NBER WORKING PAPER SERIES

REDESIGNING THE LONGITUDINAL BUSINESS DATABASE

Melissa C. Chow
Teresa C. Fort
Christopher Goetz
Nathan Goldschlag
James Lawrence
Elisabeth Ruth Perlman
Martha Stinson
T. Kirk White

Working Paper 28839
http://www.nber.org/papers/w28839

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
May 2021

 We thank Kimberly Blair, Marina Krylova, and Gerald McGarvey for their programming expertise. We also thank Trey Cole for sharing programs and expertise from the Statistics of US Businesses program. Finally, we thank Emek Basker, Shawn Klimek, William Davie, and John Haltiwanger for helpful comments.

Redesigning the Longitudinal Business Database
Melissa C. Chow, Teresa C. Fort, Christopher Goetz, Nathan Goldschlag, James Lawrence, Elisabeth Ruth Perlman, Martha Stinson, and T. Kirk White
NBER Working Paper No. 28839
May 2021
JEL No. D0,E0,F0,J0,L0

ABSTRACT

In this paper we describe the U.S. Census Bureau's redesign and production implementation of the Longitudinal Business Database (LBD) first introduced by Jarmin and Miranda (2002). The LBD is used to create the Business Dynamics Statistics (BDS), tabulations describing the entry, exit, expansion, and contraction of businesses. The new LBD and BDS also incorporate information formerly provided by the Statistics of U.S. Businesses program, which produced similar year-to-year measures of employment and establishment flows. We describe in detail how the LBD is created from curation of the input administrative data, longitudinal matching, retiming of economic census-year births and deaths, creation of vintage consistent industry codes and noise factors, and the creation and cleaning of each year of LBD data. This documentation is intended to facilitate the proper use and understanding of the data by both researchers with approved projects accessing the LBD microdata and those using the BDS tabulations.

Melissa C. Chow
Center for Economic Studies
U.S. Census Bureau
4600 Silver Hill Rd.
Washington, DC 20233
melissa.c.chow@census.gov

Teresa C. Fort
Tuck School of Business
Dartmouth College
100 Tuck Hall
Hanover, NH 03755
and U.S. Census Bureau
and also NBER
teresa.fort@tuck.dartmouth.edu

Christopher Goetz
Center for Economic Studies
U.S. Census Bureau
4600 Silver Hill Rd.
Washington, DC 20233
christopher.f.goetz@census.gov

Nathan Goldschlag
Center for Economic Studies
U.S. Census Bureau
4600 Silver Hill Road
Washington, DC 20233
Nathan.Goldschlag@census.gov

James Lawrence
U.S. Census Bureau
4600 Silver Hill Rd.
Washington, DC 20233
James.Lawrence@census.gov

Elisabeth Ruth Perlman
Center for Economic Studies
U.S. Census Bureau
4600 Silver Hill Road
Suitland, MD 20746
elisabeth.perlman@census.gov

Martha Stinson
Center for Economic Studies
U.S. Census Bureau
4600 Silver Hill Rd.
Washington, DC 20233
martha.stinson@census.gov

T. Kirk White
Center for Economic Studies
U.S. Census Bureau
4600 Silver Hill Rd.
Washington, DC 20233
thomas.kirk.white@census.gov

A data appendix is availael at http://www.nber.org/data-appendix/w28839
Information on accessing the LBD through the FSRDCs is available at
https://www.census.gov/about/adrm/fsrdc.html

Contents

List of Tables

List of Figures

1 Introduction

The Longitudinal Business Database (LBD) is a confidential historical tracking system for private, non-farm business establishments with employees that operated in the United States between 1976 and 2018.[1] This database links establishment records over time to identify the beginning and end of economic activity at business locations. One of the primary goals of creating the LBD is to measure year-to-year changes in private-sector employment and the entrance and exit of businesses within and across geographic locations, industries, and firm and establishment groups defined by size and age. To this end, the Census Bureau created a formal production system to generate the LBD and its companion public-use statistical product, the Business Dynamics Statistics (BDS). This paper provides details about that production system to users of both the LBD and the BDS in order to facilitate best practices in using these data and to inform future research on ways to improve these products.

Since the mid-1970s, the Census Bureau has maintained a listing of private businesses operating in the United States, originally called the Standard Statistical Establishment List (SSEL) and today called the Business Register (BR). The basic building block of the BR database is an establishment, or place of business. Each establishment has an address, industry code, payroll, employment, and a firm identifier that groups it together with any other jointly owned establishments. The Census Bureau continuously updates the BR by surveying businesses, and by utilizing administrative records from business tax filings, shared with the Census Bureau by the Internal Revenue Service (IRS). The Census Bureau uses the BR to publish statistics, such as the annual County Business Patterns (CBP), that give a complete picture of private, non-farm business activity in the United States at a given point in time. The main purpose of the LBD is to link annual snap shots of the Business Register over time in order to enable measurement of changes in business activity. How many new firms and establishments open each year? How many die? How much employment do these firms and establishments create or destroy? These are all questions that the LBD is designed to answer.

The complications inherent to tracking establishments and firms in the Census Bureau data make the creation of the LBD much more complex than simply matching unique establishment identifiers between years. For example, businesses sometimes change the employer identification number (EIN) they use to file payroll taxes. These types of changes make it difficult for Census Bureau staff who maintain the BR to tell whether a business continues to operate from one year to the next. When the BR staff receives a tax record for a new EIN that has not previously existed in the BR, they cannot be sure whether it is a new business or an old business that has simply changed its EIN. Similarly, if they fail to receive a tax report for an already existing EIN, they cannot be sure whether the business has died or filed under a new EIN. This problem is especially pronounced for small, single-establishment firms, which are surveyed once every five years as part of the quinquennial Economic Census (EC). Without additional linking, the BR will overstate the number of small firms with one establishment that are born and die each year.

A firm may re-organize and change its associated establishments. Those changes can take the form of acquisitions, taking ownership of existing establishments, divestitures, ceding ownership of continuing establishments it once owned, or opening and closing establishments. To track these types of changes to firm organizational structure, the Census Bureau conducts the Report of Organization annually, supplementing the organization information collected in the Economic Census. Importantly, only relatively large firms are surveyed in the Report of Organization, formerly known

[1] The LBD is updated annually, usually with a two year lag from the current year. 2019 data is expected to be available in the fall of 2021.

as the Company Organization Survey (COS). Hence for many firms, the organizational structure recorded in the BR is only updated every five years. Without additional processing, the BR will show large spikes in the number of establishments that are born and die in Economic Census years. The creation of the LBD addresses this challenge with algorithms that retime the birth and death of establishments associated with small and medium size firms that are not covered by the Report of Organization.

Another longitudinal challenge is industry coding. Industry classification systems have changed substantially over time, making it hard to compare the composition of industry activity in the 1980s to that in the 2010s. In order to produce longitudinally comparable measures at the industry-level we need a single-vintage, or vintage-consistent, industry code for every establishment across all years of the LBD.

To summarize, the major contributions of the LBD include: the harmonization of many versions of BR data across more than 40 years; establishment name and address matching for small, single-establishment firms to address broken links; the longitudinal establishment identifier `lbdnum`; the re-timing of excess establishment births and deaths in Economic Census years; vintage-consistent NAICS codes; creation of all variables needed to publish the BDS. We begin with a brief history of the creation of the LBD. Section 2 outlines the steps of the production process and Section 3 provides a description of the micro-data files available for use in the Federal Statistical Research Data Centers (FSRDCs). Sections 4, 5, and 6 describe the underlying business frame, matching, and longitudinal linking. Section 7 describes the retiming of establishment birth and death in Economic Census years and Section 8 describes the creation of vintage consistent NAICS industry codes. Section 9 describes the creation and use of multiplicative noise factors in the BDS tabulations and Section 10 describes how the LBD microdata files are created, combining cross sectional establishment attributes, longitudinal links, retiming, vintage consistent industry codes, and the noise factors. Section 11 describes how the measures of firm, establishment, and employment flows in the BDS are computed and Section 12 describes how the BDS tabulations changed between the legacy LBD processing and the new LBD production process. We conclude with a list of issues that the LBD-BDS production team hopes to research in the future in order to continue to improve the production process and invite interested researchers to join us in these efforts.

1.1 LBD History

The first longitudinal business establishment database created at the Census Bureau, the Longitudinal Research Database (LRD), was developed at the Center for Economic Studies (CES) in the 1980s (McGuckin and Pascoe, 1988). The inputs to the LRD were cross-sectional plant level data from the quinquennial (five-year) Censuses of Manufactures and the Annual Survey of Manufactures (ASM), augmented with administrative data. These data were linked longitudinally at the plant level using numeric identifiers from the input datasets. These longitudinal linkages allowed researchers to measure how the number of businesses changed–birth and death as well as net changes–and how individual businesses were growing or shrinking over time. The LRD was used to conduct original empirical research on business dynamics in the manufacturing sector such as Dunne, Roberts, and Samuelson (1988) and Davis, Haltiwanger, and Schuh (1996). The academic interest in business dynamics measures stemmed from the fact that these statistics allowed for the examination of the relationship between establishment characteristics such as size, age, industry, and geography, and job creation and destruction (Davis, Haltiwanger, and Schuh, 1996).

In the late 1990s, CES began developing an economy-wide establishment-level longitudinal database, the LBD. The creation of the LBD was spurred by the need to see if results obtained

with the LRD applied to other sectors of the economy outside manufacturing, and by the fact that manufacturing's share of activity in the U.S. economy was decreasing. The essential element of the longitudinal linking was the use of name and address matching to link establishments over time that had different numeric identifiers but were in fact still the same business. The development of the first vintage of the LBD is described in Jarmin and Miranda (2002). One of the most important innovations stemming from the longitudinal linking was the measurement of firm age, information not generally available in other firm-level data. Subsequent research published in Davis, Haltiwanger, Jarmin, and Miranda (2007) formalized the definition of firm age as the age of the oldest establishment in the firm at its inception and developed an algorithm to more accurately date the birth and death of establishments first observed in Economic Census years. Interest in LBD research findings led to a grant from the Kauffman Foundation to assist the Census Bureau with LBD development. In particular, the grant called for the creation of public use tabulations from the LBD to shed light on the role of entrepreneurship in job creation. The resulting BDS data products were first published in 2009 and covered the time period 1977-2005. The LBD has since been used in numerous microeconomic analyses including Haltiwanger, Jarmin, and Miranda (2013), Jarmin, Klimek, and Miranda (2005), and Davis, Faberman, Haltiwanger, Jarmin, and Miranda (2010). The BDS has also developed a large constituency of users, including policy makers, the business community, and researchers.

While the early work producing longitudinal linking of business and research on business dynamics was done at CES, by the early 1990s, interest in these statistics had also developed at the U.S. Small Business Administration (SBA). This led the SBA to request that the Census Bureau develop and publish statistics on business dynamics using establishment-level data from the economy-wide CBP program. The resulting program, the Business Information Tracking Series (BITS), was similar to the LBD in that it was based on linked establishment data and tabulated statistics on business dynamics. However the focus of BITS and the resulting public-use data product, Statistics of U.S. Businesses (SUSB), was to measure year-to-year changes rather than a longitudinal time series. The SUSB tables also never included tabulations by firm age.

1.2 Integration of LBD and BITS

The popularity of both the BDS and SUSB data products led the Census Bureau to make plans in 2015 to integrate the two products and produce one longitudinal business tracking system with a single set of public-use tables. As part of the integration process, every aspect of both systems was re-examined, re-programmed, and documented in order to institute best practices for data management, linking, imputation, and measurement. In September 2020, the Census Bureau published an expanded set of tables using this new production system. This set of tables was larger than what was previously produced and more detailed in terms of firm age categories within detailed geographic areas and industry segments. In November 2020, the new LBD microdata files became available in the FSRDCs for approved research projects. This paper seeks to provide details about both the LBD and BDS creation process in order to support the use of these data products.

2 Summary of Steps

The LBD and BDS production process is designed to produce both microdata files for use by researchers working with the confidential data and summary files for publication on the Census Bureau's website. The primary goal of the process is to link establishments over time and edit key variables necessary for creating the public use tables.

Step 1: Creating CBPBR (see Section 4)
The LBD spans over 40 years, beginning in 1976 and adding new data every year. One of the most significant challenges in the creation of a longitudinal database that spans such a significant number of years is consistency of the data items over time. The LBD production system makes such consistency a top goal. The first step in the production process is to standardize variables over time, changing things like geography codes, payroll values, and processing flags to have standard definitions. Even variables as fundamental as establishment identifiers have changed over time.

The other major challenge addressed at the beginning of the processing is the integration of accurate historical files. Many of the SSEL files from the late 1970s and early 1980s were known to have serious problems, missing large numbers of establishments for various reasons. One of the major contributions of this re-design was the discovery, cleaning, and integration of new CBP microdata files from 1976 to 1984. These files were recovered from an obsolete Census server and were transformed from fieldata and binary to ascii using file-specific conversion code. As a result of incorporating this information, the BDS time series had fewer spikes.

Finally, the earliest part of the process combines the SSEL and BR files with the CBP files and reconciles establishments found in only one of the files as well as differences in employment and payroll among overlapping establishments. Some records are determined by this processing to not represent establishments and are dropped before the year-to-year matching in Step 2. These include sub-masters, ghost records, and records mistakenly created when a firm transitions from a multi-unit enterprise to a single-unit one. At the end of Step 1 we have clean annual establishment information that can be used to perform matching and to provide establishment attributes for the LBD. These data files are called the CBPBR files.

Step 2: Year-to-Year Matching (see Section 5)
The second step in the LBD production process is year-to-year matching. The year pair matches created in this step form the foundation on which we build longitudinal linkages in Step 3. We begin with the simplest type of matching: BR/SSEL identifiers. These include both standard establishment IDs and other identifiers such as `ppn` (permanent plant number) and `oldid`, which are meant to match over time. We search records that are not matched using identifiers for cases that transition from single to multi-unit enterprises (SU-MU) and vice versa (MU-SU). This involves matching by EIN and determining whether an SU establishment can be matched to one of the MU establishments based on zip code and street address. After SU-MU and MU-SU transitions are addressed, we identify single-unit reorganizations using name and address matching to link establishments with different IDs. These types of matches can happen both within a year and across years. At the end of the year-to-year matching process we have year pair establishment links that are used to create the longitudinal links file in Step 3.

Step 3: Longitudinal Links (see Section 6)
After the year-to-year matching is completed, the next step builds a wide link file by combining year-to-year matches. The wide link file contains all annual identifiers, if any, associated with an establishment for all years covered by the LBD. A single observation in this wide link file represents a single, longitudinally consistent establishment across the entire time series. This file is built up one year pair at a time, beginning with 1976-1977, and then adding additional year pairs. Each time a new pair is added, discrepancies in the matching across year pairs are resolved.

Step 4: Pre-retime LBD (see Section 10)
The fourth step of the process creates an initial version of the LBD establishment and firm files. This step uses the wide link file to create a unique longitudinal identifier, `lbdnum`, for each establishment

in the data. We then create annual establishment and firm files by combining the longitudinal links and annual attributes from the CBPBR files. These annual files will contain all of the employment, scope, and classification information used to create the BDS. These initial annual LBD files are necessary to identify establishments whose birth or death needs to be re-timed.

Step 5: Re-timing (see Section 7)
The fifth step re-times the birth and death of establishments at multi-unit firms in Economic Census years. This step is necessary because the more complete reporting in Economic Census years by small firms that open or close establishments generates bunching of establishment births and deaths in time. These firms are not surveyed in intercensal years and hence all closures and openings are captured at one point in time. We build a model to predict what year a given establishment birth or death occurred and then create or remove records in intercensal years to reflect these imputations. Total employment reported for an EIN is allocated across all establishments either reported or imputed to be active in intercensal years. The re-timing algorithms generate modified linkages in the wide link file and modified attributes in the CBPBR files.

Step 6: Post-retime LBD (see Section 10)
The sixth step repeats the fourth step, using the link and attribute files modified by the re-timing algorithms. The annual establishment and firm files made by Step 6 reflect imputed establishment births and deaths.

Step 7: Vintage Consistent NAICS (see Section 8)
The seventh step creates vintage-consistent NAICS codes for each `lbdnum`. This process currently creates a 2012 NAICS code for all establishments that have positive payroll in any year between 1976 and the end of the time series. This process will be updated on a regular basis to push forward to the latest vintage of NAICS.

Step 8: Noise Factors (see Section 9)
Step 8 creates and applies multiplicative noise factors to the LBD microdata in order to protect confidential data in the published BDS tables. The CBP began using establishment-level multiplicative noise factors as a disclosure protection method in reference year 2007. To prevent possible revelation of the parameters of the noise distribution by taking independent draws, the LBD multiplicative noise factors are identical to the CBP multiplicative noise factors for all establishments that are in both the CBP and LBD microdata files from 2007 to the end of the time series. Moreover, we carry the CBP multiplicative noise factors backward in time such that establishments maintain their multiplicative noise factors to the greatest extent possible. For establishments that were no longer in existence by 2007, meaning no CBP multiplicative noise factors are available, new multiplicative noise factors are generated following the Hybrid Balanced Multiplicative Noise Infusion procedure used by the CBP. At the completion of Step 8, multiplicative noise factors are added to the annual LBD establishment files.

Step 9: Tabulation (see Section 11)
The final step in the process, Step 9, generates the public-use BDS tables. We compute measures of the count and flows of firms, establishments, and employment by both firm and establishment characteristics. Those characteristics include firm and establishment size and age, geography, and industry. This process also suppresses cells with fewer than three firms for confidentiality reasons.

3 Data Files

This section describes the microdata files created by the LBD production process as an FSRDC user might find them. We describe the firm and establishment identifiers in each file and describe how to link the files to each other as well as how one might link them to other business microdata.

3.1 cbpbr{year}

The *cbpbr*{*year*} files are annual data sets that contain all records with positive payroll in either the BR or CBP microdata files. The purpose of these files is to bring together information from the BR and the CBP and to standardize variable names across years in order to facilitate matching establishments across time. The CBPBR contains the essential linking variables and establishment attributes used to create the LBD annual establishment files. These include name, address, industry, legal form of organization, Census-assigned identifiers for both the establishment and the firm, `ein`, annual and quarterly payroll, March 12th employment, and county and state FIPS codes.

Researchers that wish to use business name and address matching to link external files should use the *cbpbr*{*year*}. The creation of the *cbpbr*{*year*} files addresses the myriads of data discrepancies in the source BR and CBP data over time. However, these files do not contain an exhaustive set of all variables from any of the BR or CBP files and hence, for some special purposes, these files will not be an adequate replacement for matching to either the BR or CBP directly.

The BR source files are the SSEL MU and SU files from 1976-2001 and the *empunits* and *einunits* files from 2002 forward. The CBP source files are SUSB load files from 1988-1995 and 2000, the *estab* files from 1996-1999 and 2001 forward, and recovered historical files for 1976-1984. There are no CBP source files available for 1985-1987.

There are three main types of records: submasters, payroll active establishments, and ghost establishments. Only active establishments are included in the final *lbd*{*year*} files, but the *cbpbr*{*year*} contains submaster and ghost records for research purposes. Submasters provide combined payroll and employment reports for all establishments that share an `ein` and have their own addresses. Ghosts are establishments that link to another establishment within the same year using `newid` on the Ghost record and `oldid` on the new record. These ghost records only exist prior to 2002 and were created as part of BR processing when BR analysts changed the identifier of an existing establishment due to some type of re-organization.

Many of the variables on the *cbpbr*{*year*} files have two versions: one from the BR and one from the CBP. Variables from both sources have been kept to facilitate research into differences between the BR and CBP files and also for use in the LBD creation process. An initial reconciliation between BR and CBP payroll and employment values is performed as part of the creation of the *cbpbr*{*year*} files resulting in the variables `emp_final` and `ap_final`. This processing is described in detail in Section 4.

The primary identifier on the *cbpbr*{*year*} files is `id`, which is the original identifier from the BR and CBP files. Prior to 2002, this is the `cfn` (Census File Number). From 2002 forward, this is `empunit_id_char`. This variable links to any Census Bureau data set that uses BR establishment identifiers.

In addition to `id`, the *cbpbr*{*year*} files also contain additional establishment identifiers used to match establishments over time and identify reorganizations. Permanent Plant Number, `ppn`, is used to link across years for multi-unit establishments that re-organize (populated on the *cbpbr*{*1982*} to *cbpbr*{*2001*}). `oldid` is populated for establishment records that were created by changing the ID on existing establishment records due to a multi-unit re-organization. It contains the prior ID

so the link across time can be maintained even though the establishment now has a new ID. `oldid` on the new (i.e. re-organized) record links to ID on the record for the establishment in the prior year. It also links to ID on the obsolete record within the same year, see `newid` below for further details (found on the *cbpbr*{*1976*} to *cbpbr*{*2001*}). `newid` is populated for establishment records that became obsolete (`actv_stat`=G) because they were part of a multi-unit firm re-organization. `newid` for the obsolete (i.e. ghost) record links to ID on the replacement (i.e. re-organized) record within the same year. It is the counter-part to `oldid` (found on the CBPBR 1976-2001).

cbpbr{*year*} can be linked to:

- *lbd*{*year*} by matching *cbpbr*{*year*}.`id` to *lbd*{*year*}.`estabid`.
- For reorganizations, *lbd*{*year*} by matching *cbpbr*{*year*}.`id` to *lbd*{*year*}.`estabid_rorg`.
- Any Census Bureau dataset with establishment-level identifiers (`cfn` or `empunit_id_char` or `survunit_id`) by *cbpbr*{*year*}.`id`.

Not all establishments found in *cbpbr*{*year*}.`id` match to establishments in *lbd*{*year*}.`estabid`. There are several reasons for this. First, the *cbpbr*{*year*} files made available to researchers have not been modified by retiming. As described in Section 7, the LBD retiming algorithms modify the *cbpbr*{*year*} data to add retimed establishment births, remove retimed establishment deaths, and modify continuers associated with a retimed birth or death. Matching the *cbpbr*{*year*} without retiming modifications to *lbd*{*year*} will cause some mismatch between the files. The second reason the files will not match perfectly is the treatment of reorganizations, described in Section 5. Within-year reorganizations will involve two records in *cbpbr*{*year*} for the same establishment. As described in Section 10, when this occurs we select a single establishment identifier to associate with the `lbdnum` and store the unused establishment identifier in *lbd*{*year*}.`estabid_rorg`. Hence, many establishments in *cbpbr*{*year*} that do not match to *lbd*{*year*}.`estabid` will instead match to *lbd*{*year*}.`estabid_rorg`. Similarly, the firm associated with the unused establishment is stored in *lbd*{*year*}.`firmid_rorg`.

Finally, it is useful to note nuances created by retiming changes when matching external data at the firm-level to the *cbpbr*{*year*} files. When an establishment transitions from SU to MU in an Economic Census year and the birth of additional establishment(s) are retimed to occur in the intercensal period, the retiming algorithms, in addition to creating new records for those births, will modify data of the continuing SU. In such cases we assign the MU `alpha` to the continuing SU. Therefore, the continuing SU will appear as an MU in the *lbd*{*year*} but an SU in the *cbpbr*{*year*} files made available to researchers. If the researcher matches that continuing SU's `firmid` ("0"+`ein`) to the *lbd*{*year*}.`firmid`, they will find that the SU does not exist in *lbd*{*year*}. Matching by `ein`, or at the establishment-level for SUs, will circumvent this issue.

3.2 lbd{year}

The *lbd*{*year*} files are establishment level files that contain one record for every establishment with positive annual payroll in the current and/or prior year, after dropping the submaster and ghost records and combining records for any establishments within a year that were deemed to be the same entity operating under different ownership. Thus each *lbd*{*year*} file contains fewer records than each CBPBR annual file. The primary innovation of the LBD files is the creation of `lbdnum`, a longitudinal identifier that is consistent across time and allows an establishment to be tracked

as it undergoes re-organizations within or across years (i.e. expansion to multi-unit, contraction to single-unit, new ownership).

The *lbd*{*year*} files serve two main purposes. First, they are the main files that researchers will use for microdata analyses of establishments. Second, they are the source files used to tabulate the public-use BDS tables. Thus in addition to standard variables used for research (employment, payroll, industry, geography), there are variables created specifically for BDS tabulation that apply edits to employment, geography, and industry and determine which establishments are in scope to be tabulated for the BDS.

The primary identifiers on the *lbd*{*year*} files are `lbdnum`, `lbdfid`, `ein`, `estabid`, and `firmid`. `lbdfid` is the firm counter-part to `lbdnum` and is the linking key for firms in the LBD. As described in Section 10, the initial version of `lbdfid` is not a true longitudinal ID and does not yet resolve the complex longitudinal changes to an enterprise over time.[2] `ein` is carried on the *lbd*{*year*} files to facilitate matching to EIN-level data sources (e.g. W-2 records). `estabid` is the source establishment identifier used to collect attributes of the establishment from the *cbpbr*{*year*} files. The legacy firm identifier, `firmid`, is also stored on the LBD. For single-units, `firmid` is equal to "0" followed by the establishment's `ein`. For multi-units, the `firmid` is equal to the `alpha` followed by "0000".

In addition to these identifiers there is a suite of "rorg" identifiers that are meant to facilitate matching to data that does not perform the same re-organization matching as the LBD. The LBD processing selects a single establishment identifier from the wide file links file for each `lbdnum` each year. When a re-organization occurs, multiple establishment identifiers may link to a given `lbdnum` in a single year. In such cases, we store the establishment identifier not selected for the `lbdnum`s attributes in `estabid_rorg`.[3] Similarly, `firmid_rorg` and `ein_rorg` contain the `firmid` and `ein` associated with the establishment identifier not selected for the `lbdnum`. When matching *lbd*{*year*} to other Census Bureau data sources, or any data that contains `ein`, data users should attempt to match unmatched records to the "rorg" fields.

lbd{*year*} can be linked to:

- Other annual *lbd*{*year*} files by matching by `lbdnum`.
- *lbdfirm*{*year*} by matching by `lbdfid`.
- *cbpbr*{*year*} by matching *lbd*{*year*}.`estabid` to *cbpbr*{*year*}.`id`.
- Any Census Bureau dataset with establishment-level identifiers (`cfn` or `empunit_id_char`) by matching by *lbd*{*year*}.`estabid` (or `estabid_rorg`).
- Any Census Bureau dataset with ein-level data by matching by *lbd*{*year*}.`ein` (or `ein_rorg`).
- Any Census Bureau dataset with firm-level data by matching by *lbd*{*year*}.`firmid` (or `firmid_rorg`).

3.3 lbdfirm{year}

The annual *lbdfirm*{*year*} files contain firm-level characteristics shared by all establishments that belong to a given firm. These firm-level attributes include firm size and firm age. Storing firm-level

[2]The initial version of `lbdfid` does address when an enterprise is mistakenly assigned the same `firmid` due to `alpha`s being recycled over time. Future research will expand the functionality of `lbdfid` to link firms over time that re-organize. See Section 10 for details.

[3]See Section 10 for details on how the establishment identifier is selected.

characteristics in a separate file allows us to avoid repeating firm-level attributes across establishments in large multi-unit firms.

The primary identifiers on the *lbdfirm*{*year*} files are `lbdfid` and `firmid`.

lbdfirm{*year*} can be linked to:

- *lbd*{*year*} by matching by `lbdfid`.
- Other annual *lbdfirm*{*year*} files by matching by `lbdfid`. This should be done with caution since `lbdfid` is not fully longitudinally consistent.
- Any Census Bureau dataset with firm-level data by matching by *lbd*{*year*}.`firmid`.

3.4 naics{year}

The *naics*{*year*} files are annual data sets that contain vintage consistent (VC) NAICS codes for all establishments that are payroll active in the current year. All naics{year} files contain NAICS 2012 codes. We assign NAICS 2002 industry codes for establishments in 1976 to 2009 and we assign NAICS 2007 industry codes to establishments in 1976 to 2014. The *naics*{*year*} files also contain information on industry code splits. The variable `fk_indcode_splits` denotes the potential number of new vintage codes to which the old vintage code could have been assigned. Splits variables are available every time there is a change in the industry code.

The primary identifier on the *naics*{*year*} files is `lbdnum`.

naics{*year*} can be linked to:

- Other annual *naics*{*year*} files by matching by `lbdnum`.
- *naics*{*year*}_*flags* by matching by `lbdnum`.
- *naics*{*year*}_*aux* by matching by `lbdnum`.
- *lbd*{*year*} by matching by `lbdnum`.

3.5 naics{year}_flags

The *naics*{*year*}_*flag* files are annual data sets that provide more information related to the VC NAICS codes. In addition to the variables in the *naics*{*year*} files described above, the *naics*{*year*}_*flag* files include variables that indicate the source, year, and methodology for the VC NAICS variables.

naics{*year*}_*flag* can be linked to:

- Other annual *naics*{*year*}_*flag* files by matching by `lbdnum`.
- *naics*{*year*} by matching by `lbdnum`.
- *naics*{*year*}_*aux* by matching by `lbdnum`.
- *lbd*{*year*} by matching by `lbdnum`.

3.6 naics{year}_aux

The *naics*{*year*}_*aux* files are a subset of the *naics*{*year*} files that consist of auxiliary establishments for the years 1976 through 2001. These are establishments that primarily serve other establishments of a firm. In the VC NAICS process, we reassign all the auxiliary establishments to their appropriate auxiliary industry in the SIC years so that the VC NAICS codes reflect true changes in economic activity. The NAICS 2002 codes in the auxiliary files represent the industry that the auxiliary establishments serve. These files also include splits variables.

naics{*year*}_*aux* can be linked to:

- Other annual *naics*{*year*}_*aux* files by matching by `lbdnum`.
- *naics*{*year*} by matching by `lbdnum`.
- *naics*{*year*}_*flag* by matching by `lbdnum`.
- *lbd*{*year*} by matching by `lbdnum`.

4 Input Files

The first step in the LBD production processing involves combining BR and CBP data to create the *cbpbr*{*year*} files. As described in Jarmin and Miranda (2002), the earliest versions of the LBD were constructed from the 1976-2001 SSEL files, which were annual snapshots of the BR, the sampling frame for the quinquennial Economic Census and other economic surveys. Following the redesign of the BR in 2002, the SSEL files for 2002 forward were constructed as part of LBD processing using SAS datasets that were themselves snapshots of the BR's Employer Units and EIN Units tables. In the final years of the legacy LBD, these files were supplemented with edits from the CBP programs. For 2013-2016, the CBP's edited files were incorporated directly as inputs to the legacy LBD's processing stream.

The CBP program uses as inputs the same BR files used by the LBD. However, before producing its tabulations, the CBP program edits the geography, industry, and in some cases, employment and payroll values of a large number of establishments. Especially for Economic Census reference years, these edits can affect national totals. In order to make the quality of the LBD more consistent over time, the production LBD uses CBP microdata for every year of the time series — except 1985-1987 — as input files.

Some SSEL input files from the 1970s and 1980s were either missing a significant number of records, or were missing entirely (Jarmin and Miranda, 2002). The 1976 and 1981 SSEL SU files were completely missing, as was the 1988 SSEL MU file. The 1978 SSEL SU file was missing data for all establishments with an EIN beginning with 7 or 8. The legacy LBD dealt with these missing data issues in several ways. In place of the 1976 and 1981 SSEL SU files, the legacy LBD used prior-year values for employment and payroll from, respectively, the 1977 and 1982 Economic Census control files.[4] However, prior-year data from the subsequent year did not completely replace the missing data. Establishments that exited in 1981 were not necessarily in the 1982 Economic Census control file. For the new LBD we were able to fill in many of the missing records by taking advantage of some recently recovered historical microdata–including CBP microdata files–from tape

[4]To a lesser extent, the legacy LBD also used prior-year data from the subsequent-year files to fill in missing data in 1983, 1984, 1986, 1989, and 1991.

backups. CBP microdata enhances the LBD in two ways: by incorporating the many CBP edits and filling in for missing SSEL records in early years of the LBD time series.

4.1 Recovering Lost Data Files

In 2009, the Census Bureau's Center for Economic Studies led an effort to recover historical microdata and associated paper record layouts from over 2,500 data tapes before the Census Bureau Unisys mainframe computer was decommissioned (Atrostic, Becker, Gardner, Grim, and Mildorf, 2010). Included on these tapes were CBP microdata files for 1976-1984, as well as SSEL-like EI universe files for 1978-1980 and 1982-1986.[5] Although the data was successfully extracted from the tapes in 2009, they were stored in a combination of formats including binary integer, ASCII, and fieldata format.[6] In the new production system, using the record layouts associated with each file as a guide, we wrote SAS code to convert the recovered files from their various formats to usable SAS datasets. This effort paid off in several ways. Not only were we able to fill in hundreds of thousands of missing records in 1976, 1978, and 1981, but we were also able to take advantage of the many thousands of edits done by the CBP program in 1976-1984 as well as some metadata that were not used in previous vintages of the LBD.

Here we briefly describe how the conversion process worked for one of the files that had the largest impact on the BDS tabulations: the 1981 CBP file containing data for MUs and large SUs. Other files had different record layouts and different data formats, requiring different conversion programs, but the overall recovery process was similar. The 1981 CBP file had two different record layouts, depending on whether the record was an SU or an MU. The record layout shows which variables were stored at each position in the file, and the number of characters taken up by each variable. The data were read off the Unisys tapes in 2009 under the assumption that the entire file was stored in fieldata format and then converted to ASCII text. However, the record layout shows that the data in this particular file were stored in two formats, depending on the data type: character variables were stored in fieldata form, and numeric variables were stored in binary format (represented by a B on the record layout). Fortunately there is a one-to-one mapping from binary to fieldata — every 6 bits represent a single fieldata character, which in turn represents a (base-10) integer from 0 to 63. Using this mapping, we converted the numeric variables from base-64 numbers to base-10 integers.[7]

4.2 Business Register Files

4.2.1 1976-2001 Standard Statistical Establishment List Files

Prior to the 2002 Business Register redesign, annual snapshots of the BR were provided to CES in the form of the Standard Statistical Establishment List Single Unit and Multi-Unit files. These files, created by the BR staff for 1976 to 2001, were the main input–in most years, the only input — for the legacy LBD. The content and structure of the files is described in Jarmin and Miranda

[5] These EI, or employer identification, universe files are analogous to what later became known as EIN universe files.

[6] See https://en.wikipedia.org/wiki/Fieldata for a description of the fieldata system of encoding alphanumeric characters.

[7] See White (2014) for a more detailed description of the conversion algorithm. The code for reading in the 1981 CBP MU/large SU file and for doing the conversion from fieldata format to base 10 numbers can be made available to researchers with access to the LBD and CBPBR files. This code may be useful for improving the LBD if, for example, CBP microdata files for 1985-1987 are found among the many unlabeled Unisys data tapes that were saved as part of the CES historic data recovery effort.

(2002). Detailed definitions of every variable in either the SU file or the MU file are in the SSEL Glossaries, which are available via the Federal Statistical Research Data Centers (FSRDC). For continuity, after the 2002 Business Register redesign, CES developed code to convert the new BR files to the old SSEL SU and MU formats. With the BR redesign now 18 years in the past, it makes less sense to continue converting the newer BR files back to the old format. In the new LBD production process, we do not reproduce the legacy SSEL files for years after 2016. Instead, the first step of LBD processing creates the *cbpbr*{*year*} files for all years. The *cbpbr*{*year*} files will include much of the information formerly stored in the SSEL files along with additional information that was not available in the SSEL files.

4.2.2 2002-2018 Business Register Database and Files

The post-2001 Business Register is a relational database consisting of over 70 tables and hundreds of variables (a.k.a. data elements). It is the sampling frame for the vast majority of economic surveys conducted by the Census Bureau. Here we briefly summarize how the BR is updated, with a focus on the data and processes that are most important for the LBD. For a more detailed description of the BR and its functions, see DeSalvo, Limehouse, and Klimek (2016).

The BR is updated from three categories of data sources: administrative records data, survey/census response data, and interactive edits by Census Bureau staff. The administrative data comes from the IRS, the Social Security Administration (SSA), and the Bureau of Labor Statistics (BLS). IRS data relevant for the LBD include business names, mailing addresses, physical addresses, and filing requirements (all updated monthly), industry code (updated annually) and payroll tax data (updated weekly). The SSA provides monthly updates of industry codes for newly assigned EINs, and BLS provides quarterly updates of industry codes from unemployment insurance records. Updates to legal form of organization (i.e., corporation, partnership, sole proprietorship, etc.) can come from IRS, BLS, or SSA. Administrative records for employer businesses are provided at the EIN level. For SUs this is equivalent to the establishment and the firm. However, since MU firms can report for multiple establishments under one EIN, establishment-level updates to MU data must come either from survey responses or imputation.

The BR receives updates from three surveys: the (annual) COS, the Annual Survey of Manufactures, and the quinquennial Economic Census. The COS, also known as the Report of Organization, is sent to all of the largest MU enterprises and a probability sample of smaller (but not necessarily small) MUs and SUs. Its primary objective is to capture changes in organization and operational status. The ASM and COS sample designs are integrated, so that firms that are in the ASM are not in the COS and vice versa. Both have certainty samples for firms above a certain size. Importantly for the LBD, these size thresholds and the sampling rates for firms below those thresholds have both changed over time.[8] Between quinquennial Economic Censuses, all updates to the organization of MU enterprises in the BR–mergers and acquisitions and establishment births and deaths–come from the COS and ASM. Because of the COS/ASM sample design, this means that between Economic Censuses, the BR does a better job of keeping track of these changes for large MU enterprises than it does for smaller MUs. The Economic Census, conducted for reference years ending in '2' or '7', is the most comprehensive survey of the non-farm private economy.[9] For most smaller MU enterprises as well as SU enterprises that transition to MUs, births of new establishments and deaths of old

[8]More details on the ASM sample design are available on the survey program's website: https://www.census.gov/programs-surveys/asm/technical-documentation/methodology.html (accessed 03/03/2021).

[9]Despite the name, the Economic Census is not actually a census. In the manufacturing sector, establishments with fewer than 5 employees are not surveyed at all. In other sectors, all size classes are surveyed with some positive probability, but it is still the case that the majority of SUs are not surveyed.

ones are only captured in the BR in Economic Census years. For this reason, we see spikes in the numbers of establishment births and deaths in the BR in these years. LBD processing uses imputation models to retime some of these births and deaths, as described in Section 7.

Several times each year, the Business Register staff creates extracts from a subset of the BR tables in the form of SAS dataset files. For the year 2002 forward, the LBD is constructed from these files along with edited versions of the files after CBP processing. These BR files, the EINUNITS files and the EMPUNITS files, can be thought of as counterparts to the the SSEL SU and MU files for 2002 and later years. Here we briefly describe these files. A complete listing of all the variables in these files and definitions of these variables (as well as all other variables in BR) are provided in the annual Business Register glossaries, which are accessible to researchers with approved access in the FSRDCs.

4.2.3 EINUNITS Files: 2002-2018

As the file name suggests, all of the information on the EINUNITS file is at the EIN level. For SU enterprises, this is equivalent to establishment-level data. MU firms may file payroll taxes for multiple establishments under one EIN. For MUs, the record on the EINUNITS table is called a submaster, and in general does not correspond to a single establishment. All of the data on the EIN_UNITS BR table derives from administrative records. The LBD only uses data for SUs from the EINUNITS file, but the *cbpbr*{*year*} files also keep the submaster records. For both the CBPBR and the LBD, variables used from the EINUNITS file include the EIN, current-year and prior-year March 12 employment, quarterly and annual payroll, and (for 2004 forward) 2nd-4th quarter employment.

It is important to note that although the payroll data in particular are considered high quality, the BR data are not perfect. IRS employment data are considered somewhat lower quality than its payroll data, and employment is sometimes missing or zero even though payroll are positive in the same quarter. With some exceptions, BR processing imputes employment that is missing or zero if payroll is positive for the corresponding quarter. BR processing also replaces employment values from IRS records with imputations if the implied wage from the reported data (payroll over employment) is not within industry-size-geography-specific upper and lower limits.

4.2.4 EMPUNITS Files: 2002-2018

The EMPUNITS files contain establishment-level data for all non-farm employers, including both SUs and MUs. Like the EINUNITS file, the EMPUNITS file contains the EIN and current-year and prior-year March 12 employment and quarterly and annual payroll variables. The EMPUNITS file also contains many additional variables. These include:

- `empunit_id` — the primary establishment-level identifier
- `alpha` — firm-level identifier
- establishment name (two variables)
- indicator for active status
- NAICS industry code
- indicator for government operation

- indicator for foreign status
- legal form of organization
- physical and mailing address, including street, ZIP code, census tract and block
- latitude and longitude of the establishment (for 2007 forward)

For SU firms, the data on the EMPUNITS file is at the same level as the EINUNITS file. However, for MUs, the EMPUNITS file is the primary source for establishment-level data in the LBD.

4.3 CBP Files

One of the major improvements in the new LBD is the incorporation of edits from the CBP program for every year of the LBD time series except 1985-1987. The CBP files fall into four categories: (1) CBP universe files, (2) CBP "long record" files (MUs and large SUs), (3) CBP "estab" files (tabulated cases only), and (4) CBP complete "estab" files. We briefly describe each type. The variables included in each file are listed in Table 1 and the record counts for each type of file are listed in Table 2.

The 1976 CBP file is unique in that it includes all the "long record" variables for the universe of the establishments from that year's CBP. For 1977 and 1979-1984 we have the CBP universe files, which include every establishment in the CBP, but only a few current-year variables.[10] The CBP "long record" files, available for the years 1978-1984, include only MUs and large (payroll greater than $125,000) SU establishments. These files include both current-year and prior-year March 12 employment and quarterly payroll, as well as more detailed geography. The CBP estab files for 1988-1995 and 2000 contain only the records that were tabulated in the CBP Annual Data Tables for those years. Finally, the CBP estab files for 1996-1999 and 2001-2018 contain all records and variables used to tabulate the CBP for these years, as well as some records that were not tabulated (with an indicator variable to distinguish tabbed and non-tabbed cases). Due to the 2002 BR redesign, some of the variables on the estab files change starting with the 2002 file. For the years 2007 forward, the CBP estab files also include the establishment-level CBP noise factor that was used in the tabulations for disclosure protection. See Section 9 and Evans, Zayatz, and Slanta (1998) for a description of how these noise factors are generated.

[10] When the last 1977 CBP universe file was recovered from the Unisys mainframe data tapes, the last four digits of the CFN variable were corrupted. Although the 1977 file includes data for the universe of CBP establishments in that year, in Table 1 we report only the number of unique firm identifiers (the first 6 digits of the CFN), since these are the only records included in the 1977 CBPBR file.

Table 2: Variables in County Business Patterns Microdata Files

	File Type				
Variable	Universe 1977, 1979-1984	MU & Large SU 1976, 1978-1984	"Estab" 1988-1995, 2000	"Estab" 1996-1999,2001	"Estab" 2002-2018
CFN	X	X	X	X	
empunit ID (2002 forward)					X
EIN		X		X	X
alpha (firm id)	X	X			X
prior-year alpha		X			
March 12 employment	X	X	X	X	X
quarter 1 payroll	X	X		X	X
quarters 2-4 payroll		X		X	X
subsequent year quarter 1 payroll					X
annual payroll	X	X	X	X	X
prior-year payroll q1-q4		X			
prior-year March 12 employment		X			
prior-year annual payroll		X			
data flags for March 12 employment		X		X	X
data flags for Q1-Q4 payroll		X		X	X
data flag for annual payroll				X	X
data flags for prior-year March 12 employment	X				
data flags for prior-year Q1-Q4 payroll		X			
4-digit SIC code	X	X	X	X	
6-digit NAICS code				X	X
type of operation code	X	X	X		X
state	X	X	X	X	X
county	X	X	X	X	X
ZIP code		X	X	X	X
Census place		X		X	X
central administrative office indicator	X	X			
legal form of organization		X			X
active status indicator		X		X	X
predecessor/successor ID		X			
Census Processing Division code				X	X
employer unit type					X
noise factor (2007-forward only)					X
CBP tabulation flag				X	X

Notes: The table shows which variables are included in each type of CBP microdata file and the years for which those files are available.

Table 1: County Business Patterns Microdata Record Counts

File Type	Years	Record Counts
CBP Universe	1976	4.2 million
	1977	147,000
	1979-1984	4.2-5.2 million
CBP MU & Large SU	1978-1984	1.7-2.5 million
CBP Tabulated Estabs	1988-1995	6.0-6.6 million
	2000	7.1 million
CBP Estab Files	1996-1999, 2001-2018	6.8-8.4 million

Notes: The table shows the number of usable records in each CBP microdata file. The CBP establishment-level files that were available for 1988-1995 and 2000 only include records for establishments that were tabulated in the CBP Annual Data Tables. The record counts for the CBP files in these years are just the published national establishment counts from the CBP. Since these establishment counts come from an official Census Bureau data product (unlike the other record counts in this table), these record counts had already been cleared for release by the Disclosure Review Board when the corresponding CBP tabulations were published.

4.4 Creating the *cbpbr{year}* Files

For every year except 1985-1987 we now have up to three potential source files for each establishment-year: the current-year CBP universe file, the CBP MU/large SU file, and the BR file. Furthermore, for employment and payroll variables, for any given year t, we have current-year variables from the year t files and prior-year variables from the year $t+1$ file. Because of differences in their creation dates and the scope of the CBP, even the CBP and BR files for the same year do not have the same set of establishment records or the same attributes for common establishment records. To create the *cbpbr{year}* file for each year, we first merge the files within each year.

4.4.1 Merging Establishment Files within Year

For each of the years 1979-1984 we have two CBP files — a "universe" file and a "long record" file which includes only MUs and large SUs. For each of these years we combine the two files, merging on `cfn`. For the set of variables that overlap, we keep both versions of the variable. For use in subsequent processing we also choose one value as "the" CBP value for the given establishment-year. If the establishment is found in both the "long record" file and the universe file, we choose the values from the long record file, otherwise we choose the values from the universe files. Since the long record files have many variables that are not in the universe files, for records that only exist in the CBP universe files, in the merged file we set the values of those variables to missing.

After merging CBP files within a year, we merge the combined CBP with the BR file for the same year, again merging on `cfn` or (for 2002 forward) empunit ID (`id` and `empunit_id_char` depending on the file). We keep all records in either file (CBP or BR) with the following exceptions:

- Records that are in the CBP but not the BR and have an activity status equal to "G" or "D". These are "ghost" or "delete" records, which means they are inactive or duplicates of other records.
- For 2002 forward, CBP SU records that do not match a BR record on empunit ID but do

match a BR SU record on `ein`. These are duplicate records.

For this merge we create a flag variable indicating whether the record was found in both files (CBP and BR), only the BR file, or only the CBP file.

4.4.2 Merging Establishment Data from Year t+1 Records

Due to late filing, payroll tax data for year t sometimes arrives at the Census Bureau after BR processing of year t is complete and BR processing for year $t+1$ has begun. In some cases these are for establishments that are born in year t. If they are new establishments, then the year $t+1$ BR file will have prior-year employment and payroll data for these establishments even though there is no corresponding record for these establishments in the year t BR file. In other cases, these late-filing establishments are not new, but (because of the late-filing) the year t BR file had imputed employment or payroll and the prior-year variable in the year $t+1$ has reported data. To make sure we have the most up-to-date information for year t, we merge the year t and year $t+1$ files. We do this only after doing all the merges described in the previous subsection for every year.

Prior to merging the year t and $t+1$ files, we select records that appeared in the year $t+1$ BR file — that is, we exclude CBP-only records from the $t+1$ file. We do this because after 1984 there are no prior-year variables on the CBP files. We also add a "_brt1" suffix to the variable names (other than the establishment identifier) to indicate that the variable comes from the $t+1$ file. After preparing the year $t+1$ file for the merge, we merge the year t and $t+1$ files using the appropriate establishment identifier, `cfn` or empunit ID, depending on the year.[11] After the merge we create a flag variable, `flag_match_yrt1`, to indicate if the record was found in both the year t file and the year $t+1$ file, only in the year t file, or only in the year $t+1$ file.

After merging the year t and $t+1$ files, we deduplicate records that are CBP-only records or BR-only records by merging records that have an `oldid` in the CBP-only record that matches to the establishment ID variable of a BR-only record or vice versa. Some of these are establishments that either changed ownership or changed SU/MU status between the creation of BR file and the CBP file. For records that appear for the first time in the year $t+1$ file but have year t data in the prior-year variables, we also attempt to match to year t-only records using alternative identifiers: deduplicated `ein`, `ppn`, `oldid`, `oldei`, and predecessor/successor id (`ps_id`). Unmatched year $t+1$-only records we treat as potential year t establishment births. For these records, we drop the _brt1 suffixes from the variable names. These records are eligible for name-and-address matching, as described in Section 5.

Prior to creating the intermediate files that are passed to the name and address matching step, we drop records with annual payroll less than or equal to 0 or with a missing establishment ID. We also fill in blank or all-zeros EINs where possible. Finally, we create several "spurious birth" variables to flag cases where it appears that the record's employer-unit type erroneously switched from being a submaster in one year to being an SU in the next year.[12] The intermediate files that emerge from this process, which include data from both the CBP files and the BR files, are the *cbpbr*{*year*} files.

4.4.3 Scope and Purpose of the CBPBR files

The *cbpbr*{*year*} files exclude records from the original input files that have missing IDs (`cfn` or `empunit_id_char`) or for which annual payroll is less than or equal to zero in every input file in

[11] For the special case of the 2001-2002 merge, we use a `cfn`-empunit ID crosswalk for this merge.

[12] These will be used in the BDS tabulation step to drop what appear to be very large SU births.

which that record exists — so-called inactive records.[13] However, the *cbpbr*{*year*} files include many records that are not included in the *lbd*{*year*} files. While the *lbd*{*year*} only includes establishment-level records, the *cbpbr*{*year*} includes submaster records — the EIN-level records which summarize payroll for multiple multi-unit establishments. Prior to 2002, both the BR and the CBP files also included so-called "ghost" records. With the exception of some records in the 1980 files, these are duplicate records in the sense that they represent an establishment for which another active record (i.e., with positive payroll) exists in the same year. In many cases they exist because a firm transitioned from being an SU to an MU. For such transitions, the ghost record represents the original SU establishment and the corresponding active record for that same establishment is part of the MU firm.

The *cbpbr*{*year*} files also include many variables that are not in the LBD. The extra variables that researchers who use the *cbpbr*{*year*} files will probably utilize mostly frequently are the name and address variables, including the street address and zip code for physical as well as mailing addresses. These are useful for matching the *cbpbr*{*year*} files to external datasets via name and address matching. Keeping some records, such as ghost records and submaster records — which are excluded from the LBD — also facilitates matching to external datasets. In addition to business name and address variables, the other variables included in the *cbpbr*{*year*} but not in the *lbd*{*year*} files generally fall into three categories: additional versions of variables that are in the LBD, metadata related to the sources of the data, and metadata related to processing of the data.

Except for the years 1985-1987, *cbpbr*{*year*} has at least two source files for the vast majority of establishments — the CBP file and the original BR or SSEL file. In addition, since the BR files (and in 1978-1984, also the long-record CBP files) have prior-year data, the year $t+1$ BR files provide an additional source of employment and payroll data for year t. In the vast majority of cases, the employment and payroll values for a given establishment agree in all three data sources (CBP year t, BR year t and BR year $t+1$). However, because of CBP analyst edits and the fact that the source files were created at different times, the employment and payroll values for a given establishment are sometimes not the same in all three files. Section 10 describes how LBD processing chooses only the mostly likely value for each variable for the LBD. In order to make this process more transparent and also to facilitate potential future improvements to the LBD, the *cbpbr*{*year*} files keep all of the employment and payroll values from the original input files as separate variables.

As mentioned above in Section 4.2.2, response data from the COS/ASM surveys and the Economic Census are used to populate the employment and payroll data for MUs in the employer units table of the BR. All of these surveys suffer from unit and item non-response. The Census Bureau imputes these missing data. BR processing also uses data from prior year surveys or the previous Economic Census to impute employment for MUs that are not in the COS/ASM survey between Censuses. For researchers who are interested in assessing the affect of these imputations on data quality or changes in data quality over time, the *cbpbr*{*year*} files also include flag variables for each of the payroll and employment variables indicating whether the value was reported, edited, or imputed, and if imputed, the method of imputation.

The final category of variables in *cbpbr*{*year*} are data about processing or the status of the establishment. These include BR variables and variables created as part of LBD processing. The former group includes the following:

- codes for whether or not the record represents a new establishment

[13] Researchers who would like to use the inactive records from the SSEL files or the post-2001 BR files can also access those via the FSRDCs. The post-2001 BR files will need to be requested separately.

- codes to indicate if a previously-existing establishment was sold to another firm, closed down, purchased from another firm, or added to or deleted from the ASM
- indicator for if an establishment was in the ASM/COS sample
- indicator for if the establishment is classified as a government operation
- indicator for foreign location
- alternative establishment-level indicators (in addition to the standard ones on the LBD) useful for longitudinal linking.[14]

The final group of variables in the *cbpbr*{*year*} that are not in the LBD are those added as part of LBD processing. These include a flag for the record's source file(s) — BR file, CBP file, or both — and indicators for records that appear to have erroneously switched classification from submaster record to SU establishment record.

After we create the *cbpbr*{*year*} files we perform name and address matching to create year-to-year establishment links that will form the basis of our longitudinal establishment identifier, `lbdnum`. We turn to that matching process next.

5 Matching Across Years

This section describes the process of linking establishment IDs across year t and $t-1$ pairs. These year-pair links will be combined, as described in Section 6, to create longitudinally consistent establishment identifiers. Year-to-year matching proceeds in several steps, first using identifiers, and eventually attempting business name and address matching to avoid spurious births and deaths due to business re-organizations that sever identifier linkages.

5.1 Matching with Identifiers

The first type of longitudinal matching across year pairs that we attempt involves using establishment identifiers that already exist on the SSEL/BR files. Prior to 2002, the main establishment identifier was `cfn`, which was equal to "0+`ein`" for SUs. For MUs, the `cfn` was equal to `alpha` plus four digits that uniquely identified each establishment. After 2002, the establishment identifier was `empunit_id_char`, a unique ten-character string assigned by the Census Bureau to each establishment of a firm. We used the 2002 BR file to create a crosswalk from the 2001 `cfn` to the 2002 `empunit_id_char`.

Before matching, we drop inactive establishments. This includes submaster records, which summarize employment within an EIN and hence represent duplicate employment; ghosts, which are duplicate records of establishments that have been re-organized and received new `cfn`s; and various cases where standard CBP processing determined the record was erroneous and should be dropped.[15]

We perform matching in year pairs, starting by linking the 1976 list of payroll-active establishments to the 1977 list of payroll-active establishments. After all the year pairs have been matched

[14] Section 6 describes how these identifiers are used in LBD processing.

[15] One of the most important types of error identified by CBP processing is spurious births. These are SU firms that appear to be births with a large number of employees but in reality are old submaster records that no longer connect to an active firm, often due to a re-organization. CBP processing sets the CBP tab flag to S (submaster) or D (old submaster) for these cases.

independently, we combine the matches in Step 3 of the LBD processing, reconciling longitudinal discrepancies. In Step 2 of the LBD processing, however, we focus only on matching year pairs. For ease of exposition in this section, the first chronological year in each pair will be referred to as year1 and the second year will be year2.

The most basic match is performed first: `cfn` to `cfn` for 1976-1977 to 2000-2001, `cfn` to `empunit_id_char` for 2001-2002, and `empunit_id_char` to `empunit_id_char` for 2002-2003 to the end of the time series.

After matching by the current establishment identifier, we make use of the `ppn`, available for year pairs beginning with 1982-1983 and running through 2000-2001, to match year1 and year2 records that were not initially matched by establishment ID (`cfn`). `ppn` is intended to serve as a permanent plant number for establishments that had their regular identifier changed due to some type of re-organization. We next try to match the establishment ID (`cfn`) from year1 to `oldid` in year2 for year pairs 1976-1977 to 2000-2001. `oldid` on the year2 file records the identifier used by that establishment in year1 and is meant to preserve a record of the change. Finally we match establishment ID (`cfn`) to `ps_id` for year pairs 1978-1979 to 1983-1984 where `ps_id` serves the same purpose as `oldid` and simply comes from a CBP file instead of an SSEL/BR file. These matching steps that use establishment identifiers account for the overwhelming majority of longitudinal links in a given year pair across the entire time series.

5.2 SU-MU and MU-SU Transitions: EIN Matching

After establishment identifier matching, we next turn to handling cases where a firm changes from a single-unit to a multi-unit (SU-MU transition) and vice versa (MU-SU transition). Prior to 2002, this type of change automatically altered the `cfn`s of existing establishments at the transitioning firm. For example, if a firm was originally an SU in year1 and opened new establishments in year2, becoming an MU firm, the `cfn` of the original establishment would change from being "0+`ein`" to "`alpha`+estab counter." Thus, no establishment ID link would be possible for the original establishment. The converse problem occurs when a firm transitions from MU to SU. In this case, the continuing establishment changes its ID from "`alpha`+estab counter" to "0+`ein`" and again the establishment-level link breaks.

Absent any link between a year1 establishment and a year2 establishment, an SU-MU or MU-SU re-organization will appear as the birth of a new age 0 firm, as both the establishment ID and the firm identifier appear for the first time. To prevent this, we use the firm's `ein` to link each SU estab from year1 to the pool of establishments at the new MU in year2. We treat this group of `ein` matches as "potential matches" and among this group, we look for unique and exact matches on zip code. If one is found, we link this new establishment at the year2 MU firm to the old establishment at the year1 SU firm. If no unique or exact matches are found on zip code (perhaps there are multiple establishments at the MU firm in the same zip code), we next use street address to identify exact and unique matches. After street address, we use name of the establishment and then county FIPS code. If, after all these comparisons, we still have not found a unique and exact match, we use a string comparator function on name and street address to rank all potential matches and choose the highest ranked match. This ensures that, if we find an EIN match across year1 and year2, we will link either the SU estab from year1 to a single MU estab in year2 or the SU estab from year2 will link to a single MU estab in year1.

5.3 Name and Address Matching

After establishment ID and EIN matching, we perform name and address matching for the remaining unlinked establishments. As was the case in the legacy LBD, name and address matching is only done for single-units. The first step of the name and address matching is to identify candidate establishments. We begin by identifying all the establishments in year1 that do not match using an identifier to an establishment in year2. These appear to be establishment deaths but may in fact be re-organizations of some kind that caused the identifiers to change but did not change the business operation. We next identify all the establishments in year2 for which there was no establishment identifier match in year1. These appear to be establishment births but may be the other half of a re-organization event.

There are two types of re-organizations. The first type is one that crosses the year boundary. This occurs when a "new" establishment appears in year2 that should match to an existing establishment in year1 but has no common identifiers. We identify these types of re-organizations by matching the potential deaths from the year1 file to the potential births from the year2 file.

The second type is a mid-year re-organization. This type of event can happen in either year1 or year2. A mid-year1 re-organization occurs when a new establishment appears in year1 that is in fact the same as another establishment in the same year. For these mid-year re-organizations, we require the new establishment to continue into year2 (i.e. link by establishment id to year2) and the existing establishment from year1 to not link to an establishment in year2. We identify potential year1 births that continue into year2 by looking for establishments that had no first quarter payroll in year1 and then matched by establishment id to year2, explicitly excluding establishments that are born and die in the same year. These establishments are candidate year1 births that continue into year2.

A mid-year2 re-organization occurs when a new year2 establishment is in fact the same as another existing establishment in year2. We require that the new establishment be a birth in year2 (i.e. not link by establishment id to year1) and we require the continuing establishment to have linked by establishment id back to year1 and to have zero fourth quarter payroll in year2 and zero first quarter payroll in the year following year2 (i.e., year3). Thus the continuing establishment appears to die but is actually re-organized into the new establishment, which appears to be born in year2.

While conceptually the two types of mid-year re-organizations are the same, the differences in timing have important implications for LBD processing. We rely on some data from outside the year1-year2 window to identify mid-year2 re-organizations (i.e. first quarter payroll of year3). The group of re-organizations identified as having taken place during year2 is usually incomplete due to the fact that the year3 data available on the year2 file are still preliminary. When we move forward to link the next pair of years, we will re-label year2 to be year1 and will look again for mid-year re-organizations, now using the mid-year1 set of rules to identify cases during what was previously year2. This process will produce two sets of mid-year re-organizations for each data year, with the second set presumably being more accurate. This linking methodology requires a reconciliation between different batches of processing the same year, which we describe in Section 6. Finally, it is important to note that the quality of linking in the final year of the LBD time series will usually be lower than other years because the mid-year2 re-organizations cannot yet be validated and/or augmented by an additional round of matching.

We run two separate name and address matching processes concurrently. The first was developed as part of the Business Integrated Tracking System (BITS) and mostly involves exact matches. The second comes from the legacy LBD system and involves mostly probabilistic linking. Both systems

have been refined as part of the production integration process. After an initial match by company name to identify records that match exactly, we take the residual year1 and year2 non-matches and feed them into the BITS name and address matching AND into the LBD probabilistic name and address matching. The result is two separate sets of matching results, one from the BITS process and one from the LBD process. The sections below describe each process in detail and explain how we reconcile the output into one set of final matches.

5.3.1 BITS Name and Address Matching

The BITS name and address matching process relies on three different versions of the business name, where the business name has two parts: `name1` and `name2`. The first version is the first 28 characters of either `name1` or `name2` as it appears on the BR. We call this version the "exact name." The second version, which we refer to as the "pseudo name," makes a few changes to the original `name1` and `name2` fields. It removes all non-alphabetic, non-numeric characters, as well as common words, and concatenates the remaining words into one string to remove any blanks, keeping 28 characters in total for each name field. The third version, which we refer to as the "standardized name," removes non-alphabetic, non-numeric characters, deletes `name2` if it begins with "%" or "ATTN," replaces common words with specified abbreviations, removes other common words (e.g., "and," "company") and one-character strings, and abbreviates city if it appears in the name. The remaining characters are saved to a 12 character string field with no spaces. If the final string has fewer than 5 characters, the standardized `name1` or `name2` is set to blank.

To create an address field that is useful for matching, the process keeps the first 12 numeric values from the street address field from the BR. This generally corresponds to the house/building number portion of the establishment's address. Hence this address matching is relatively simple and relies on an exact match between the building number of two establishments in order to identify a re-organization.

Using the three versions of name and the simplified version of address, we begin by matching on exact `name1` and exact `name2`. For each type of re-organization, we make three attempts: `name1` to `name1`, `name1` to `name2`, `name2` to `name1`. Thus for year to year re-organizations, we attempt to match `name1` from year1 "deaths" to `name1` year2 "births," followed by matching `name1` to `name2`, and `name2` to `name1` for this same group of establishments. Next, for mid-year1 re-organizations, we match `name1` from year1 "deaths" to `name1` from year1 estabs that are missing quarter 1 payroll and that continue into year2, again followed by matching `name1` to `name2` and `name2` to `name1` for the same group. Finally, for mid-year2 reorganizations, we match `name1` from year2 "births" to `name1` from a year2 continuing establishment that has missing quarter 4 payroll, after which we again match by `name1` to `name2` and `name2` to `name1`.

We then move to matching by pseudo names and repeat the process outlined above. In order, we match year-to-year re-organizations, mid-year1 re-organizations, and mid-year2 re-organizations by the same three combinations of pseudo name as we used for exact name. After this is finished, we repeat the process using the standardized names. After all the name matching is completed, we match using the standardized address field described above.

At each stage of the name and address matching process, records that are determined to match are removed from the pool of potential matches and only unmatched records are passed to the next stage. Thus matches are prioritized based on our priors about quality. Exact name matches are deemed to be the highest quality (i.e., most likely to be true matches) and those matches are identified and removed first. Matches identified using our simplified address are deemed to be the lowest quality and hence are saved for last, after all other forms of matching have been exhausted.

5.3.2 LBD Name and Address Matching

The LBD name and address matching process is run concurrently with the BITS pseudo and standardized name matching and BITS address matching. After the exact name match described in Section 5.3, we feed the residual unmatched year1 and year2 records into the LBD probabilistic name and address matching.

This process begins by doing name and address standardization. To standardize `name1`, `name2`, and address, we use the SAS Data Quality Server function DQSTANDARDIZE and call on the database ENUSA. We also remove common words such as "the," "of," "company," or "LLC" from `name1` and `name2`. For address we convert numbers written as words to numeric values and drop standard words such as "street," "road," "floor," or "num."

The next step is to create "fuzzy" versions of the name and address that will allow us to match across establishments in spite of small spelling differences. To accomplish this, we use the SAS Data Quality Server function DQMATCH, also with the database ENUSA, to create less precise versions of our linking variables, where the level of precision (or conversely the level of fuzziness) is ranked by a numeric score. Variables with precision of 95 are essentially unchanged from their original format. Variables become increasingly fuzzy relative as the precision score decreases, with the lowest possible level being 50. As with BITS, we utilize both `name1` and `name2` and create several different fuzzy versions of these variables, including: standardized `name1` and `name2` at sensitivity level 65; standardized `name1` and `name2`, compressed to remove spaces and special characters but left unfuzzed; standardized `name1` and `name2` appended together to create one name variable and fuzzed at a sensitivity level of 50. At the same time, we also utilize fuzzy versions of the mailing street address and the physical street address. We create versions of street that mirror those created for names: standardized street at sensitivity levels 50 and 65; standardized and compressed mailing and physical street with spaces and special characters removed but left unfuzzed. Finally, we create both a physical and mailing zip3 variable which contains the first 3 digits of the establishment's zip code.

As with the BITS matching, we do multiple rounds of matching using increasingly fuzzy match variables. Establishments that do not match in any given pass are moved forward to the next pass. In each pass, we directly match two name variables, one street variable, and one zip variable across establishments and identify matches where the fuzzy variables agree. For each set of fuzzy `name1` and `name2` variables, we do four passes where we match physical street to physical street, physical street to mailing street, mailing street to physical street, mailing street to mailing street and likewise for zip code. For example, the first pass matches on the compressed versions of the standardized `name1`, `name2`, and physical street name and the first three digits of the physical zip code. The subsequent second, third, and fourth passes match on the same name variables but change the street address and zip code so that physical is matched to mailing in both directions and mailing is matched to mailing. These steps ensure that if the BR mistakenly switches how addresses are stored, our matching algorithm will not fail. We also cross-match the name variables to account for the possibility of changes in the name of the business as stored in the BR. In our fullest cross-matching, we attempt to match the following combinations.

- `name1-name1` and `name2-name2`
- `name1-name2` and `name2-name1`
- `name1-name1` with `name2` missing
- `name2-name2` with `name1` missing

- `name1`-`name2` with the other names missing
- `name2`-`name1` with the other names missing

After we finish the probabilistic matching with fuzzy names and addresses, we perform four additional passes that match on common words in business names. In these passes, we first create a set of potential matches based on records with the same physical or mailing state, standardized physical or mailing street at sensitivity level 70, and physical or mailing five-digit zip code. Within this group of potential matches, we count the number of common words between the non-standardized versions of `name1` and `name2` at a sensitivity level of 95 for each pair of establishments and keep matches with at least two common words that are unique within the group of potential matches. In other words, we will declare establishment A to be a match to establishment B if they are both in the potential match block (defined by state, street70, and zip5), have at least two words in common, and no other establishments also have two of those words in common with establishment A or establishment B.

Choosing which variables to attempt to match and what levels of precision to use required some experimentation. Our goal was to implement a system that maximized the number of correct links (true positives) while minimizing the number of false matches (false positives) and the number of missed matches (false negatives).

To assess the quality of our probabilistic match passes we performed clerical review of 2,900 match candidates with three reviewers per record. In the production processing, we used five main types of probabilistic matches:

- standardized, compressed, and unfuzzed `name1`, `name2`, and street name (24 total passes)
- standardized, compressed, and unfuzzed `name1`, `name2` and street name with sensitivity level 65 or 50 (16 total passes)
- fuzzy `name1` and `name2` with sensitivity level 65 and standardized, compressed, unfuzzed street name (24 total passes)
- fuzzy `name1`, `name2`, street name with sensitivity level 65 (4 total passes)
- oncatenated and fuzzy `name1` and `name2` with sensitivity level 50 and standardized, compressed, unfuzzed street name (four total passes)

In each of these groups, we also match on the first three digits of the zip code. The Step 2 LBD programming specifications provide complete details on the matching variables used in each of the 72 probabilistic passes plus the four common word passes.[16]

5.3.3 Reconciliation of BITS and LBD Matches

Since the BITS and LBD matching processes run in parallel, they have the potential to produce different establishment matches. There are three types of disagreement between the BITS and LBD matches. First, the BITS process finds a match and the LBD process does not. Second, the LBD process finds a match and the BITS does not. Third, both BITS and LBD find matches but they do not agree. The most common disagreement is that the LBD linking finds a match while BITS

[16]Final versions of LBD programming instructions are stored in the Census Bureau data warehouse, together with the LBD microdata, and available to researchers with FSRDC projects authorized to use the LBD.

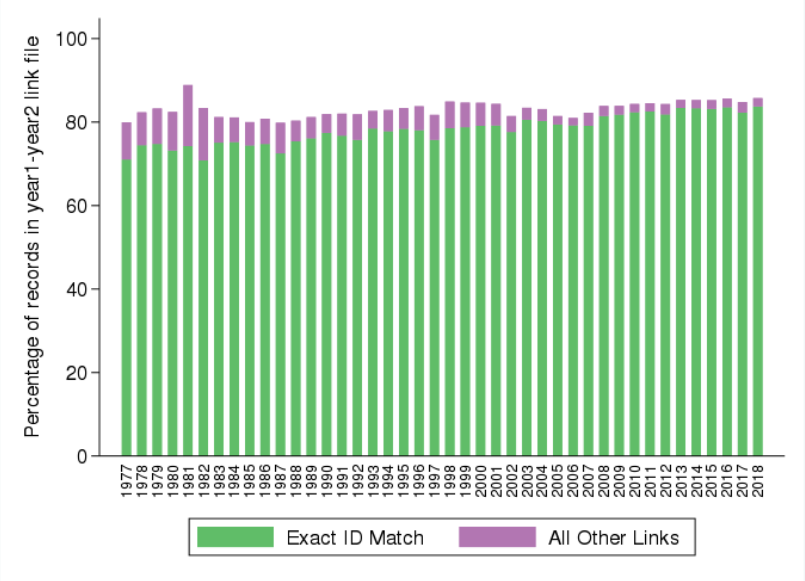
does not. The next most common outcome is that the BITS process finds a match while the LBD does not. Conflicts between the two matching approaches are rare. We reconcile these discrepancies by taking all unique BITS matches, all unique LBD matches, and preferencing the BITS match, as the more conservative of the two, when both processes find a link.

The final output from the matching process is a set of files that contain links across year pairs. Each link file contains `id1_{year1}`, `id2_{year1}`, `id1_{year2}`, `id2_{year2}`. These four variables capture all possible links within a given year or across a year pair.

5.4 Matching Summary Statistics

In Figure 1, we show the percentage of establishment records in a two-year pair that linked by an exact ID match using the BR ID (either `cfn` or `empunit_id_char` depending on the year) and the percentage that linked using one of our other methods described above over our full time series. Each year bar is labeled with the second year of the year pair. In 1977, 71.05% of records that had positive payroll in either 1976 or 1977 were matched to each other using `cfn`. An additional 4.46% were matched using alternative IDs such as `ppn`, `ein` in an SU-MU or MU-SU transition, or name and address in a single-unit re-organization. The year pair 1980-1981 has the highest percentage of alternative links (7.33%). This percentage declines to between 2% and 3% in the 1990s and continues to fall to around 1% in the 2000s. The increasing height of the exact ID match bar over time suggests two trends. First, there are more continuing establishments over time, as the number of entrants and exits in the economy relative to the number of continuers has decreased over time, and this leads to more exact ID matching. Second, the BR has gotten better at tracking firms over time, as evidenced by the decreasing size of the all other links bar. In 2018, only 1% of establishments needed to be linked by an alternative method.

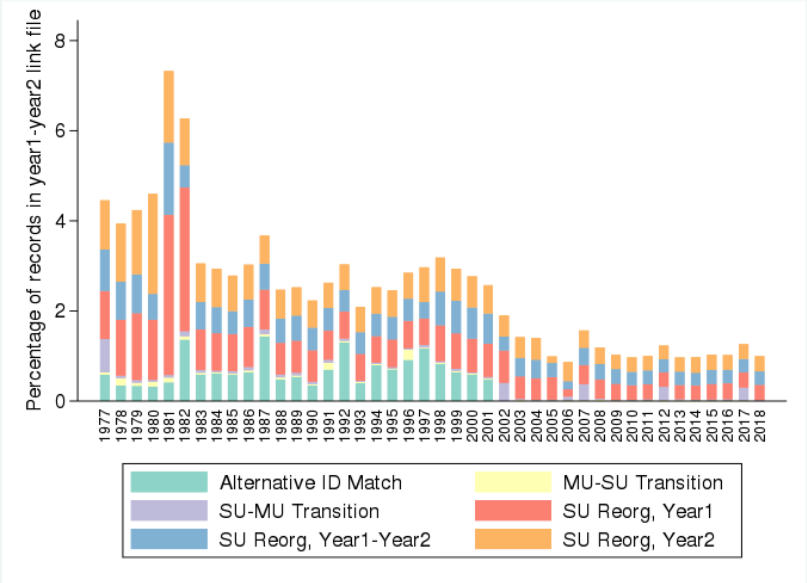
Figure 1: Year-to-Year Match Type Distribution



Source: LBD Production Year Pair Match Files 1976-1977 through 2017-2018
Note: Percentage of records that match between a given pair of years. Each year on the graph represents the second year in the pair. Matches are divided into those made using the Business Register establishment ID field and those made using an alternative method. Records that do not match represent births and deaths.

Next in Figure 2, we show the breakdown of "all other links" into its component parts: alternative ID match (`ppn`, `oldid`, `ps_id`), MU-SU transition or SU-MU transition identified by EIN and geography, or name and address matching either within year1, across year1 and year2, or within year2. Single-unit re-organizations are split roughly equally across the three types, a feature of the data that remains fairly constant over time. We see the most re-organizations in the late 1970s and early 1980s, most likely a feature of data quality as opposed to an indicator of economic trends. In contrast, SU-MU transitions peak in Economic Census years, as expected, a feature that continues to the end of the series. The alternative ID matches vary in importance prior to 2002 and then vanish completely as the BR adopted `empunit_id_char`, a new type of identifier that did not need to be replaced due to changes in multi-unit status and which generally tracked establishments better over time.

Figure 2: Composition of non-Exact Year-to-Year Matches



Source: LBD Production Year Pair Match Files 1976-1977 through 2017-2018
Note: Percentage of records in a pair of years that match by alternative methods. Each year on the graph represents the second year in the pair. Categories of matches include alternative BR establishment IDs, EIN matches that span firm organizational changes (SU-MU or MU-SU), and name and address matching to identify single-unit re-organizations.

Finally, in Figure 3, we compare the alternative links from the new production LBD to alternative links in the legacy LBD. Here we see that for establishments linked between two years using something other than the main BR ID, between 30-40% of them have this type of link in both the production and legacy LBD. An additional 40-60% of these links are made only by the production LBD and 5-20% are made only by the legacy LBD. The highest percentages of legacy-only alternative links are found in more recent years, although the audited linking system of the production LBD does not corroborate them all. The years 1978-1981 have the lowest percentages of alternative matches found in the legacy LBD, with the legacy-only comprising 6% or less of total re-organizations and matches made in both the legacy and new production systems being less than 30%. This is consistent with the new production system standardizing the matching across the full time series to push improvements in the matching algorithms backwards.

Figure 3: Legacy and Production LBD Alternative linking Comparison

Source: Legacy LBD and LBD Production Year Pair Match Files 1977-1978 through 2014-2015
Note: Division of all alternative links from either the legacy or production LBD into those found in either or both versions of the database.

The linking procedures described in this section represent one of main contributions of LBD processing, namely repairing artificial breaks in the histories of individual establishments. The linking of records with different establishment ids over time serves as the foundation for creating `lbdnum`. We now turn to a description of how the year-to-year match files are reconciled and combined to create a file that captures the history of each establishment over the entire time series.

6 Longitudinal Links

In this section we describe the construction of a "linkwide" file, which captures the combination of all longitudinal links from 1976 to the end of the time series. Each observation in this linkwide file represents a unique establishment, with the variables containing the establishment identifier(s) for the given establishment in each year. This file is constructed by iteratively merging the year pair link files described in Section 5. This process begins with the 1976-1977 pair and adds subsequent year-pairs chronologically one at a time. During each iteration of the merging, certain discrepancies and inconsistencies between the two-year link files are resolved.

6.1 Combining Year Pair Matches

The output of this process is referred to as the "linkwide" file, which is named *lbd1976*{*t*} for the current year *t* of processing. The program takes as its inputs the linkwide file from the previous year of processing *lbd1976*{*t-1*}, and the current two-year link file to be added (*link*{*t-1*}{*t*}). During the very first iteration, the 1976-1977 year pair link file is used as *lbd19761977*, and the year pair link file from 1977-1978 is merged on to it. From then on, the linkwide file created in the previous iteration is used as the input file in the current year's processing.

The merging of each year pair into the previous year's linkwide file is based on the shared information that the two files have about the previous year. As described in the previous section, mid-year re-organizations in a given year are identified twice–once when that year is processed as year1 and once when it processed as year2. Since the year pair link file contains an `id1` and potentially an `id2` for each $t-1$ and t, the primary merging variables are `id1_{year-1}` and `id2_{year-1}`. The goal of the merging is to attach the new information for `id1_{year}` and `id2_{year}`, and reconcile any possibly conflicting information contained in the ID fields.

The process consists of several rounds of merging, each representing different types of cases. The most straightforward merging occurs when the year $t-1$ IDs are perfectly consistent between the prior year's lbdwide and the current year's year pair link file. These represent the large majority of cases, and are simply included in the final lbdwide file without any further processing. However, because year $t-1$ data is processed twice – both in the $t-2$ to $t-1$ link file and in the $t-1$ to t link file – the potential exists for there to be conflicting `id1`s and `id2`s for $t-1$, particularly in the case of re-organizations. Therefore, we make various comparisons of `id1_{year-1}` and `id2_{year-1}` between the two input files in order to identify the common establishment. Next we apply a set of rules to reconcile the remaining ID fields which contradict one another, either in their existence, their order, or in the IDs themselves. These rules follow the ones initially developed by the BITS program, and emphasize a preference for longitudinal consistency and for new data over old. As a result of this algorithm we identify certain ID fields as erroneous, either in the prior year's lbdwide file or the new link file, which requires links to be broken. We put the newly uncoupled establishments through a final round of ID merging that may result in new links.

6.2 Reactivations in the Linkwide File

The year pair link files are unable to detect whether an establishment birth is actually a reactivating establishment that was inactive in the prior year. We use the linkwide file to search backward up to seven years in order to determine whether an apparent year t birth was, in fact, active in a previous year. We classify such cases as having been temporarily inactive, and record their activity in year t on the same row as the previously active establishment. This process retains longitudinal linkages spanning gaps in activity up to seven years, conceptually assigning the same `lbdnum` before and after the gap in activity. Table 3 reports the average number of reactivations during the time-series broken down by the number of years (2-7) the establishment was inactive, and the share of establishment births that each type of reactivation represents. We observe that a substantial fraction of entry is attributable to establishments that were once active in a year prior to $t-1$. Approximately 10% of all births are in fact reactivations over the period of 1978 to 2018, with around 5% representing establishments who were last active only two years before. The share of births that are reactivations from longer ago in the past declines the further back we look, but even the 7-year reactivation category numbers in the thousands per year on average.

Table 3: Establishment Reactivations 1978-2018

Type of Reactivation	2-year	3-year	4-year	5-year	6-year	7-year
Avg. Num. of Establishments per year	39,751	17,522	8,373	5,552	3,943	2,936
As a Pct. of Establishment Births	5.29%	2.33%	1.11%	0.74%	0.52%	0.39%

Notes: This table shows the average yearly number of establishments that reactivate after a spell of inactivity, categorized by the number of years since they were last observed as active. The fraction of establishment births that each category represents is also reported. The sample comprises all establishment births from 1978-2018 pooled together across all years.

6.3 Linkwide Match Flags

Besides the `id1` and `id2`s for each year of the time series, the *lbd1976{t}* file also contains a set of flags called `flag_match_pass_{year}`, which describe how each yearly link was created for the given establishment. In most cases, where the matching in the *link{t-1}{t}* was not deemed to be erroneous, the flag simply passes along the information about how the match was created in the year pair matching process, described in the previous section. In the cases where IDs had to be decoupled, deleted, or re-matched by the linking process, the flag is edited to report the type of match. A description of the major categories of these kinds of edits, and the way that they are flagged, is included in Appendix B Table 3. A complete list of the values that the variable `flag_match_pass_{year}` can take is reported in Appendix B Table 4.

The complete time series of links for a given establishment, along with gaps in activity, provides the basis for many of the core concepts in the LBD and BDS. Entry, reactivation, re-organization, continuation, and exit are all readily observable in these data. Since each observation represents a unique establishment, `lbdnum`s are assigned according to the row of data on which each establishment is recorded in the linkwide file. This file represents an important piece of the data infrastructure, and is a key input for many of the subsequent processing steps described in the following sections. In the next section, we describe how we address the spurious entry and exit observed in Economic Census years.

7 Retiming Births and Deaths

This section describes the algorithms used to address the spurious bunching in the administrative data of entry and exit in Economic Census years. To understand the challenge of accurately determining when an establishment is born, it is helpful to first consider the two types of establishment births that appear in Census Bureau records.

The first type of birth occurs when a new single-unit establishment begins to operate and files payroll taxes for the first time using a new EIN. This type of birth is recognized immediately by the Business Register staff when the new EIN fails to link to any already existing EIN in the BR. Once we have determined that this new EIN is not a re-organization of an old establishment that operated under a different EIN (see Section 5), we know the exact year that the establishment was born.

The second type of establishment birth occurs when an existing firm opens a new establishment. The establishment may be the second establishment to begin operations within that firm, transitioning the firm single-unit to multi-unit. Alternatively, the new establishment may be added to a

firm already classified as a multi-unit. These types of births are difficult for the BR to track because they are not readily observable in tax records. Small multi-unit firms often file taxes for multiple establishments on a single form (EIN), in which case the additional establishment will appear as a change in the number of employees.[17] The reverse is true for deaths. When a single establishment firm goes out of business, its EIN ceases to report taxes to the IRS and the BR recognizes the death. However, when a multi-unit firm closes an establishment, the tax records do not reflect the change in firm structure–the BR will only reflect a change in employment at the firm.

To determine and update the structure of firms and where they are doing business, the Census Bureau relies on annual surveys such as the COS, the Annual Survey of Manufacturers, and the quinquennial Economic Census. In years that end in "2" or "7", most establishments are sent an Economic Census form and asked to report on the number of operating establishments they have. In other years, only large multi-unit companies and a sample of small multi- and single-unit companies are sent the COS or ASM. We can accurately time establishment births and deaths at multi-units only if they are in a survey or census two years in a row. For example, if a single-unit fills out an Economic Census form in 2002 and then in 2003 it is sampled by the COS and reports a new establishment birth, we can accurately date this birth and the transition to multi-unit status to 2003. However, if the single-unit is not surveyed again until the next Economic Census in 2007, any new establishments reported in this year will have uncertain birth years, with the possibilities ranging from 2003-2007.

An important complication of the changes in firm structure observed in the COS is that the sample structure has changed over time. Since 2005, a small number of single-units have been surveyed by the COS each year, but prior to 2005, none were surveyed. Even among single-units that are surveyed in the COS, they are unlikely to be surveyed twice in intercensal years, making it difficult to accurately time their births and deaths.

An innovation of the LBD is the use of dates of birth and death reported in the COS and ASM by known multi-unit firms to build an imputation model for establishments with unknown beginning and end dates. We call this imputation process "re-timing" because we change the year of an establishment's birth or death. The fundamental assumption behind this model is that establishments with known dates of operation have patterns of opening and closing that are similar to those of establishments with less precisely known dates of operation, conditional on observable firm characteristics such as industry and geography. One reason this assumption may not hold is that the training data for the imputation model consist almost entirely of establishments born to large multi-unit firms whereas the set of establishments with missing data comes almost entirely from small multi-unit or single-unit firms.[18]

Researchers should exercise caution when using re-timed births and deaths in any analysis of the LBD microdata. While the results from the imputation model smooth the BDS aggregate

[17] If one establishment opens and all the other establishments maintain their employment levels, there will be an increase in total firm employment. However if other establishments shrink at the same time a new one opens, overall firm employment could remain constant or even decline.

[18] We have made some efforts to link data from the Longitudinal Employer Household Dynamics (LEHD) infrastructure files on establishment births and deaths to Business Register data as the State Unemployment Insurance (UI) systems that provide LEHD source data generally update their records on the number of operating locations for a firm at least once a year. While augmenting our training data in this manner has shown promise for improving the imputation model, difficulties in firm and establishment linking across the two separate business databases have prevented us from implementing this approach in production. Instead, we have chosen to rely on the Business Register data and leave integration with LEHD data as a topic for future research. The focus of current efforts has been to build a modeling system that is flexible enough to utilize future training data as it becomes available and to smooth the current time series of establishment births in a plausible way.

time series, they are unlikely to be highly reliable estimates of birth and death dates for individual establishments. Without appropriate treatment of the uncertainty introduced by the birth and death imputations, analysis of correlations between establishment characteristics and opening and closing patterns could be biased. We advise researchers to use either changes between Economic Census years when studying the number of establishments opening or closing over time or to use firms that were surveyed by one of the Census Bureau's annual surveys, such as the COS or ASM.

Our re-timing efforts have some other limitations as well. Some establishments begin or end in intercensal years but have an uncertain birth/death date because they were not sampled every year by the COS/ASM. We do not use these establishments in our training data nor do we impute new birth/death dates for them. We take a conservative approach in the imputation model, focusing only on the spurious entry and exit in Census years. Some uncertainty about dates remains even among establishments that appear or disappear in intercensal years.

It is important to note that our imputation algorithm does not, conceptually, impact the timing of firm births and deaths. This is intentional since for firms, the time series of births and deaths does not generally display substantial bunching in Economic Census years. The anomalies that exist for firm entry and exit seem to be caused by factors different than those addressed by the establishment re-timing model.

7.1 Re-timing Imputation Model

In the implementation of our imputation model, we utilize the Sequential Regression Multivariate Imputation (SRMI) method of Raghunathan, Lepkowski, van Hoewyk, and Solenberger (2001), which has been implemented in other production settings at the Census Bureau such as imputation of missing survey responses in the Survey of Income and Program Participation (see SIPP 2014 Panel Users' Guide) and the creation of the SIPP Synthetic Beta (see Benedetto, Stinson, and Abowd (2013) and Benedetto, Stanley, and Totty (2018)). SRMI involves running regressions using a set of observations with no missing data as training data to estimate the relationships between observable, non-missing characteristics and the variable to be imputed. These relationships are then used to predict a value for the observations that are missing this information. The accuracy of imputations improves as the set of establishments used to estimate the statistical relationships becomes representative along more dimensions of the set of establishments which require an imputed value. To increase similarities and reduce heterogeneity, it is common to stratify the training data and run regressions on subsets of establishments that share a set of common characteristics.

For our application of SRMI, we first stratify our sample of establishments that were born or died in a single Economic Census year. This entails splitting the training data into segments of births/deaths that were reported in EC year t but could have occurred within the previous four years, $t-1$ to $t-4$, with 1982 being the first EC year and 2017 being the last. This allows us to estimate time-varying relationships between our predictor variables and year of establishment birth/death. We further stratify births/deaths by characteristics of the EIN to which they belong: EIN size categories created with respect to establishment count, EIN size categories created with respect to employment, and two-digit industry.[19] Finally we categorize establishments based on the year during the intercensal period when the EIN to which they belong had the largest employment growth or decline. For example, using the 2013-2017 intercensal period, establishment births belonging to EINs whose employment grew the most between 2014 and 2015 are grouped together and establishments belonging to EINs whose employment grew the most between 2015 and 2016 are grouped together.

[19]Establishment count categories are 2-5, 6-10, 11-25, or 26+ establishments; employment categories are 1-100 or 101+ employees

Taken together, establishments within the same stratum were (1) first/last reported in the same Census year, (2) belonged to EINs whose largest employment growth/decline happened at the same point in the intercensal period, (3) had similar numbers of operating establishments and total employees in the EC year, and (4) operated in similar industries.

Within these strata, we regress year of birth or death on various measures of employment growth or decline using our training data, establishments with known birth and death dates. We use estimated coefficients from those regressions to predict the missing year of birth/death for the uncertain births/deaths. Specifically we use predictor variables specific to a given EIN such as the rate of change of total EIN payroll and total EIN employment and we use state-level measures of job creation and destruction in the same industry as the EIN to which the establishment belongs. We include these measures for every year in the 5 year window ending with the EC year. We also include state and two-digit industry dummies and actual EIN employment and number of establishments. The regression we estimate is a multinomial logit where the dependent variable can take on the following values: EC year, EC year - 1, EC year - 2, EC year - 3, and EC year - 4. A minimum birth year or death year is set so that an establishment cannot be imputed to be born in a year before the firm was known to have positive employment.

The largest challenge we face with our stratification method is that many cells become too small to estimate meaningful statistical relationships. To combat this problem, we require that a stratum have at least 2000 observations from the training data before we run regressions for this group of establishments. Strata smaller than this threshold are iteratively coarsened. After stratifying our sample by our full list of categorical variables, we check the size of each stratum and drop those that are not large enough. We estimate regressions for the remaining groups that meet the size requirements and perform imputations for records with missing data that belong to these strata. After the first round of regressions, we drop one stratifying variable and re-stratify, creating coarser groups of establishments. We check observation counts, identify sufficiently large strata, estimate regressions on those strata, reduce the stratifying variables, and re-stratify. This process continues until all necessary imputations are done or until we have dropped all the stratifying variables except EC year. The stratifying variables are dropped in the following order: two-digit industry, firm employment category, firm establishment count category, intercensal year of maximum employment growth. For further details on the implementation of the SRMI method, see Benedetto, Stanley, and Totty (2018).

7.2 Re-timing Modifications to *cbpbr{year}* and *lbd1976{finalyear}*

The results from birth/death re-timing are used to modify the wide link file described in Section 6. For establishments that have their birth year set to an earlier point in time than the originally reported EC year, we push their identifiers to earlier year fields in the linkwide file (*lbd1976{finalyear}*) and create new records in the *cbpbr{year}* files. For deaths, we remove identifiers from year fields in the linkwide file and drop records from the *cbpbr{year}* files. Total employment for an EIN in a given year is never changed. In years where a new establishment was imputed to be in operation, we re-allocate employment across all the establishments in operation that year. For births and continuers, we use the share of total employment in the EC year to allocate employment in intercensal years. For deaths, we use the share of employment in the prior EC year to allocate employment in intercensal years prior to the imputed death date. While total EIN employment is unchanged in every year, BDS calculations of job creation and destruction will change across the years as employment is reassigned from an entering establishment to an existing establishment. For example, if EIN employment is equal to 40 in EC year 2007 and also equal to 40 in year 2006, but

there were two establishments in year 2007 and only one establishment in 2006, then the employment associated with the new establishment in 2007 would be classified as job creation. However, if this new establishment was imputed to be born in year 2005, then there would be no job creation between year 2006 and 2007 as both employment and the number of establishments would stay constant between 2006 and 2007.

8 Industry Classification in the LBD

In this section, we provide an overview of the industry classification systems used to categorize an establishment's primary activity in the LBD, and provide details on the industry code variables in the data. The Census Bureau uses industry classification systems to assign individual establishments to a primary industry. Industry-level statistics, including those for employment, are therefore based on an establishment's industry code.

The Standard Industrial Classification (SIC) system, first developed in the 1930s, is used to classify establishments prior to 1997. This system was updated numerous times to account for changing economic activity, for example with the emergence of new computer and software activities. We refer to each update within the SIC system as a new "vintage." The LBD contains SIC codes on SIC1972, SIC1977, SIC1987, SIC1992, and SIC1997 vintages.[20] Official SIC codes consist of four digits.

By the 1990s, the SIC system faced several significant limitations. First, its existing categories could not adequately classify new types of activities, especially those in services and new technologies. Second, it classified activity based on a number of different concepts, including both production and demand-based definitions. Third, it was not easily comparable to international classification systems. In 1997, after years of cross-agency discussions and analyses, US statistical agencies addressed these issues by initiating a transition from SIC to the North American Industrial Classification System (NAICS).

Under NAICS, all industry categories are based on production concepts. Establishments are classified into industries based on the activities performed at the establishment, rather than as a function of the products they sell or the customers they serve. The NAICS production-process approach was chosen to provide consistency across different industries, and to facilitate comparisons across the US, Canada, and Mexico at the four-digit level. NAICS was also designed to be flexible and is updated every 5 years, during Economic Census years (i.e., years that end in 2 or 7) to a new vintage. The 2018 LBD contains NAICS codes on NAICS1997, NAICS2002, NAICS2007, and NAICS2012 vintages. Official NAICS codes consist of six digits.

We recommend that researchers use the system most proximate to the time period of their analysis. For example, using SIC to study economic activity during the 1980s is likely to induce the least amount of measurement error. Alternatively, using SIC during the 2000s will lead to considerable noise, especially in terms of capturing the importance of new activities, such as Computer Systems Design. Research that spans the transition across systems in 2001 must address a number of significant changes between SIC and NAICS. We discuss these issues, along with a vintage-consistent industry code that spans this transition, in Section 8.2 below.[21]

[20] SIC1992 and SIC1997 are not official vintages with published descriptions or concordances. Instead, these are codes developed internally with additional detail that we exploit when concording the data.

[21] See Fort and Klimek (2018) for an analysis of the differences between SIC and NAICS.

8.1 Industry Codes Over Time

The LBD contains three industry variables: `sic`, `naics`, and `bds_vcnaics`. In this subsection, we describe the `sic` and `naics` variables and how to use them. The `sic` and `naics` variables in the LBD are raw industry codes derived from either the BR or CBP microdata. The initial information source for the BR industry codes is tax returns, but this is supplemented with information from sources such as the Bureau of Labor Statistics. The Economic Censuses collect direct, and generally more reliable, information on establishments' activities and industry in Economic Census years. The EC staff apply their expertise to assign the most accurate industry code to establishments. Industry code information from the EC data is usually incorporated into the BR two years after the EC (i.e., it is incorporated in years ending in 4 and 9).

The `sic` variable in the LBD is populated for the majority of establishments from 1976 to 2001. The `naics` variable is populated for the majority of establishments from 2002 to the present. While continuing establishments that existed prior to 2001 often have values for `sic`, those values simply consist of their last reported SIC code during the SIC era, and do not necessarily represent the establishment's current activity. The default source for the `sic` and `naics` variables is the BR but the CBP value may be selected for several reasons. In cases in which the BR industry code is missing, or appears to be a partial code (i.e., it has one to five leading digits that correspond to a valid code but is padded with zeros at the end such that the full six-digit code does not correspond to a valid code), we use the CBP industry code for a particular record if available. We also select the CBP value if we suspect that a mid-year update occurred. As described in Section 4, the CBP data contain mid-year updates to the BR that are missing from our BR data. We identify mid-year updates, selecting the CBP value, whenever the BR and the CBP industry codes differ in year t, but agree in year $t+1$, and the CBP value remains unchanged. This is the only longitudinal change we make to the raw `sic` and `naics` codes in the LBD. Table 4 provides the years in which each of the `sic` and `naics` variables is available in the LBD, as well the primary vintage of each variable in each year.[22]

The same establishment may have different `sic` or `naics` codes in different years for three reasons. First, a code may change due to a change in the source (and potentially the quality) of information for that code. For example, the BR updates industry codes when additional information on an establishment is collected in the EC. Second, a code may change due to a change in the SIC or NAICS vintage used in year t versus $t+1$. Third, an establishment's industry code may change due to an actual change in the establishment's primary activity. These types of changes are most likely to be collected in EC years.

As noted above, we recommend that research focused solely on the SIC time period (e.g., prior to 2001) rely on the first four digits of the `sic` variable.[23] As evident in Table 4, this may require mapping the various SIC vintages to a consistent vintage. Electronic versions of the SIC1972 to SIC1977 and SIC1977 to SIC1987 concordances are available in Fort and Klimek (2018). It is also important to note that SIC codes in the EC data are likely to be the most accurate, and may take two years to appear in the LBD. Research focused on industry code changes may choose to rely most heavily on that information.

[22]It is important to note that the raw `sic` and `naics` variables often contain a mix of vintages. Table 4 reports the primary vintages. In years in which the BR and CBP industry vintages differ, the LBD industry variable contains a mix of those vintages.

[23]The last two digits in the `sic` variable are used by Census Bureau staff, but do not correspond to official codes. For example, in the 1997 to 2001 period, these digits can be used with internal concordances to map uniquely from SIC to NAICS.

Table 4: Classification systems and vintages in the LBD and data sources

Year	BR	CBP	Economic Census	LBD
1976-1978	SIC72	SIC72	SIC72, SIC77	SIC72
1979-1986	SIC77	SIC77	SIC77	SIC77
1987	SIC77	SIC77	SIC87	SIC77
1988-1996	SIC87	SIC87	SIC87	SIC87
1997	SIC97	SIC97, NAICS97	SIC97 NAICS97	SIC97
1998	SIC97	SIC97, NAICS97		SIC97
1999	SIC97	SIC97, NAICS97		SIC97
2000	SIC97	SIC97, NAICS97		SIC97
2001	SIC97	SIC97, NAICS97		SIC97
2002	NAICS97	NAICS97	NAICS97 NAICS02	NAICS97, NAICS02
2003	NAICS97	NAICS02		NAICS97, NAICS02
2004	NAICS97, NAICS02	NAICS02		NAICS02
2005	NAICS02	NAICS02		NAICS02
2006	NAICS02	NAICS02		NAICS02
2007	NAICS02	NAICS02	NAICS02, NAICS07	NAICS02
2008	NAICS02	NAICS02		NAICS02
2009	NAICS02, NAICS07	NAICS07		NAICS02, NAICS07
2010	NAICS07	NAICS07		NAICS07
2011	NAICS07	NAICS07		NAICS07
2012	NAICS07	NAICS12	NAICS07, NAICS12	NAICS07
2013	NAICS07	NAICS12		NAICS07, NAICS12
2014	NAICS07, NAICS12	NAICS12		NAICS12
2015	NAICS12	NAICS12		NAICS12
2016	NAICS12	NAICS12		NAICS12
2017	NAICS12	NAICS17	NAICS12, NAICS17	NAICS12
2018	NAICS12	NAICS17		NAICS12, NAICS17

Notes: Table describes the primary vintages for the `sic` and `naics` variables in the input data sources and the LBD. SIC97 is an unofficial and internal SIC code used in the SIC-NAICS transition in the BR. Note that a particular vintage may be present in any year, most commonly in years after the vintage was replaced.

8.2 Vintage-consistent NAICS codes

There are substantial differences between SIC and NAICS that must be addressed prior to conducting a longitudinal analysis that spans the transition across systems. First, the rise of codes classified as services is evident in the fact that the transition led to a 36 percent increase in the share of employment classified as Services in 1997 (Fort and Klimek, 2018). This increase highlights the fact that changes occurred not only within detailed industries, but also across broad sectors, and demonstrates that the transition cannot be addressed by aggregation. Since many SIC codes map to multiple NAICS codes, a simple concordance cannot be used to assign a NAICS code. Second, SIC and NAICS differ in their treatment of headquarter and auxiliary establishments. Auxiliaries are defined as those establishments primarily serving other establishments of the same enterprise. Under SIC, auxiliaries are classified in the primary industry of the establishments they serve. In contrast, NAICS classifies auxiliaries based on what they do. For example, an R&D establishment that primarily conducts biotechnology research for the manufacturing establishments in its firm

would be classified: a) in the two-digit SIC industry code of the plants that it served under SIC, but b) in Scientific Research & Development (541711) under NAICS.[24]

To calculate BDS statistics over the entire LBD time period, and to facilitate research spanning the SIC-to-NAICS transition, we incorporate vintage-consistent NAICS codes, first developed by Fort and Klimek (2018), into the LBD. The main LBD dataset includes the variable `bds_vcnaics`. This variable contains the most recent vintage NAICS code (NAICS2012 in the 2018 version of the LBD) for all payroll denom active establishments in the LBD. This includes establishments with positive payroll in t or $t-1$, with the `bds_vcnaics` code carried forward for exits. For auxiliary establishments, `bds_vcnaics` contains the most recent vintage NAICS code for what the establishment does.

The *naics*{*year*} files contain additional vintage consistent industry codes that will aid researchers' use of the vintage most closely tied to their period of analysis. The files contain `fk_naicsZZ` codes, where `ZZ` refers to the vintage of the industry code.[25] Each vintage code is available through `ZZ`+7. For example, `fk_naics02` contains NAICS2002 codes and is available through 2009. To minimize noise due to changes in the NAICS vintages, researchers may use these files to access the vintage that provides full coverage for their time period, but does not go beyond it. For example, research spanning 1999 to 2009 can be conducted using NAICS2002 codes, rather than relying on the NAICS2012 vintage in the LBD files. This will avoid two additional industry mappings: from NAICS2002 to NAICS2007 and from NAICS2007 to NAICS2012, in which some noise will enter the data due to random assignments, as described in the next section.

During SIC years, the files *naics*{*year*}*_aux* contain a NAICS2002 code that corresponds to the industry served by auxiliary establishments. Researchers interested in a similar variable during NAICS years can obtain the contemporaneous NAICS code of the industry served by an auxiliary from the SSL files through 2016, using the variable `naics_aux`. We plan to include this `naics_aux` variable in future versions of the CBPBR.

The *naics*{*year*}*_flags* files contain information on the sources and assignment methods for the vintage-consistent NAICS codes. The method flags are described in Appendix Section C.

8.2.1 VC NAICS assignment methods

Here we provide a brief overview of the goals and construction of the vintage-consistent NAICS codes in the LBD. To construct the `bds_vcnaics` variable, we first use the methods from Fort and Klimek (2018) to assign a consistent, NAICS2002 industry code to every establishment in the LBD from 1976 to 2009. Those codes were designed to: 1) improve the accuracy of the industry codes in the LBD; 2) replace missing, partial, and erroneous codes; 3) provide a continuous industry code basis for the entire LBD and; 4) minimize "industry switching" that might be induced by random assignments of NAICS codes that do not map uniquely from SIC codes. We build on that work by mapping the 2002 codes to the most recent vintage of NAICS codes, by improving the method to identify and assign auxiliary establishments, and by correcting potential spurious industry switching for imputed births.

Note that our goal is not to identify cases that might be considered errors. For example, if an establishment is classified in industry A then B then back to A, where A and B are legitimate

[24] In 1997, the largest number of auxiliary establishments and associated employment were classified in NAICS 551114, "Corporate, Subsidiary, and Regional Managing Offices." Other auxiliaries include "Warehousing and Storage," "Truck Transportation," and "Accounting, Tax Returns, and Payroll Services." When an establishment performs more than one of these activities for its firm, it is classified under 551114. See Fort and Klimek (2018) for additional details on auxiliaries, including their distribution of employment across activities in 1997.

[25] In future versions of the LBD, we plan to replace the `fk` prefix with `vc`.

industries, we do not change its code, though assignment to B may be an error. Industry codes in `bds_vcnaics` are allowed to change over time for reasons that include changes in activity, as well as errors in the underlying data.[26]

To assign a vintage consistent industry code, we first construct detailed empirical concordances from various sources, including the EC data. For transitions across SIC vintages, we use the official published concordances.[27] In 1997, the EC collected information on an SIC and a NAICS basis. We use these data to build size-dependent concordances on the share of establishments that map from a particular SIC code to the corresponding NAICS code(s), by establishments' industry payroll quintile. We restrict these concordances to mappings in the official concordance to ensure their accuracy.[28] A limitation of our approach is thus that we use a concordance based on the 1997 distribution of economic activity to map from SIC to NAICS in all the SIC years. Although this may lead to biases in earlier years, (see Bayard and Klimek, 2003, for a discussion of this issue in manufacturing), we believe it is the least problematic approach when mapping from SIC to NAICS for all sectors. Starting in 2002, the EC data collected information on the old NAICS vintage and the new NAICS vintage. We exploit that information to build comparable size-dependent establishment shares by industry payroll quintile for each NAICS transition.

In our second step, we retrieve industry code information for all establishments in the EC. When the LBD and EC disagree, we use the EC code. In those instances, we check whether the establishment's EC industry code in year t matches the LBD code in $t+1$ or in $t+2$, in which case we replace the LBD code with the EC code in the intervening years. If the EC code matches the LBD code in $t+2$, we change the LBD code to the EC code in t and $t+1$. We incorporate the EC code since these are generally more reliable, but researchers should note that this may induce some spurious industry switching in the data. This re-timing is the only longitudinal data correction we make to continuing establishments' existing industry codes when constructing the vintage-consistent codes.

In our third step, we use the CBPBR post-retiming files to exploit the longitudinal information available for every establishment.[29] We begin by identifying erroneous and partial codes and, when available, use longitudinal information for a particular establishment to fill in missing or incomplete codes. We then use the detailed concordances to assign all NAICS codes that map uniquely from an establishment's SIC code. For each establishment that exists in the SIC and NAICS eras, we check whether its earliest NAICS code is one (of possibly many NAICS) to which its SIC code maps in our concordances. If so, we assign that NAICS code back in time to all years for which its SIC codes map to that NAICS code. We apply the same longitudinal techniques to map across vintages within each system.

After exploiting all available longitudinal information for an establishment, we perform random assignments. The random assignment method draws a probability from a uniform distribution for

[26] A growing body of work documents systematic patterns of establishments switching their industry in response to economic shocks (Bernard, Jensen, and Schott, 2006; Bloom, Handley, Kurman, and Luck, 2019), so allowing for legitimate industry switching is important to measure the evolution of US economic activity.

[27] We have no information about the number of establishments that map from one SIC vintage industry to another because the SSL files do not include two SIC vintages at once, so we construct shares based on the number of mappings for each relationship. In practice, the differences across these SIC vintages are fairly small.

[28] In practice, there are discrepancies between the official concordance and what was implemented in practice and thus appears in our concordances constructed from the EC data. The NASS memo in the LBD documentation directory discusses these implementation issues in more detail. The 2018 version of the LBD contains partial NAICS2012 codes for establishments in industries that are out of scope of the EC (e.g., industries related to agriculture). We aim to provide full codes for these establishments in future versions on the LBD by constructing concordances from the BR data for these industries.

[29] Note that the RDC only contains the pre-retimed CBPBR files.

each establishment within an SIC code that maps to multiple NAICS codes. The method then ensures that the establishments in that SIC code are assigned to the corresponding NAICS codes so that the ensuing shares of activity from that SIC to NAICS matches the share of activity in the detailed concordance. We begin random assignments in the latest year of the data for a particular vintage (e.g., in 2001 for the SIC97 to NAICS97 mapping). After making all the required random assignments in one year, we again exploit longitudinal information to assign the randomly assigned code to all other years in which the establishment exists and for which it does not have contradictory industry information. For example, if we randomly assign NAICS code A to an establishment that has SIC code B in year t, we also assign NAICS code A to the establishment in all other years in which it has SIC code B, in any years in which it has missing industry information, and in any years in which it has another SIC code that could legitimately map to NAICS code A. This "rolling" of the randomly assigned code reduces spurious industry switching in the data.

Auxiliary establishments are not included in the concordances described above since they are considered to be distinct from other activity under SIC. We identify auxiliaries in the SIC era using the Census of Auxiliaries (also referred to as the Enterprise Support Survey (ESS)) for the years 1977, 1982, 1987, 1992, and 1997. We also identify auxiliaries using the `toc` variable in the LBD, which contains an establishment's type of operation. The ESS and `toc` information generally allows us to assign auxiliaries a partial NAICS code (i.e., the first 3-4 digits of a NAICS code). To assign auxiliaries a full NAICS code, we construct size-dependent concordances using the 1997 Census of Auxiliaries and apply the same random assignment method described above. As for all our assignments, we employ all longitudinal information available for an establishment to assign the code from the latest year back (or forward) in time whenever there is not information in the other year that conflicts with that code.

The steps described above provide a NAICS2002 code for all active establishments in the LBD from 1976 to 2009. We perform comparable steps for each transition across NAICS vintages. We first use the EC data to construct size-dependent NAICS-to-NAICS vintage concordances by establishments' industry payroll quintiles. We then use all longitudinal information to roll codes as far as possible. We then perform a random assignment in the latest year, roll all randomly assigned codes as feasible, and repeat for every year until all establishments are assigned that vintage code.

When performing the random assignments, we create a variable `fk_{indcode}_splits`, which denotes the potential number of new vintage codes to which the old vintage code could have been assigned. There is a splits variable for each vintage transition available in the *naics*{*year*}*_flags* files. Researchers can sum across these variables to obtain the total number of splits for an establishment with one or more random assignments. Note that this variable only captures one dimension of randomness. For example, the random assignment of a SIC code for which half the establishments map to NAICS A and half to NAICS B might be considered more random than one for which 99 percent map to NAICS A and 1 percent map to B. We note that in the SIC years, some establishments never have any industry information. We randomly assign those establishments an industry code to match the underlying share of total US activity. Researchers may choose to drop those establishments, which will have a sum of split count greater than 200, from an analysis since the probability that their industry codes are correct is close to zero.

The final step in the creation of the `bds_vcnaics` variable is a correction for intercensal births, as described in Section 10.3. Intercensal births captured by the COS may have lower quality information in their first several years of existence before they are first measured in the Economic Census. We therefore "push back" the establishment's first high-quality observation of industry classification from the EC data. Researchers should note that this correction is only applied to the `bds_vcnaics` variable in the 2018 LBD.

9 Noise Factors

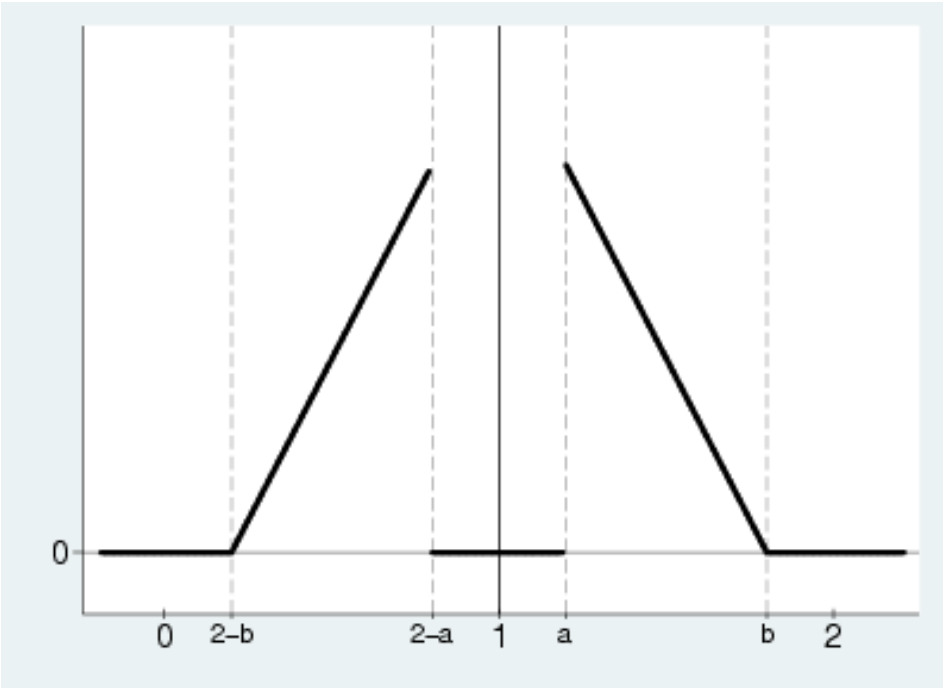
In order to avoid disclosure of confidential data in the published BDS tables, the BDS follows the CBP in using Hybrid Balanced Multiplicative Noise Infusion as described by Evans, Zayatz, and Slanta (1998). In implementing this methodology, we associate a noise factor with every establishment in the LBD. This noise factor, which is near but not equal to one, is used to distort continuous attributes, such as employment, associated with the establishment. The CBP has released tabulations using noise-infusion as the primary disclosure avoidance methodology since 2007 (Massell and Funk, 2007).[30] For the purposes of the public-use BDS tabulations, we aim to not only apply the noise infusion methodology as implemented by the CBP, but also, to the greatest extent possible, assign establishments the same noise factor assigned by the CBP.

The BDS faced two challenges in implementing the noise infusion methodology to the BDS tabulations. First, the 2018 BDS release includes data for years starting in 1978, where some, but not all, of the underling establishments have been assigned a noise factor by the CBP. Establishments that died before 2007 are never assigned a noise factor in the CBP. For these cases we assign new noise factors using methods similar to how CBP noise factors are generated. Second, the BDS includes statistics on both the stock and flow of employment. The CBP, in contrast, publishes data only on stock measures such as the count of establishments and employment. We address this by following the noise infusion procedures used for the Quarterly Workforce Indicators (QWI) when calculating BDS flow statistics (Abowd, Stephens, and Vilhuber, 2006).

Multiplicative noise infusion perturbs data values for each establishment prior to tabulation by applying a random noise multiplier to the magnitude data (i.e., characteristics such as number of employees). These noise multipliers are longitudinally consistent—an establishment retains the same noise factor for each year in which it contributes to published totals. Noise multipliers are drawn from a split triangular distribution, illustrated in Figure 4. Each establishment's noise factor value is perturbed by at least a given amount (a in Figure 4). All establishments in a firm are given a multiplicative noise factor on the same side of one, so the firm's employment is also protected by a minimum distortion; this later condition may force a break in the longitudinal consistency of an establishment's noise factor, which we discussed below.

[30] See Massell, Zayatz, and Funk (2006) for a description of a test on the Commodity Flow Survey.

Figure 4: Distribution of Multiplicative Noise



Source: Abowd, Stephens, and Vilhuber (2006).
Notes: The parameters a and b are confidential.

9.1 Assigning Noise Factors in the LBD

The algorithm for noise assignment proceeds in several steps. We begin by assigning an initial noise factor to establishments by rolling backwards in time the CBP-assigned noise factors. The CBP-assigned multiplicative noise factors are moved backwards in time if the establishment appears in the LBD in t and $t-1$. Because the BDS tabulates some establishments that are not tabulated in the CBP, we do this roll back in a pair-wise manner starting in 2018. However, it is in the last year for which the CBP was published without noise-infusion, 2006, that these rolled back factors are applied to a large number of establishments. This step carries noise factors from one year to the former, backwards until 1976. We then push noise factors forward in time, starting in 2007 and continuing through 2018.

After assigning initial noise factors, we evaluate those initial noise factors in each year and assign new ones, iterating backwards from 2018. At the end of this step all multi-unit establishments and almost all tabulated single-unit establishments will have associated noise factors. The first part of this step imposes firm-level "directional" consistency. As noted previously, we require that all of a firm's establishments have noise factors either above or below one. After the initial noise factor assignment, merger and acquisition activity may result in firms having establishments in a given year with factors both above and below one. To address this we select a "direction" for each multi-unit with establishments with noise factors with the objective of keeping noise factors unchanged for the largest amount of payroll among tabulated establishments. We make no other effort to maintain longitudinal consistency of noise factors at the firm level. After this direction is chosen (above or below one), we generate new noise factors for establishments that are inconsistent with the firm-level direction and we create new noise factors for establishments that still missing them.

When the previous year is processed these newly- and re-assigned multiplicative noise factors will be carried backwards. An establishment may change its multiplicative noise factor if it changes firms, or if its firm acquires new establishments. Next in this step, multi-unit firms that had no establishments in existence the following year and thus no establishments with assigned noise factors are randomly chosen to have multiplicative noise factors greater than or less than one with equal probability. Establishment noise factors are then assigned, so that all multi-unit establishments in this year have associated noise factors.

To reduce cell-level distortion we use a balancing algorithm that minimizes noise for particular table cells when assigning noise to single-unit firms.[31] As in the CBP, the BDS noise assignment algorithm considers the distortion of state-county-detailed NAICS cells created by the noise factors already assigned to tabulated establishments in that cell when picking which side of one a newly assigned noise factor will be on. Thus, the algorithm determines if a new (not previously given a noise multiplier in an adjacent year) tabulated single-unit establishment's noise multiplier is greater or less than one by considering if the existing noise in its cell distorts the cell value above or below its true value. New single-unit establishments in cells with fewer than three tabulated firms are not balanced in this fashion, but are (later) randomly chosen to have multiplicative noise factors greater than or less than one with equal probability.

Once the multi-unit and balancing single-unit noise factor assignments are completed for a year, the algorithm carries the assigned noise factors back one year and again assigns new noise factors when necessary as described above. Once this step is complete for all years 2018 to 1976, assigned noise factors are then pushed forward in a pair-wise fashion. This assures that an establishment that enters the data as a multi-unit or as a tableted single unit but later becomes a non-tabulated single unit retains its noise factor.

As described above, we assign noise new noise factors in 2018 (the current processing year of the 2018 BDS release) working backward. After all balanced single-unit noise is applied and multiplicative noise factors have been carried to other years (backwards and then forwards), the algorithm randomly chooses multiplicative noise factors greater than or less than one with equal probability for any unassigned establishments. These randomly chosen noise factors are also carried backward in a pairwise manner.

Thus, an establishment will receive a noise factor as part of a multi-unit firm or via balancing in the last year in which it is tabulated in a cell with three or more tabulated firms if it is ever observed in one of these states. If neither ever occur an establishment's direction will be randomly chosen. However, as multiplicative noise factors are carried only pairwise across years (following the CBP), an establishment will be reassigned a new multiplicative noise factors, by the applicable above processes, if there is a gap in which it does not appear in the LBD.

In data years after 2018, CBP noise factors for each processing year will be applied first, then previously assigned multiplicative noise factors will be pushed forward, and finally newly observed establishments will get new multiplicative noise factors in the first year they appear. Next, we describe how noise factors are used in the BDS tabulations.

[31] Randomly drawn noise multipliers would cancel out in expectation. Massell and Funk (2007) tests the quality improvement of balancing over random noise. Massell and Funk (2007) also shows how balancing at a lower level of aggregation improves the quality of more aggregated cells. In the BDS published statistics that do not consider industry or geography, noise is generally larger in these cells.

9.2 Use of Noise Factors in the BDS

Applying a small multiplicative noise factor to small integer values will have little or no effect on the data if standard rounding is applied to the resulting noisy values. The minimum observed employment for any establishment is one, so establishments of size one are not distorted to size zero. Distorting a fraction of one-employee establishments adds bias as we are only able to increase (and not decrease) the size of these establishments. To address these issues, we apply a probability rounding technique at the establishment-level that ensures some small values change some of the time, and that values of two or more are more often rounded down than up. The algorithm keeps track of how many ones it has changed to twos, and uses this count to increase the chance that larger establishment sizes are rounded down. Larger values are also always rounded away from their true value. Rounding is applied at the establishment-level for all stock variables (e.g, employment).

Noise infusion for flow and rate variables in the BDS follows Abowd, Stephens, and Vilhuber (2006). For these statistics, noise and rounding are applied at the cell-level. Because uncertainty propagates (and increases) through arithmetic operations, flow values are expected to have greater error than the corresponding stocks from which they are calculated. Applying noise at the cell-level reduces this propagated error. For instance, job creation for establishments i in group s is calculated as,

$$JC_{s,t} = \left(\sum_{i \in s; g_{i,t} \geq 0} E_{i,t} - \widetilde{E_{i,t-1}} \right) \times \left(\frac{\sum_{i \in s; g_{i,t} \geq 0} X^n_{i,t}}{\sum_{i \in s; g_{i,t} \geq 0} X_{i,t}} \right) \tag{1}$$

Where $E_{i,t}$ is establishment i's employment at time t, and the tilde indicates that is is rounded as described. Establishment i's denom ($X_{i,t}$) and noisy denom ($X^n_{i,t}$) are defined as,

$$X_{i,t} = \frac{(E_{i,t} + E_{i,t-1})}{2} \tag{2}$$

$$X^n_{i,t} = \frac{(\widetilde{E_{i,t} * noise_{i,t}} + \widetilde{E_{i,t-1} * noise_{i,t}})}{2} \tag{3}$$

Finally, $g_{i,t} = \frac{(E_{i,t} - E_{i,t-1})}{X_{i,t}}$, conditioning on this being greater than zero means we are only considering establishments with positive employment changes as contributers to job creation. Note that we apply the noise factor in year t to the employment value in $t-1$ and t and that rounding allows the effective multiplicative noise factor for some establishments or cells to be outside the bounds of assigned establishment-level multiplicative noise factors. Distorted cells are rounded, probabilistically for cells with small values (as there is no theoretical reason that these flow variables may not take on the value of zero, rounding to zero is allowed and downward compensation is not used). Ratios, such as the job creation rate, are created from the protected, published, values.

After the BDS tables are computed, cells in these resulting tables with fewer than three firms are suppressed.

10 Generating the LBD

The creation of annual LBD datasets involves integrating multiple years of CBPBR data with the linkwide file described in Section 6. We use the CBPBR data to recover annual attributes for each

establishment, for a given year t, from $t-2$ to $t+1$.[32] The linkwide file provides the establishment identifier linkages (`cfn` or `empunit_id_char`) necessary to create `lbdnum`, the core longitudinal establishment identifier used in the LBD.

The creation of the *lbd*{*year*} and *lbdfirm*{*year*} files proceeds in several steps. First, we create `lbdnum` by selecting a single establishment identifier for each `lbdnum` in each year from the linkwide file. Second, after using `lbdnum`-establishment identifier links to combine four years ($t-2$ to $t+1$) of CBPBR data, we edit establishment attributes. Third, we compute establishment age variables. Fourth, we define the "bds_" variables, which are used to generate the BDS tabulations. Finally, we create a file of firm-level attributes such as firm size and firm age. Each of these steps is described below.

10.1 Creating lbdnum

The first step in generating the LBD is the creation of `lbdnum`. Conceptually, each observation in the linkwide file represents a single `lbdnum`, stringing together cross sectional establishment identifiers over time. We create `lbdnum` as a unique identifier on the linkwide file by adding "10000000000" to the (arbitrary) observation number in the linkwide. `lbdnum` cannot be used to link establishments across separate executions of the LBD production system. Any change to the underlying CBPBR data or links causes `lbdnum` to change arbitrarily for any given establishment.[33] As described in Section 6, for each year in the linkwide file an `lbdnum` may have zero, one, or two establishment identifiers. To link establishment attributes from the CBPBR to an `lbdnum` we must select a single establishment identifier, or `estabid`, for a given year. When present, the "second" establishment identifier (`id2`) captures reorganizations. The reorganizations identifier may or may not have positive employment in the year it appears in the `id2` field. The algorithm that selects which establishment identifier on the linkwide is associated with the `lbdnum` in a given year proceeds as follows, with `id1` and `id2` being the link file establishment identifiers for a given year and `id1emp` and `id2emp` being the employment associated with `id1` and `id2` in the given year (collected from the CBPBR).

1. `estabid` = `id1` if `id2` is missing.
2. `estabid` = `id2` if `id2` is not missing, `id1emp` is missing, and `id2emp` is not missing.
3. `estabid` = `id1` if `id2` is not missing, `id1emp` is not missing, and `id2emp` is missing.
4. `estabid` = `id1` if `id2` is not missing, `id1emp` and `id2emp` are not missing, and `id1emp`>`id2emp`.
5. `estabid` = `id2` if `id2` is not missing, `id1emp` and `id2emp` are not missing, and either (`id1emp`=0 and `id2emp`>0) or (`id1emp`>0 and `id2emp`>0) or (`id1emp`=0 and `id2emp`=0).
6. `estabid` = `id2` if `id2` is not missing and both `id1emp` and `id2emp` are missing.

When we select an `estabid` for the `lbdnum` we record information about why it was selected in `estabid_flg`. This flag has four characters. The first character captures what id was selected

[32] The first two years do not have lagged attributes and the final year of the data does not have lead values for $t+1$. For these years we limit processing to steps that can be done with the subset of years available.

[33] This implies that `lbdnum` should only ever be used to link establishments within the LBD files. As described in Section 3, `estabid` must be used to link to other establishment-level datasets such as the Economic Census or economic survey data such as the ASM. Similarly, `firmid` must be used to link to enterprise level information from other Census data sources.

("1" = `id1`, "2" = `id2`). The second character captures whether `id2` is present ("1" = present, "0" = missing). The third and fourth characters capture whether `id1emp` and `id2emp` are present ("11" both present, "10" `id1emp` present and `id2emp` missing, etc.). For example, an `estabid_flg` value of "2101" means `estabid` is equal to `id2`, `id2` is present, `id1emp` is missing, and `id2emp` is non-missing.

When two ids are present, the id that was not selected is stored in `estabid_rorg`. This may be helpful to users linking files with establishment identifiers that may be associated with reorganizations. When linking by `estabid`, users can attempt to match records that do not link by `estabid` using `estabid_rorg`. We also retain the `link_flg` from the linkwide, which describes the type of longitudinal link made (see Section 6).

10.2 Building and Editing the LBD

With the `estabid` selection made, for each year t, we use the `lbdnum-estabid` links to collect establishment attributes from $t-2$ to $t+1$. This window of attributes allows us to use longitudinal information to edit the data and include exits in the final LBD files. Another challenge unique to the new LBD infrastructure is the possibility of conflicting information coming from the BR and CBP. As described in Section 4, where the legacy SSEL files primarily held BR data, the CBPBR files contain both BR and CBP microdata. This introduces the possibility for disagreement in the attributes coming from the BR and CBP. For example, the CBP may flag an establishment as a C-corporation while the BR has the same establishment flagged as a sole proprietorship. One reason for this disagreement is that by the time the CBP microdata is finalized for a given reference year, it is often more up-to-date than the final BR extract.

Our approach to dealing with BR/CBP disagreements is, on a variable-by-variable basis, and in some cases for groupings of variables, to identify cases where it appears the CBP had more up-to-date information that was later fed back into the BR in the following year. For all variables, we first preference the BR value. We then identify cases where the BR and CBP values disagree in t, agree in $t+1$, and the BR value switched to match the CBP between t and $t+1$. In such cases, which we designate "BR-CBP switchers", we take the CBP value in t.

One of the most fundamental attributes of an establishment in the BR, especially as it relates to identifiers, is whether it is a single-unit or multi-unit. The first variable on which we identify BR-CBP switchers is multi-unit status. We flag a multi-unit switcher when its multi-unit status (`mu`) in the BR and CBP disagree in t, agree in $t+1$, and the BR value switches to match the CBP. For multi-unit switchers we also take, if available, the CBP value for `alpha`, `ein`, `lfo`, `pdiv`, and `toc`.[34] After resolving disagreements in multi-unit status, we perform a similar resolution of BR-CBP differences on `alpha`, `ein`, `lfo`, `toc`, `pdiv`, `act`, `state`, `county`, `zip`, `naics`, and `sic`. We also perform some basic editing of these variables including (1) taking non-missing CBP values when the BR value is missing, (2) removing a limited set of invalid values (e.g. a state code equal to "XX"), (3) using similar codes across years.[35] For each variable we store the final source of the value (BR or CBP) in variables with the suffix "_src".[36]

[34] `lfo` is the legal form of organization code, `pdiv` is the processing division code, and `toc` is the type of operation code.

[35] An example of the edits to ensure similar codes across years is the county code for Dade county in Florida. In the late 1990s Dade county's county code switched from "086" to "025". We are able to make this change because it was nominal and did not also entail changes to boundaries. For a full list of variable-level edits see the "Step 4" program specifications, which are available to FSRDC researchers with a project approved to use the LBD.

[36] For 2017 and 2018 we perform a special cleaning step for `naics`. In the 2017 Economic Census, cell phone stores were originally coded as retail but were later corrected to be in to be in the services sector. For a set of cell phone

For establishments with zero payroll in t we are unable to recover establishment attributes such as geography, `lfo`, `pdiv`, and even `estabid`. For such cases we carry forward attributes from $t-1$. This allows us to categorize exits for the BDS tabulations.

In 2002 we perform several processing steps to deal with duplicates created in the linkwide file. A number of important changes were made in 2002 as part of the BR redesign including the switch of establishment identifier from `cfn` to `empunit_id_char`. We bridge between these identifiers but not every case cleanly maps across the 2001 and 2002 barrier. As a result, a few `estabid`s map to two `lbdnum`s. For those `estabid`s, only one of the `lbdnum`s can be found in 2003. We keep the `estabid` for records with an `lbdnum` that matches to 2003 and replace `estabid` for these cases with "dup" followed by a zero padded integer unique to the `lbdnum`. This maintains the uniqueness of `estabid` within the annual LBD files and allows users to identify the problematic establishments by looking for `lbdnum`s with "dup" contained in `estabid`.[37]

10.3 Creating Variables Used in the BDS

The final LBD files contain a series of variables with the prefix "bds_".[38] These variables are used to produce the BDS tabulations and may be of interest to researchers due to the processing embedded within them. The "bds_" geography variables such as state and county are time invariant at the establishment level. These variables contain the most recent Economic Census year value observed for each `lbdnum` or the modal value if the `lbdnum` is never observed in an Economic Census year. Ties in mode are broken randomly. After we assign vintage consistent NAICS codes we create `bds_vcnaics`. For inter-censal establishment births we replace `bds_vcnaics` with the vintage consistent code assigned in the establishment's first Economic Census year for years before the first Economic Census year. This "pushes back" the first high quality observation of industry classification for inter-censal births.

`bds_emp` and `bds_emp_tm1` are employment values for t and $t-1$ that have been subjected to smoothing algorithms. Prior to applying the smoothing algorithms we replace all missing employment values with zero. Missing employment values occur when an establishment has no payroll in a given year. The smoothing algorithms reduce the impact of large, transient positive or negative employment changes. We apply smoothing algorithms symmetrically to employment in $t-1$ and t in years from 1977 to the second to last year in the data. An example of a rule aimed at smoothing negative employment spikes is as follows.[39]

IF ($t-2$ employment is $>$ 500) AND (t employment $>$ 500) AND (DHS employment growth between $t-2$ and $t-1$ is < -1.67) AND (DHS employment growth between $t-1$ and t is > 1.67)

THEN $t-1$ employment is replaced with the average of $t-2$ and t employment

With the edited employment and lagged employment values we compute establishment-level employment flows. Positive and negative employment changes are stored in `bds_pos` and `bds_neg`

store establishments we replace `naics` with the correct 2012 service industry code.

[37] The `estabid` linked to these `lbdnum`s in 2003 can be used to identify specific records on the *cbpbr*{*year*} file in 2002.

[38] Many of the "bds_" variables are missing in *lbd1976* because year-to-year employment changes are first observed in 1977.

[39] DHS employment growth rates follow Davis, Haltiwanger, and Schuh (1996) in computing employment growth as $\frac{emp_t - emp_{t-1}}{denom}$, where *denom* is equal to $\frac{emp_t + emp_{t-1}}{2}$. For the full list of smoothing edits see the "Step 4" program specifications, which are available to FSRDC researchers with a project approved to use the LBD.

respectively. Establishment exits are flagged in `bds_exit`, set to 1 when `bds_emp_tm1` > 0 and `bds_emp` $= 0$. Similarly, establishment entrants are flagged in `bds_entry`, set to 1 when `bds_emp_tm1` $= 0$ and `bds_emp` > 0. Positive and negative employment changes associated with entry and exit are stored in `bds_bflow` and `bds_dflow` respectively. The establishment-level employment denominator, average employment between $t-1$ and t, is stored in `bds_denom`.

In addition to these establishment-level employment measures we also retain firm-level size measure used in the BDS, `bds_firmsize_emp` and `bds_ifirmsize_emp`. One of the firm size measures used in the BDS, fsize, is the average firm size associated with an establishment in $t-1$ and t. Since an `lbdnum`'s firm identifier may change from $t-1$ to t the `bds_firmsize_emp` is unique to a combination of $t-1$ and t firm identifiers. As such, we store this average firm size measure at the establishment-level in the LBD files. For convenience, we also provide the initial, or $t-1$ firm size, `bds_ifirmsize_emp`, which is the lagged size of the firm identifier associated with the establishment in $t-1$.

One of the most consequential variables used in generating the BDS tabulations is the scope flag, `bds_tab`. There are many establishments included in the BR that are inappropriate to include in the BDS, either because of industry scope (e.g. governments, pension funds) or out of scope geography (e.g. Island areas, Puerto Rico). The creation of `bds_tab` is a two-stage process. First, we use a set of conditions to flag each establishment as in or out of scope in $t-1$, t, and $t+1$ based on contemporaneous attributes. Second, we pool those flags, setting establishments in scope if they are in scope in any of those three years or were flagged as in scope for the CBP (`cbp_tab`="Y") in any of those three years. We do this pooling because high quality observations about the attributes of an establishment, especially smaller establishments not covered by the COS, vary over time. The highest quality information about an establishment is collected in Economic Census years. Finally, we apply a series of "override" conditions that set establishments as out of scope regardless of the three years of scope information. For example, even if an establishment is found to be in scope in $t-1$, if the `pdiv` is equal to "G" (governments) in t, it is set out of scope in t.

Initially, all establishments are set in scope (`bds_tab`=1), then each set of conditions sets certain cases out of scope. The rules for assigning scope are modeled after those used in the CBP data. The conditions remove inactive cases, those with invalid geographic codes, government owned or operated establishments, or those in out of scope industries such as crop production. The year-specific contemporaneous conditions are as follows, setting `bds_tab`=0 if any of the following conditions are met:

- `pay` $= 0$
- `st` in [00, AA, XX, YY, ZZ]
- `cty` in [AAA, XXX, YYY, ZZZ]
- `pdiv` in [A, O, N, U, W, V, Z, 9, P, M, G, S][40]
- `toc` $= 90$[41]
- `sic` in [01, 02][42]
- `naics` in [111, 112][43]
- `emp` > 30000

[40] For value definitions see Appendix B Table 5.
[41] `toc`=90 are government-owned and operated establishment, except liquor stores or liquor warehouses.
[42] SIC 01 and 02 capture crop and livestock production respectively.
[43] NAICS 111 and 112 capture crop and livestock production respectively.

- `cbp_tab` $\neq$ Y AND not missing `cbp_tab`

The cross-section override conditions that set cases out of scope regardless of the scope flag in $t-1$, t, or $t+1$ computed above are as follows.

- `bds_denom` $= 0$
- `pdiv` in [M, G] OR `pdiv` in $t-1$ or $t+1$ is M
- `toc` $= 90$
- `bds_exit` $= 1$ AND `bds_tab` in $t-1 = 0$
- `bds_emp` > 30000 OR `bds_emp_tm1` > 30000
- `act` $=$ G
- `bds_entry` $= 1$ AND `bds_emp` > 5000 AND mu $= 0$
- `pdiv` $\neq$ H AND `lfo` $=$ G AND (`naics` NOT in [622, 4248, 7132, 44531, 552120, 522130, 511130, 453991, 521110, 721120] AND sic not in [60, 592, 518, 611, 806])[44]
- sic in [01, 02]
- naics in [111, 112]

10.4 Creating lbdfid

One improvement incorporated into the new LBD database is the introduction of a unique firm-level linking identifier analogous to `lbdnum`, which we call `lbdfid`. In addition to establishment-level information, the Census Bureau also captures information at the enterprise or firm level. A firm is defined as an economic unit comprising one or more establishments under common ownership or control. `lbdfid` is a platform for building longitudinal linkages at the enterprise-level that would consistently assign a single enterprise unit the same unique identifier over time.[45] The initial version of this variable does not implement longitudinal enterprise linkages. We leave the disambiguation of enterprises over time to future research.

Historically, the LBD used a variable called `firmid` to uniquely identify firms and compute firm age and firm size. For single-units, `firmid` is equal to "0" followed by the establishment's `ein`. For multi-units, the `firmid` is equal to the `alpha` followed by "0000". `firmid`, similar to `cfn` and `empunit_id_char`, is created and validated only in the cross-section and cannot be treated as longitudinally consistent. An enterprise may change `firmid` for any number of reasons unrelated to its ownership structure. To see this, consider a single-unit firm that changes its `ein`. Businesses may change the `ein` they use to file payroll taxes for any number of reasons that are unrelated to the business's ownership; switching to a new accountant might trigger an `ein` change. If the owner(s) of a single-unit establishment decide to file under a different `ein` from one year to the

[44] NAICS codes: 622 - Hospitals, 4248 - Beer, Wine, and Distilled Alcoholic Beverage Merchant Wholesalers, 7132 - Gambling Industries, 44531 - Beer, Wine, and Liquor Stores, 522120 - Savings Institutions, 522130 - Credit Unions, 511130 - Book Publishers, 453991 - Tobacco Stores, 521110 -Monetary Authorities-Central Bank, 721120 - Casino Hotels. SIC Codes: 60 - Depository Institutions, 592 - Liquor Stores, 518 - Beer and Ale, Wine and Distilled Alcoholic Beverages, 611 - Federal and Federally-Sponsored Credit Agencies, 806 - General Medical and Surgical Hospitals, Psychiatric Hospitals, Specialty Hospitals, Except Psychiatric.

[45] Note that the goal of `lbdfid` is not to assign establishments a longitudinal firm identifier. Establishments change their associated firm for many reasons including mergers and acquisitions. Instead, `lbdfid` aims to ensure that a single enterprise does not change its firm identifier over time for reasons unrelated to the definition and organization of the firm.

next, the LBD links the underlying establishments, giving them a common `lbdnum`, but the `firmid` will change.

The creation of `lbdfid` provides the infrastructure to link `firmid`s over time. Using this infrastructure, future iterations of the LBD can provide a more accurate approximation of a longitudinally consistent enterprise identifier. As our understanding of `firmid` improves, linkages can be built into `lbdfid`, improving its ability to capture unique enterprises over time. Though relatively little longitudinal editing is currently done on `lbdfid`, it does address `firmid` "recycling," which occurs when multi-unit `firmid`s are re-used for different enterprises after extended periods of inactivity. `lbdnum` resolves a similar issue with establishment identifiers, which may be reused after a long period of inactivity. If an establishment identifier has more than seven years without payroll and then re-appears with positive payroll, we assign it a new record in the linkwide file and in turn a new `lbdnum`. We use a similar seven year window for `lbdfid`, giving a `firmid` a new `lbdfid` if it is payroll inactive for more than 7 years.[46]

Creating `lbdfid` requires two pieces of information. First, we need year-to-year `firmid` links, which capture `firmid`s that we consider the same enterprise across years. Second, we need to know in each year when a given `firmid` was last payroll active (associated with at least one establishment with positive payroll).

In the initial version of the `lbdfid` we only utilize exact `firmid` matches across years. Much in the way that improvements in establishment-level linkage technologies can be incorporated into LBD processing, creating more accurate `lbdnum`s, these `firmid` link files can be augmented to address any number of firm identifier changes. One candidate for such an improvement would be to bridge across the SU-to-MU transition, when `firmid`s break by construction. Additional research will be necessary to distinguish between SU-to-MU transitions that occur simultaneously with unobserved ownership changes.[47]

10.5 Firm File

The *lbdfirm*{*year*} files contain firm-level information such as firm size and firm age. We use `lbdfid` to compute both firm size and firm age. We calculate both employment and payroll-based firm size and firm age. We carry `firmid` on the *lbdfirm*{*year*} files to facilitate linking to the other administrative datasets where `firmid` is available. Users should note that `lbdfid`-to-`firmid` is a 1-to-many relationship *across years*. For example, if a `firmid` is pay inactive for 9 years before becoming active again, it will be assigned two different `lbdfid`s, one before and one after the period of inactivity.

Payroll-based firm age is the payroll age of the oldest establishment in the year an `lbdfid` first appears with positive payroll. Establishment payroll age is the number of years since the establishment first appears with positive payroll. Similarly, employment-based firm age is the employment age of the oldest establishment in the year an `lbdfid` first appears with positive employment.

[46] Analysis of the distribution of reactivation lags for both establishment and firm identifiers exhibit similar patterns, justifying the application of a similar heuristic.

[47] It is often difficult to know whether `firmid` changes represent changes in ownership. In some cases, we are much more confident that changes are not related to ownership changes. For example, establishments that transition from single to multi-units, by construction, experience a `firmid` change regardless of conditions of ownership and control. Moreover, if the `ein` used for the single unit persists into the multi-unit we may be comfortable treating the SU and MU establishments as under a common firm identifier because IRS requires that businesses file for a new `ein` when ownership changes.

`firmsize_emp` and `firmsize_pay` are the total employment and payroll associated with the `lbdfid` in a given year. We also include on the *lbdfirm*{*year*} files lagged firm size measures, both employment and payroll, in variables with a "_tm1" suffix. We also include a "bds_" firm size measure, which captures the sum of `bds_emp`.

`firm_firstyear_pay` and `firm_firstyear_emp` contain the first year an `lbdfid` is observed with positive `pay` or `emp` respectively. `firm_initialyear_pay` and `firm_initialyear_emp` contain the `firstyear_pay` or `firstyear_emp` of the oldest establishment in the `lbdfid`'s first year with positive `pay` or `emp`. These are the implied `pay` and `emp` first year assigned to the `lbdfid`. `firmage_pay` and `firmage_emp` are then calculated as the current year minus `firm_initialyear_pay` and `firm_initialyear_emp`, respectively.

The firm death variables (`firmdeath_pay` and `firmdeath_emp`) are defined as the `lbdfid`'s last year of positive pay or emp in $t-1$ if all associated estabs in $t-1$ have their last year of positive pay or emp in $t-1$ (as captured by the establishment's `lastyear_pay` and `lastyear_emp` respectively). When computing firm death we identify the `lbdfid`'s last year with positive payroll and employment, which are stored in `firm_lastyear_pay` and `firm_lastyear_emp`. When processing that last observed year in the LBD, we also include subsequent-year first-quarter payroll (`qp1_sy`) as a proxy for whether the firm is likely to have employment in $t+1$.

Finally, the use of `lbdfid` and the treatment of firm reactivations also affects our measure of firm deaths. If a firm identifier reactivates after a long spell without employment, the reactivated firm is assigned a different `lbdfid`, which allows for the possibility that the firm would be classified as a firm death at the beginning of its inactivity.

11 Tabulating Business Dynamics Statistics

In this section we describe how the *lbd*{*year*} and *lbdfirm*{*year*} files are used to generate BDS statistics. As described in Section 10, the *lbd*{*year*} files contain a suite of variables used to generate the public use BDS tabulations, all of which contain the prefix "bds_". All of the BDS statistics are employment-based, meaning that entry is defined as having zero Q1 employment in $t-1$ and positive Q1 employment in t. The employment measure used is number of employees at the establishment in the payroll period including March 12. Employment changes are therefore March-to-March net changes. Firm size and firm age measures are also employment-based, with firm age computed based upon the oldest establishment in the first year a firm identifier (`lbdfid`) is observed with positive employment.[48]

11.1 BDS Tables

The initial release of the redesigned BDS includes 78 tables that cover economy-wide tabulations and combinations of one to four by-variables. Year is a by variable in all tables. See Appendix B Table 6 for a list of tables included in the initial release of the redesigned system. The by-variables include firm characteristics such as firm size, initial firm size, and firm age. The BDS tables also incorporate by-variables based upon establishment attributes including establishment size, initial establishment size, establishment age, industry, and geography. For geography we publish several levels of disaggregation including metro/non-metro, state, metropolitan statistical areas (MSA),

[48]Note that a firm's first year with positive Q1 employment may differ significantly from a firm's first observation with positive payroll. A transient business that appears with positive Q3 payroll in one year and then first appears with Q1 employment several years later will have a gap between its payroll-based and employment-based ages.

and county. The industry dimension includes 18 sectors, 3-digit, and 4-digit NAICS codes. The industry codes used for the initial release from the redesigned data are the vintage consistent 2012 NAICS codes. As described in Section 8, a unique 2012 NAICS code has been assigned to all establishments in the data between 1976 and the last year of the data. In some tables by-variables are coarsened to have fewer groups. For example, the firm age table includes twelve firm age groups while the MSA-sector-firm age table has only five firm age groups. Providing less detail on some dimensions allows us to publish more granular measures overall.

The redesigned BDS tables begin with 1978 whereas the legacy BDS tables began with 1977 (showing changes between 1976 and 1977). In the redesigned tables we chose to begin with 1978 because 1977 is the first Economic Census year covered by the LBD. The Economic Census year births and deaths in 1977 cannot be retimed in the same way as birth and deaths in 1982 because the full intercensal period (1973 to 1977) is not covered by the LBD. The first, high-quality observation year-to-year changes in the LBD are in 1978. Establishment and firm age continue to use LBD data back to 1976. The "Left Censored" firm age will capture firms first observed in 1976.

11.2 Creating BDS By-Variables

To produce the BDS tables we create the by-variable groupings using the "bds_" variables in the *lbd*{*year*}. Some firm characteristics are stored on the *lbd*{*year*} files while others must be recovered from *lbdfirm*{*year*}. Firm size, as described in Section 10.3, is stored at the establishment-level since establishments may change firm identifiers between $t-1$ and t. Thus, the BDS by-variable `fsize`, the average size of the firm associated with an establishment in $t-1$ and the firm associated with the establishment in t, is generated from `bds_firmsize_emp` from *lbd*{*year*}. Similarly, we create the initial firm size BDS variable `ifsize` using `bds_ifirmsize_emp` from *lbd*{*year*}. We collect firm age (`firmage_emp`), firm initial year (`firm_initialyear_emp`), and the firm death indicator (`firmdeath_emp`) by matching *lbdfirm*{*year*} to *lbd*{*year*} by `lbdfid`. The BDS by-variable `fage` is created from `firmage_emp`. The left censored category is defined for establishments with `firm_firstyear_emp` = 1976 or {year}-`firmage_emp` = 1976. These establishments are associated with firms first observed in 1976 and therefore we do not observe the true birth year and thus the true age.

Establishment `age`, `eage`, is defined as {year} minus `firstyear_emp`, which is the first year the establishment is observed with positive employment. Again, left censored establishments are defined as those where year-`eage` = 1976. Establishment size, `esize`, is defined by bins in `bds_denom_noise`, the noised establishment-level `denom`. Initial establishment size, `iesize`, is based on groupings of `bds_emp_tm1_noise`. Establishment size groups are based upon noisy employment values while firm size groups are based on employment totals without noise.

As described in Section 10, geographic variables used in the BDS are time invariant–establishments are disallowed from changing their geographic attributes. The time invariant assignment algorithm uses either the most recent Economic Census year or the modal (breaking ties randomly) when the establishment is never observed in an Economic Census year. The same is done for county (and implicitly MSA and metro/non-metro). Note also that the BDS includes a "nationwide" state code 97, which captures establishments that are not assigned to a single location.[49] The BDS also include statewide county code (999), which implies a statewide metro/non-metro category.

The industry codes used in the BDS tabulations are the vintage consistent industry codes developed by Fort and Klimek (2018). The BDS allows an establishment's industry code to change

[49]For example, mining establishments without a fixed location.

over time. Industry switching in some cases represents the lumpy nature of the detail the Census Bureau collects over time, with much more complete collections occurring in Economic Census years. In other cases, however, industry switching represents an important channel of reallocation (Fort, Pierce, and Schott, 2018; Bloom, Handley, Kurman, and Luck, 2019). As described in Section 10.3, for inter-censal establishment births we push back the first economic census year's code into the inter-censal period. This allows us to utilize the first high-quality signal of the establishment's industry.[50]

11.3 Computing BDS Statistics

Here we summarize the formal definitions of the employment, firm, and establishment stock and flow measures used in the BDS, which have become standard in the analysis of business dynamics (Davis, Haltiwanger, and Schuh, 1996; Tornqvist, Vartia, and Vartia, 1985; Haltiwanger, Jarmin, and Miranda, 2013).

The two central employment change measures in the BDS are job creation and job destruction. Job creation is the sum of all positive employment changes from establishments that saw net employment increases from $t-1$ to t including establishments with zero employment in $t-1$. Similarly, job destruction is the sum of all negative change for establishments that saw net employment decreases from $t-1$ to t including establishments that exited to zero employment in t.

Employment growth $g_{i,t}$ for establishment i at time t is defined as follows.

$$g_{i,t} = \frac{(E_{i,t} - E_{i,t-1})}{X_{i,t}} \tag{4}$$

Where $E_{i,t}$ is establishment i's employment at time t. This employment growth measure shares properties of log differences but also naturally accommodates entry and exit (Davis, Haltiwanger, and Schuh, 1996; Tornqvist, Vartia, and Vartia, 1985). $X_{i,t}$ is given by:

$$X_{i,t} = \frac{(E_{i,t} + E_{i,t-1})}{2} \tag{5}$$

Job creation and job destruction for group s (e.g. group defined by firm age, firm size, industry) is defined as:

$$JC_{s,t} = \sum_{i \in s; g_{i,t} \geq 0} E_{i,t} - E_{i,t-1} \tag{6}$$

$$JD_{s,t} = \sum_{i \in s; g_{i,t} < 0} |E_{i,t} - E_{i,t-1}| \tag{7}$$

The net employment change, $NET_{s,t}$ then is given by:

$$NET_{s,t} = JC_{s,t} - JD_{s,t} = \sum_{i \in s; g_{i,t} \geq 0} E_{i,t} - E_{i,t-1} - \sum_{i \in s; g_{i,t} \geq 0} |E_{i,t} - E_{i,t-1}| \tag{8}$$

The analogous rate measures are given by:

[50] Research distinguishing true industry switching from spurious changes due to varying data sources would provide clear benefits to the Census Bureau. This includes issues with auxiliary establishments, and in particular the noticeable time-series break in 1997, identifying auxiliaries during both the SIC and the NAICS era, the industries served by the auxiliaries, and clear input on how to measure both.

$$JCR_{s,t} = \sum_{i \in s; g_{i,t} \geq 0} \frac{X_{it}}{X_{s,t}} g_{i,t} = \frac{JC_{s,t}}{X_{s,t}} \tag{9}$$

$$JDR_{s,t} = \sum_{i \in s; g_{i,t} < 0} \frac{X_{it}}{X_{s,t}} |g_{i,t}| = \frac{JD_{s,t}}{X_{s,t}} \tag{10}$$

$$NETR_{s,t} = \sum_{i \in s} \frac{X_{it}}{X_{s,t}} g_{i,t} = \frac{(JC_{s,t} - JD_{s,t})}{X_{s,t}} \tag{11}$$

Where

$$X_{s,t} = \sum_{i \in s} X_{i,t} \tag{12}$$

$X_{s,t}$ is the sum of average establishment employment between $t-1$ and t for all establishments in group s. In the tabulations this measure is called `denom`. To convert employment changes to rates we simply divide by the appropriate denom measure. Note that `denom` cannot be recovered by simply averaging the total employment in group s in the two periods because as establishment attributes change they may enter and exit groups. As stated above, establishment entry (exit) is defined as having zero (positive) employment in $t-1$ and positive (zero) employment in t. Though the majority of establishment openings are true "greenfield" entrants, some of the establishments flagged as entrants and exits are reactivations or temporarily closings.

Because the BDS tabulations are based upon a combination of administrative and survey collections, they have no associated measure of sampling error. Instead, the BDS is subject to non-sampling error such as reporting errors in business filings and late filings. The data generation algorithms attempt to mitigate as many known and understood sources of nonsampling error as possible. For example, late filings are often captured in the subsequent year's data in lagged $(t-1)$ variables. Wherever possible, we utilize these variables to improve measurement in t. However, the longitudinal processing of the LBD and BDS cause the final year to change as additional subsequent years of data are added. This is particularly true when the final year is an inter-censal year, where establishment entrant and exits are not observed until the following Economic Census, at which point they are re-timed into those inter-censal years (see Section 7).

11.4 BDS Quality Assurance

We use several different approaches to ensure the quality of the new BDS data series. One approach involves comparing the new BDS data series to other similar programs, including: County Business Patterns, Statistics of US Businesses employment change data, and the US Bureau of Labor Statistics' Business Employment Dynamics (BED) data. We also compare to the legacy BDS series previously released by the Census Bureau. When we find large discrepancies between the new BDS data versus these other programs, we investigated the underlying microdata. In some instances, we identified errors in LBD/BDS processing, which we corrected, and in others cases the BDS data was found to be correct. We also reviewed each unique time series in the data looking for odd patterns. These odd patterns often involved time series with very large changes across years (i.e. spikes). We identify spikes where a time series saw a large change in one direction (positive or negative) in a given year, followed by a similarly large change in the opposite direction in the following year. We define large year-to-year changes in both absolute value and percentage terms, as well as their relative significance in the context of the time series. Identified spikes were similarly investigated

and either an error in LBD/BDS processing was identified and corrected, or the data was verified.

In spite of extensive quality assurance work on the new data, there remained some odd patterns in some time series that caused concern about measurement error. When we deemed any statistic in some of the highly aggregated tables–i.e., only tables with 1 by-variable or fewer–to be too uncertain because of very large changes across years, we suppressed this statistic and reported only "(S)" in the cell. The table below lists each table that contains (S) suppressions, as well as the number of cells suppressed in each table:

Table 5: Data Quality Suppressions in the BDS Tables

Table Name	No. of Cells with Data Quality Suppressions
Firm Age	22
Establishment Age	32
Establishment Size	4
Initial Estab Size	1
State	8
MSA	290
County	0
Sector	18

Notes: The table shows the number of cells that were suppressed for data quality reasons in each of the highly aggregated BDS tables (tables with just one "by" variable).

The firm and establishment age tables, in the firm and establishment age 2 category in 1978, for each variable in this row, have been suppressed with an (S). This is because while these cells are structurally missing–that is, it is impossible to have an establishment or firm age of 2 in the BDS in 1978–an error in LBD/BDS processing led to some establishments being placed in this cell. These cells are suppressed until the error can be fixed, which will be done with the next release of the BDS data series.

12 Comparing to Legacy LBD

Below we provide a brief summary of the improvements in processing between the legacy LBD and redesigned LBD. We then describe how the redesigned BDS tabulations differ from earlier releases and how those differences relate to the aforementioned improvements.

12.1 Summary of Improvements

- One fundamental improvement to the LBD is the improved documentation and transparency that resulted from the transition to formal production processing. The underlying code and code specifications for the LBD are available to researchers with approved projects via the FSRDC research network, allowing microdata users to better understand and help improve the processing of the LBD.

- Improved scope and data quality by combining microdata from both the BR and the CBP throughout the entire time series, 1977-2018. Previous releases of the BDS only included

CBP-edited data for the last few years of the time series. See Section 4.

- Combined and augmented the longitudinal establishment matching algorithms from both BITS and LBD. This includes applying previous improvements made to later years of the legacy LBD to all years of the LBD. See Section 5.

- New algorithms to retime the bunching of measured establishment entrants and exits within multi-unit firms in Economic Census years. The new algorithms flexibly incorporate information on inter-censal establishment births and deaths from the COS into a formal statistical model used to impute first or last year of operation for establishments that appear to be entrants or exits in Census years at multi-unit firms that were not surveyed by the COS. These imputations improve the allocation of the timing of establishment expansions, contractions, and entry and exit for multi-unit firms in inter-censal years. See Section 7.

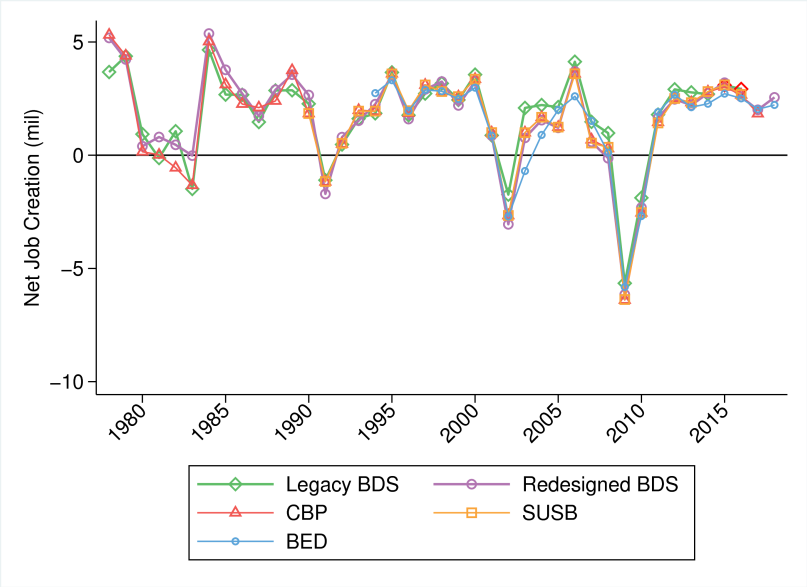
- Assigned vintage consistent NAICS codes (single classification vintage, e.g. 2012 NAICS codes) to all establishments for the entire time series. Using these vintage consistent NAICS codes we are able to produce business dynamics measures for the entire span of the data at significantly more detailed industry breakouts. Since the 1970s, several vintages of industry classification codes have been used including the SIC and several versions of the NAICS. The vintage consistent algorithms, developed by Fort and Klimek (2018), apply official and derived concordances between different vintages of industry coding schemes (including Standard Industrial Classification codes that were phased out beginning in 1997) to assign a more recent vintage of NAICS codes to establishments observed as far back as 1976. The vintage consistent codes also leverage both partial and longitudinal information to make the most accurate assignment possible when administrative information is limited. See Section 8.

- Following the SUSB and CBP data products, the new BDS tables use multiplicative noise to avoid the disclosure of sensitive information (Massell and Funk, 2007).

12.2 Implications for BDS Tabulations

Levels and trends in firm, establishment, and employment flows in the new BDS tables are very similar to the legacy tabulations. Patterns in net job creation, shown in Figure 5, line up well across a number of different Census Bureau data products including the legacy BDS tables, new BDS tables, CBP, and SUSB, as well as the Bureau of Labor Statistics Business Employment Dynamics data. All of these measures are highly correlated. The correlation between the legacy and new BDS measures of net job creation is over 0.98. The average correlation between the new BDS and each of the other Census data sources is over 0.99 and the correlation between the new BDS and BED measures of net job creation is over 0.97.

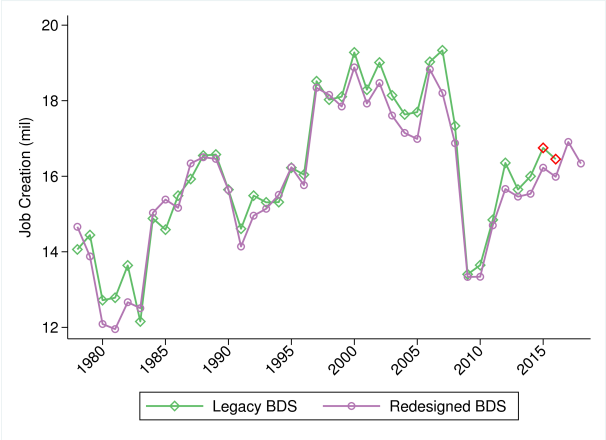
Figure 5: Net Job Creation



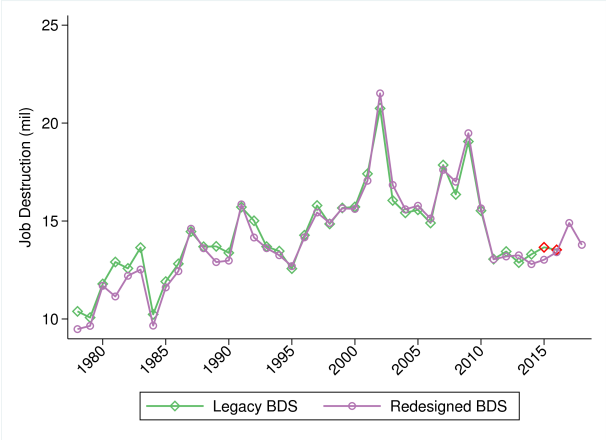
Source: 2018 BDS, CBP, SUSB, and BLS Business Employment Dynamics Data

Economy-wide gross employment flows are also very similar in the legacy and redesigned BDS data. Figure 6 shows job creation (a) and job destruction (b). On average, the new BDS has about 267 thousand fewer jobs created each year on a base level of job creation of about 16 million per year–a difference of about 1.6% of total job creation in the legacy data. Similarly, job destruction is about 176 thousand less each year, about 1.5% lower. The time series pattern of gross employment flows is very similar, rising and falling together over time. Improvements in the measurement of gross employment flows in earlier years is apparent even at the economy-wide level. The changes between 1980 and 1983 are smoother in the new data, rising less in 1982 and falling less in 1983 with an increase in 1985 rather than a decrease.

Figure 6: Job Creation and Destruction



(a) Job Creation

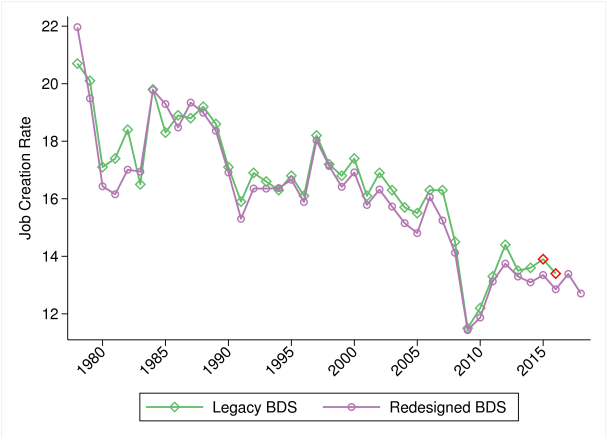


(b) Job Destruction

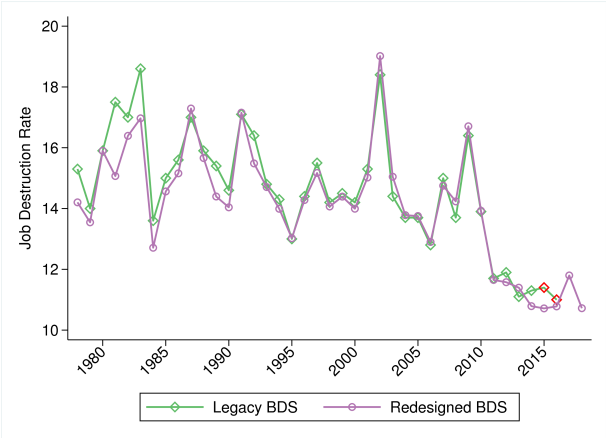
Source: 2018 BDS
Notes: The red points in the legacy BDS series are drawn from the limited, single-year BDS releases made in 2015 and 2016. These single-year releases, executed with the legacy system, were appended to the 2014 BDS release for the purposes of comparing the legacy and new BDS data. Subsequent figures append the 2015 and 2016 releases to the 2014 release for comparison purposes.

We find a similar pattern in rates. Figure 7 shows the job creation rate (a) and job destruction rate (b).[51] The figures exhibit the well documented secular decline in employment flows, with the job creation rate falling by 7.3 points in the legacy and 9.1 points in the new BDS between 1978 and 2016. The job creation rate in the redesigned BDS is on average about 0.31 points lower each year but again follows a similar pattern over time. The difference in the the job destruction rate is very similar at 0.29 points lower in the new release, again, with a very similar time series pattern.

Figure 7: Job Creation and Destruction Rates



(a) Job Creation Rate



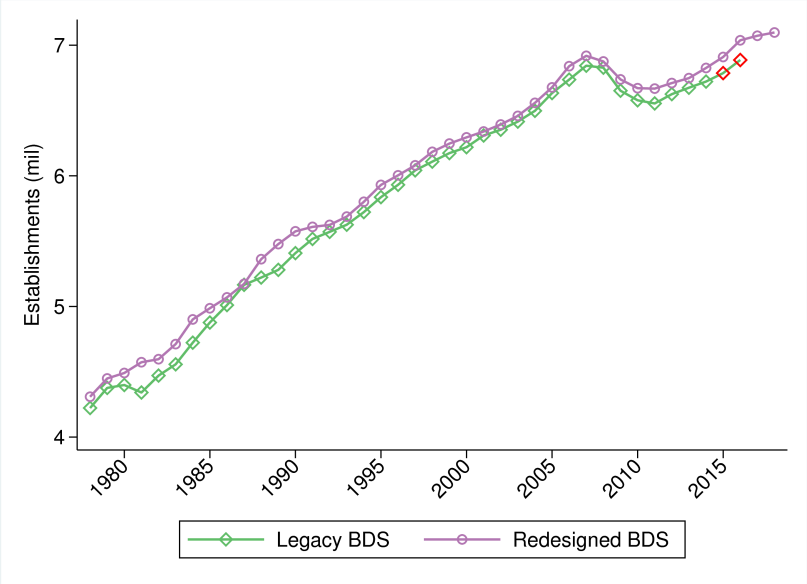
(b) Job Destruction Rate

Source: 2018 BDS

[51] These rates capture job creation and destruction as a fraction of total employment. See Section 11.

As mentioned above, the new LBD incorporates information from both the BR and the CBP microdata files. In some cases we have added establishments that were entirely missing from the legacy LBD.[52] In other cases we have incorporated additional information allowing more accurate identification of in-scope establishments. Both of these changes tend to increase the number of establishments contributing to the BDS. Figure 8 shows the count of establishments in the legacy and redesigned BDS tables. The establishment count in the new tables is always higher, and on average the new data include about 93 thousand (1.7%) more establishments each year. The number of additional establishments varies dramatically by year, ranging from an additional 7 thousand establishments in 1987 to an extra 232 thousand (5.3%) establishments in 1981. As explained in Section 4, the most successful data recovery efforts were for data in the 1980s.

Figure 8: Establishments



Source: 2018 BDS

These additional establishments are often flagged as in-scope in the CBP, meaning those establishments contribute to CBP data products.[53] To see this, we match all establishments in the new LBD flagged as in-scope for the BDS to the set of establishments used to tabulate the legacy BDS by the common establishment identifier (CFN or EMPUNIT_ID_CHAR).[54] We then examine the characteristics of establishments only found in the new BDS. Note that the CBP tabulation flag is one of the highest quality signals of whether an establishment ought to be considered in-scope.[55] On average, between 1978 and 2014, about 94% of establishments found only in the new BDS are

[52]This includes using "prior year" fields for a given year t to fill in missing records in $t-1$. This process of using prior year information to "fill in" holes in the data is done more systematically in the new LBD production system relative to the legacy system. See Section 4.

[53]The CBP and BDS use very similar but independent scope criteria, as shown in Section 10. The BDS uses additional longitudinal information from surrounding years while the CBP uses strictly cross sectional information to define scope.

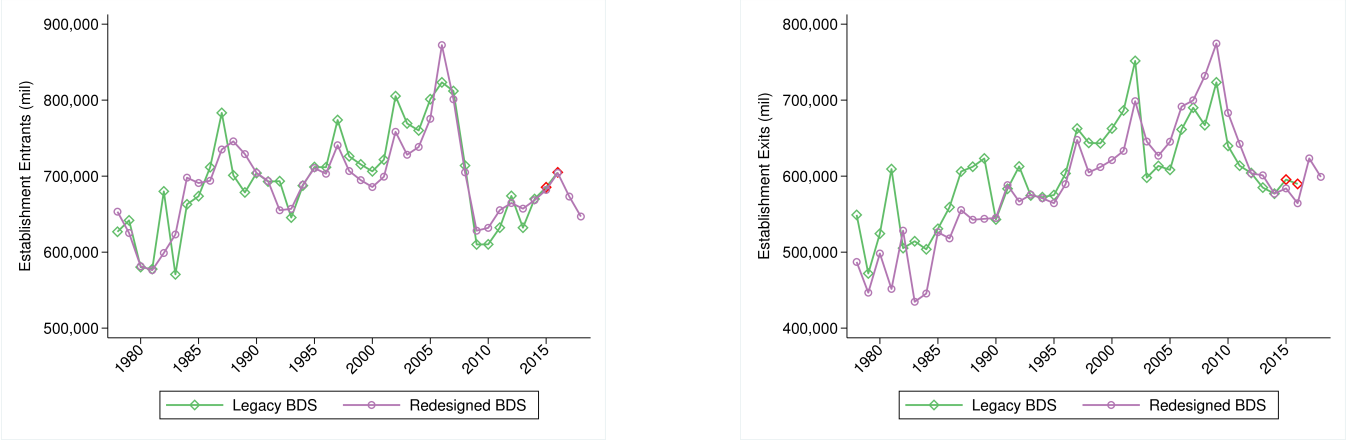
[54]For this exercise we use the microdata for the 2014 reference year BDS release.

[55]Note also that the CBP tabulation status derived from CBP microdata was not incorporated in the legacy BDS for years before the mid 2000s.

flagged as in-scope to the CBP.[56] This share is very similar for both continuing establishments as well as establishment entrants. Moreover, these establishments tend to be relatively small, with the average having fewer than 13 employees whereas the average size of all establishments in the BDS ($\frac{emp}{estabs}$ in the economy-wide table) is roughly 17 employees.

The counts of establishments entering and exiting each year, shown in Figure 9, shows a smaller gap between the new and legacy data than that found in the total establishment count. Prior to 2002, the year of the BR redesign, the count of establishment entrants in the new BDS was greater than in the legacy for only 11 of the 24 years in that period. Across all years, the count of establishment entrants, which averages about 690 thousand per year, is about 3 thousand lower in the new data than in the legacy data, or about 0.3% lower than the count of entrants in the legacy BDS. The differences in establishment exit counts is larger. The new data has 16 thousand (2.8%) fewer exits on average. Since the total number of establishments rose more than the count of entering and exiting establishments, we see a greater share of activity among continuing establishments.

Figure 9: Establishment Entry and Exit



(a) Establishment Entry

(b) Establishment Exit

Source: 2018 BDS

As expected, given the additional microdata and matching improvements, there are establishments flagged as entrants in the legacy BDS that are considered continuers in the new BDS. Due to the complexity of the data and processing, the inverse also occurs–some establishments are flagged as entrants in the new data but continuers in the legacy data. This can happen if the legacy data and new production systems make different linkages. The relative frequency of these two cases is informative for thinking about the differences in establishment entry and exit activity. On average, among the overlapping establishments for which the entry designation disagrees, 72% are entrants in the legacy that are now continuers. Of those, roughly 69% have link flags indicating a reorganization was identified.[57] Another 29% of these cases only in the new data have link flags indicating that they existed in both t and $t-1$ and were linked using establishment IDs.[58]

[56] For this calculation we exclude 1985, 1986, and 1987, the years for which no CBP microdata are available.

[57] We identify reorganization as flag_match_pass of D, E, F, G, and H. See Appendix B Table 4.

[58] Establishment ID matches are the most basic type of linkage in the LBD. One might ask why the legacy data did not make the same match. This may be another consequence of the additional establishments. Establishments may have been incorrectly classified as entrants in the legacy data because of missing records in $t-1$.

The combination of a relatively stable count of entrants and exits and a rising count of total establishments mechanically decreases establishment entry and exit rates. Figure 10 shows the establishment entry rate (a) and exit rate (b).[59] The entry rate in the new data is lower for the entire series but exhibits a very similar secular decline. In the legacy data the establishment entry rate fell 4.7 points (31.3%) from 15 in 1978 to 10.3 in 2016. In the redesigned BDS the entry rate falls 5.4 points (35%) over the same period from 15.5 to 10.1. Each year the entry rate is about 0.2 points (1.6%) lower in the new data. The entry rate in the new data is consistently lower in the 1990s through the mid 2000s. With the exception of 2006, the entry rate series appears smoother in the new BDS. This is especially true before 1990.

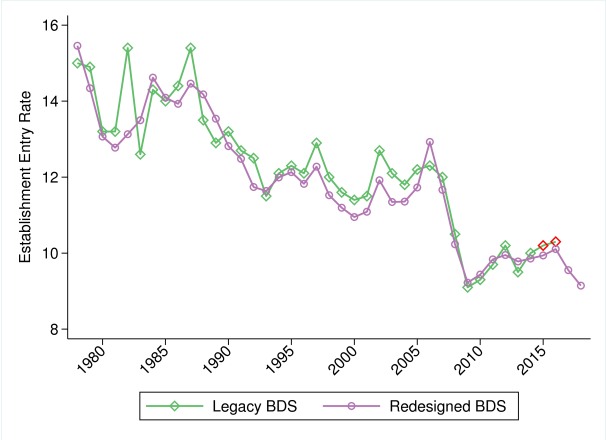
As was the case for exit counts, the gap is larger for the exit rate than for the entry rate. The establishment exit rate is about 0.5 points (4.2%) lower in the new BDS. Again, we see that the gap is bigger in earlier years. During the 1980s the gap in establishment exit rate was 1.4 points (11.2%). The time series pattern is much more flat during the 1990s and 2000s, with the 1990s being lower and late 2000s being a bit higher. The exit rate in the legacy data fell from 4.5 points (34.4%) from 1978 to 2016. The new BDS shows a decline of 3.4 points (29.7%) over the same period but nearly all of that decline occurs after 2010.

We can also see in Figure 10 that though the model for retiming Economic Census year births appears to perform better in the new data (see Section 7), there is still room for improvement as there are upticks in the entry and exit rates in most EC years. Nonetheless, the exit and entry spikes in 1982, 1987, 1997, and 2002 are noticeably lower in the redesigned BDS compared to the legacy version.[60] The new series also tracks a larger rise in the entry rate leading up to the Great Recession from 2005 to 2006. Critical to understanding the differences in entry and exit in the new BDS data is patterns in job creation from entrants and job destruction from exits, which we turn to next.

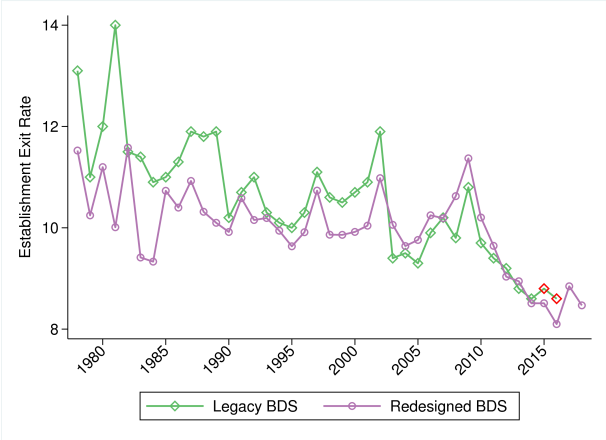
[59] These rates capture establishment entry and exit counts as a fraction of the longitudinally consistent total establishments active in t andor $t-1$. See Section 11. To "back out" the longitudinally consistent establishment denom we compute
$estabs\_denom_t = 0.5 * ((estabs_t + estabs\_exit_t - estabs\_entry_t) + estabs_t)$.

[60] From 2001 to 2002, in the legacy data, the entry rate rose 10% (1.2 points) and in the new data the entry rate rose 7.5% (0.8 points).

Figure 10: Establishment Entry and Exit Rates



(a) Establishment Entry Rate

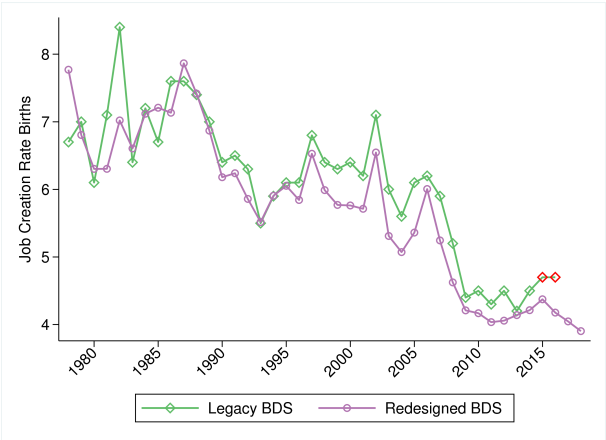


(b) Establishment Exit Rate

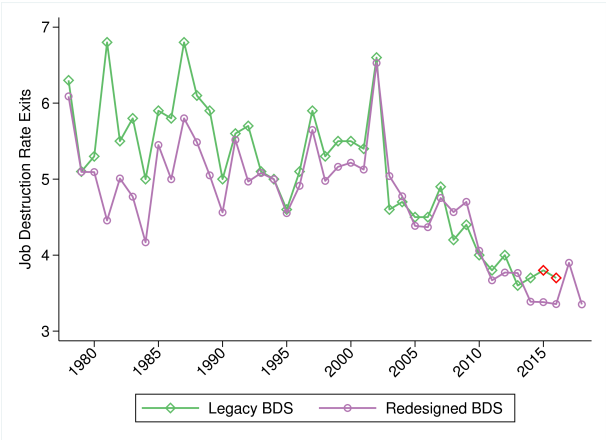
Source: 2018 BDS

Figure 11 shows the job creation rate for establishment entrants (a) and job destruction rate for exits (b). These capture the share of jobs created by establishment entrants relative to the sum of average employment in t and $t-1$ (`denom`). These employment weighted entry and exit rates exhibit similar patterns to the establishment weighted entry and exit rates shown in Figure 10. Each year, on average, both the job creation rate births and job destruction rate deaths are about 0.3 points lower in the new data. In employment flows from entry, the new data again are lower in the 1990s but that pattern continues through the end of the series. The job creation rate from of establishment entry is lower in every year after 1995 in the new data (on average about 0.4 points (7.6%) lower). Similar to the establishment exit rate, the job destruction from establishment exits is notably lower throughout the 1980s. After 2002 the legacy and redesigned job destruction exit series appear very similar.

Figure 11: Job Creation from Entry and Job Destruction from Exit



(a) Job Creation Entrants Rate

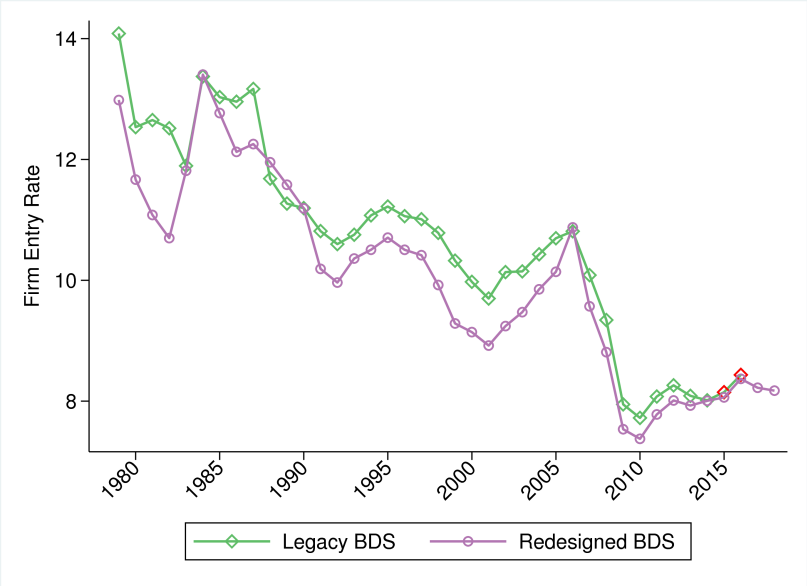


(b) Job Destruction Exits Rate

Source: 2018 BDS

Many times we are more interested in *firm* entry rates as opposed to establishment entry rates. Establishment entry rates capture transitions from zero to positive employment. Those transitions may represent either reactivations of old establishments or the creation of new establishments. Firm entry, on the other hand, captures the share of age zero (new) firms in the economy. By construction, age zero firms have only new establishments. Additional microdata and higher quality links tend to decrease firm entry rates in a way similar to establishment entry rates. The firm entry rate, shown in Figure 12, is lower in the new BDS, particularly in the 1990s and early 2000s. On average, the share of firm entrants is about 0.5 points (4.7%) lower in the new data. After 2007, the firm entry rates appear very similar.

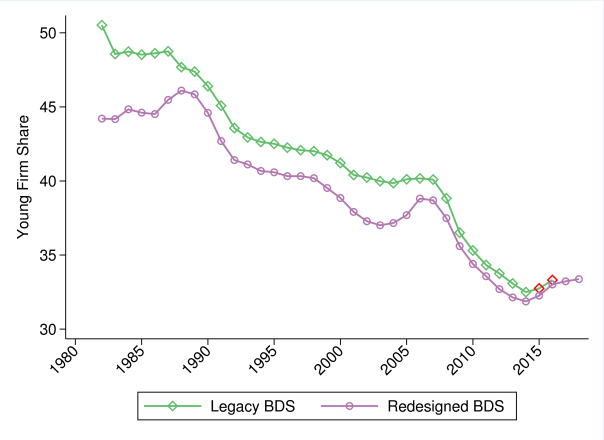
Figure 12: Firm Entry Rate



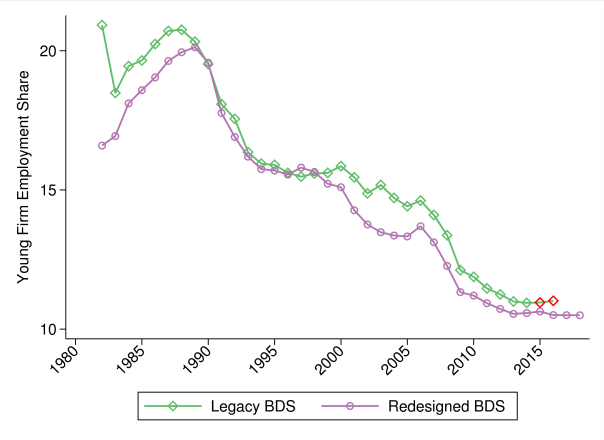
Source: 2018 BDS

The lower firm entry rate translates into a lower share of activity associated with young firms (which we define here as firms age ≤ 5). Figure 13 shows the share of young firms (a) and the share of employment at young firms (b). The gap in young firm share is greatest in 1982 at 6.3 points (12.5%) but quickly declines to about 2 points in 1990. The gap steadily declines from 2005 ending at about 0.6 points lower in the new data in 2014. On average, the young firm employment (denom) share is about 0.8 points (5%) lower in the tables.

Figure 13: Young Firm Activity Shares



(a) Young Firm Share



(b) Young Firm Denom Share

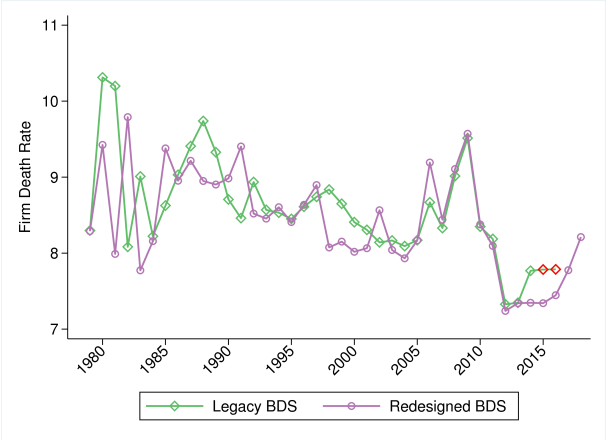
Source: 2018 BDS

The complement to firm entry is firm exit, something that is particularly difficult to measure in the underlying administrative data. Described in greater detail in Section 10, the definition of a firm death in the BDS requires that both the firm identifier and all its associated establishments cease to have positive employment. Nearer the end of the time series this measure becomes more noisy as future reactivations cannot be observed past the final year of data.[61] Figure 14 shows the firm death rate (share of all firms averaged in t and $t-1$) classified as firm deaths. Each year, roughly 8 to 9 percent of firms exit and 2 to 3 percent of employment is destroyed by firm exit. The firm death rate in the new data is relatively more volatile in the late 1990s and early 2000s with notable spikes in 2002 and 2006. The 2002 spike is even more dramatic on an employment-weighted basis, rising and falling by about a third.[62] The differences can at least partially be accounted for by the more effective identification of multi-unit firm deaths in the new BDS. The right panel of Figure 14 shows the ratio of firm death firm counts to firm death establishment counts. This ratio is a proxy for the amount of multi-unit activity among firm deaths. If all firm deaths were single-units this ratio would be equal to one. The ratio of firm death firms to firm death establishments is lower in every year in the new data, suggesting that more multi-unit firm deaths are being identified in the new data.

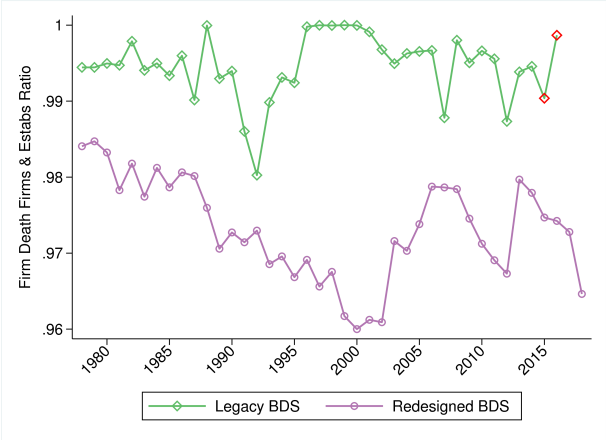
[61] As described in Section 10, we leverage subsequent-year first-quarter payroll in the last year of the data as a proxy for whether a firm is observed with positive employment in the following year and therefore should not be classified as a firm death.

[62] From 2.37 in 2001 to 3.17 in 2002 and back to 2.25 in 2003.

Figure 14: Firm Death Rate and Firm Death Firm to Firm Death Establishment Ratio



(a) Firm Death Rate



(b) Firm Death Firm-to-Estab Ratio

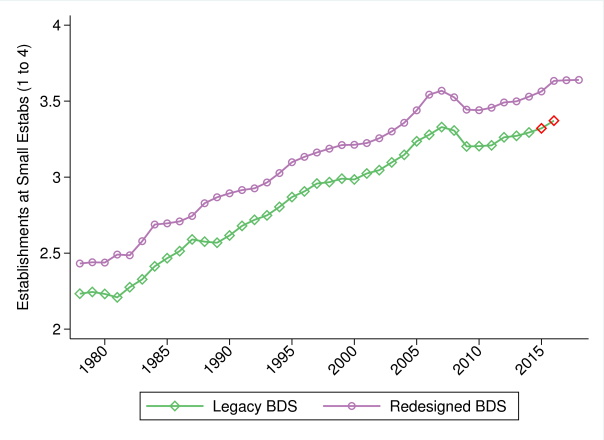
Source: 2018 BDS

Finally, another notable difference between the legacy and new BDS data is how average establishment and firm size categories are defined (`esize` and `fsize` respectively). These groupings capture the average establishment size between $t-1$ and t and the average size of an establishment's associated firm in $t-1$ and t. The rounding scheme used in the legacy data would place establishments (firms) with a size value in [4.5,9.5] into the "b) 5 to 9" group. In the new BDS data all establishment with average establishment (firm) size ≤ 5 are assigned to the "a) 1 to 4" group. This difference affects other groups as well. For example, establishments (firms) size [9.5,10) will be assigned to the "b) 5 to 9" whereas in the legacy data they would have been assigned to the "c) 10 to 19" group. This has implications for the distribution of activity by establishment (firm) size bins but to a great extent those changes do not reflect a shift in the underlying size distribution of establishments and firms. It is important to note that there has been a relatively small shift towards small establishments, which can be seen in comparisons of *initial* establishment size groups, which remain comparable between the legacy and new BDS tables.

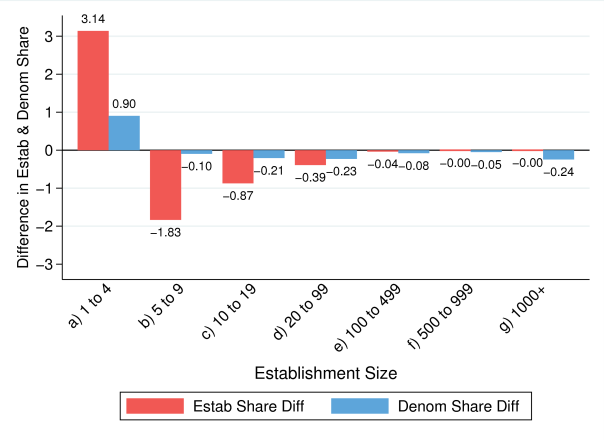
Panel (a) of Figure 15 shows the count of establishments in the "a) 1 to 4" `esize` group each year. Panel (b) of the figure shows the average difference across years in the share of establishments and employment (`denom`) by `esize` groups. On average, the new BDS data assigns 229 thousand additional establishments to the "a) 1 to 4" group. In percentage terms, this is about an 8.1% increase in the number of establishments in the smallest average size group with an accompanying 14.5% increase in employment in that group. The right panel of Figure 15 captures how the differing rounding scheme affects the size distribution across all establishment size categories. The share of establishments in the 1 to 4 bin rose by 3.14 points on average each year (49.5 to 52.6). The share of `denom` for this group rose less (0.9 points). By construction, since we are computing shares of total establishment counts and employment, other groups declined in share. Most of the decline was concentrated among establishments with less than 100 employees with the biggest decline (1.8 points) in the `esize` "b) 5 to 9" group. The movement of activity among these groups will be driven by the relative frequency of establishments sitting in the range of values where different classifications are made. The most numerous set of establishments at risk for a different classification are those in the [4.5,5) range. Data users should not make direct comparisons of the `esize` and `fsize` data between the legacy and new BDS data due to the differences in rounding.

Initial establishment and initial firm size (`iesize` and `ifsize`), however, remain comparable, which we turn to next.

Figure 15: Small Establishments and Percent Difference by Establishment Size



(a) Small Establishment Count

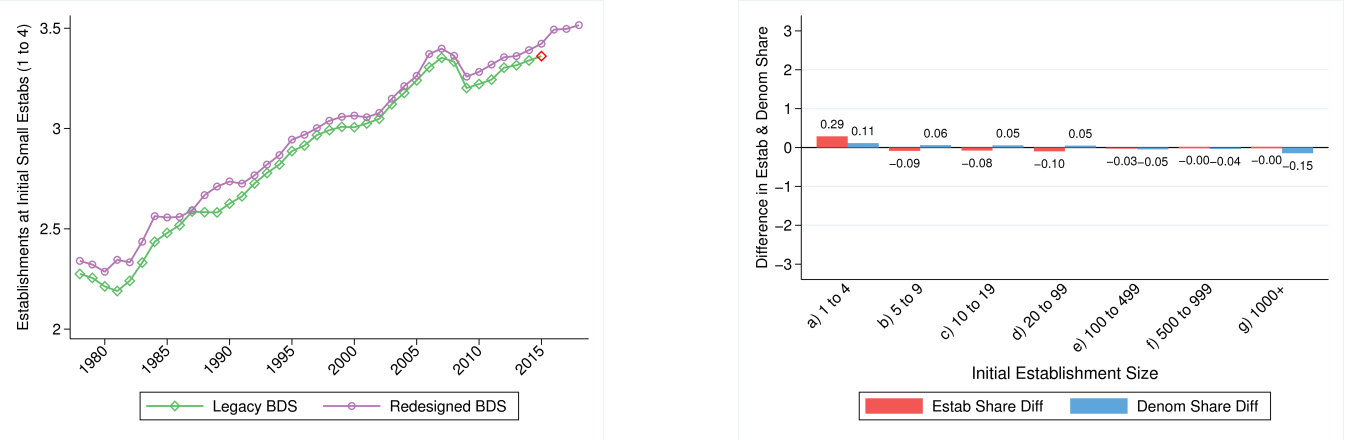


(b) Estab and Denom Share Difference

Source: 2018 BDS
Notes: Establishment and denom differences computed as the redesigned value minus the legacy value.

Initial establishment and initial firm size captures the size of the establishment or associated firm in $t-1$. Since initial establishment and firm size are based upon integer counts, they are not affected by the difference in rounding method. The initial size measures in the legacy and new data are directly comparable and any differences we see represent shifts in the size composition of the underlying microdata. Figure 16 replicates the exercises in Figure 15 but using `iesize` data rather than `esize` data. Figure 16 confirms the finding that there are more establishments in the new BDS data and the additional establishments tend to be small. The new data has about 61 thousand (2.3%) more establishments in the smallest initial establishment size group each year. In the right panel (b) of Figure 16 we see that the share of initial establishment size "a) 1 to 4" rose 0.29 points (49.7 to 50) and their share of employment rose 0.11 points (6.8 to 6.9). The remaining groups saw changes in establishment shares that were much closer to zero. Interestingly, the employment shares of groups less with than 100 employees rose while their establishment share declined, suggesting a slight increase in the average size of establishments in those groups.

Figure 16: Initial Small Establishments Size and Percent Difference by Initial Establishment Size



(a) Small Establishment Count

(b) Estab and Denom Share Difference

Source: 2018 BDS
Notes: Establishment and denom differences computed as the redesigned value minus the legacy value. To ease comparison, the y-scale of panel (b) is the same as the y-scale of panel (b) in Figure 15.

Overall, the BDS tabulations exhibit similar long term trends and time series patterns. However, even at the economy-wide level, the effects of the improvements made to the LBD are visible. Net job creation rates and gross job creation and job destruction levels are very similar between the legacy and new BDS tables. Job creation and destruction rates are smoother over time and slightly lower. The new tables also include many more establishments each year, the vast majority of which are found to be in-scope to the CBP and tend to be relatively small. The biggest differences between the new and legacy tables are on the entry and exit margins–establishment and employment weighted entry and exit rates are lower in the new data. Firm entry and young firm activity is also lower in the new BDS tables. We also find a more volatile firm death series, which is driven by the more accurate identification of mulit-unit firm death. Finally, BDS users will find mechanical differences in the composition of activity by `esize` and `fsize` due to differing rounding methodology. That mechanical change makes those tables incomparable, but in the still comparable initial size series (`iesize`, `ifsize`) we find a slight increase in small establishment activity.

13 Future Improvements

After almost three decades of research on business dynamics with Census Bureau data, a great deal has been learned about the longitudinal linking of establishments. At the same time, Census staff continue to investigate ways to improve the LBD and BDS through the use of additional linking, additional data, or additional modeling tools. We conclude this paper with a discussion of current research topics in hopes of providing ideas to researchers interested in collaborating with the Census Bureau on LBD-BDS projects.

An important area requiring additional research is the longitudinal linking of firms. The LBD currently contains the traditional firm identifier, `firmid`, as well as a new firm linking identifier called `lbdfid`. `lbdfid` is meant to serve as a platform for linking firms over time but it does not yet incorporate rules to bridge breaks in firm identifiers over time. `firmid`s change due to ownership

or restructuring changes. The best known reason for a change in `firmid` is a transition from single to multi-unit status. Although the firm as an economic entity continues, in the LBD we assign a different `firmid` when a SU-MU transition occurs because of how we define `firmid`. Research is needed to establish what constitutes meaningful economic changes that should be labeled as a new firm, and what types of changes are inconsequential to the conceptual definition of the firm and hence should not alter the `firmid`. This work would involve investigating merger and acquisition activity and potentially name and address matching of establishments belonging to multi-unit firms, something currently not done in the LBD production linking system. The lessons learned from these analyses could be built into the `lbdfid` field, allowing for more accurate measurement of firm characteristics and longitudinal changes in firm outcomes.

Closely tied to longitudinal firm linking is the need to improve the measurement of firm death. The BDS defines firm death as (1) the firmid ceases to be observed with positive payroll or employment and (2) all establishments associated with the firmid in its last year of activity also cease to have positive activity. Firm death is difficult to measure for small entities because we cannot always distinguish between an exit from economic activity and missing data due to late or negligent tax returns. In fact, it is not uncommon for small firms to have some years of non-employment followed by a return to employer status in a later year, thus making initial counts of firm deaths in a year inaccurate until subsequent years of data are added to the time series. Given the economic impact of the COVID-19 pandemic, more accurately measuring firm deaths in the LBD and BDS will be an essential task.

The BDS is based on March 12th employment and all reported changes are between March 12th of consecutive years. However since 2004, the BR has captured June 12th, September 12th, and December 12th employment as well. In the future it is possible that quarterly employment changes could be reported or annual changes between quarters 2, 3, or 4. In order to facilitate this new type of BDS publication, more research is needed on the quality of these additional quarters of data, the cyclicality of employment tax reports, and the distribution of quarterly data across multi-unit establishments. Likewise research on the quality of the quarterly and annual payroll variables on the Business Register could facilitate inclusion of payroll-based measures in future BDS publications.

Another area in which the LBD-BDS team hopes to improve measurement is through the use of W-2 tax filings. These person-level records can be linked to business data through the EIN identifier and offer another source of employment information in addition to the Form 941 tax filings that figure prominently in the construction of the Business Register. The W-2 records could be used to impute missing employment reports in the BR or to confirm lack of employment. W-2 records may also be useful in linking single-unit firms by tracing large groups of workers who move from one EIN to another, signaling a re-organization of an existing firm instead of the entry of a new one. Finally, W-2 records could be used to create additional worker-level detail for inclusion in the BDS, such as the age/sex distribution of workers at firms in various categories or even the average 90-10 earnings differential within a group of firms. Currently the the Census Bureau has W-2 records dating from 2005 to 2019 and is actively seeking for additional past years in order to facilitate this research.

As a longitudinal database, the LBD is particularly concerned with measuring the timing of changes at firms. However, this is often difficult given the data collection cycle, with most detailed information coming only every 5 years as part of the Economic Census. As described in Section 7, considerable effort is spent in the LBD production process on re-timing the births and deaths of establishments that belong to multi-unit firms. However other changes such as geography and industry are not re-timed and hence exhibit large spikes in Census years. In particular, the number of firms that report a change in their industry classification that is not vintage-related (i.e., not merely a function of an updated NAICS coding system), is much larger in Census years than the years before

or after. This is almost certainly caused by the Business Register "catching up" with cumulated changes that had gone unobserved. Likewise, changes in address are more frequent in Census years as both mailing and physical addresses are updated. More research is needed to understand these changes and how to predict when they truly occur. At the same time, consideration is being given to how to improve the birth-death retiming algorithm. While the current model provides a definite improvement to the time series by spreading births and deaths across the intercensal period, the timing of single-unit establishments growing into multi-unit firms remains difficult to impute accurately. Future modeling efforts will consider using LEHD data to try to measure this transition more precisely. Research into what predicts firm growth along this dimension would be helpful to the Census Bureau in evaluating the quality of modeling outcomes.

Some researchers using the LBD to study a particular industry or geography will happen upon cases that appear to be missed links. Researchers can examine underlying Business Register data for a large entrant or exit to look for evidence of a re-organization of an existing business. This type of attention to special cases is very helpful to the Bureau as it supplements internal quality assurance work. Census staff have begun to populate a database with links identified in this manner that will then be used as part of the production process to repair false breaks in establishment histories.

Finally, the Census Bureau is actively seeking to add information about firms to the LBD in order to produce BDS tables with more detail. Plans are underway to add a goods-trader designation (importer, exporter, or both), a patenting firm designation, and a High Tech industry designation (Kamal and Ouyang, 2020; Graham, Grim, Islam, Marco, and Miranda, 2018; Dreisigmeyer, Goldschlag, Krylova, Ouyang, and Perlman, 2018; Goldschlag and Miranda, 2020). These additional firm and establishment characteristics will allow us to produce BDS tables of firm, establishment, and employment flows for globally-engaged, patenting, High Tech businesses.

Perhaps the most important continuing addition to the LBD and BDS will be further years of data. By integrating its two longitudinal business products through a formal production process, the Census Bureau has created a system that will be able to provide meaningful information about business dynamics for many years to come.

References

Abowd, J. M., B. E. Stephens, and L. Vilhuber (2006) "Confidentiality Protection in the Census Bureau Quarterly Workforce Indicators," Longitudinal Employer-Household Dynamics Technical Papers Technical Paper No. TP-2006-02, Center for Economic Studies, U.S. Census Bureau.

Atrostic, B., R. A. Becker, T. Gardner, C. Grim, and M. Mildorf (2010) *2009 Research Report: Center for Economic Studies and Research Data Centers*chap. Recovery of Historical U.S. Census Bureau Microdata: Success to Date. Available at https://www2.census.gov/library/publications/2010/adrm/ces/data-recovery-extract-annual-report-2009.pdf.

Bayard, K., and S. Klimek (2003) "Creating a historical bridge for manufacturing between the Standard Industrial Classification System and the North American Industry Classification System," *The Proceedings of the Annual Meeting of the American Statistical Association*.

Benedetto, G., J. C. Stanley, and E. Totty (2018) "The Creation and Use of the SIPP Synthetic Beta v7.0," Sipp working paper.

Benedetto, G., M. H. Stinson, and J. M. Abowd (2013) "The Creation and Use of the SIPP Synthetic Beta," Sipp working paper.

Bernard, A. B., J. B. Jensen, and P. K. Schott (2006) "Survival of the Best Fit: Exposure to Low-Wage Countries and the (Uneven) Growth of U.S. Manufacturing Plants," *Journal of International Economics*, 68, 219–237.

Bloom, N., K. Handley, A. Kurman, and P. Luck (2019) "The Impact of Chinese trade on US Employment: The Good, the Bad, and the Debatable," mimeo, Stanford University.

Davis, S., J. Haltiwanger, and S. Schuh (1996) *Job Creation and Destruction*. MIT Press.

Davis, S. J., R. J. Faberman, J. Haltiwanger, R. Jarmin, and J. Miranda (2010) "Business volatility, job destruction, and unemployment," *American Economic Journal: Macroeconomics*, 2(2), 259–87.

Davis, S. J., J. Haltiwanger, R. Jarmin, and J. Miranda (2007) "Volatility and Dispersion in Business Growth Rates: Publicly Traded vs. Private Sector Firms," *NBER Macroeconomics Annual 2006*, 21, 107–180.

DeSalvo, B., F. Limehouse, and S. Klimek (2016) "Documenting the Business Register and Related Economic Business Data," CES Working Paper CES-WP-16-17.

Dreisigmeyer, D., N. Goldschlag, M. Krylova, W. Ouyang, and E. Perlman (2018) "Building a Better Bridge: Improving Patent Assignee-Firm Links," Center for Economic Studies Technical Notes CES-TN-2018-01, Center for Economic Studies, U.S. Census Bureau.

Dunne, T., M. J. Roberts, and L. Samuelson (1988) "Patterns of firm entry and exit in US manufacturing industries," *The RAND journal of Economics*, pp. 495–515.

Evans, T., L. Zayatz, and J. Slanta (1998) "Using Noise for Disclosure Limitation of Establishment Tabular Data," *Journal of Official Statistics*, 14(4), 537551.

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

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(2), 47–72.

Goldschlag, N., and J. Miranda (2020) "Business dynamics statistics of high tech industries," *Journal of Economics & Management Strategy*, 29(1), 3–30.

Graham, S. J., C. Grim, T. Islam, A. C. Marco, and J. Miranda (2018) "Business dynamics of innovating firms: Linking US patents with administrative data on workers and firms," *Journal of Economics & Management Strategy*, 27(3), 372–402.

Haltiwanger, J., R. S. Jarmin, and J. Miranda (2013) "Who creates jobs? Small versus large versus young," *Review of Economics and Statistics*, 95(2), 347–361.

Jarmin, R., S. Klimek, and J. Miranda (2005) "The Role of Retail Chains: National, Regional, and Industry Results," CES Working Paper CES-WP-05-30.

Jarmin, R., and J. Miranda (2002) "The Longitudinal Business Database," CES Working Paper CES-WP-02-17.

Kamal, F., and W. Ouyang (2020) "Identifying U.S. Merchandise Traders: Integrating Customs Transactions with Business Administrative Data," CES Working Paper CES-WP-20-28.

Massell, P. B., and J. M. Funk (2007) "Recent Developments in the Use of Noise for Protecting Magnitude Data Tables: Balancing to Improve Data Quality and Rounding that Preserves Protection," in *Proceedings of the 2007 FCSM Research Conference*.

Massell, P. B., L. Zayatz, and J. M. Funk (2006) "Protecting the Confidentiality of Survey Tabular Data by Adding Noise to the Underlying Microdata: Application to the Commodity Flow Survey," in *Privacy in Statistical Databases*, ed. by J. Domingo-Ferrer, and L. Franconi, pp. 304–317, Berlin, Heidelberg. Springer Berlin Heidelberg.

McGuckin, R. H., and G. A. Pascoe (1988) *The longitudinal research database (LRD): Status and research possibilities*. US Department of Commerce, Bureau of the Census.

Raghunathan, T., J. Lepkowski, J. van Hoewyk, and P. Solenberger (2001) "A Multivariate Technique for Multiply Imputing Missing Values Using a Series of Regression Models," *Survey Methodology*, 27, 85–96.

Tornqvist, L., P. Vartia, and Y. Vartia (1985) "How Should Relative Changes Be Measured," *The American Statistician*, 39(1), 43–46.

White, T. K. (2014) "Recovering the Item-Level Edit and Imputation Flags in the 1977-1997 Censuses of Manufactures," CES Working Paper CES-WP-14-37.