The Longitudinal Business Database

July 16, 2002

Ron S. Jarmin[1]
Center for Economic Studies, U.S. Census Bureau
rjarmin@ces.census.gov

And

Javier Miranda
American University and
Center for Economic Studies, U.S. Census Bureau
jmiranda@ces.census.gov

[1] Disclaimer: This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review by the Census Bureau than its official publications. This report is released to inform interested parties and to encourage discussion. Any findings, conclusions or opinions are those of the authors. They do not necessarily reflect those of the Center for Economic Studies or the U.S. Census Bureau.

We would like to thank Cathy Buffington, Lucia Foster, Shawn Klimek, C.J. Krizan, James Monahan and Al Nucci, each of whom provided crucial assistance and guidance on this project. We also wish to thank Randy Becker, John Haltiwanger, Paul Hanczaryk, Brad Jensen and Richard Moore and seminar participants at Census, CAED 2001 and BLS for helpful comments.

I. Introduction

As the largest federal statistical agency and primary collector of data on businesses, households and individuals, the Census Bureau each year conducts numerous surveys intended to provide statistics on a wide range of topics about the population and economy of the United States. The Census Bureau's decennial population and quinquennial economic censuses are unique, providing information on all U.S. households and business establishments, respectively.

The censuses and most surveys are static snapshots of the populations they are intended to describe. Researchers and statisticians, however, have long known the value of longitudinal data that follow the same survey units over time (see McGuckin and Pascoe, 1988). There are two ways to construct such datasets. First, statistical agencies can explicitly design longitudinal surveys (examples include the SIPP, NLSY and PSID). Alternatively, records from successive survey years can be linked together. For the vast majority of the statistical programs at the Census Bureau, the latter is the only alternative.

The Center for Economic Studies (CES) at the Census Bureau has a mandate to construct, maintain and conduct research with longitudinal datasets. CES maintains these and other micro datasets for use by economists, and other social scientists. These data sets primarily contain information received from respondents to Census Bureau censuses and surveys, and may contain data on individuals, households or businesses. Traditionally, however, CES has focused on business data, mostly from the manufacturing sector.

The first longitudinal dataset created at CES was the Longitudinal Research Database (LRD)[2]. The LRD contains longitudinally linked plant level data from the Censuses and Annual Surveys of Manufactures. Over the years, a large and successful research program has been carried out using the LRD (for reviews see Bartelsman and Doms 2000; Caves 1998).

This paper describes recent efforts at CES to create a new longitudinal research dataset: the Longitudinal Business Database (LBD). The LBD is a major improvement over existing longitudinal establishment datasets. Unlike the LRD, which covers only manufacturing, the LBD covers nearly all the non-farm private economy, as well as some public sector activities. Also, research using the LRD found problems with broken longitudinal linkages that lead to spurious establishment births and deaths (Dunne 1992). We supplemented the longitudinal numeric identifiers assigned by the Census Bureau with name and address matching to repair broken linkages. Other economy wide longitudinal files, such as the Bureau of Labor Statistics' Longitudinal Database (LDB) (Spletzer 1997; Pivetz, Searson and Spletzer 2001) and the Small Business

[2] The LRD actually evolved from the Longitudinal Establishment Database (LED) that was constructed in the early 1980's. Attempts to longitudinally link plant level data from the Annual Survey of Manufactures were initiated as early as the late 1950's (Monahan 1992).

Administration sponsored Business Information Tracking Series (BITS) at the Census Bureau (Robb 1999) only extend back into the early 1990's. The LBD, by contrast, contains data back to 1975.

There has been talk of creating the LBD for several years (Nucci 1993), but due to various constraints, serious work was delayed. The LBD provides longitudinally linked data for all employer (i.e., those with paid employees) establishments contained in the Census Bureau's business register, the Standard Statistical Establishment List (SSEL). Currently, the core linkage files of the LBD have been constructed and work continues to add additional components that will contain basic data items, such as payroll, employment, location, industrial activity and firm affiliation.

The LBD will be invaluable to researchers examining entry and exit, gross job flows and changes in the structure of the U.S. economy. The LBD will also be useful to aid the Census Bureau in examining how its census and survey programs describe the U.S. economy. While the LBD is useful as a stand-alone research dataset, we intend it to be used in conjunction with other Census Bureau establishment and firm level micro data. The LBD describes how establishment units in various Census Bureau censuses and surveys are linked over time. Identifiers on the LBD facilitate linking to other datasets.

In this paper, we will focus on describing the longitudinal linkages that are the core of the LBD. We first discuss the source data we used. Next we describe how we linked establishment records over time. We then discuss some of the features of the LBD and how we think it can be useful to the Census Bureau and the research community.

II. Source Data for the LBD

Various components of the LBD will contain information from a variety of sources including the Business Register (or SSEL), Economic Censuses and surveys. To construct the longitudinal linkages, however, we relied on only one source: the business register.

Since 1972, the Census Bureau has maintained a general-purpose business register for use by the Federal Statistical System. The business register is a database of U.S. business establishments and companies. In 1968 the Office of Management and Budget, which oversees all Federal Executive Branch statistical activities, directed the Census Bureau to develop and maintain the Standard Statistical Establishment List (SSEL) on behalf of all federal statistical agencies. The SSEL is authorized under Title 13, U.S.C. (for more see U.S. Bureau of the Census, 1979).

The SSEL is used as the frame for Census Bureau firm and establishment surveys. It is also the source data for employment and payroll data summarized

by industry and geographic area in the County Business Patterns (CBP) Program. Longitudinal business demographics and summary statistics derived from the SSEL have been used by multiple government agencies and private organizations.[3] The SSEL is continuously updated with administrative data from other federal agencies, as well as data collected by the Census Bureau. It has undergone significant changes and enhancements over time. For example, the SSEL has seen two major revisions in industry coding: the 1987 SIC revision, and the adoption of NAICS in 1997. The Census Bureau is currently undertaking a major redesign of the SSEL for the 2002 Economic Census. While this redesign will have important implications for the LBD going forward, we do not discuss it here. Instead, we focus on the characteristics of the SSEL over the period of the current LBD.

Coverage of the SSEL[4]:

The SSEL covers legally operating entities operating in the U.S. and its territories. Entities engaging in illegal or "underground" activities are not covered. The SSEL covers only entities with paid employees. Before 1994, nonemployers that were subject to Federal income tax were included in the SSEL in Economic Census years. Since 1994, these are kept in a separate nonemployer file. Note that the SSEL files stored at CES contain employer entities only.

Industry coverage in the SSEL depends on whether entities are privately or government owned or controlled. For private entities, the SSEL covers all industries except private households. However, entities in industries outside the scope of the Economic Census are not broken into establishment units. The industrial scope of the Economic Census has changed over time[5]. Currently, out of scope industries include: Agriculture, Forestry and Fishing (SIC Division A), railroads (SIC 40), U.S. Postal Service (SIC 43), Certificated Passenger Air Carriers (part of SIC 4512), Elementary and Secondary Schools (SIC 821), Colleges and Universities (SIC 822), Labor Organizations (SIC 863), Political Organizations (SIC 865), Religious Organizations (SIC 866) and Public Administration (SIC Division J). Most government owned or operated entities (SIC Division J) are outside the scope of the Economic Census and their establishments are not broken out on the SSEL (these entities are, however, in scope for the Census of Governments and are included on the Government Integrated Directory). The only exceptions are Wholesale Distributors of Beer, Wine and Distilled Alcoholic Beverages (SIC 518), Liquor Stores (SIC 5912), Central Reserve Depository Institutions (SIC 601), Federal and Federally-sponsored Nondepository Institutions (SIC 611) and Hospitals SIC (806).

Statistical Units on the SSEL

[3] A brief overview can be found in http://www.census.gov/econ/overview/mu0600.html
[4] The section draws heavily on Walker (1997).
[5] Finance, Insurance and Real Estate (FIRE) and Transportation, Communications and Utilities (TCU) were out of scope prior to 1992.

The SSEL is constructed from a variety of data sources including administrative records and Census surveys. As a result the SSEL covers different statistical units. The primary unit on the SSEL is the business establishment. An establishment is a single physical location where business is conducted. The establishment is the economic unit used in the Quinquennial Economic Censuses and in many of the Census Bureau's other business surveys.

However, there are other statistical units in the SSEL that may or may not be equivalent to the business establishment. These are important when constructing the LBD. First, the EIN unit is an administrative construct. The IRS assigns Employer Identification Numbers for tax reporting purposes. An EIN unit is comprised of one or more establishments. Second, the enterprise is an economic unit comprising one or more establishments owned by the same legal entity. That is, the enterprise is the highest-level parent company that controls more than 50% interest in its affiliated establishments.

The relationship between establishment, EIN and enterprise units

By far the most common entity on the SSEL is the *single unit* establishment. This occurs when the parent enterprise does business at only a single physical location. In this case the establishment, EIN and enterprise units will refer to the same single economic entity. The organization of the single unit enterprise facilitates the use of administrative data for statistical purposes. Namely, the enterprise files with the IRS under a single EIN, which in turn represents a single physical location or establishment. In this case, the enterprise (or firm) is identical to the tax reporting entity, which is in turn identical to the establishment, which is the statistical unit for the Economic Census.

The other primary entity on the SSEL is the *multi unit* establishment. These units represent establishments owned by enterprises conducting business in multiple locations. The organization of multi unit enterprises complicates the use of administrative data for statistical purposes.

The Census Bureau wants to provide data that describes the economy in as much industrial and geographic detail as possible. The most logical way to achieve the Census Bureau's goals of providing both industrial and geographical detail is to break the activities of multi unit enterprises up by establishments. This also has the important advantage that it matches the reporting unit for single unit enterprises.

Sources of administrative data, such as the IRS, do not share the Census Bureau's need for establishment level data. Multi unit enterprises report to the IRS under at least one EIN. The Census Bureau must then break the enterprise and its EINs into their constituent establishments. This is done primarily via the Economic Census and the annual Company Organization Survey (COS).

Processing issues with the SSEL

There are a number of processing features of the SSEL that affect how accurately it describes the universe of employer establishments in the U.S. The primary purpose of the SSEL is to provide a mail-out frame for the Economic Census. Understandably, the Census Bureau devotes more resources to the SSEL when it is preparing to do an Economic Census and when it uses the results from the census to update the SSEL. This introduces a 5-year cycle that manifests itself in the data in several ways that users of the LBD should be aware.

First, the quality of the single/multi-unit breakout declines after an Economic Census and then improves again with the next census. Administrative data describe EI units. New EI units will enter as single units. Most of these will, in fact, be single units, but some will not. The Census Bureau often does not learn this until an Economic Census. This results in spikes in the number of new multi-unit establishments in Economic Census years.

Also, the coverage of the COS varies over the SSEL processing cycle and due to budgetary considerations. This effects how well information on multi-unit establishments is updated between Economic Censuses. This impacts both the number of multi-unit establishments, and updates to their data items, such payroll, employment, industrial activity and firm affiliation.

III. Creation of the LBD linkages

The core of the LBD is a set of links that describe how establishments in a particular annual SSEL file relate to those in the preceding year. These links flag establishment records as births, deaths or continuers. We construct these linkages using numeric establishment identifiers along with name, address and other information. Before creating the links, however, we tried to ensure that the SSEL files we used were as complete and accurate as possible.

Preparation of SSEL files

1. Assembling the files

As mentioned above, the SSEL is constantly updated. The SSEL files maintained at CES, however, are annual snap-shots. In recent years, CES obtains files for a given reference year in fall of the following year, once COS processing is complete. These files are obtained by CES from the Economic Statistical Methods and Programming Division (ESMPD), which is responsible for maintaining the SSEL file electronically. For each year, CES data staff creates SAS® datasets for both the multi and single unit files.

Prior to 1995, CES did not directly acquire annual snapshot files from ESMPD. For the years 1974 through 1994, CES obtained annual SSEL datasets from tape backups or County Business Patterns (CBP) micro-data files. This work was done in the mid nineties, by CES researcher Al Nucci, and involved translating the data from a proprietary Census Bureau binary format to ASCII. To our knowledge, the original archived data are no longer available.

Once at CES, the data were stored in compressed ASCII format, in as many as 50 files per year. Due to space limitations, the data were stored and used on various media and computers. To facilitate the use of these files for this and other research projects, CES data staff recently assembled all the data together on a single computer and in SAS® datasets.

There were problems with several of the files that required attention before we could begin matching. For example, there were tapes missing when the 1978 Single Unit data were brought to CES. Therefore, we are missing those establishments whose data was stored on the missing tapes. We are also missing the entire 1988 and part of the 1989 Multi Unit files. For these, we were able to supplement with data from County Business Patterns files. Several other fixes had to be made to other files before we could begin constructing links. These are documented in Miranda (2002a).

2. Defining Active records

Not all records in the SSEL pertain to active establishments. There are several reasons why this is so. SSEL records can represent non-establishment entities or recently closed or otherwise defunct establishments (e.g. for single-units, the Census Bureau keeps records up to 9 quarters after establishments stop reporting data to the IRS). In addition, there are typically records on both the single and multi-unit files that pertain to the same EIN. Before successive years of the SSEL could be matched, we excluded inactive and duplicate records.

The procedures we used to arrive at the final annual SSEL datasets used to construct the LBD are based on Foster (1999)[6]. These are somewhat complicated and would involve discussing Census confidential material. Details on the procedures used to delete inactive and duplicate records are available in Miranda (2002a). The following is a summary of those procedures.

- We drop all records with zero annual payroll.
- We drop all records with flags indicating inactivity or that the record does not pertain to an establishment.[8]

[6] Foster (1999) worked in close consultation with the Economic Planning and Coordination Division, which is responsible for the construction and maintenance of the SSEL.

[8] An exception to this is out of scope EIN level records that represent multi location entities. These occur primarily in public administration.

- We unduplicated records pertaining to the same EIN across both the single and multi unit files. These usually arise when a multi-unit firm reverts back to a single unit. In this case, we keep the MU record as the information is more reliable than the administrative data on the single unit file.

Table 1. SSEL Establishment Counts
By Year and Single / Multi-Unit Status

Year	All Single-Units	All Multi-Units	Total	Active Single-Units	Active Multi-Units	Total Active
74	7,537,208	NA				
75	7,777,714	1,119,935	8,897,649	3,866,226	822,906	4,689,132
76	6,335,307	984,980	7,320,287	4,120,205	854,077	4,974,282
77	6,580,974	1,221,155	7,802,129	4,124,203	1,029,573	5,153,776
78	5,653,200	1,318,752	6,971,952	3,516,614	1,032,373	4,548,987
79	8,193,678	1,429,334	9,623,012	4,311,722	1,051,010	5,362,732
80	8,519,356	1,531,623	10,050,979	4,252,325	1,059,400	5,311,725
81	4,672,851	1,634,420	6,307,271	4,200,275	1,072,311	5,272,586
82	5,227,582	1,543,325	6,770,907	4,154,743	1,165,171	5,319,914
83	4,958,975	1,564,158	6,523,133	4,110,086	1,168,247	5,278,333
84	5,282,458	1,694,606	6,977,064	4,367,954	1,194,318	5,562,272
85	5,738,143	1,279,839	7,017,982	4,734,395	1,169,512	5,903,907
86	5,791,491	1,424,844	7,216,335	4,789,675	1,197,080	5,986,755
87	5,974,577	1,909,018	7,883,595	4,860,589	1,336,888	6,197,477
88	6,058,104	1,353,827	7,411,931	4,918,356	1,327,763	6,246,119
89	6,470,451	1,665,868	8,136,319	4,996,037	1,306,561	6,302,598
90	8,724,368	1,860,663	10,585,031	5,309,701	1,354,733	6,664,434
91	9,061,948	1,634,273	10,696,221	5,302,086	1,351,572	6,653,658
92	9,605,164	1,958,000	11,563,164	5,293,965	1,493,705	6,787,670
93	10,561,452	2,019,632	12,581,084	5,401,127	1,471,281	6,872,408
94	10,733,629	2,231,134	12,964,763	5,498,328	1,498,450	6,996,778
95	11,430,782	2,356,256	13,787,038	5,584,717	1,515,521	7,100,238
96	10,231,525	1,781,654	12,013,179	5,705,858	1,484,490	7,190,348
97	10,656,372	2,150,213	12,806,585	5,728,271	1,604,417	7,332,688
98	11,701,778	2,324,255	14,026,033	5,760,682	1,612,765	7,373,447
99	12,778,708	2,549,522	15,328,230	5,791,404	1,655,470	7,446,874

We do not impose any industry or geographic scope restrictions. The Census Bureau's County Business Patterns program, which also uses the SSEL as its primary data source, makes several restrictions of this nature. Therefore, the LBD universe is larger than the CBP universe and our establishment counts will exceed those in the CBP based BITS/LEEM file.

Table 1 lists the establishment counts for the SSEL. We list both the total records on the CES SSEL files and the number of "active" establishments. The total number of records on the SSEL fluctuates considerably from year to year. However, the number of active establishments is much more stable and trends up over the period.

It is possible to see the effects of missing SSEL data in table 1. This is especially true for 1978. Fortunately, we are able to exploit the longitudinal properties of the data to fill most of these gaps. This will be discussed further below.

Creating the Link Files

Serious work on constructing the LBD began at CES in 1999. Krizan (1999) developed a methodology for linking adjacent years of the SSEL that was based on Trager and Moore (1995). We have further modified the Krizan methodology here (see Miranda 2002b and 2001).

The goal is to construct link, or pointer, files that describe how a record in one year is related to records in the previous year. For each year of the LBD, we want to be able to flag a record as a:

- Birth – a record in the current year that does not match to a record in the preceding year,
- Death – a record in the preceding year that does not match to a record in the current year, or
- Continuer – a record in the current year that matched to a record in the preceding year.[10]

The Census Bureau assigns numeric establishment identifiers that make creating longitudinal linkages a straightforward task for the majority of establishments. However, for several years we have only a subset of these identifiers. In addition, there are problems with the longitudinal characteristics of the Census assigned numeric identifiers for a significant number of cases. These problems necessitate augmenting the numeric matches with character-based matches using information, such as establishment name and address. We used Automatch®, sophisticated commercially available software originally developed at the Census Bureau, to perform statistical record linkage using character based information. Like Trager and Moore (1995) and Krizan (1999), we first matched using numeric identifiers and then submitted the unmatched residual records to a name and address matching procedure using Automatch®.

1. Numeric Matching

The Census Bureau maintains several numeric identifiers that can be used for linking establishments longitudinally. These identifiers have different purposes and characteristics that affect their usefulness for creating longitudinal linkages.

[10] The LBD actually has a much richer set of flags that detail the nature of the linkage across two years. For more information on how the flags were constructed and what information they convey about the linkages, see Jarmin (2002a).

The following is a brief discussion of these characteristics and more detailed explanations can be can be found by referring to the SSEL Glossary and Chapter 3 of the LRD Documentation (U.S. Census Bureau, 1998 and Center for Economic Studies 2002).

- CFN: Census File Number. This is the primary processing ID used by the Census Bureau to identify establishments in Economic Censuses and surveys. The CFN for a given establishment can change over time for several reasons (including ownership changes, changes in single and multi unit status and changes in legal form of organization). Because it can change over time, the CFN has limited usefulness as a longitudinal identifier. The CFN is available over the entire period covered by the LBD.

- PPN: Permanent Plant Number. The PPN was introduced to the SSEL in 1982 to facilitate longitudinal linkages. As it does with the CFN, the Census Bureau assigns each establishment a unique PPN. Unlike the CFN, however, the PPN for a given establishment is designed to remain unchanged as long as the establishment remains active at the same location. The PPN is the best available longitudinal identifier on the SSEL. Unfortunately, it is not available for all years and sectors on the LBD. It was not introduced to the business register until 1982. PPNs exist prior to 1982 only for manufacturing establishments on the LRD. In addition, there was a change in the method by which PPNs were assigned in 1985.

- EIN: Employer Identification Number. This is the establishment's taxpayer ID assigned by the IRS. For single unit establishments, the EIN is a unique identifier and is equal to the CFN (less a leading 0 attached to create the 10-digit CFN from the 9-digit EIN). This is not the case, however, for multi-unit establishments. A multi-unit firm will have at least one EIN. Each EIN that a multi-unit firm reports under will be contained on the Single Unit file and flagged as a "sub master" record.

- NEWID/OLDID: The SSEL tracks one change in the CFN. The OLDID is the predecessor CFN and NEWID is the current CFN.

We use the same numerical linking methodology as Krizan (1999). When they were available, we matched first by PPN, flagged the matches and set aside the residuals. If PPNs were not available, we matched by CFN first. After CFN, we matched by EIN and then finally by OLDID/NEWID.

2. Name and Address Matching.

There are two reasons why we are interested in augmenting the numeric matches with name and address matches. First, the PPN is the only true longitudinal identifier, and it is missing for the early years of the LBD. Second, numeric identifiers, including the PPN, are subject to errors. We want to make

sure we are finding as many valid year-to-year establishment linkages as possible, without picking up erroneous ones. Missing matches inflate the number of establishment births and deaths.

The first problem is pretty straightforward and it's easy to see why we would want an additional matching method when PPNs are not available. In years where PPNs are missing we simply pass the unmatched records from the two adjacent years through the statistical matching algorithm we developed in Automatch®.

Dealing with PPN errors is more complicated and, therefore, requires a little more discussion. Although the PPN for an establishment is not supposed to change as long as business is conducted continuously at a given location, there are reasons why a PPN might change. First, there is the valid reason of splitters and combined reports. In the case of splitters, the Census Bureau may ask an establishment to break out reporting at a single establishment into two or more establishments. Likewise two adjacent establishments may be merged into one combined report. In both cases, no establishment entry or exit has occurred, but the number of records and, hence, PPNs on the SSEL changes. These types of changes are rare, however.

The most common reason why PPN linkages are broken is due to processing errors triggered by changes in establishment status. These are caused by events such as changes in ownership due to mergers and acquisitions, changes in the establishment's legal form of organization, and changes in single/multi-unit status. Trager and Moore (1995) term these "reorganizations." Depending on when the reorganization is coded to the SSEL, there are different ways of dealing with this problem.

The simplest cases to handle are year-to-year reorganizations. These arise when the break in the PPN and other numeric identifiers occurs across two annual versions of the SSEL. That is, continuing establishments in the year t SSEL will not link numerically to records in the year t+1 file. We would record false deaths between year t and t+1. Likewise, continuing establishments in year t+1 would not link numerically to records in the year t file and would be incorrectly identified as births.

The remaining classes of reorganizations are more complicated. Mid-year and birth reorganizations arise when the break in the numeric linkages occurs within an annual SSEL file. In the case of mid-year reorganizations, an establishment continuing from t-1 through t+1 would undergo some change in status during year t that causes a break in the numerical linkages. This change, however, results in duplicate records for the establishment on the year t SSEL. One of these records links numerically to year t-1, but not to year t+1 and the other links numerically to year t+1 and not year t-1. Likewise birth reorganizations occur when a new establishment undergoes a change in status in its initial year. Again there will be duplicate records for the establishment in year t, but only one of

them will link numerically to a record in the subsequent year. Finally, in addition to splitters/combined reports and reorganizations, numeric establishment links can be broken through keying and coding errors.

Name and address information on the SSEL can be used to repair broken numerical linkages. However, these fields are difficult to compare using standard computer matching techniques. Matching fields in SAS® or other commonly used statistical or database software requires an exact match. There are many ways that names and addresses can vary across two lists and still contain the same information. To the human eye a match may be obvious, but to a computer if the fields being matched differ even slightly, no match will be made.

The problem of linking records when exact matching is not possible has received considerable attention by a small cadre of researchers (see Internal Revenue Service 1985). The state of the art in statistical record linkage has advanced such that there is now commercially available statistical matching software. We use Automatch®, a software package based on code originally developed at the Census Bureau.[12] A key feature of the software is its standardization routines. These take the name and address fields that we supply from the SSEL, remove extraneous information and output standardized fields for matching. The software assigns weights to potential matches. Those above a user-defined cutoff are matches.

We did considerable testing before settling on the Automatch® algorithm we employed to repair broken numerical linkages for the LBD. A summary of these tests and a description of the final algorithm are available in Miranda (2002b).

3. Post-Matching Edits and Coping with Extended Periods of Establishment Inactivity

Discussion of our longitudinal linking algorithms so far has focused solely on how active establishment records in a particular annual snap-shot of the SSEL link to active records in the previous annual snap-shot. There are two situations where we need to link records over longer time periods: i) to fill in missing source data, and ii) to allow for extended periods of establishment inactivity.

In the first case, we simply exploit the longitudinal nature of the LBD to infer establishment records in cases where we know we are missing SSEL source data. This occurs for the years 78, 83-86, 89 and 91 (see Miranda 2002a, 2002c). To see how we fill "holes" in the LBD, imagine an active establishment whose record is on one of the missing 1978 SSEL tapes. We simply match 78 deaths (active in 77 but not in 78) to 79 births (active in 79 but in 78). If a match

[12] Automatch® is now available under the trade name Vailty® (see www.vality.com). For details on Automatch® see MatchWare Technologies (1997).

is made, we infer that the establishment was active in 78. We also can do this for missing 78 births. Namely, we look positive prior year payroll on 79 births that were not active in 77. The vast majority of the cases that required post-matching linkage flag edits, were cases where we needed to correct for missing source data. In a smaller number of cases, we needed to fix incorrectly assigned linkage flags. All post-match edits are flagged accordingly to give researchers more information about the nature of the longitudinal links. More details are available in Jarmin (2002a).

In the second case, we consider establishments that undergo periods, in excess of a year, with no payroll for reasons other than missing source data. For example, consider an establishment that was active from 1988 to 1994, then was inactive (no payroll) for two years and then became active once more in 1997. The linking algorithms we have described thus far would classify the establishment as a death in 1995 (active in ’94 but not in ’95) and as a birth in 1997 (active in ‘97 but in ’96) with a new longitudinal identifier. That is, the LBD would contain two records for the same physical establishment.

Following the methodology we have outlined thus far, we constructed a prototype LBD that did not allow establishments to undergo extended periods of inactivity. Jarmin (2002b) matched this first prototype LBD to a longitudinal version of the Annual Survey of Manufactures (ASM) and found that these periods of inactivity are much more prevalent than we had previously thought. Many establishments leave the active SSEL universe for two or more years and re-enter later, often with the same numeric IDs.

There are several explanations for why we might observe this type of behavior in the data. At this point, we have not done a detailed analysis, but a likely explanation is that we are observing businesses that transition back and forth between the employer (positive payroll) and non-employer universes. Other explanations include processing errors and re-tooling.

We have modified our linking algorithms to take account of establishments with extended periods of inactivity. The details of this are given in Jarmin and Miranda (2002). When we compare the revised LBD with the first prototype, we find that accounting for extended periods of establishment inactivity reduces the number of unique establishments we have in the file by over 1.2 Million. This reduces the number of births and deaths in the current version of the LBD by over 40,000 per year as compared to the first prototype. Because we are still not sure of the status of many of these establishments we have separately identified them in the LBD so that researchers can treat them differently if they choose.

4. Matching Results

The current version of the LBD contains data for all employer establishments from 1975 through 1999. We use data from the 2000 SSEL, but the nature of the

algorithms we used to deal with reorganizations means we can't compute links for the last year with available data. The first prototype LBD contained data only up to 1998. Thus, we have already updated the file once and will continue to do so on an annual basis when we receive new SSEL data.

Table 2 summarizes the results from our numeric and name and address matching. It also shows the results from post-matching fixes we implemented to correct for missing data and miscoded linkages and to accommodate temporarily inactive establishments.

The first two columns in table 2 present, by year, the number of active establishments on the archived SSEL files at CES (see table 1), and the number of active establishments on the LBD after all matching and editing was completed. For most years the number LBD establishments is lower than the number of SSEL establishment because we eliminate duplicate records introduced by reorganizations. For the years 78, 83-86, 89 and 91, we use the longitudinal nature of the LBD to correct for missing source data on the SSEL (see Jarmin, 2002a for more details). Thus, the LBD has more records than the (active) SSEL for these years.

The next three columns break out the LBD by continuers, births and deaths. Recall that the linkages on the LBD for a given year describe how establishments link to records in the prior year. Therefore, when interpreting the results in table 2, one should remember the number of continuers in, for example, 1994 is the number that were present in both 1993 and 1994. Likewise the number of births in 1994 refers to establishments that were active in 1994, but not in 1993, and the number of 1994 deaths refers to establishments that were active in 1993 and not in 1994. We list "true" births, deaths and continuers. That is we exclude from deaths those establishments that exit the active employer establishment universe only to re-enter at some later point. Likewise, we exclude from births those establishments that are re-entering after an extended period of inactivity. The numbers of establishments that enter and exit the active employer universe due to extended periods of inactivity are listed separately in columns under the "Temporarily Inactive" header.

Table 2: Match Results

Year	Establishment Counts		Nature of Linkages[13]					
	Active SSEL	Active LBD	Continuing	Births	Deaths	Temporarily Inactive		Name & Address Matches
						Exit	Enter	
76	4,974,282	4,945,528	4,120,965	824,563	580,697	*56,238*		47,538
77	5,153,776	5,125,942	4,244,170	844,422	658,862	*42,496*	*37,350*	47,808
78	4,548,987	5,152,243	4,437,114	683,598	660,576	*28,252*	*31,531*	39,204
79	5,362,732	5,330,266	4,633,798	681,813	480,788	*37,657*	*14,655*	48,625
80	5,311,725	5,283,897	4,632,827	610,991	621,769	*75,670*	*40,079*	45,309
81	5,272,586	5,244,139	4,526,994	649,292	690,877	*66,026*	*67,853*	56,296
82	5,319,914	5,294,765	4,570,950	702,036	579,573	*93,616*	*21,779*	44,832
83	5,278,333	5,586,606	4,711,849	755,528	548,637	*34,279*	*119,229*	31,575
84	5,562,272	5,833,945	5,012,940	779,039	550,405	*23,261*	*41,966*	32,831
85	5,903,907	5,981,692	5,181,467	771,830	618,645	*33,833*	*28,395*	41,532
86	5,986,755	6,098,536	5,305,934	763,103	635,823	*39,935*	*29,499*	41,862
87	6,197,477	6,174,220	5,298,607	851,033	729,996	*69,935*	*24,580*	39,299
88	6,246,119	6,228,218	5,442,626	717,030	613,470	*118,129*	*68,562*	35,206
89	6,302,598	6,388,877	5,482,319	797,117	709,458	*36,443*	*109,441*	36,715
90	6,664,434	6,645,560	5,673,200	933,622	643,935	*71,751*	*38,738*	32,611
91	6,653,658	6,729,082	5,853,353	799,454	745,541	*46,670*	*76,286*	34,016
92	6,787,670	6,759,906	5,939,778	787,850	704,433	*84,872*	*32,289*	43,189
93	6,872,408	6,860,000	6,026,126	746,635	654,895	*78,918*	*87,239*	31,970
94	6,996,778	6,973,457	6,128,377	760,594	661,737	*69,912*	*84,486*	39,509
95	7,100,238	7,077,456	6,245,314	754,795	658,473	*69,695*	*77,347*	39,057
96	7,190,348	7,167,943	6,314,604	766,265	705,050	*57,867*	*87,075*	38,942
97	7,332,688	7,305,127	6,330,006	894,978	790,222	*47,896*	*80,146*	44,380
98	7,373,447	7,351,196	6,511,988	754,708	793,139		*84,799*	20,073
99	7,446,874	7,405,245	6,577,081	828,164	774,793			16,427
Total	147,840,006	148,943,846	129,202,387	18,458,460	15,811,794	*1,283,351*	*1,283,324*	928,806

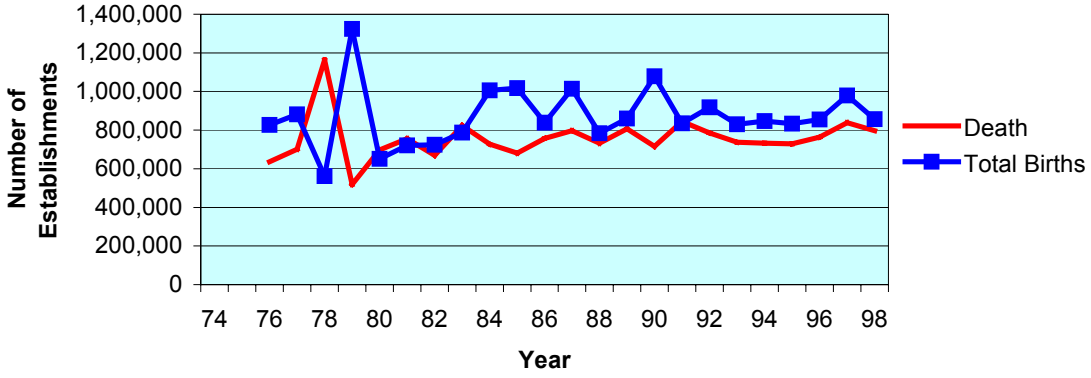
The 1976 through 1999 LBD (which also includes data, but not links for 1975) contains 166,087,537 longitudinal (active LBD + deaths) linkages for 23,259,023 unique establishments. The last column of table 2 lists the number of name and address matched by year. Name and address linkages account for a small portion of the total number of linkages. However, at some point in their tenure in the LBD, approximately 3.68% of the establishments required name and address matching, at least once, in order to preserve their longitudinal linkages.

[13] Note that the categories "births", "deaths" and "continuers" are agglomerations of more refined linkage flags contained in the LBD linkage file. Thus, the information contained in the linkage flags is more extensive than what in presented here. Jarmin (2002a) and Jarmin and Miranda (2002b) provide more details.

A key concern is to separate true births and deaths from spurious ones generated by the processing systems that generate the data underlying the LBD. We have tried to ensure that the data in the LBD are as accurate as possible. We have used both numeric identifiers and name and address matching to try to find all the linkages between records on successive years of the SSEL. We have also used the longitudinal nature of the LBD to fill in for missing SSEL source data and account for establishments that enter and exit the LBD due to periods of inactivity. The impact of this "post-match" processing can be seen in figures 1 and 2. Figure 1 shows the births and deaths obtained after linking the SSEL files, as they exist at CES. The most striking feature of figure 1 is the huge spike in deaths in 1978 (i.e., establishments active in 1977 and not in 1978) followed by a similar spike in births in 1979 (i.e., establishment active in 1979 and not in 1978). This results from the fact that the 1978 SSEL file at CES is incomplete.

Missing source data for 1978 and other years, miscoded LBD linkages and temporary periods of inactivity create problems when computing establishment birth and death rates in the LBD. That is why we go to such care to correct, or at

Figure 1
LBD Establishment Births and Deaths Before Post-Match Fixes

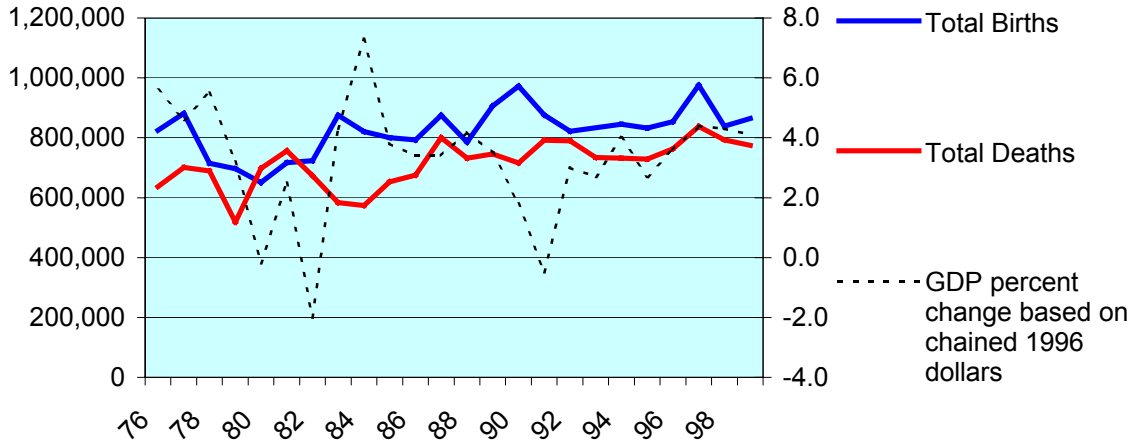


least take account of, these problems. Figure 2 shows the total numbers of establishment births and deaths after we perform our post-matching edits.[15] The figure clearly shows that the post-match fixes smooth the birth and death series' considerably. We juxtaposed the LBD birth and death series' in figure 2 with GDP growth. As expected, net entry (the difference between births and deaths) appears to be pro-cyclical. However, we are still concerned that processing drives some of the results. We are suspicious of the birth spikes in 1987 and

[15] "Total Births" is births plus re-entry of temporarily inactive establishment. Likewise, "Total Deaths" is deaths plus temporary exits. We report totals since the algorithms understate the number of re-entrants towards the end of the LBD since we do not yet know whether establishments that exit in the last few years, might yet re-enter.

1997. The spike in births in 1990 can be attributed to changes in the processing of agricultural establishments.

Figure 2
LBDv2 Establishment Births and Deaths
After Post-Match Fixes



IV. The LBD

The LBD covers over 23 million establishments between 1975 and 1999 and contains limited information on their characteristics and activities. Additional establishment level information is available, but we chose only to include what information was available consistently over time and across different sectors. The limited detail on the LBD is mitigated by the ease with which it can be linked to other establishment and firm data from the Census Bureau and other sources. In this section, we discuss several of the data items we have included on the LBD.

Establishment Age and Tenure

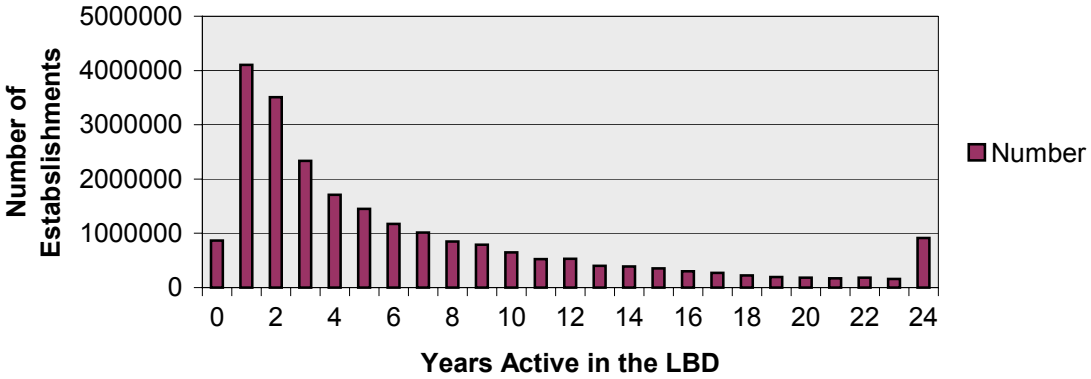
One of the big advantages of the LBD is better establishment age and tenure information. Static datasets based on Economic Census or survey files typically contain no information on establishment age.[17] The LRD can be used to measure the age of manufacturing plants opened after 1963. Researchers using the BITS file have utilized a variable (syr) that indicates the year the record was added to the SSEL (Robb 1999; Acs and Armington 1998; Nucci 1999). However, that variable is not completely reliable for the same reasons that the PPN is not.

[17] The exception is the 1975 And 1981 ASMs where samples of manufacturing plants were asked to list when the plant began operations at the current location.

We can easily compute establishment tenure in the LBD. For all but those establishments active in 1975, we can compute age as well. Establishment tenure is just the difference between the last and first year for which an establishment has non-missing linkages in the LBD. For example, an establishment born in 1978 and dying in 1988 would have tenure of 11 years since it would have valid linkage flags from 1978 (birth) to 1989 (death, recall that the flags describe the nature of the linkage to the prior year). Computing age is similar. However, we can't compute a true age for establishments with left censored data (i.e., those in the LBD in 1975). This problem is most acute in the early years of the LBD and steadily diminishes over time as the 75 cohort thins and is replaced by new entrants. For example, we can compute age for only 17% of the establishments active in 1977 and for 88% in 1998.

Figure 3 shows the tenure profile for establishments in the LBD. Clearly, most establishments are in the LBD for only a short time. Tenure is the same as age for all but the 75 cohort, and, as expected, we see a monotonic decrease in tenure. The blip at tenure equal to 24 years represents establishments active in

Figure 3
LBD Establishment Tenure

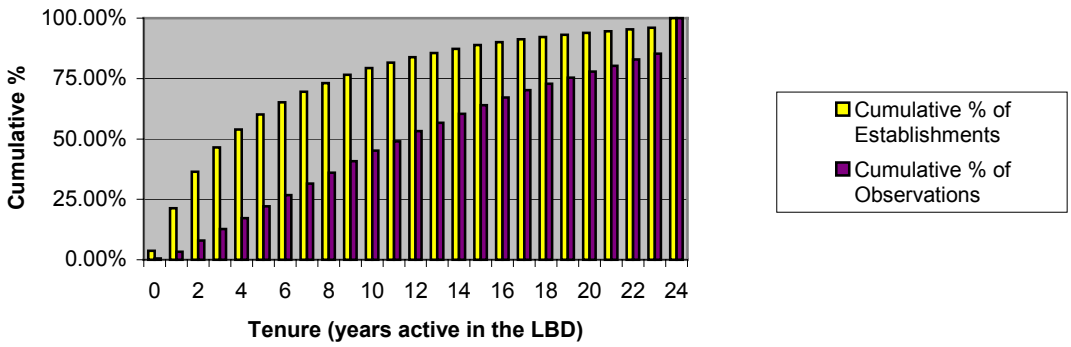


every year of the LBD. Note that establishments with tenure equal to 0 are 1999 entrants and, therefore, have the same initial and terminal year in the current version of the LBD.

We are concerned that data quality may be poor for establishments with few observations in the LBD. This is especially the case for establishments never surveyed by the Census Bureau, and for which all the information we have is derived from administrative sources. While this is clearly a problem when looking at individual establishments, it is less so when looking at the number of plant-year observations in the LBD. Figure 4 shows the cumulative tenure distributions for establishments, and for establishments weighted by the number of years we observe then in the LBD. Whereas 50% of establishments are in the LBD for three years or less, more than 75% of the plant-year observations are contributed

by establishments that are active in the LBD for 5 or more years. These establishments are much more likely to have been canvassed by the Economic Census or some other Census Bureau collection and, therefore, have better industry, geographic, employment and other data. We will be able to test this more precisely once we begin testing the linkages between the LBD and census and survey data. The LBD will be an excellent tool to assessing whether current data collections are of sufficient detail and frequency to accurately measure the economy.

Figure 4
Cumulative Tenure Distributions



Industry and Geographic Classifications on the LBD

The LBD contains industry and geographic classifications similar to those on most Census Bureau establishment data sets. Currently, however, these are limited to contemporaneous codes and they can be missing or inconsistent over time. We are currently working on improving both the geographic and industry data contained in the LBD. One of our biggest challenges is to provide consistent industry coding over the entire history of the LBD. This is complicated by the fact that the period of the LBD spans two major SIC (Standard Industrial Classification) regimes and the switch to NAICS (North American Industrial Classification System). Since these classifications (especially geographic) vary little within establishments over time, we can exploit the longitudinal nature of the LBD to fix missing and inconsistent codes.

Basic Economic Variables in the LBD

In addition to geographic and industry information, the LBD contains basic data on payroll and employment. These provide researchers with basic measures of the scale of economic activity taking place at establishments. In the current prototype LBD, these items come from the SSEL. Once we link the LBD to the Economic Censuses and other surveys, we will edit the payroll and employment data in the LBD to reflect the best available information.

Unlike industry and geography codes, payroll and employment data are much more sensitive to data quality problems, since they vary more within establishments over time. This is particularly important in reference to missing source data and SSEL processing cycles.

To get an idea of the scope and quality of our current employment measures we compare LBD employment with numbers from CBP. Figures 5 and 6 plot employment, by year in major industry groups, and compare CBP numbers (red) with LBD employment numbers (green) and with the LBD using CBP type in-scope restrictions (brown). We see that, especially in agriculture, TCU, FIRE and services, the scope of the LBD is larger than the scope of CBP resulting in higher LBD employment for these sectors. On the whole, however, the LBD tracks CBP employment very well for most sectors once we restrict attention to establishments that are in scope for the CBP.

Firm Affiliation

Currently there is basic firm affiliation information for all establishments in the prototype LBD. This comes from the CFN. We intend to make several improvements and enhancements to the information on firm affiliation in the LBD. We want to track ownership changes and create flags that describe the nature of the linkages between the ownership statuses of establishments over time.

V. Summary

This paper documents recent efforts at the U.S. Census Bureau to construct a new longitudinal establishment data set. Several aspects of the LBD set it apart from existing longitudinal establishment datasets. First, it covers all industrial sectors of the economy. Widely used longitudinal establishment files, such as the LRD, have been limited to the manufacturing sector. However, this sector is declining in importance and it may not be representative of other industries. The LBD will allow research into expanding sectors of the U.S. economy. Second, the LBD covers all employer establishments (with a minimum of one employee). The wide coverage offers possibilities for research into the life cycle of small firms. Third, it contains detailed geographic data allowing research on the dynamics of business location and structure. It will be possible to explore the interactions between local communities and businesses and also amongst types of businesses. Fourth, establishment data go back 25 years and covers multiple business cycles. Other establishment data sets are typically limited to the 1990's. Finally, linking the LBD to other Census Bureau establishment data is relatively trivial making it possible to considerably enhance the scope and depth of information for LBD establishments. The new data has the potential to enhance our understanding of such topics as job creation and destruction, firm turnover, the life cycle of establishments and changes in the structure of the U.S. economy amongst others.

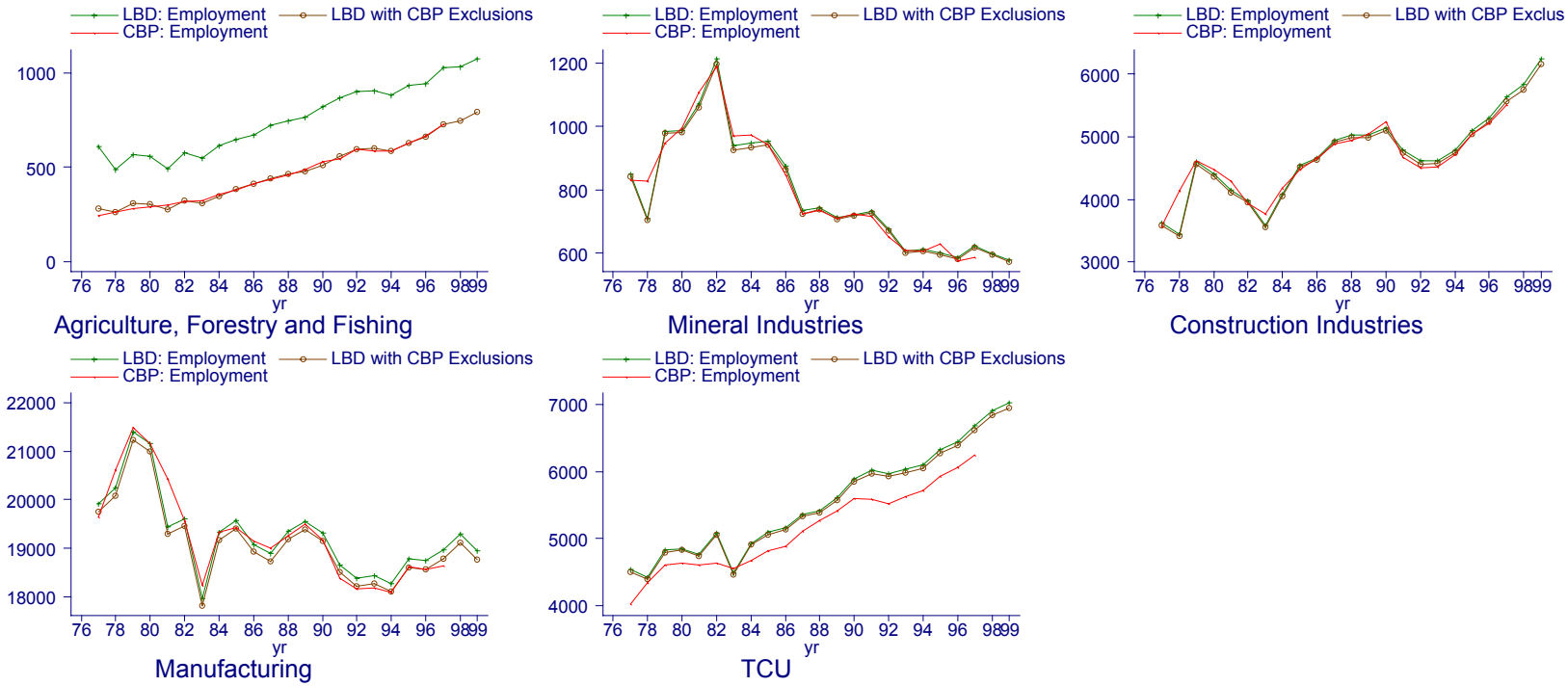


Figure 5 Employment (x1000): by SIC 87 Major Division

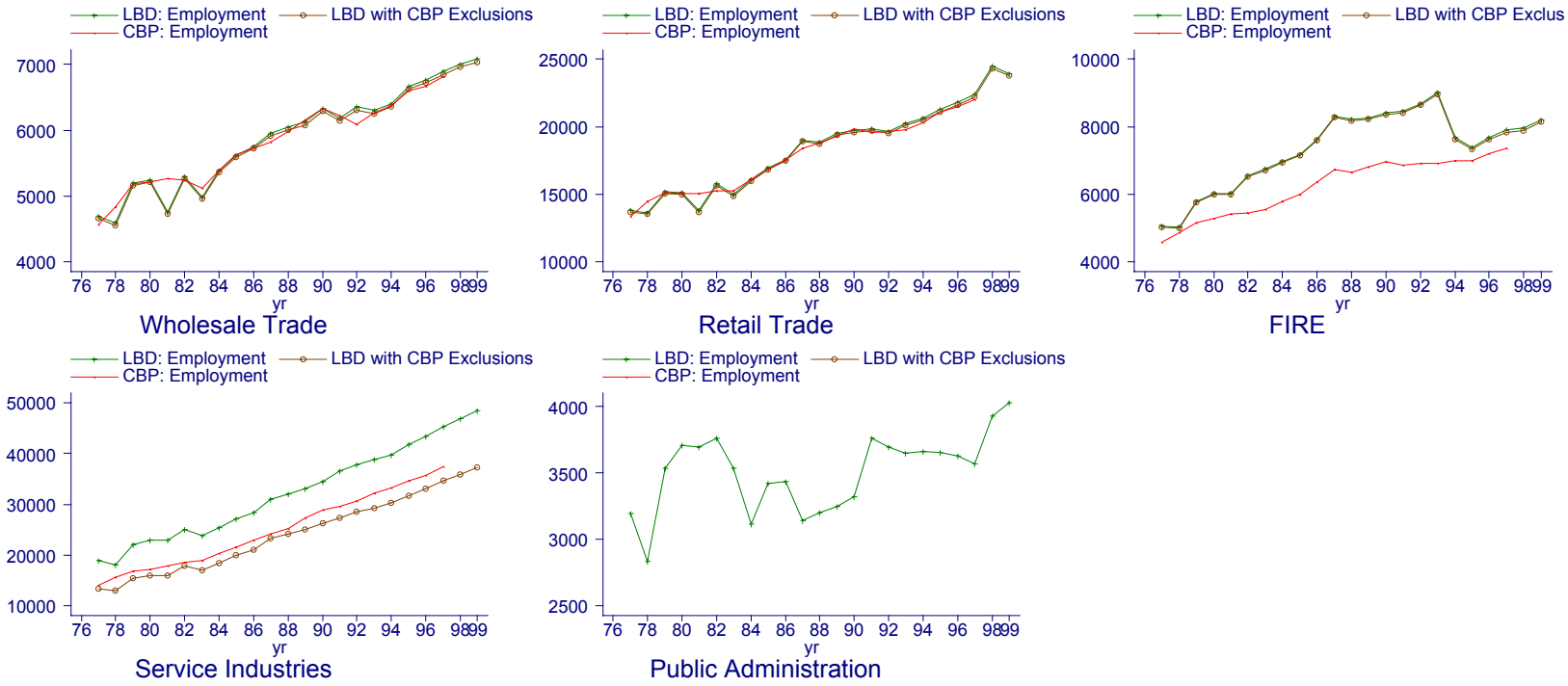


Figure 6: Employment (x1000): By SIC 87 Major Division

References

Acs, Z., and C. Armington, (1998), "Longitudinal Establishment and Enterprise Microdata (LEEM) Documentation," Center for Economic Studies Working Paper CES 98-1.

Bartelsman, E.J, and M. E. Doms (2000), "Understanding Productivity: Lessons from Longitudinal Microdata," Journal of Economic Literature, 38(3), pp. 569-94.

Caves, R. E., (1998), "Industrial Organization and New Findings on the Turnover and Mobility of Firms," Journal of Economic Literature, 36, pp. 1947-82.

Center for Economic Studies, (2002), "Establishment and Firm Definitions," in Longitudinal Research Database Documentation Manual, Ch3.

Dunne, T., (1992), "Documentation Note of Corrections made to PPN's During Correction Process on Plant Linkage Errors on the LRD.

Foster, L., (1999), "Documentation for the ES202/SSEL Joint Project," internal memo, Center for Economic Studies.

Internal Revenue Service, (1985), Record Linkage Techniques – 1985: Proceeding of the Workshop of Exact Matching Methodologies, Statistics of Income Division, Internal Revenue Service, Publication 1299 (2-86).

Jarmin, R., (2002a), "Post-Matching Linkage Edits on the Longitudinal Business Database," Center for Economic Studies Technical Note, *CES-TN-2002-03.*

Jarmin, R., (2002b), "Improved Age Information for ASM Establishments," Center for Economic Studies Technical Note, *CES-TN-2002 [Forthcoming].*

Jarmin, R. and Miranda, J., (2002), "LBD Documentation: Broken Longitudinal Links from Missed Reorganizations and Inactive Spells," Center for Economic Studies Technical Note, *CES-TN-2002 [Forthcoming].*

Krizan, C.J., (1999), "SSEL Linkage Project Documentation," Internal Document, Center for Economic Studies, U.S. Census Bureau.

MatchWare Technologies, (1997), Automatch, Generalized Record Linkage System: User's Manual, Burtonsville, MD.

McGuckin, R., and G. Pascoe, (1988), "The Longitudinal Research Database (LRD): Status and Research Possibilities," Survey of Current Business, Nov. 1988, pp. 30-37.

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

Miranda, J., (2002b), "LBD Documentation: Evaluation of Probabilistic Linking Algorithms in the Auto Repair Shop Industry". Center for Economic Studies Technical Note, *CES-TN-2002-02.*

Miranda, J., (2002c), "LBD Documentation: Geography". Center for Economic Studies internal document.

Miranda, J., (2001), "Statistical Linking With AUTOMATCH®: Diagnosis of the SSEL 1996 to 1997 Full Linking Procedure". Center for Economic Studies, Internal Memo.

Monahan, J., (1992) "Longitudinal Research Database (LRD) Technical Documentation Manual," Internal Document, Center for Economic Studies, available at http://weasel.ces.census.gov.

Nucci, A., (1993), "Longitudinal Business Database: Project Proposal," Internal Document, Center for Economic Studies.

Nucci, A., (1999), "The Demography of Business Closings," Small Business Economics, 12, pp. 25-39.

Pivetz, T., M. Searson and J. Spletzer, (2001), "Measuring Job and Establishment Flows with BLS Longitudinal Microdata," Monthly Labor Review, 124(4), pp. 13-20.

Robb, A. (1999), "New Data for Dynamic Analysis: The Longitudinal Establishment and Enterprise Microdata (LEEM) File," Center for Economic Studies, CES 99-18.

Spletzer, J., (1997), "Longitudinal Establishment Microdata at the Bureau of Labor Statistics: Development, Uses, and Access," Proceedings of the Section on Survey Research Methods, American Statistical Association.

Trager, M. L., and R. A. Moore, (1995), "Development of a Longitudinally-Linked Establishment Bases Register, March, 1993 Through April, 1995, mimeo, U.S. Census Bureau.

U.S. Census Bureau, (1979), "The Standard Statistical Establishment List Program," Technical Paper #44.

U.S. Census Bureau, (1998), “Business Register: Glossary of Data Elements,” Internal Document.

Walker, E., (1997), “The Census Bureau’s Business Register: Basic Features and Quality Issues,” paper presented at the Joint Statistical Meetings, Anaheim, CA.

Appendix

Table A.1: LBD files and variables		
File	Format	Data Items
LBD_LINK7599_V2	SAS 8 Dataset	LBDNUM; CFN75-CFN99; MU75-MU99; RECNUM75-RECNUM99; FLAGA75-FLAGS99
LBD_PAY_EMP_7599	SAS 8 Dataset	LBDNUM; EMP75-EMP99; PAY75-PAY99
LBD_GEO	SAS 8 Dataset	LBDNUM; STGEO75-STGEO98; STGEO275-STGEO298; CTYGEO75-CTYGEO98; CTYGEO275-CTYGEO298; ZIP75-ZIP98
LBD_GEOCTE	SAS 8 Dataset	LBDNUM; STATE; COUNTY, FLAG, FLAG2
LBD_LINK7599_V2_IND	SAS 8 Dataset	LBDNUM; YR; RECNUM; MU; SICRL; SICRL87; SICYR; SICYR87; SIC; SIC72; SIC87