows:

- In 1946, the CBP industry *21 Tobacco Manufactures* corresponds exactly to official SIC industry 21.
- In 1946, CBP defines an industry *099 for Other Agriculture, Forestry and Fisheries*. More on this in Section B.2.
- In 1951, the CBP defines the industry *337 Primary Metal Industries (Nonferrous)*, which comprises four official SIC industries (333, 334, 335, and 336). More on this in Section B.2.

TABLE A.2: Excerpts of CBP county reference files

CBP	type	fipstate	fipscty	ctyname	fipscty_ref	ctyname_ref	stname
1946	FF	13	011	BANKS	011	BANKS COUNTY	GEORGIA
	PF	13	003,065	ATKINSON & CLINCH	003	ATKINSON COUNTY	GEORGIA
	PF	13	003,065	ATKINSON & CLINCH	065	CLINCH COUNTY	GEORGIA
1947	FN	13	999	STATEWIDE			GEORGIA

Notes: This table presents example records from CBP county reference files. Each record represents a concordance between CBP county observation and the official county classification. Variables `fipscty` and `ctyname` are the county code and county name used by CBP. Variables `fipscty_ref` and `ctyname_ref` are the official 3-digit county FIPS code and county name. Variable `type` indicates the nature of each concordance record.

TABLE A.3: Excerpts of CBP industry reference files

CBP	type	code_cbp	new_name	code_ref
1946	FF	21	TOBACCO MANUFACTURES	21
	FN	099	OTHER AGRICULTURE, FORESTRY AND FISHERIES	
1951	PF	337	PRIMARY METAL INDUSTRIES (NONFERROUS)	333
	PF	337	PRIMARY METAL INDUSTRIES (NONFERROUS)	334
	PF	337	PRIMARY METAL INDUSTRIES (NONFERROUS)	335
	PF	337	PRIMARY METAL INDUSTRIES (NONFERROUS)	336

Notes: This table presents example records from CBP industry reference files. Each record represents a concordance between CBP industry and the official industry classification standard. Variables `code_cbp` and `name` are the industry code and name used by CBP. Variable `code_ref` is the official industry classification code. Variable `type` indicates the nature of each concordance record.

A.3 Classification Files

In addition to the CBP data, we also collected the official industry classification codes used by the CBP and the concordance tables between the consecutive vintages of industry codes where available. We also tabulate historical changes of county boundaries, FIPS codes, and names. These files are archived in Classification.zip and are listed in Table A.1.

Industry Vintages. During the sample covered in this paper, the CBP has used several vintages of industry classification standards, including one ICC code and six SIC codes. The data files share the same structure and contain two variables, one for the numerical code (`icc` or `sic`) and the other for the industry name (`title`).

Concordances & Supplements. The concordance tables are constructed from information in the publication of each new edition of the industry classification standard or from the U.S. Census Bureau website.[19] We provide the following two concordance tables:[20]
SIC1957_vs_SIC1967,
SIC1967_vs_SIC1972

The nature of these changes is recorded using the `type` variable, similar to the county/industry reference files (see Section A.2). Besides the major revisions, there have been minor amendments to the SIC through the supplement edition SIC1963_Supplement (to amend SIC1957). This supplement data file thus has a similar structure to the other concordance tables. Note that these concordance tables and supplement files only record industries with at least one change in their code, name or content. We omit industries that did not experience any changes.

Examples. Some examples from the industry concordance or supplement files are shown in Table A.3. `old` and `new` are the old and new editions of the industry classification codes, respectively. `new_name` exists only in supplement files to show industry names after amendments. `note` contains notes about the changes. For example, the new edition of SIC1967 contain the following changes compared to SIC1957:

- An old industry code 3619 was deleted.
- A new industry code 4619 was created.
- Part of industry 0119 was split apart to form a new industry code 0114, while the remaining part maintains the original code.

[19] https://www.census.gov/naics/

[20] Concordances for later SIC and NAICS vintages are provided by EFSY. The BEA also provides concordances for the earlier SIC industry classifications before 1957.

- Industry 1982 was combined into 1081.

Amendments made by SIC1963_Supplement on SIC1957 contain the following:

- Deletion of industry code 3619.
- Creation of new industry code 4619.
- Part of industry 3651 changed to 3679 while the remaining kept the original code.
- Part of industry 3679 split apart to form a new industry code 3674, while the remaining part extended to incorporate part of 3651.

The FIPS change file tabulates changes of county boundaries, names, and FIPS codes. Each entry is also recorded using the "FPN" system. We plot some examples from these files in Table A.4. `date` is the date on which the change became effective. `old`, `old_name`, `new` and `new_name` are the old/new county FIPS code and name. Variable `cbp_year` is the year when the CBP incorporated the change. See Section B.1 and Eckert et al. (2022b) for additional details. The example records in Table A.4 can be interpreted as follows:

- On Dec 15, 1979, Ste. Genevieve County, Missouri changed FIPS code from 29193 to 29186. The CBP incorporated this change starting in 1983.
- On Jun 6, 1981, Cibola County, New Mexico (FIPS35006) was created from Valencia County (FIPS35061). The CBP incorporated this change in 1989.
- On Feb 9, 1988, Charlottesville City, Virginia (FIPS51540) annexed part of Albemarle County (FIPS51003). The CBP incorporated this change in 1989.

TABLE A.4: Excerpts of FIPS change file

date	type	old	old_name	new	new_name	cbp_year
1979-12-15	FF	29193	STE. GENEVIEVE COUNTY	29186	STE. GENEVIEVE COUNTY	1982
1981-06-19	PF	35061	VALENCIA COUNTY	35006	CIBOLA COUNTY	1983
1981-06-19	PF	35061	VALENCIA COUNTY	35061	VALENCIA COUNTY	1983
1988-02-09	PP	51003	ALBEMARLE COUNTY	51540	CHARLOTTESVILLE CITY	1989
1988-02-09	PF	51003	ALBEMARLE COUNTY	51003	ALBEMARLE COUNTY	1989
1988-02-09	FP	51540	CHARLOTTESVILLE CITY	51540	CHARLOTTESVILLE CITY	1989

Notes: This table presents example records from the FIPS change file. Variable `date` is the date of change. Variables `old`, `old_name`, `new` and `new_name` are the old/new county FIPS code and name. Variable `cbp_year` is the year when CBP dataset incorporated the change.

B. County and Industry Classifications

The primary identifier for file containing the CBP data is a combination of county classification code and industry classification code. Counties are assigned their FIPS code. Industries are classified according to ICC or SIC. However, the correspondence between CBP counties/industries and the official US Census classifications is not always one-to-one. At times, the CBP groups small counties together for publication purposes or defines its own industry group by combining several closely related industries. There are also special codes assigned to designate unknown counties or industries. In this section, we document these discrepancies and the evolution of classification standards over time. The changes in the total number of observations in the CBP we document in Figure 1 largely reflect changes of county and industry classifications.

B.1 County Classification

Counties are assigned 5-digit FIPS (The Federal Information Processing Standards) codes. The first 2 digits designate the state while the last 3 designate the county.

Business units with an unidentified county location in each state are classified under "Statewide" by the CBP, and are assigned the county code "999". We assign the FIPS code "72998" to the defunct municipality Rio Piedras, Puerto Rico.

In 1950, counties with no manufacturing establishments were omitted from the CBP. We manually add these counties with entries of 'NaN' for Total employment and establishments, and zero for Manufacturing employment and establishments to the datasets to maintain continuity of county observations.

B.1.1 State Coverage

All years of the CBP cover 50 states. The District of Columbia (FIPS11) is covered in all years except in 1946, where it is missing. Data on Puerto Rico (State FIPS 72) are reported starting in 1959, but are missing after 1969 (and until 2002 in the more recent CBP files provided by Eckert et al. (2020a)). Coverage of the US Virgin Islands (State FIPS 78) started in 1964 but is missing from 1969 onwards.

B.1.2 County Grouping

For the period 1946–1962, the CBP groups together counties for publication purposes. Significant groupings happen in States with more than 100 counties: Georgia, Illinois, Kansas, Kentucky, Missouri, North Carolina, Texas, and Virginia. The total number

of county observations in each of these states is no more than 101. Grouping also happens in the State of New York but only for counties comprising New York City: Bronx (FIPS 36005), Kings (FIPS 36047), New York (FIPS 36061), Queens(FIPS 36081), and Richmond (FIPS 36085). In addition, Hawaii, Montana, New Mexico, and South Dakota also have one or two incidences of county grouping starting in 1950s. For these county groups, county FIPS codes are created by combining all FIPS codes of the component counties, separated by commas. For example, New York City is assigned the county FIPS code "005,047,061,081,085".

Virginia. Independent cities in Virginia are usually combined with adjacent counties in earlier years but are largely missing in the 1947 publication. For example, "Alleghany County and Clifton Forge City" (County FIPS 005,560) in years 1946 and 1948 only appear as "Alleghany County" in 1947. We believe that the statistics recorded for these counties also include the independent cities as in other years, so we deliberately modify the county names in 1947 to be compatible with those in 1948 before assigning FIPS codes, so that the actual number of counties remains comparable to adjacent years. We also standardize a handful of county names for Virginia in 1946 for the same reason.

B.1.3 County Boundary Changes

County boundaries change over time. New counties are usually created from parts of existing ones. Counties become extinct by merging into others. Sometimes counties exchange territories. At times, there are also changes in county names or FIPS codes without a change in boundaries.

The CBP tends to incorporate these changes with 1–2 years of delay. Because of potentially different county classifications in the CBP, especially during the earliest years, these changes may not materialize in the same manner for CBP county equivalent entities. For example, Virginia Beach City, VA (FIPS 51810) was created from Princess Anne County, VA (FIPS 51151) on Feb 14, 1952. The latter then merged into the former on Jan 1, 1963. The CBP incorporated these changes in 1953 and 1964, respectively. However, during 1953–1962, these two counties are displayed as a county group (Figure A.1). As a result, this change in the CBP appears as if these county-equivalent entities are recoded or renamed. We classify these changes as "Recode/Rename" to better reflect their impact on the CBP's data structure.

In a technical note (Eckert et al., 2022b), we document such county changes in chronological order according to when the CBP incorporated them. We describe each event and provide a graphical illustration of changes in employment and establishment counts for the relevant counties around it.

Alaska. County classification in Alaska is the most unconventional and has changed considerably over time. During 1946–1967, the state is divided into four Judicial Divisions. During 1968–1970, geographical divisions hardly conform to any standard classification, and we thus created "artificial" FIPS codes. The period 1971–1981 used 29 census divisions, after which the system changes substantially again to boroughs and census areas, which is the classification used today (with occasional minor adjustments).

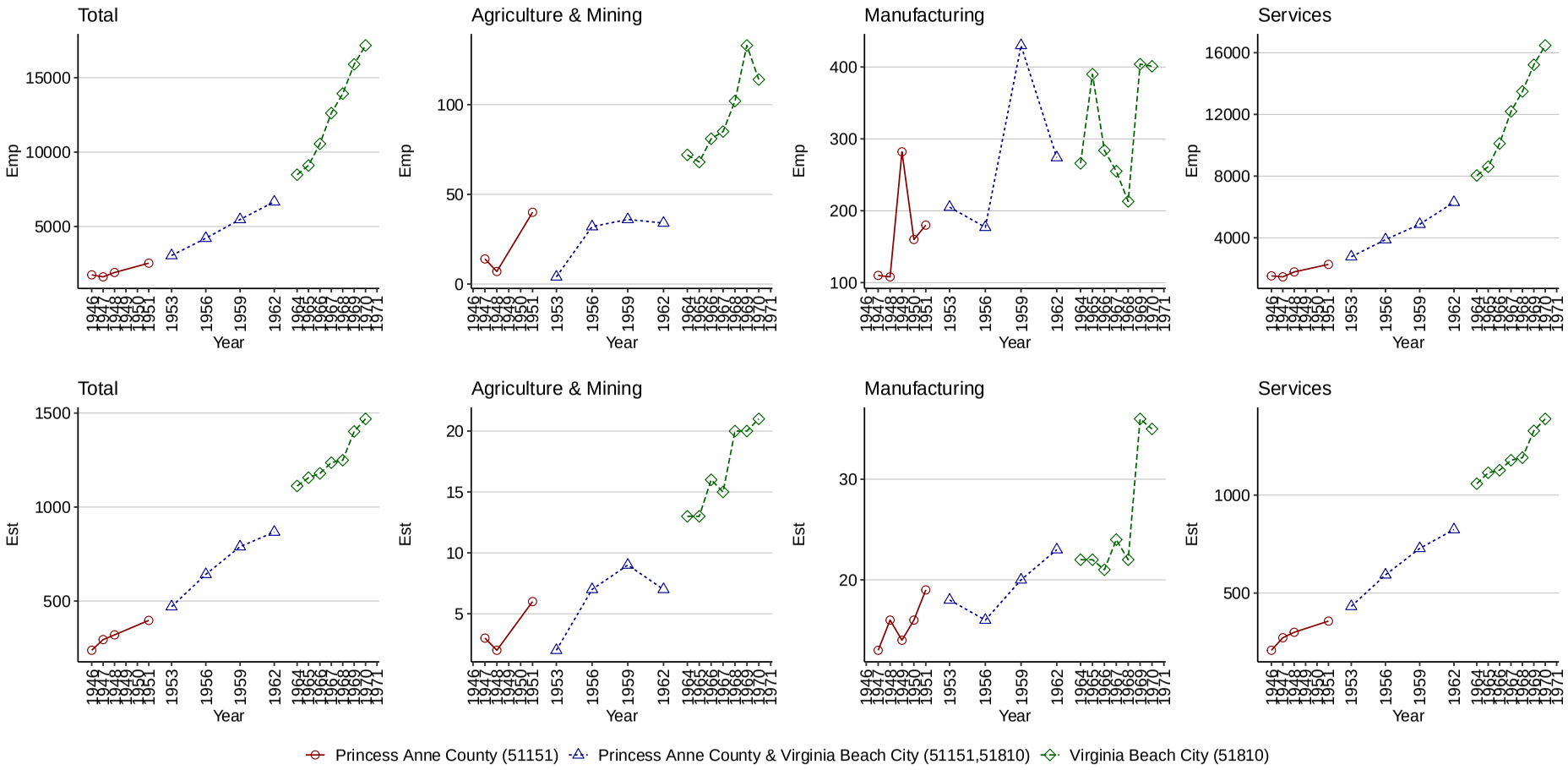


FIGURE A.1: County boundary change—Princess Anne County & Virginia Beach City, VA

Notes: This figure plots the number of employment and establishments in Princess Anne County and Virginia Beach City of Virginia. Virginia Beach City was created from Princess Anne County on Feb 14, 1952. They merged again on Jan 1, 1963. CBP incorporated these changes in 1953 and 1964 respectively. However, during 1953–1962, these two counties are combined and displayed as a county group.

B.2 Industry Classification

The industry classification used in the CBP has evolved with changes in the official US Census industry classification standards. The CBP sometimes defines its own industry codes for county/sector aggregates and other industries that are not classified by the US Census. In this section, we introduce the industry classification system employed by the CBP and its evolution over time. We have also gathered the US Census industry classification codes that are referenced by the CBP as well as the concordance tables between consecutive editions whenever they exist. These could be helpful in building consistent time series at the industry level.

Section B.2.1 lays out the official industry classification standards referenced by each CBP dataset; Section B.2.2 discusses special industry codes defined in the CBP.

B.2.1 Industry Classification Standards

Table A.5 lists the official industry classification codes used in the CBP and the most detailed industry level in slightly more detailed than Table 1 in the main text.

During 1946–1948, manufacturing industries are classified according to the Standard Industrial Classification (SIC) 1945 while non-manufacturing industries are classified according to the Social Security Industrial Classification Code (ICC) 1942 developed by the Federal Security Agency.

In 1949 and 1950, the CBP only reports data on manufacturing industries, classified by SIC1945.

During 1951–1956, manufacturing industries are still classified according to SIC1945 while non-manufacturing industries are classified according to SIC1949.

During 1959–1974, all industries are classified according to SIC. There are three editions of the SIC (1957, 1967 & 1972), with major changes in each new edition. A supplement was also published in 1963, which makes minor changes to the previous edition.

B.2.2 Special Industry Codes

Some industry codes in the CBP do not conform to the official US Census classification standards. We describe each of these cases below.

County totals. Observations for county totals are assigned industry codes "----".

Top level industries. Before NAICS was adopted in 1998, the SIC/ICC do not have proper codes for the top level industry divisions. We assign these according to Table A.6.

TABLE A.5: Detailed CBP industry classifications

CBP Year	Classification	Digits
1946	SIC1945 & ICC1942	2
1947–1948	SIC1945 & ICC1942	3
1949–1950	SIC1945	3
1951–1953	SIC1945 & SIC1949	3
1956	SIC1945 & SIC1949	4
1959–1962	SIC1957	4
1964–1967	SIC1957 & SIC1963	4
1968–1973	SIC1967	4
1974–1977	SIC1972	4

Notes: The Industry Classification Code (ICC) 1942 only contains non-manufacturing industries. The Standard Industry Classification (SIC) 1945 only contains manufacturing industries. CBP1949 & 1950 contain only manufacturing industries. SIC1963 is a supplement edition that make amendments to SIC1957.

TABLE A.6: CBP top level industry codes, 1946–1974

Code	Industry Name
07--	Agricultural Services, Forestry and Fisheries
10--	Mining
15--	Contract Construction
20--	Manufacturing
40--	Transportation and Public Utilities
50--	Wholesale Trade
52--	Retail Trade
60--	Finance, Insurance and Real Estate
70--	Services
99--	Nonclassifiable Establishments
00--	Unclassified Establishments

Notes: These codes are assigned to top level industry divisions during 1946–1974.

TABLE A.7: CBP administrative and auxiliary industries

Code	Industry Name
098/	Auxiliary to agricultural services
149/	Auxiliary to mining
179/	Auxiliary to construction
399/	Auxiliary to manufacturing
499/	Auxiliary to transportation and utilities
519/	Auxiliary to wholesale trade
599/	Auxiliary to retail trade
679/	Auxiliary to finance and banking
899/	Auxiliary to services

Notes: These codes are assigned to administrative and auxiliary industries during 1949–1974.

Administrative and auxiliary industries. These industries represent central administrative office and auxiliary activities, such as warehouses, research laboratories, and maintenance. They appear for top level industry divisions and are assigned industry codes according to Table A.7. Statistics for these observations were presented only for the manufacturing division during 1949–1956. During 1959–1974 they were displayed for each industry division.

Other non-standard industries. The CBP defines several industries to represent those not classified to other industries or a group of closely related industries. We list them in Table A.9. We provide details for these cases below:

- In 1946, although only data on 2-digit industries were reported, there were a few CBP-defined 3-digit industries representing those that could not be classified into other 2-digit industries under each top industry division. For example, industry 099 represents business units that are not classified into any sub-industries under industry division Agriculture, Forestry and Fisheries.
- In 1947–1950, the CBP defines 3-digit industry codes for a few values not reported as part of 2-digit industries. For example, industry 330 represents business units that are not classified into any sub-industry under SIC 33 Primary Metal Industries.
- In 1951–1953, the CBP defines industry *337 Primary metal industries (nonferrous)* to include four SIC industries (333, 334, 335 & 336).
- In 1959–1973, the CBP defines industries *3717 Motor Vehicles and Parts* to include three SIC industries (3711, 3712 & 3714), *4211 Trucking without Storage* to include

two SIC industries (4212 & 4213), *453 Air Transportation* to include two SIC industries (451 & 452), *4781 Transportation Services, Not Elsewhere Classified* to include three SIC industries (4783, 4784 & 4789), and *6791 Oil Royalty And Commodity Traders* to include two SIC industries (6791 & 6792).

- In 1974, the CBP defines industry *8242 Vacational Schools* to include two SIC codes (8244 & 8249).

B.2.3 Level of Industry Details

Figure 2 in the main text plots the number of unique industries in each CBP dataset together with the timing of major classification changes. We explain some of the apparent structural breaks shown by the figure:

- In 1946, only data on 2-digit level industries were reported. During 1947–1953, the CBP contains data on 3-digit level industries. Starting from 1956 (and more so from 1959), the CBP reports data for the most detailed industries (4 digits for SIC).
- The CBP versions for 1949 and 1950 only contain manufacturing industries.
- Until 1953, the CBP explicitly distinguished between large counties and small ones based on the total number of reporting units. Only top level industry divisions were reported for small counties, while more detailed statistics were made available for large counties.

TABLE A.8: Excerpts of industry classification concordance or supplement files

File	type	old	new	new_name	note
SIC1957_vs_SIC1967	FN	3619			
	NF		4619		
	PF	0119	0114		
	PF	0119	0119		
	FP	1081	1081		Change effective from January 1, 1963.
	FP	1082	1081		Change effective from January 1, 1963.
SIC1963_Supplement	FN	3619			Deleted
	NF		4619	Pipe lines, not elsewhere classified	New industry
	PP	3651	3679	Electronic components and accessories, not elsewhere classified	Change Magnetic recording tape to 3679
	PF	3651	3651	Radio and television receiving sets, except communication types	Change Magnetic recording tape to 3679
	PF	3679	3674	Semiconductor (solid state) and related devices	Change part to 3674
	PP	3679	3679	Electronic components and accessories, not elsewhere classified	Change part to 3674, incorporate Magnetic recording tape from 3651

Notes: This table presents example records from CBP industry classification concordance or supplement files. Variables `old` and `new` are old and new editions of the industry classification codes respectively. Variable `new_name` exists only in supplement files to show industry name after amendments. Variable `note` contains some descriptions about the change.

TABLE A.9: Other non-standard industry codes

CBP	Code	Industry Name
1946	099	Other Agriculture, Forestry and Fisheries
	149	Other Mining
	189	Other Contract Construction
	399	Other Manufacturing
	499	Other Transportation, Communication and Public Utilities
	529	Other Wholesale Trade
	599	Other Retail Trade
	699	Other Finance, Insurance, and Real Estate
	939	Other Service Industries
1947–1950	330	Primary Metal Industries, Unclassified
	340	Fabricated Metal Products, Unclassified
	370	Transportation Equipment, Unclassified
	430	Miscellaneous Transportation, Unclassified
	510	Miscellaneous Distributors, Unclassified
	520	Wholesale and Retail Trade Combined, Unclassified
	530	Retail General Merchandise, Unclassified
	540	Retail Food and Liquor Stores, Unclassified
	550	Retail Automotive, Unclassified
	560	Retail Apparel and Accessories, Unclassified
	570	Miscellaneous Retail Trade, Unclassified
	620	Miscellaneous Finance Agencies, Unclassified
	700	Lodging Places, Unclassified
	720	Personal Services, Unclassified
1951–1953	337	Primary Metal Industries (Nonferrous)
1959–1973	3717	Motor Vehicles and Parts
	4211	Trucking without Storage
	453	Air Transportation
	4781	Transportation Services, Not Elsewhere Classified
	6791	Oil Royalty And Commodity Traders
1974	8242	Vacational Schools

Notes: These codes are defined by CBP and may not have correspondence to the standard industry classification codes.