Upcoming Improvements to the Longitudinal Business Database and the Business Dynamics Statistics[1]

Martha Stinson*

T. Kirk White*

James Lawrence**

*Center for Economic Studies, U.S. Census Bureau

**Economy-Wide Statistics Division, U.S. Census Bureau

Abstract:

The Business Dynamics Statistics (BDS) provide annual measures of business dynamics (such as job creation and destruction, establishment births and deaths, and firm startups and shutdowns) for the economy and aggregated by establishment and firm characteristics. The BDS is created from the Longitudinal Business Database (LBD), a confidential database available to qualified researchers through secure Federal Statistical Research Data Centers. The use of the LBD as the BDS source data permits tracking establishments and firms over time. As part of the LBD Initiative, the Census Bureau is making several improvements and enhancements to the LBD and BDS, including: (i) reconstructing the LBD/BDS using a longitudinal linking methodology that is as consistent as possible over the entire time series; (ii) filling data gaps and improving data quality by incorporating nearly four decades of data from the Census Bureau's County Business Patterns (CBP) program as well as recently recovered data from the Census Bureau's Business Register; (iii) integrating the LBD with another Census Bureau data product, the Business Information Tracking Series (BITS), incorporating the best features of each program; (iv) streamlining and documenting the LBD's code base to make the LBD/BDS easier to maintain and improve in the future; (v) publishing the entire BDS on a NAICS basis; (vi) implementing a new disclosure avoidance methodology for the BDS.

Background on the LBD/BDS and the BITS

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)

[1] Any opinions and conclusions expressed herein are those of the authors and do not necessarily represent the views of the U.S. Census Bureau. All results have been reviewed to ensure that no confidential information is disclosed.

in the early 1980s. The inputs to the LRD are cross-sectional plant level data from the quinquennial Censuses of Manufacturers and the Annual Survey of Manufacturers, augmented with administrative data. These data are linked longitudinally at the plant level using numeric identifiers from the input datasets. These longitudinal linkages allow researchers to measure how the number of businesses change—entry and exit as well as net changes—and how individual business are 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), Davis, Haltiwanger, and Schuh (1996) and Haltiwanger, Jarmin, and Miranda, (2013). The academic interest in business dynamics statistics 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 besides manufacturing, and by the fact that manufacturing's importance as a source of jobs 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). The LBD was subsequently utilized in numerous microeconomic analyses including Jarmin, Klimek, and Miranda (2005), and Davis, Haltiwanger, Jarmin, and Miranda (2007). Interest in LBD research findings generated requests for special tabulations. This led to the creation and release of the Business Dynamics Statistics (BDS) data tables in the late 2000s. While still technically a research product, the BDS has developed a wide constituency of users, including policy makers, the business community, and researchers. One of the most important innovations of the BDS is that it includes a measure of firm age, which is not available for any other dataset that covers the entire non-farm employer economy.

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 SBA to request that the Census Bureau develop and publish statistics on business dynamics using establishment-level data from the economy-wide County Business Patterns (CBP) program. This was the Business Information Tracking Series (BITS) program. Similar to the LRD and later the LBD, BITS linked establishments over time and based on the linked micro data, tabulated statistics on business dynamics. The BITS tabulations continue to be published today by a production Division within the Census Bureau as part of the Statistics of U.S. Businesses (SUSB) program.

Since 2002, the Bureau of Labor Statistics (BLS) has published the Business Employment Dynamics (BED).[2] The BED is a set of statistics generated from the Quarterly Census of Employment and Wages (QCEW) program.[3] These quarterly data series consist of gross job gains and gross job loss statistics from 1992 forward. Although the data sources and firm definitions are different, the BED and BDS series generally track each other closely at the national level.

1. Overview of Transition, Redesign, and Integration of the LBD/BDS

The popularity of the BDS led the Census Bureau to make plans to adopt the program as an official data product. This required that it be transitioned from CES, which is in the Research and Methodology Directorate, to a production division in the Economic Directorate. Part of this transition has involved applying production standards for data processing, programming, documentation, and dissemination which has resulted in a redesign of LBD/BDS processing. At the same time, because the LBD and BITS data programs are so similar, it was decided to integrate the two programs and produce both the BDS and the relevant SUSB tables from the same underlying datasets. The LBD Initiative is providing additional funding to the Census Bureau for resources to move production of the LBD to the Economy-Wide Statistics Divsion, allowing CES to focus research on enhancements to the LBD and developing new data products.

The transition includes six major types of improvements to the LBD/BDS/BITS products: (i) reconstructing the LBD/BDS using a longitudinal linking methodology that is as consistent as possible over the entire time series; (ii) filling data gaps and improving data quality by incorporating nearly four decades of data from the Census Bureau's County Business Patterns (CBP) program as well as recently recovered data from the Census Bureau's Business Register; (iii) integrating the LBD with another Census Bureau data product, the Business Information Tracking Series (BITS), incorporating the best features of each program; (iv) streamlining and documenting the LBD's code base to make the LBD/BDS easier to maintain and improve in the future; (v) publishing the entire BDS on a NAICS basis; (vi) implementing a new disclosure avoidance methodology for the BDS. We briefly describe each of these improvements.

1.1 Applying a Consistent Longitudinal Linking Methodology Over the Entire Time Series.

The LBD and BDS are research databases, which are updated annually as each new year of data becomes available. As described in Jarmin and Miranda (2002), the vast majority of longitudinal linkages in the LBD are created by matching numeric identifiers. However, establishment-level numeric identifiers sometimes change over time. For example, prior to the 2002 redesign of the Census Bureau's business register, the establishment identifier automatically changed when a firm with a single establishment (single-unit or SU) became a

[2] See https://www.bls.gov/bdm/.
[3] See https://www.bls.gov/cew/.

multi-unit (MU) firm. Other changes happen because a single-establishment firm changes ownership and gets a new tax identifier (EIN) and the Census Bureau mistakes this re-organization for a birth and assigns a new Business Register establishment identifier. To fix the resulting broken longitudinal linkages, the LBD uses various kinds of name and/or address matching between consecutive years. As CES researchers have developed new methods of name and/or address matching, these additional methods have been applied to the LBD. In general these improvements have tended to reduce the number of establishments identified as births and deaths in a given pair of years and increase the number of establishments identified as continuers. In most cases these improvements to the longitudinal linking methodology have been applied to the most recent year of the LBD, but they have not been used to revise the entire time series. This means that more recent years of the LBD and BDS time series are more likely to have establishments linked longitudinally, other things equal. Of course, the underlying "true" numbers and rates of establishment births and deaths also change over time, so it is impossible to know exactly how much the changes in methodology affect the measured numbers without applying the same methodology to the entire time series. However, the vast majority of longitudinal linkages are made using numeric identifiers, so the overall levels of continuers vs. births and deaths are unlikely to change dramatically.

1.2 Incorporating Additional Data

The redesign of the LBD/BDS will incorporate a number of files which have not been used in previous vintages of the LBD/BDS. Here we give a brief overview of these data and the motivation for using them. We provide a detailed description of how we incorporate the new data in section 2.

The Census Bureau's County Business Patterns (CBP) program uses the same Census Bureau Business Register (BR) data as an input. However, after Business Register processing is completed for a given reference year, CBP analysts make significant edits to some records. For example, an analyst may determine that a record classified as a single-unit establishment in the BR is in fact a multi-unit enterprise. In some cases, these edits cause large changes to establishment-level employment, as the record's employment is allocated to each of the multi-unit enterprise's separate establishments.

The BITS has always been processed downstream from the CBP edits, and the earliest vintages of the LBD used the CBP-edited files for selected years for which the original BR files were missing. Beginning with the 2013 vintage of the LBD, CBP edits from 2013 forward were incorporated into the LBD. However, as with changes to the linking methodology, the CBP edits from earlier years were not incorporated into the LBD time series. Until recently, it was thought that the CBP edits to the microdata prior to 1988 (when the BITS time series begins) were no longer available. However, a recent effort by CES recovered tens of thousands of data

tapes used by a 1970s era Unisys mainframe.[4] These data include the analyst edits to the CBP microdata for 1976-1984. We are now incorporating these CBP data into the integrated BITS-LBD product.

The first vintage of the LBD was constructed in the late 1990s using annual snapshots of the BR, known as the Standard Statistical Establishment List (SSEL). In each year the data is divided into a file of single-establishment (SU) firms and establishment-level data for multi-establishment (MU) firms. CES is the official data archive for the Census Bureau's Economic Directorate, but unfortunately a few of the SSEL files from the early years of the LBD time series did not make it to CES in complete form. The 1976 and 1981 SSEL SU files were missing, and had to be reconstructed using prior-year variables from the 1977 and 1982 SSEL files. The 1978 SSEL SU file was also missing data for a large number of establishments. Until now the LBD has included processing to address missing longitudinal linkages due to this missing data, but the data were still missing. The recovered CBP microdata files include all of the active 1976, 1978 and 1981 SUs, so we are able to fill in gaps due to previously missing data in those years.

The Census Bureau's Business Register has included fields for both a mailing address and a physical address at least since 1976. However, due to data storage constraints, the versions of the SSEL files that were archived at CES for the SSEL years prior to 1986 included only one address. Prior to 1983 only the mailing address was kept in CES's SSEL files. For 1983 until 2001, if the record included a physical street address, that address was kept in the mailing address field. Until now, LBD processing has only used the "mailing" street address (which may or may not be the same as the physical address) for longitudinal linkages using name-and-address matching. If the physical address is very different from the mailing address and an establishment changes from reporting a mailing address to a physical address (or vice versa), address matching will have a hard time finding the correct longitudinal linkage. The recovered CBP files do not include names and addresses. However, CES has also recovered the original SSEL SU files for 1977-1980 and 1982-1986 from the Unisys tapes, all of which include street, place, state and ZIP code for both a mailing address and a physical address. For a very large number of SUs in each of these years, the physical street address is different from the mailing street address. Furthermore, CES's existing 1987-2015 SSEL files already include separate variables for mailing and physical addresses. In the redesign of the LBD, we are incorporating both sets of addresses as part of the name-and-address matching for longitudinal linkages in every year of the time series.

1.3 Integrating the BITS and LBD.

Although both the BITS and the LBD use the Business Register as their primary data sources, there are a number of differences between the BITS and the LBD and BDS. One important example is that the BITS methodology has changed relatively little over time, while the LBD and

[4] For descriptions of this data recovery effort see https://www.census.gov/ces/dataproducts/recovered/ and Atrostic et al. (2009).

BDS methodologies have continued to evolve. This is the result of the LBD and BDS being research products, while the BITS is used almost exclusively to produce statistical tables. Another key difference between the programs is that the LBD includes a number of edits that are done to improve the longitudinal consistency of the employment time series, while the BITS program relies strictly on CBP edits to employment. One of the key innovations of the BDS is a measure of firm age, but this measure is not produced as part of the BITS program. Finally, the LBD/BDS time series begins in 1976, and the BITS time series begins in 1989.

The vast majority of longitudinal linkages in both the BITS and the LBD are made using numeric identifiers, which are the same in both datasets. To fix broken linkages due to changes in numeric identifiers, both the BITS and the LBD use name and address matching but with somewhat different methodologies. We describe both matching processes in detail in section 3.

Unsurprisingly, these two different sets of matching algorithms do not produce the same sets of matches and non-matches. In some cases, the BITS algorithm calls two records a match and the LBD does not. In other cases, the LBD algorithm call two records a match, and the BITS does not. As part of the integration of the LBD and BITS programs we are developing a machine learning algorithm to determine which passes of the BITS and LBD matching algorithms produce better matches so that the integrated product can include the best types of matches from each program.

1.4 Streamlining and Documentation.

Although the LBD/BDS processing and data products have several advantages relative to the BITS program, the BITS has two key advantages for an official data product of the Census Bureau: (1) the code has changed very little over time, making the entire time series replicable; and (2) in part because of (1), the code is easy to follow and well documented.

In contrast to the BITS, the LBD has always primarily been a research database. As such, over time CES researchers have made many updates to the code that produces the LBD, for example, after developing improvements to the name-and-address matching methods. Typically when improvements were made to the code, the updated code was only run on the most recent years of the BR files. Thus different years of the LBD time series were created using different sets of code, and the changes in the code over time were not well-documented. As research products, until now the LBD and BDS were not required to adhere to Census Bureau and Office of Management and Budget (OMB) standards for official data products, including specifications and documentation that make it possible to replicate the product. Furthermore, the nature of the existing code made it difficult to follow and to make changes to the product. As part of the LBD Initiative, several enhancements to the BDS are planned. These enhancements, together with the transition of the core BDS to a production environment, make documentation and maintainability of the code a high priority. As part of the transition of the LBD/BDS to a production environment, we are writing detailed specifications for all of the LBD and BDS

processing and code is being written that follows these specifications in accordance with Census Bureau production standards. We have also streamlined the processing to make it more efficient and easier to follow. This new code will be run for the entire time LBD/BDS time series so that the entire time series is replicable and uses a consistent methodology.

1.5 From SIC to NAICS.

Most of the Census Bureau's Economic programs switched to publishing industries on a North American Industrial Classification System (NAICS) basis beginning with reference year 1997. Although NAICS codes are available in the confidential establishment-level LBD data, until now BDS tables including industry classification have used Standard Industrial Classification (SIC) codes. Beginning with the 2016 vintage of the BDS, the entire BDS time series will be published on a 2012 NAICS basis. In the future the entire BDS time seires will be revised using final Economic Census data to update to the latest vintage of NAICS codes. The current plan is to update to the 2017 NAICS codes in reference year 2019 or 2020 of the LBD/BDS.

1.6 Disclosure Protection

In the past, BDS tables were protected from the risk of disclosing the identity of individual firms by suppressing employment totals for cells with too few establishments or firms. However, research is currently underway to change the BDS disclosure protection methodology to be differentially private (Dwork 2006). Describing the disclosure methodology in detail is beyond the scope of this paper. Intuitively, there is a tradeoff between the accuracy of the released statistics and the confidential protections (privacy). Differential privacy uses the concept of a *privacy-loss budget,* which allows us to make choices that move us along an accuracy-privacy Production Possibilities Frontier. In slightly more technical terms, this will involve synthesizing the actual published numbers in a manner that preserves the usefulness of the data content while putting an acceptable bound on the probability that an individual firm or establishment would be re-identified by the published characteristics.[5]

2. Reading In and Integrating CBP and SSEL/BR Input Files.

As mentioned above, the earliest vintages of the LBD primarily used annual snapshots of the Business Register, known as the Standard Statistical Establishment List (SSEL). The SSEL consists of two files for each year—one for single units (SU) and one for mult-unit (MU) establishments. When the LBD was first constructed, in certain years one of the two SSEL files was either partially or completely missing. For example, the 1988 SSEL MU file was completely missing and the 1989 SSEL MU was missing all establishments with a CFN beginning with 1. The early vintages of the LBD used the 1988 and 1989 CBP establishment files to fill in the missing data for these years. The 1976 and 1981 SSEL SU files were also completely missing and were reconstructed for the LBD using prior-year data from the 1977 and 1982 SSEL SU files,

[5] For more details on modernizing statistical disclosure limitation at the Census Bureau, see https://www2.census.gov/cac/sac/meetings/2017-09/statistical-disclosure-limitation.pdf.

respectively. Finally, the 1978 SU file was missing all establishments with CFNs starting in 7 or 8. Miranda (2002) and Jarmin (2002) provide more details on how the earlier vintages of the LBD dealt with these missing input data.

In addition to inconsistencies in the input files (CBP-edited vs. original SSEL files) there have been changes in the format and content of the files that are inputs to the LBD. The Business Register files that were used for program area processing included separate fields for both physical and mailing address. Due to space constraints, a decision was made to keep only the mailing address fields on the SSEL files archived at CES for 1976-1982. For the 1983-1986 SSEL, the mailing street address field was populated with the physical street address if available and otherwise only the mailing address was kept. Beginning with reference year 1987, the SSEL files include separate variables for street, city, state, and zip code for both mailing and physical address. However, until now the physical address variables have not been used as part of name-and-address matching in the LBD.[6]

2.1 Using the Newly Acquired Data Input Files and Variables

As part of its historical data recovery project, CES recovered CBP files for 1976-1984 that were not previously available for use in the LBD.[7] In addition, the BITS program has archived versions of the CBP files from 1988 to 2014. Thus we now have the edited CBP files for every year of the LBD time series except 1985-1987, and all of these files will be used as inputs to the integrated BITS-LBD product. In addition to providing CBP edits for almost the entire LBD time series, these files are also being used to fill in data that was missing from the 1976, 1978, and 1981 SSEL files.

CES has also recovered versions of the SSEL SU files for 1978-1980 and 1982-1986. In some years (1982-83) the recovered files contain exactly the same set of establishments as the SSEL SU files previously used as inputs to the LBD. In other years, the recovered files contain only active (payroll>0) SU establishments or a subset of the records in the SSEL SU files used previously. However, all of these files have separate variables for physical street, place, state zip code which are not available on the SSEL files for these years used to create previous versions of the LBD. In a large number of cases in every year, the physical address is populated and differs from the mailing address for the same establishment.

---

[6] In most cases either the physical address variable is blank or it is the same as the mailing address. However, for a large number of establishments in every year, the physical address is present and different from the mailing address.

[7] The files were read from the Unisys data tapes under the assumption that they were stored in Fieldata or XS-3 formats (see https://en.wikipedia.org/wiki/Fieldata and https://en.wikipedia.org/wiki/Excess-3). In fact, many of the files were stored either as a combination of Fieldata and binary representation or ASCII and binary. As part of the LBD-BITS integration effort, these files were read and translated character-by-character (or in the cases where 9-bit ASCII characters had been assumed to be 6-bit Fieldata, bit-by-bit) into SAS-readable formats. For a description of a similar process used for a different set of recovered files, see White (2014).

As part of the LBD transition, we plan to use the physical address variables that are now available for 1978-2015. These are useful for two reasons. First, for many research purposes (including production of the BDS), the address we care about is the physical address of the establishment. To the extent that the physical address differs from the mailing address, using the physical address fields for 1978-1982 will give us a more accurate measure of the establishment's location. Second, if the physical and mailing addresses for a given establishment are different, having both addresses is useful for longitudinal linking. Every year, especially in Economic Census years,[8] many establishments' addresses on the BR change.[9] To the extent that these changes are the result of a firm reporting the physical location of the establishment, instead of, e.g., a P.O. Box, or vice versa, having both addresses is useful for longitudinal linking. By separately matching each type of address in year t to each type of address in year t-1, we hope to fix some broken longitudinal linkages that result simply from address changes. We plan to do this for 1978-2015, thus improving the longitudinal linking and using a consistent methodology for almost the entire LBD time series.

2.2 New File Formats after the 2002 Business Register Re-design

Beginning with the 2002 reference year, the Census Bureau's Business Register was completely redesigned and transitioned to an Oracle database. Describing all of the changes to the BR in the 2002 redesign is beyond the scope of this paper.[10] Here we focus on the changes that were most important for the LBD.

First, instead of the SSEL SU and MU files, the new BR produces two files each year for the LBD: an empunits file and an einunits file. These files are extracts from the corresponding tables in the Oracle database, and contain data for the current processing year as well as the prior year. The einunits file contains EIN-level data, including March 12 employment and quarterly payroll from administrative records. For single-unit establishments, these data are equivalent to establishment-level (and firm-level) data. The empunits file contains *establishment*-level data for all employer establishments--including individual establishments that are part of multi-unit enterprises. Multi-unit enterprises can report payroll for multiple establishments under the same EIN. In years that end in "2" or "7" most establishments are mailed a survey and asked to report on the number of operating establishments they have and employment at those establishments. In other years, large multi-unit companies and a sample of small multi- and single unit companies are sent the Census Bureau's annual Company Organization Survey[11] (COS) or the Annual Survey of Manufactures (ASM). For multi-unit enterprises that report

[8] The Economic Census is the census of businesses, which Census Bureau conducts every five years (collecting data on years ending in 2 and 7).

[9] In a given year, most establishments' addresses on the BR do not change, and for a given establishment, the address stays the same most years.

[10] See DeSalvo, Limehouse and Klimek (2016)---available at https://ideas.repec.org/p/cen/wpaper/16-17.html--for a detailed description of the Business Register.

[11] The COS is also known as the Report of Organization. See https://www.census.gov/programs-surveys/cos.html for more information about this survey.

payroll taxes for multiple establishments under the same EIN, establishment-level data comes from the COS/ASM and/or from the latest Economic Census. If establishment-level data for an MU is not available in a given year—e.g., because the MU was not in the COS/ASM sample that year, or it failed to respond in time—the establishment-level payroll and employment are allocated (imputed) from the EIN-level data and the most recent establishment-level survey data (which may be from the COS/ASM or the Economic Census). In addition to establishment-level payroll and employment, the empunits file includes firm and establishment numeric identifiers, the EIN, establishment characteristics, such as business name, physical and mailing address, industry and geography codes, legal form of organization, and BR processing codes that are useful for identifying births and deaths.

As part of the 2002 BR redesign, the primary establishment-level identifier also changed from Census File Number (CFN) to the Employer Unit Identifier Number (empunit_id). Prior to 2002 there were also other numeric identifiers available for longitudinal linking. These identifiers went away in the BR redesign. These changes have important implications for longitudinal linking, which will be discussed in more detail in section 3.

In addition to the LBD microdata files, researchers at CES and in the FSRDCs have also used the SSEL files, which contain variables such as business name and address that are not on the final LBD microdata files. Up to and including the 2015 vintage, LBD processing has continued to create these SSEL files for research use by translating variable names in the empunits and einunits files to the old SSEL variable names. This facilitates research by having a consistent set of variable names across the entire time series. In vintages of the LBD through 2015, these post-2001 SSEL files were also used as the primary inputs to the LBD. This made it easier to update the LBD processing code because the variable names used in the processing code could stay the same.

Beginning with the 2016 vintage of the LBD, we will use the empunits and einunits files in LBD processing for the years 2002 forward. We still plan to make SSEL SU and MU files available to researchers with approved projects who want to use those files in the FSRDCs. Using the "new" files for processing facilitates maintenance and enhancement of the LBD production code going forward, both because it makes the processing easier to follow and because the current EWD staff are more familiar with the post-2001 file and variable names. In addition, some of the variables themselves--not just the names--changed completely in the transition from the old SSEL files to the empunits and einunits files. For example, the set of possible values of the flag variable used to identify imputed annual payroll changed from the SSEL files to the post-2001 BR files. Using the old SSEL variable names for LBD processing made this change less apparent, which led to a bug in the code used to identify and retime deaths of multi-unit establishments in 2002, 2007, and 2012. Using the post-2001 BR files and variable names in the LBD processing code makes these sorts of changes in the input data variables more transparent and thus facilitates maintenance and enhancement of the LBD going forward.

2.3 Reading in and determining which input data source to use for a given establishment.

In the redesigned LBD-BITS we have two or more input files for nearly every year (1985-1987 are the exceptions) and for most establishments. In the vast majority of records, the employment and payroll for a given establishment-year observation are the same in both the SSEL files and the CBP file. However, in some cases records with the same establishment identifier have different values for employment and/or payroll in the CBP file versus the SSEL file. This can happen for at least two reasons. First, the current-year variables in BR are frozen—the BR is "closed out"—towards the end of the calendar year (usually in November, but earlier in the year before an Economic Census) so that the Economic Census or surveys that use the BR for sampling have a fixed frame. CBP analysts get repeated updates to their BR data after closeout and edit the extracted establishment-level data when they find anomalies or mistakes. The edits are in the year t CBP file, and may show up in the prior-year variables in the year t+1 BR, but they are not in the year t BR close out files. Second, firms that are late in filing their IRS payroll taxes for year t ("late filers") sometimes do not show up in the BR until year t+1. In these cases, the prior year employment and payroll variables in the year t+1 BR file may have different values than the current-year employment and payroll variables in the year t BR file. For example, the year t file might have only two quarters of payroll data, while the year t+1 file has 4 quarters of prior year payroll data. In the case of *new* establishments (births) that are also late filers, the record might not show up at all in year t, but shows up in year t+1 with prior year data.

To address these discrepancies in the data we use the following hierarchy to decide which file to use as input data for a given establishment. If a record (for a given establishment ID) exists in year t and t+1 BR files and in the year t CBP file, and prior year employment is not missing in year t+1, then we use that value for year t employment in the LBD. If prior-year employment for that establishment is missing in the year t+1 file but present in the year t CBP and BR files, then we use the value from the year t CBP file for year t employment in the LBD. We follow similar logic for other combinations of input data sources—e.g., the establishment id exists in year t and t+1 BR files, but not in the CBP file. We use non-missing year t+1 prior-year employment when available; if not, we use non-missing year t CBP current-year employment; finally if neither year t+1 BR or year t CBP employment are available, we use year t BR employment as the LBD value. We follow the same logic for choosing which value of annual payroll to use in the LBD.

The integrated LBD-BITS code also has logic to handle the special case where an establishment identifier appears in year t+1 with positive prior-year payroll but the establishment ID appears in neither the year t BR file nor the year t CBP file. We have separate logic for reference years 1976-2001 and 2002-2016, since the establishment-level numeric identifiers differ pre- and post-BR redesign. For 1976-2001, we first attempt to match the year t+1 record on EIN to a list of unduplicated EINs from the merged year t BR and CBP files. Records that match on EIN are *not* kept as separate year t+1 establishments—the logic here is that the year t+1 record may be

an SU establishment in year t+1 that was an MU establishment in year t or vice versa. Prior to the 2002 BR redesign, the main establishment identifier, CFN, automatically changed when a firm switched from being a SU to an MU or MU to SU, but in most cases the EIN stayed the same. For pre-2002 year t+1 records (with prior-year data) that don't match on either CFN or EIN to year t, we then attempt to match the year t+1 OLDID variable to the year t CFN. Any matches are not carried forward as separate year t records. Finally, for records that failed to match in the previous two steps, we attempt to match year t+1 OLDEI to year t EIN. Any records that fail to match in all three steps are carried forward as year t establishments. For 2002 and going forward, we no longer have the CFN, OLDID, or OLDEI variables. We have only the establishment-level identifier, empunit_id and the EIN. For these years, after attempting to match year t+1 records to year t on empunit_id, we attempt to match on EIN. Any year t+1 records with prior-year data that fail to match to year t on both empunit_id and EIN are carried forward as year t establishments.

3. Linking across years in the integrated LBD-BITS process

After all the different sources of data for each year have been merged, as described in Section 2, the next step is to link pairs of years to each other. For each year pair, we will refer to the earlier year as year 1 and the later year as year 2. The goal of this linking process is to determine which establishments are potential births, i.e. they appear in year 2 but not year 1, and which establishments are potential deaths, i.e. they appear in year 1 but not year 2. At this phase we cannot fully determine births since re-activations are also a possibility. An establishment that appears in year 2 but not in year 1 could have been active in a year prior to year 1. Likewise, an establishment that appears in year 1 but not in year 2 could become active again in a later year and consequently not be a true death in year 2. Hence, in this first phase of matching we will only determine potential births and deaths and then will reconcile across years in the next step of the integrated BITS-LBD process.

The year-pair matching consists of two main parts: ID matching and name and address matching. ID matching makes use of the main establishment identifier, either cfn or empunit_id, and also historical identifiers such as PPN, OLDID, and PY_ID. In addition, it looks for matches between single-units and multi-units using EIN and OLDEI. The name and address matching uses the Business Register name1 and name2 fields as well as physical and mailing address to match.

Section 3.1.1 Establishment ID Matching

The first step is to attempt to match all establishments either by CFN (1976 – 2001) or EMPUNIT_ID (2002-present). These are the main Business Register establishment identifiers and we make the assumption that if establishments match between year 1 and year 2 using either CFN or EMPUNIT_ID, then in fact they are the same establishment.[12] This assumption

[12] When year 1=2001 and year 2=2002, we use a crosswalk that maps all CFNs in the SSEL to all empunit-ids in the redesigned BR.

means that we take as given any edits/decisions by the Business Register staff about how to assign establishment identifiers.[13]

Prior to 2002, there are several other identifiers available for matching. For the year pairs 1982-1983 through 2000-2001, we match establishments by PPN (permanent plant number), an identifier created by BR analysts to attempt to track establishments over time. For the year pairs 1976-1977 to 2000-2001, we match an identifier called OLDID on the year 2 file to the CFN on the year 1 file. When populated, this OLDID field contains a prior (i.e. historical) CFN that was changed for some reason. Matching it to CFN from the year 1 file has the potential to fix broken links due to CFN changes caused by business re-organizations. For year pairs from 1978-1979 to 1983-1984, we do a similar merge using the PS_ID field.

Section 3.1.2 Single-unit to Multi-unit and Multi-unit to Single-unit matching by EIN

After all the establishment identifier matching has been completed, we next turn to the case of establishments that switch from being single-units to being part of a multi-unit enterprise, and vice versa. These cases are particularly problematic prior to 2002, when the CFN for single-units was created using the EIN but the CFN for multi-units was an identifier assigned by the BR staff. Thus any switch from single-unit to multi-unit status (or vice versa) automatically caused a change in the establishment's primary identifier and hence a break in the linking of that establishment over time. Even after 2002, when the EMPUNIT_ID was assigned in a similar manner for establishments regardless of whether they were single units or belonged to multi-unit enterprises, breaks still happen when a single establishment splits into multiple places of operation.

To solve this problem of identifier changes due to multi/single unit status changes, we make use of the EIN and establishment geography data. For single-units from year 1 that convert to multi-units in year 2, we use the EIN to match to all the year 2 establishments that are part of the new multi-unit enterprise. The next step is to determine which of these year 2 establishments is the continuation of the original single-unit establishment from year 1. We first compare the five digit zip code for the year 1 establishment to the 5 digit zip code for each of the year 2 establishments, and if we find a unique pair, we declare that pair to be a match and link these establishments. We next compare the street address, followed by name, and then county FIPS code, and repeat the process we used for zip code. Any time we find a unique, exact match between the information from the year 1 single-unit record and one of the year 2 multi-unit establishment records, we link the two establishments. After all these matches have been completed, we check for cases where the year 1 single-unit matched by EIN to only one year 2 establishment. These are cases where the new establishments that form the multi-unit enterprise have a different EIN from the original establishment. In this case we can

[13] There are times when an establishment changes location but the BR staff keep the same identifier, for reasons that are not always clear. However, we do not attempt to change any assigned identifiers due to address or any other changes.

link the year 1 single-unit to the correct year 2 establishment by default since there is only one year 2 establishment with the same EIN.

This process is repeated for establishments that are originally part of multi-unit enterprises in year 1 but become single-units in year 2. We use the EIN to match the full set of establishments from year 1 to the single-unit establishment in year 2 and then use zip code, street address, name, and county FIPS to identify unique matches, which we then link. For both types of status transitions, if we do not find any unique matches, we form no links between year 1 and year 2 establishments. This is not an entirely satisfactory outcome since we believe a single-unit establishment that becomes part of a multi-unit enterprise (or vice versa) most often will continue to exist and should link to another establishment in the alternate year. Indeed, we see some evidence of this existence due to the EIN link. More research is needed to determine whether there are additional methods that could be used to match these problematic cases. At the moment, in cases where we see an EIN match across years for a group of establishments but cannot link specific establishments to each other within the group, the year 2 establishments will appear to be establishment births and the year 1 establishments will appear as establishment deaths.

Section 3.2.1 Process Description for Name and Address Matching

As was the case for previous vintages of the LBD, name and address matching is only done for single-units. The first step is to identify candidate establishments. We begin by identifying all the establishments in year 1 that did not match using an identifier to an establishment in year 2. These appear to be "deaths" but could 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 year 2 for which there was no establishment identifier match in year 1. These appear to be "births" but alternatively could be the other half of a re-organization event.

There are two possible types of re-organizations. The first type is one that crosses the year boundary. This happens when a new establishment appears in year 2 that is in fact the same as an existing establishment in year 1 but with a different establishment identifier. We identify these types of re-organizations by matching the potential deaths from the year 1 file to the potential births from the year 2 file.

The second type is a mid-year re-organization. This type of event can happen in either year 1 or year 2. A mid-year 1 re-organization occurs when a new establishment appears in year 1 that is in fact the same as another existing establishment in year 1. For these mid-year re-organizations, we require the new establishment to continue into year 2 (i.e. link by establishment id to year 2) and the existing establishment from year 1 to die in year 2 (i.e. not link by establishment id). We identify potential year 1 continuing establishments by looking for establishments that had no first quarter payroll in year 1 and then matched by establishment id to year 2. These are potentially births in year 1 and they continue into year 2.

A mid-year 2 re-organization occurs when a new establishment appears in year 2 that is in fact the same as another existing establishment in year 2. We require that the new establishment to be a birth in year 2 (i.e. not link by establishment id to year 1) and we require the continuing establishment to have linked by establishment id back to year 1 and to have fourth quarter payroll in year 2 and first quarter payroll in the year following year 2 (i.e. year 3) to be zero. Thus the continuing establishment appears to die but it is really re-organized into the new establishment which was born in year 2.

Currently there are two separate name and addressing matching processes that are run concurrently. After an initial match by company name to identify records that match exactly, we take the residual year 1 and year 2 non-matches and feed them into further 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 plan to reconcile the output into one set of final matches.

Name and address matching is done by year pairs but this has the potential to introduce discrepancies over time. Establishments that are labeled as re-organizations during one year pair, could have a contradictory status in the next year pair. Thus after all the year pairs are created, there is a reconciliation process to compare the year pairs and make decisions about which re-organizations to keep and which to drop. This process is described in more detail in Section 3.3.

Section 3.2.2 Details of BITS Name and Address Matching Process

The BITS name and address matching process relies on three different versions of the business name. The first version is the first 28 characters of the name as it appears on the Business Register. We will call this version the "exact name." The second version, which we will refer to as the "pseudo name," makes relatively few changes to the original name fields. It replaces all non-alphabetic and non-numeric characters with spaces, replaces common words with spaces, and concatenates the remaining words into one string to remove any blanks, keeping 28 characters in total. The third version, which we will refer to as the "standardized name," replaces non-alphabetic and non-numeric characters with spaces, deletes name2 if it begins with "%" or "ATTN," replaces common words with specified abbreviations, replaces other common words (i.e. "and," "company") and one-character strings with spaces, and abbreviates city if it appears in the name. After all these edits are made, the remaining characters are saved to a 12 character string field with no spaces. If the final string is less than 5 characters, the standardized name 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 Business Register. 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.

Once the three versions of name and the simplified version of address have been created, 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 year 1 "deaths" to name1 year 2 "births," followed by matching name1 to name2, and name2 to name1 for this same group of establishments. Next, for mid-year 1 re-organizations, we match name1 from year 1 "deaths" to name1 from year 1 estabs that are missing quarter 1 payroll and that continue into year 2, again followed by matching name1 to name2 and name2 to name1 for the same group. Finally, for mid-year 2 reorganizations, we match name1 from year 2 "births" to name1 from a year 2 continuing establishment that has missing quarter 4 payroll, after which we again match by name1 to name2 and name2 to name1.

Once we have completed the exact name matching for all three types of re-organizations, we then move to matching by pseudo names and repeat the same process as above. In order, we match year to year re-organizations, mid-year 1 re-organizations, and mid-year 2 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, only after all other forms of matching have been exhausted.

Section 3.2.3 Details of LBD Name and Address Matching Process

Currently 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 3.2.1, we feed the residual year 1 and year 2 non-matches 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.[14] We also remove common words such as "the," "of," "company," or

[14] See http://support.sas.com/documentation/cdl/en/dqclref/63171/HTML/default/viewer.htm#p0k705exnmtpgin1xppkpx5f7r30.htm for documentation.

"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 dqmatch again with the database ENUSA. We create street, name1, and name2 at sensitivity levels of 50, 55, 65. We also create versions of compressed versions of street, name1, and name2 that have spaces and special characters dropped and a version of name that is name2 appended to name1. We create fuzzy versions of this concatenated name at sensitivity levels of 50 and 65. We also create fuzzy versions of name1 and name2 that are not standardized at a sensitivity level of 95. Finally, we create a fuzzy value of street at level 70 and we standardize the city name and create a fuzzy version at level 70.

We run 34 match passes where we use the various versions of name, street, and part or all of the zip code to identify matches. We directly match four of these variables in a pass across establishments and identify matches where the fuzzy variables agree. For example, the first pass matches the compressed version of the standardized name1, name2, and street fields without any special characters and the first three digits of the zip code. Establishments that do not match in any given pass are moved forward to the next pass. We end the process with a final (35^{th}) match pass that first creates a set of potential matches based on records that link on state, standardized street at sensitivity level 70, and standardized city at sensitivity level 70. Within this group of potential matches, we count the number of common words between each pair of establishments and keep matches with at least 2 common words that are unique within the group of potential matches. In other words, establishment A will be declared a match to establishment B if both are in the potential match block (defined by state, street70, and city70) they have at least 2 words in common, and no other establishments also have 2 of those words in common with establishment A or establishment B.

Section 3.2.4 Reconciliation of BITS and LBD matches

Since the BITS and LBD matching processes run simultaneously after the exact name matching step, they have the potential to produce different establishment matches between the same pair of years. 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. Our initial results show that the most common disagreement is that the LBD linking finds a match while BITS does not. The next most common outcome is that the BITS process finds a match while the LBD does not. Only in a few cases are there conflicting matches. We are currently working on a system for deciding which matches are high enough quality to keep and how to reconcile the direct disagreements.

Section 3.3 Reconciliation across year-pairs

Once the year pair files have all been created, the next step is to create a wide file that links together all the Business Register IDs over time that belong to the same establishment. Thus the file is set up to have the fields id1_1976, id2_1976, id1_1977, id2_1977, …, id1_{finalyear}, id2_{finalyear} for each record. The file is created by combining year pair files step-wise, beginning with 1976-1977 and 1977-1978. The resulting file is lbd_bits_1976_1978. Then lbd_bits_1976_1978 to merged with the 1978-1979 year pair file and so on until the final lbd_bits_1976_{finalyear} file is complete.

Merging year pair files requires reconciling differences in decisions made about re-organizations across years. For example, if an establishment was part of a mid-year 2 reorganization because it had no 4th quarter payroll in year 2, it might nonetheless have positive payroll again in year 1 of the next year pair file, returning to activity in quarter 2 of what is essentially year 3, for example. In this case, we drop the mid-year 2 reorganization link and allow the establishment to link across year 2 and year 3. The other establishment that was part of this mid-year 2 re-organization is also unlinked and labeled as a birth in year 2. This is an example of how the wide link file can be changed retroactively when a new year pair file is merged on.

It is also possible that information from the wide link file can change decisions that were made in the most recent year pair file creation. For example, two establishments might be linked together as a mid-year 1 re-organization but then when this pair is matched to wide link file, we discover that the establishment with missing quarter 1 payroll that looked like a potential birth in year 1 was in fact a continuation of an establishment that had existed in year 2 of the prior year pair file. For some reason, it simply had no first quarter payroll in year 1 of the new year pair file. In this case we link this establishment to the record on the wide link file and dissolve the year 1 re-organization link. The other establishment in this potential re-organization is then labeled as a death.

In simpler cases, there are times when something that looks like a birth in year 2 of the year pair file turns out to be a reactivation of an establishment that was active in the year prior to year 1. Likewise, an establishment that looks like a death in year 1 can be found active again in year 1 of the next year pair. These also cause changes to either the year pair file or to the underlying wide link file. We are currently investigating how long in the past we should search for a re-activation. Current BITS processing only searches one year in the past. Current LBD processing searches 7 years in the past.

These examples demonstrate why the LBD-BITS file will change over time as new year pairs are added and why the links from the last year will always be the most uncertain. The wide link file essentially acts as the repository of all the linking decisions that have been made up to a certain point in time. When another new year of data becomes available, it is first linked to the prior year and then this year pair is compared to the wide link file to resolve any differences. In this manner we incorporate another year's worth of information into the link repository.

We deliberately only keep IDs on this file in order to prevent it from becoming overwhelmingly large. The final important function of this file is that it enables us to create a unique longitudinal ID that will track an establishment over time in spite of any changes to the BR establishment ID caused by re-organizations of some kind.

Section 4 Birth – Death Retiming

To determine and update the structure of firms and where they are doing business, the Census Bureau relies on annual surveys such as the Company Organization Survey and the Annual Survey of Manufacturers and the quinquennial Economic Census. In years that end in "2" or "7" most establishments are mailed a survey 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. Hence we can accurately time establishment births 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 a census form in 2002 and then in 2003 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 sampled by the COS until 2006, we will not know the exact birth year of any 2006 newly reported establishment. Likewise if the single unit is not surveyed until the census in 2007. Since only a small number of single-unit establishments are surveyed by the COS/ASM each year, and only beginning in 2005, there are large spikes in establishment births in Census years. Even single units that are surveyed still cannot be accurately dated because it is almost always the case that they are only surveyed once in intercensal years. An important innovation of the LBD over the CBP and BITS is that the LBD attempts to use available information to retime some establishment births and deaths that first show up in census years.

There are two types of establishment births. The first type of birth happens when a new single-unit establishment begins to operate and files 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 other 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, we know the exact year that the establishment was born.

The second type of establishment birth happens when a new establishment is born within an existing firm. The establishment may be the second establishment to begin operations within that firm, effectively changing the firm from a single-unit to a multi-unit, or it may be an additional establishment in a firm that already operates in multiple locations. These types of births are much harder for the Business Register staff to identify because they will not show up on tax records. Since multi-unit firms most often file taxes for multiple establishments on a single form, there is no record of a new place of business, just an increase in the number of employees.

This difficulty in knowing the actual first year of operation for establishments born within multi-unit firms causes difficulty in measuring and comparing the number of births across years. To deal with this challenge, we divide multi-unit births into those where we know the actual year of the birth with some degree of confidence because the firm responded to a survey or census at least two years in a row and those where we do not know the actual year of birth because of gaps in survey coverage. We then build an imputation model using the known births as our training data and predict birth years for the uncertain establishments. The major short-coming of this approach is that the training data consists almost entirely of establishments born to large multi-unit firms because these are the firms that are repeatedly surveyed. However, the set of establishments with missing data comes almost entirely from small multi-unit or single-unit firms. While we recognize that this is not ideal, the data available to us leave us little choice and we feel it is better to make some attempt to re-time the births so that Census years do not have inordinately large spikes in births rather than leave the data without further edits.

We are in the process of revising and improving the existing imputation model but intend to use many of the same predictor variables. We will first stratify the sample on industry and time period, splitting the training data into segments of births that occurred within five-year windows bounded on either side by an economic census. Within these industry/time period strata, we will use regression analysis to correlate various measures of employment growth and geography with birth year and then use these correlations to predict the missing data for the uncertain births. After modeling year of birth, we will repeat the process for year of death of establishments that have ceased operating according to reports in census years.

Our current plans are to re-time both births and deaths that occur in Census years but not to re-time those that begin or end in intercensal years but have an uncertain birth/death date because they were not sampled every year by the COS/ASM. If resources permit, we may re-visit this issue and attempt to re-time all uncertain births and deaths.

The results from birth/death retiming are used to modify the wide link file that was described in Section 3.3. Establishments that have their birth year set to an earlier point in time than the originally reported census year have their identifiers pushed to earlier year fields in the wide link file and new records are created in the characteristics files that contain an imputed employment value for these establishments in years they were imputed to be active prior to the census year.

Section 5 Assignment of consistent NAICS codes over time

Industry codes change over time for two primary reasons. First, an establishment may change its line of business and second, industry coding standards change over time, most notably switching from SIC to NAICS between 1997 and 2002. A final part of the LBD process attempts to create consistent industry codes across time for all establishments. This involves creating a crosswalk between the various vintages of SIC and NAICS coding so that a 2012 NAICS code can be assigned to every establishment in the wide link file, regardless of when that establishment

operated. Documenting this process is beyond the scope of this paper at this point. Details of the current research version of the NAICS time series crosswalk can be found in Fort and Klimek (2016).[15]

Section 6 Creation of final LBD-BITS year files with characteristics

The final step in the creation of the LBD-BITS data is to create annual files that contain the longitudinal establishment identifier, the firm identifier, and characteristics of the establishment, in particular industry, employment, payroll, and geography. These files are created by merging the wide link file, by year, with the characteristics files created at the beginning of the process and modified as part of birth/death re-timing. The LBD-BITS annual data are used by internal researchers at CES and by external researchers with approved FSRDC projects.

Section 7 Creation of BDS

The BDS is created by merging four consecutive years of LBD-BITS annual data and creating measures of change across years. Establishments and firms are then classified as births, deaths, or continuers and have positive or negative job creation. Within these categories, we sum the counts of firms and establishments within various geographies, industries, and firm size and firm age classifications and publish both the counts and the associated total employment for every cell. A detailed description of this methodology and the current version of the published tables can be found at https://www.census.gov/ces/dataproducts/bds/index.html.

During the creation of the BDS, some additional editing of first quarter employment is done to smooth what look like spurious large changes in employment across years. We also implement an algorithm to identify large single-unit births that are spurious. These happen when a new EIN record is created in the BR for various administrative reasons but the EIN does not actually represent an operating firm. If these spurious births are not dropped, they can grossly inflate the employment growth in smaller geographic areas.

Conclusion

The process of transitioning the LBD into the Census Bureau's production environment has provided a unique opportunity to improve the existing BDS product and integrate with another existing Census Bureau product, the SUSB. When completed, the new system should support not only the annual creation of the combined statistics but also the development of additional statistics of interest to researchers such as firm-level human capital measures, patent-holding information, and exporting/importing status.

References.

Atrostic, B.K., Randy Becker, Todd Gardner, Cheryl Grim, and Mark Mildorf. 2009. "Recovery of

[15] Available at http://faculty.tuck.dartmouth.edu/images/uploads/faculty/teresa-fort/fort_klimek_naics.pdf.

Historical U.S. Census Bureau Microdata: Success to Date" Chapter 4 in 2009 CES Annual Report.

Davis, Steve, John Haltiwanger, Ron Jarmin and Javier Miranda. 2007. "Volatility and Dispersion in Business Growth Rates: Publicly Traded vs. Privately Held Firms." *NBER Macroeconomics Annual* 2006, vol. 21.

Davis, Steve, John Haltiwanger, and Scott Schuh. 1996. *Job Creation and Destruction.* MIT Press.

DeSalvo, Bethany, Frank F. Limehouse, and Shawn D. Klimek. 2016. "Documenting the Business Register and Related Economic Business Data" Center for Economc Studies Working Paper CES-WP-16-17.

Dunne, Tim, Mark J. Roberts and Larry Samuelson. 1988. "Patterns of Firm Entry and Exit in U.S. Manufacturing Industries" *The RAND Journal of Economics* 19(4):495-515.

Dwork, Cynthia. 2006. "Differential Privacy." *Proceedings of the 33rd international conference on Automata, Languages and Programming - Volume Part II, pp. 1-12.*

Fort, Teresa and Shawn Klimek. 2016. "The Effects of Industry Classification Changes on U.S. Employment Composition." Working Paper.

Haltiwanger, John, Ron Jarmin, and Javier Miranda. 2013. "Who Creates Jobs? Small versus Large versus Young." *The Review of Economics and Statistics* 95(2):347-361.

Jarmin, Ron. 2002. "LBD Documentation: Final Linking and Post-Matching Linkage Edits on the Longitudinal Business Database." *Center for Economic Studies Technical Note CES-TN-2002-03.*

Jarmin, Ron, Shawn Klimek, and Javier Miranda. 2005. "The Role of Retail Chains: National, Regional, and Industry Results" *Center for Economic Studies Working Paper CES-WP-05-30.*

Jarmin, Ron and Javier Miranda. 2002. "The Longitudinal Business Database." *Center for Economic Studies Working Paper CES-WP-02-17.*

Miranda, Javier. 2002. "LBD Documentation: Defining Active Establishments and Other Data Issues." *Center for Economic Studies Technical Note CES-TN-2002-04.*

White, T. Kirk. 2014. "Recovering the Item-level Edit and Imputation Flags in the 1977-1997 Censuses of Manufactures." *Center for Economic Working Paper CES-WP-14-37.*