LONGITUDINAL ESTABLISHMENT AND ENTERPRISE MICRODATA (LEEM) DOCUMENTATION

Zoltan J. Acs
University of Baltimore, Merrick School of Business
Chief Economic Advisor, Office of Economic Research
Office of Advocacy
U.S. Small Business Administration

Catherine Armington
Consultant
Office of Economic Research
Office of Advocacy
U.S. Small Business Administration

May 14, 1998

We would like to thank Alicia Robb (Office of Economic Research, Office of Advocacy, U.S. S.B.A. and University of North Carolina at Chapel Hill) for excellent research assistance. We are also grateful for the assistance and helpful criticism of many at the Center for Economic Studies -- especially John Haltiwanger, C.J. Krizen, James Lessard, Al Nucci, and Sang Nguyen. Development of the LEEM would not have been possible without the vision and support of Bruce Phillips of the Office of Advocacy of the SBA, who initiated the Census project to link annual data on establishments, both to construct annual enterprise data and to construct longitudinal establishment data. All errors are our responsibility.

LONGITUDINAL ESTABLISHMENT AND ENTERPRISE MICRODATA (LEEM) DOCUMENTATION

Abstract

This paper introduces and documents the new Longitudinal Enterprise and Establishment Microdata (LEEM) database, which has been constructed by Census' Economic Planning and Coordination Division under contract to the Office of Advocacy of the U.S. Small Business Administration. The LEEM links three years (1990, 1994, and 1995) of basic data for each private sector establishment with payroll in any of those years, along with data on the firm to which the establishment belongs each year. The LEEM data will facilitate both broader and more detailed analysis of patterns of job creation and destruction in the U. S., as well as research on the structure and dynamics of U.S. businesses. This paper provides documentation of the construction of LEEM data, summary data on most variables in the database, comparisons of the annual data with that of the nearly identical County Business Patterns, and distributions of establishments and their employment by the size of their firms. This is followed by a simple analysis of changes over time in the attributes of surviving establishments, and a brief discussion of turnover (business births and deaths) in the population and gross changes in employment associated with both establishment turnover and with surviving establishments. It concludes with a summary of the strengths and weaknesses of the LEEM.

Table of Contents

1.0 Introduction and Summary

1.1 Longitudinal Establishment and Enterprise Microdata (LEEM) defined

The Longitudinal Establishment and Enterprise Microdata (LEEM) file has multiple years of data for each U.S. private sector (non-farm) business with employees. The current LEEM file facilitates tracking employment, payroll, and enterprise affiliation and (employment) size for the over nine million establishments that existed at some time during 1990, 1994, or 1995. This file was constructed by the Census Bureau from their Statistics of U.S. Business files and their associated Longitudinal Pointer File, which facilitates tracking establishments over time, even when they change their identification numbers.

Since 1991, the Office of Advocacy of the U.S. Small Business Administration (SBA) has been contracting with the Bureau of the Census (U.S. Dept. of Commerce) for development and production of annual comprehensive and timely aggregated data on the performance of U.S. businesses by firm size. Building on the annual County Business Patterns database (CBP), which covers all business establishments with employees, the Census Bureau constructs annual Statistics of U.S. Businesses (SUSB) Tabulation files. Data on the firm that owns each establishment (firm-wide employment, payroll, estimated receipts, primary industry and State) are appended to each establishment record. These SUSB Tabulation files have been prepared for every year from 1988 through 1995. They are the only annual federal business data supplying information classified by firm size.

Most of the establishments in the SUSB Tabulation files have the same identification number in each annual file, as long as they remain in business. For these businesses, changes in their employment can be measured by comparing their

corresponding records for different years. However, when businesses are sold, or change their legal form, or add a secondary location, their identification numbers usually change. Census has constructed a Longitudinal Pointer File to link establishment records from the SUSB Tabulation files for 1989 through 1995, so that surviving establishments can be identified even when a business changes its identification number. Using the Longitudinal Pointer File, business births and deaths can be more accurately identified, and changes in all surviving businesses can be measured consistently.

The Longitudinal Enterprise and Establishment Microdata (LEEM) file was constructed from these SUSB Tabulation files by Census' Economic Planning and Coordination division under contract to the SBA. This new composite file links three years (1990, 1994, and 1995) of data for all private sector establishments with employees in any of those years. Each establishment is represented by a record which includes the start year of the establishment and three years of annual information extracted from the 1989-1995 Longitudinal Pointer file and from the three appropriate annual SUSB Tabulation files. The annual information for an establishment includes its Census File Numbers, Standard Industrial Classification, state, county, MSA, enterprise employment, establishment employment, and annual payroll in thousands.

This document describes the process that the Bureau of the Census uses to construct the LEEM file, compares it to County Business Patterns and discusses its strengths and weaknesses. The remainder of this introduction provides a brief review of relevant terms and a summary of the construction of the LEEM data. Sections 2 through 9 go into much greater detail and provide basic tables to document various aspects of the database. Most readers will want to select from these only those sections of special

interest to their own research. Even more detailed specifications can be found by reference to the sources mentioned at the end, and by discussion with appropriate offices at the Bureau of the Census.

The LEEM may be used for non-commercial research at any location of the Census' Center for Economic Studies. Each project proposal must be approved by the CES staff, and researchers must become Special Sworn Employees of the Bureau of the Census in order to maintain the confidential nature of the data. Data from many other Census programs may be linked to the LEEM data (by CFN) to enrich the database for special analyses.

1.2 Business units and their relationships

The basic unit of the LEEM data is a business establishment. An establishment is a single physical location where business is conducted or where services or industrial operations are performed. The microdata describe each establishment for each year of its existence in terms of its employment, annual payroll, location (State, county and metropolitan area), primary industry, and start year. Additional data for each establishment identify the firm (or enterprise) to which the establishment belongs, and the total employment of that firm.

Establishments that continue their operations can usually be tracked through time, even if their identification numbers are changed. Such changes result from structural, legal, or ownership changes in the business. Establishments tend to retain the same address and industry when they change ownership or legal form, and they usually retain the same name and tax identification number when they physically move their operations. Therefore it is usually possible to clearly identify the startup of a new establishment or the

termination (death or closure) of an establishment, as distinguished from the appearance of a new identification number or the discontinuance of an old one.

Establishments are owned by legal entities, which are typically corporations, partnerships, or sole proprietorships. For tax purposes, each business legal entity is identified by a federal Employer Identification Number (EIN) if it has employees (or a similar Tax Identification Number (TIN) if it has no employees). Most legal entities conduct their business primarily at a single location, or establishment. Those that have multiple locations typically have only two establishments. But some legal entities own hundreds of establishments, and they may be located in different States and operate in diverse industries. When a business is sold, or changes its legal form, it becomes a new legal entity and gets a new EIN.

A firm (or enterprise or company) is the largest aggregation of business legal entities under common ownership or control. Most firms are composed of only a single legal entity which operates only a single establishment -- their establishment data and firm data are identical. Only 4 percent of firms have more than one establishment, and they and their establishments are both described as multi-location or multi-unit. Multi-unit firms may be composed of one or more legal entities.

All LEEM establishment records include information on the firm to which the establishment belongs. The firm's nationwide employment is available to classify each establishment that belongs to it. This information is calculated annually for each firm by aggregating the annual data from all the establishment records associated with the firm that year.

Small multi-location firms are usually fairly simple in structure – for instance, a small store with a single branch location. Some of the larger multi-location firms are very complex and diverse, including both small and large establishments in different industries and geographic areas, organized into many different legal entities. A corporation, for instance, may own other corporations or partnerships as subsidiaries, or control others as majority shareholders or in joint ventures. A few large firms are structured with each of their establishments organized as a distinct legal entity.

Since firms are primarily legal entities, any change in their tax identification number signifies a firm death and birth. This frequently occurs while all of the establishments belonging to the firm continue their operations, so there is no economic impact from the change. Even when their legal identities remain constant, multi-location firms cannot very usefully be tracked through time because of the frequency of partial firm sales, mergers, and acquisitions. There are not yet any practical general rules for defining firm continuity in cases of complex transformations of multi-location firms. The original location (frequently the headquarters) of a multi-location firm may close while the remainder of the firm continues in business. In practice , this is rare except in the very large older multi-location firms – especially those with over 10,000 employees. In many cases of merger, acquisition, and divestiture, the identity of the surviving firms appears to be determined primarily for public relations or tax considerations, rather than for organizational or economic continuity.

Therefore, the LEEM does not attempt to define the age of a firm, nor does it facilitate tracking the survival of individual firms. However, the LEEM does identify which new and closed establishments are/were single location firms (or original locations

of multi-location firms), or are/were affiliates (or secondary locations of multi-location firms). This provides a fairly good measure of births and deaths of firms and of secondary locations, except for the largest firm size categories, where firms often substantially survive the closing of their original location. However, when single establishment firms are acquired by multi-location firms and continue their operations as secondary locations, no establishment closure is involved, so that kind of termination of a single-unit firm cannot be easily observed.

1.3 Overview of LEEM microdata preparation

The Longitudinal Establishment and Enterprise Microdata are prepared from microdata – computer-based records describing individual business locations each year. The Census Bureau first assembles data from a number of sources to construct an annual Standard Statistical Establishment List (SSEL). The SSEL serves Census both as a basic business name and address register for use in drawing samples and organizing business censuses, and as the basis for annual statistics on the distribution of business establishments and their employment and payroll. Each SSEL incorporates data from the Internal Revenue Service (IRS) Business Master File (for names and addresses of business tax filers) and IRS Form 941 (for payroll and employment reported with Social Security tax payments), as well as information from Census' annual Company Organization Survey (COS) of establishments in multi-unit enterprises. Missing payroll data are imputed from prior year reported data, or other currently reported payroll data. Any missing employment data are imputed from reported first quarter payroll data, from prior year employment, or from average figures for the industry.

Census' County Business Patterns (CBP) program extracts data from the SSEL for all businesses which had any payroll payments during the year, and it further edits the key data for those, to ensure they are reported consistently with the previous year's data. CBP tables are then compiled and published showing the distributions of establishments, with their employment and payroll, by industry, State, and county.

Since 1988, annual Statistics of U. S. Business (SUSB) Tabulation files have been constructed, primarily based on records drawn from the edited CBP microdata file. These records include data on each establishment's state and county, industry, annual and first quarter payroll, and employment in the March 12 pay period. These establishment records are supplemented with estimates of annual receipts and codes for Metropolitan Statistical Areas. Values for the employment, payroll, receipts, industry, and primary State of the firm owning each establishment are calculated by aggregating the corresponding values from all of the establishments in each firm, and these values are appended to the record for each establishment.

A Longitudinal Pointer File is constructed to track the reporting of data for continuing establishments in each of the annual SUSB Tabulation files. While most establishments retain the same Census File Number (CFN) in each annual SUSB Tabulation file, the occurence of a change in ownership or legal form, or a change between multi-unit and single unit firm type, will cause a change in CFN. A variety of methods are utilized to identify continuing establishments that have changed CFNs, either between years, or within years. A small single unit establishment (single establishment firm) which changes ownership or legal form will usually appear in both its old and new form in the SSEL, since that is based on administrative data for the year. It will therefore

be double counted in the CBP tables. The dual records for such an establishment would appear to represent both the death of its old form and the birth of its new form during the year if the two forms were not correctly identified and linked in longitudinal data.

The LEEM file contains a composite record for each establishment that had any positive payroll during the years covered. This LEEM record includes an extract of data from the SUSB Tabulation file records for each year that the establishment was active. When the identity of an establishment changes during a year, the data for its newer form are used whenever they show positive employment, or if neither form has positive employment. If the older form has the only positive employment, then that form provides most of the data representing the establishment that year. Data on annual payroll is the exception to this rule, since they must be summed from the reports for both forms in order to cover the entire year. Figure 1.1 presents a schematic of the relationship between the SSEL, CBP, SUSB and the LEEM. It also shows how the Longitudinal Pointer File is utilized in producing the LEEM.

The original LEEM file for 1990, 1994 and 1995 includes up to three years of data on CFNs and possible second CFNs, establishment employment, annual payroll, SIC code (industry), Metropolitan Statistical Area (MSA), State, and enterprise (firm) employment. This has been supplemented to include county for each of the three years, and the year of first appearance in Census data (or 1973, if dated earlier).

2.0 Standard Statistical Establishment List (SSEL) data – basis for LEEM

2.1 SSEL sources

The Standard Statistical Establishment List (SSEL) is the basic business register maintained by Census. It provides the universe from which surveys are drawn and bench-marked, as well as basic data which are periodically summarized in various publications, such as County Business Patterns. The SSEL is the underlying source of the data in the SUSB.

Each annual SSEL is actually an inter-related set of data files incorporating various types of basic data on all business establishments whose existence is recognized by the U.S. government. These are compiled from a combination of administrative data and survey responses. Data for single location firms are kept in separate files from those for establishments affiliated with multi-location firms. Data on names and addresses are organized separately from those on numerical attributes. Extensive flags are maintained to track the sources of data items and any edits that may have modified them.

The primary source of SSEL data is administrative data from the Internal Revenue Service (IRS):

- the Business Master File (BMF) provides names, addresses, and tax identification numbers for businesses that file tax returns.
- the IRS Form 941 (or 943 for farmers), for filing of Social Security Tax payments for employers, provides quarterly data on payroll and March employment for legal entities.
- Non-employer data come from IRS Form 1040 Schedules C (sole proprietors) and F (farmers) and Forms 1041 (estates and trusts) and 1065 (partnerships).

In addition the Census Bureau conducts an annual Company Organization Survey (COS) to collect information on establishments in multi-location firms. Altogether, the

SSEL covers about 246,000 multi-location firms with employees, and 9.5 million single unit businesses, including sole proprietors without employees. The multi-unit firms include roughly 1.6 million individual establishments, and each year these firms acquire or start-up an additional 55,000 establishments, whose data are added to the SSEL from the COS and from other business surveys.

2.2 Company Organization Survey (COS)

The COS is conducted annually by the Bureau of the Census to collect data needed to understand the structure and identify the components of multi-location businesses. Census uses the survey responses both to identify (by name, address, company/firm number, and industry) and link together all establishments that are under common ownership, and to construct and update firm-level data. The COS' detailed data on the status, industry, payroll and employment of each location are also used to update the establishment data for multi-unit firms.

The COS is mailed to firms, and solicits information on all establishments (or locations) belonging to the surveyed firms. The response rate is 85 to 95 percent. The survey asks each company to identify establishments that have been sold, closed, started, or acquired, and those that are continuing from the prior year. For each establishment, the firm reports on first quarter and annual payroll, employment for the March 12 pay-period, and any controlling interest held by another legal entity, as well as any other business controlled by the firm.

All firms with at least 250 employees (about 30,000 firms in recent years) are surveyed annually. Medium-sized ones are surveyed on a rotating sample basis so that generally a third of them (about 50,000 in 1994, but only 20,000 in 1995) are covered in

each of three years, depending on availability of funds. A new sampling scheme was introduced in 1994 to provide flexibility with minimum impact on reliability. In addition, all except tiny firms are surveyed in the Economic Census every fifth year. Tiny businesses, with less than 5 or 10 employees (depending on the industry), are not included in the COS, but are assumed to have only one establishment, unless they are identified either by another legal entity as part of their multi-unit business, or by another Census survey which incidentally identifies them as firms with multiple units. Any new small legal entities that belong to large firms should be identified promptly in the annual survey of large firms. Any new small firms created by divestitures from large firms would also be promptly identified.

There are many possible organizational structures for multi-location businesses. The most common structure for a multi-unit firm is a single legal entity (or EIN) with 2 affiliated establishments. Very large firms typically are composed of many EIN's, some of which have multiple locations and some of which are single-unit legal entities under the same ownership. There are also both large and small multi-unit firms composed of multiple EIN's with a single unit each, but this is relatively rare. Many additional legal entities function primarily as property owners or holding companies, inserted into the structure of complex businesses to own other legal entities, but having no employees themselves, so they are not covered by CBP or SUSB.

The irregularity in the sample size of the COS causes corresponding variations in the numbers of conversions from single units to multi-unit establishments. This tends to produce surges in the relative numbers of multi-unit establishments in the years with larger samples, which are primarily prior to each Economic Census and in each Economic

Census (years ending with 2 or 7). In 1990, for instance, there were about 34,900 establishment such status changes, and in 1991 this rose to 76,700. Other years have correspondingly greater numbers of small and medium-sized single units which actually represent more than one location, but have reported their consolidated payroll and employment of all their locations together. This probably results in some distortion of the timing of individual establishment births, deaths, expansions, and contractions for affiliates of multi-unit firms, although the firms' overall employment changes are accurately reported each year.

2.3 Census File Numbers (CFN)

Census File Numbers (CFN's) are used to identify establishments consistently in all of Census' business files. Each CFN has 10 digits. For single unit establishments (neither owning nor owned by other establishments) the CFN is a zero followed by the nine-digit federal Employer Identification Number (EIN) of its legal entity. In these cases, the establishment, the legal entity, and the firm are identical, and have the same identification number.

Establishment records associated with multi-location firms have a completely different type of CFN, but store their EIN elsewhere in the SSEL files. Each multi-unit firm is assigned a six-digit number, which Census calls an Alpha number. Those Alpha numbers beginning with a number from 1 to 8 have been are randomly chosen from available numbers. Those beginning with 9 have been manually assigned. For each establishment affiliated with the firm, a CFN is constructed by appending a four-digit plant or location number to the firm's Alpha number. Plant number 0001 designates the first location of a firm, and it is often an establishment which was formerly coded as a single

unit, under a different CFN. It is frequently the headquarters of a multi-unit firm, except for large firms, which often build separate headquarters locations later.

When a business with a new EIN first registers with the IRS or withholds Social Security taxes for its employees, that EIN is assumed to represent a single location firm if it has less than 250 employees, and the business is added to the SSEL as such. If it is later found to have multiple locations itself, the entry must be revised and re-identified from being a single unit firm to being more than one unit of a new multi-unit firm. If it is found to belong to an existing multi-unit firm, the CFN of the establishment must be changed to include that firm's Alpha number. If it actually is found to belong to another formerly single-unit firm, then both establishments must receive new CFN's including a new Alpha number representing the new multi-unit firm. Most establishment births in medium or large firms would be accurately represented in the annual SUSB files, because the SUSB data incorporate the annual COS data, and because most secondary locations share EINs with other locations in the same firm, so they can easily be properly linked to their firm.

When a multi-unit firm loses all but one of its locations, it is frequently allowed to retain its multi-unit type of identification, in the expectation that it is likely to expand to multiple locations again. Therefore there are many "multi-units" that, in fact, have only a single unit currently active in the SSEL. When establishments with zero employment in March are excluded from analysis, even more multi-unit type firms appear with only a single location, because their other locations lack employees at that time.

2.4 Employment and payroll

Payroll includes all forms of compensation, such as salaries, wages, reported tips, commissions, bonuses, vacation allowances, sick-leave pay, employee contributions to

qualified pension plans, and the value of taxable fringe benefits. It includes amounts paid to officers and executives of corporations, but does not include profit or other compensation of proprietors or partners of unincorporated businesses. The SUSB Tabulation files include both annual payroll and first quarter payroll for each establishment and firm. When first quarter payroll is zero, employment must also be zero.

Employment includes all full-time and part-time paid employees who are on the payroll in the pay period including March 12, including salaried officers and executives of corporations. Those on paid sick leave and vacation are included. Proprietors and partners of unincorporated businesses are not included. This exclusion of the management level personnel (and their profit or other compensation) from the employment counts (and payroll) of unincorporated businesses affects primarily the smallest firm-size classes, and probably reduces their apparent average compensation per employee from what it would be if all workers were included.

These March employment numbers are reported along with quarterly payroll on IRS Form 941 (or 943 for farms) for the first quarter of each year. These forms are received and posted by the IRS by mid-July each year. Although the payroll reporting is required, the employment question is voluntary, and as many as 40 to 50 percent of respondents do not provide their employment data. For these businesses, employment must be imputed from the payroll numbers reported on the same form. This imputation is based on any prior year reported employment and payroll, and on the relationship of employment to payroll for those similar businesses that do respond in the current year.

This employment imputation is relatively simple for the single establishments each representing a single-unit legal entity and firm. However, when a Form 941 represents a

legal entity with multiple locations, it is much more difficult both to impute the legal entity's employment when it is not reported, and to allocate employment appropriately among its establishments. This information is provided by the COS for all firms covered by it. In 1995 the COS provided complete employment and payroll data that was consistent with the Form 941 data for 16 percent of the multi-unit firms, covering 72 percent of their employment and 80 percent of their payroll.

For those smaller multi-unit firms that are not covered by the COS or any other survey, the employment and payroll reported for each of their EINs on Form 941 must be distributed to the various locations (multi-units) by imputation, based on any previously reported payroll or employment for individual locations, or on averages. When COS data are incomplete or inconsistent with Form 941 data, the company is often called to help work out the problems. Employment and payroll data from other Census surveys of establishments frequently provide information for some establishments, and they are then imputed only for the remainder.

On the 1995 SSEL there were about 213,000 active multi-unit firms with 1.6 million establishments and 60 million employees. About 51,000 of these firms (with nearly a million establishments and 47 million employees) were surveyed in the 1995 COS. Due to non-response, 13 percent of these firms needed complete imputation to distribute their firm employment to their establishments, but this involved only 7 percent of the surveyed firms' employment.

The remaining 162,000 multi-unit firms had 640,000 establishments and 13 million employees. About 9 percent of their establishments were imputed and 41 percent of their employment was imputed, usually based on reported payroll and employment for the EINs

in the firm and often for some of its establishments (from other surveys). Prior year employment and payroll for surviving establishments are also used. When new establishments are added to a multi-unit firm without any specific employment or payroll data, they are assigned an employment factor based on the average size of reported new establishments in their industry. For 13,000 multi-unit firms there was no basis for imputation, so zeros were assigned by default, making these firms inactive. A total of 100,000 establishments in multi-unit firms had zero payroll imputed to them, rendering them inactive also.

These various estimates of establishment employment are probably very accurate in terms of their usefulness for classifying firm-size, and for calculating aggregate employment data, since they are based on administrative data on payroll of the firms. They introduce some uncertainty into the classification of establishments by type of employment change – expanding, contracting, or stable – whenever the change is calculated from the difference between estimates, or between estimates and reports. Whether this causes an overstatement or understatement of the volatility of establishment employment is not obvious. It depends on the details of both the non-response distribution and the estimation procedures, which are very complex.

2.5 Industry classification

The 1987 Standard Industrial Classification (SIC) system is used for classifying each establishment's primary industry to the 4-digit level. The classification is usually based on the industry description provided in its application for an Employer Identification Number (EIN), if that was adequate. The SIC for each of the multiple units of firms is confirmed or corrected in the Company Organization Survey (COS), or in other

establishment surveys. Additional SIC data (often greater detail) are supplied by matching establishment record data with Unemployment Insurance tax filing data collected from states by the Bureau of Labor Statistics (BLS). Some establishments are classified only to the 3-digit or 2-digit SIC level. When industry is not known, the SIC is coded as 9999. Therefore there are no records with SIC missing.

Auxiliary establishments are those whose primary activity is management or support of the activities of other establishments of the same company. Their industrial classification is based on the overall activity of the company, rather than each establishment's specific support function, such as trucking, warehousing, computer processing, or management. These auxiliary establishments account for about 0.7 percent of the establishments and 3.3 percent of all employment.

Industry classification is verified for all surveyed establishments (in enterprises with at least 5 employees) during the quinquennial Economic Censuses. It therefore tends to change primarily in these Economic Census years (years ending in 2 or 7, such as 1987). However, in the year prior to the Census, information from other surveys is used most intensively to update industry wherever possible so that the correct industry survey form will be mailed out for the Census. Changes in industry classification detected during the COS, the Annual Survey of Manufactures, the Survey of Current Business, and other periodic Census surveys are used annually to update the SSEL industry classifications.

2.6 Identification of new and closed establishments

Most data on new businesses come from the IRS when tax forms are filed under new Employer Identification Numbers (EIN's) and the IRS adds them to its Business Master File of names and addresses. If these new EIN's have employment lower than the

cutoff for the Company Organization Survey (see above), they are assumed to represent new single-establishment enterprises, unless they belong to other multi-unit firms which include them in their response to the COS. Some of the remaining new EIN's may be included in various other business surveys and identified then as multi-units, but others may not be properly identified as multi-units until the quinquennial Economic Census.

Births of new branch locations under old EIN's for the larger multi-unit firms are picked up annually by the COS. During the CBP processing, both the COS data and other establishments with large payrolls are reviewed to identify any remaining consolidated reporting for multiple establishments, and this information is fed back to the SSEL. However, some new establishments of existing small firms may not be identified for up to 5 years. During this time their employment would appear as growth of employment in another related establishment, so the firm-size classification and the change in employment would be correct, but the employment change would be wrongly classified as expansion, rather than an affiliate birth.

An establishment is assumed to have closed if it has no payroll for two consecutive years. Single establishment records are dropped from the active SSEL if they have no payroll for 8 quarters. For multi-unit firms the COS collects end-of-year status for each associated establishment in the survey, and those that are reported closed are flagged as deaths. However, establishments in multi-unit firms that are too small to be in the COS or to be reviewed by the CBP edits may continue to have employment allocated to them for several years as a result of imputations based on payroll of continuing establishments, and employment allocation algorithms based on prior year employment patterns within the

firm. This delayed recognition of small numbers of certain types of deaths is likely to be rather closely correlated with the delayed identification of similar types of births.

2.7 Economic Censuses

The Economic Censuses, which take place every five years, serve both to update the portions of the SSEL which are not supported by annual surveys or administrative data, and to collect additional types of data for large portions of the business universe. The Economic Censuses use specialized forms specific to each industry, and collect a wide variety of detailed information on operations of the establishments and enterprises in each industry. However, enterprises with less than 5 or 10 employees, depending on industry, are not surveyed. The status and basic data for these very small businesses are limited to that derived from administrative data, primarily from the IRS.

Typically, these quinquennial business censuses are held in years ending with 2 or 7, and those data are released in a sequence of reports 3 to 5 years later. Forms for the 1992 censuses, for example, were mailed out to enterprises in December of 1992, requesting data covering the calendar year 1992. These forms and their follow-up forms were received back by August 1993, and were used immediately to update the 1993 SSEL data on company organization and establishment industry.

As noted earlier, the Company Organization Survey is incorporated into the Economic Censuses, so that the organizational structure of nearly all businesses is verified for those years. This more complete updating of the affiliations among establishments usually results in a surge of status changes in the census year, primarily small firms changing from single to multi-location status. The COS program has proposed annual mailing of COS forms to selected ‘single unit’ EIN’s which have shown big employment

changes, in order to identify more of these status changes on an annual basis, but this has not yet been funded.

3.0 County Business Patterns (CBP) – annual establishment data selected and edited from SSEL

3.1 Selection of data from SSEL

The CBP program selects data from the SSEL to produce both an extensive set of tabulations of establishment, employment, and payroll data for public use, and a carefully edited microdata file for internal Census use. Its coverage is limited to private sector non-farm establishments with employees (as evidenced by positive annual payroll), excluding railroads and most government-owned establishments. It does include government-owned establishments such as liquor stores and wholesalers, depository institutions and credit unions, and hospitals. SSEL data for the current year are matched on CFN to edited CBP data for the previous year, so that large changes can be reviewed. The resulting CBP establishment edit file includes current and prior year State, county, SIC, type of organization, first quarter payroll and employment, and annual payroll.

3.2 Editing of both CBP microdata and aggregate data

A preliminary establishment edit examines all large establishments (in terms of current or prior year employment, first quarter payroll, or annual payroll) and attempts to identify and correct any errors in their data, including the reporting of multiple locations consolidated into one record and the flagging of duplicate records. Current SSEL data for employment, payroll, industry, and geo-coding, and for selected ratios, are compared to those from the prior year of CBP data and to historical averages.

A COS review team searches for and resolves cases where surveyed companies have changed the degree of consolidation in their reporting, thereby either adding apparently new establishments and shifting employment from other previously consolidated establishments, or dropping establishments and including their employment in the remaining consolidated establishment(s). They also examine large businesses which have not been treated as multi-unit companies and either verify that they are, in fact, single unit businesses, or correct their employment, or find their additional establishments.

Then a cell edit reviews aggregate data classified by current year state, county, 4-digit SIC industry, and employment size to identify cell values that are inconsistent with other current year data, prior year CBP data, or other historical data. Those cells that are flagged in this edit are reviewed by analysts to resolve or verify all big changes. Any corrections to the cell aggregates are also carried through to the establishment level in the micro-data.

4.0 Statistics of U. S. Business (SUSB) Tabulation file – CBP establishment data with firm (enterprise) data appended

4.1 Establishment data from CBP

The SUSB Tabulation file for each year is derived primarily from the County Business Patterns (CBP) on-line file, selecting all private sector establishments with non-zero annual payroll except for farms (SIC 01-02), railroads (SIC 40), Postal Service (SIC 43), private households (SIC 88), large pension, health and welfare funds (SIC 6371 with at least 100 employees), and other financial funds. Also excluded when extracting data from the CBP microdata are predecessor records for multi-unit establishments and any

other duplicate records. Records representing establishments in the 50 States, District of Columbia, and Puerto Rico are extracted, excluding Virgin Islands and other territories.

Data fields extracted for the preliminary SUSB Tabulation file have varied somewhat over the years, but always include the Census File Number (CFN); State, county, and place codes; legal form of organization; an edited form of the original SIC code; employment; and annual payroll.

4.2 Supplemental establishment data

Metropolitan Statistical Area (MSA) and CMSA codes are determined from the state, county and place codes and added to each SUSB Tabulation file. They are set to nines for establishments that are not in metropolitan areas, as defined for each year.

The industry classification of SUSB data differs from that of CBP primarily because the first step of SUSB processing after extracting the data from the CBP system is to search for missing industry classifications, and to fill them in with more current data from the following year's SSEL. The 1992 SUSB data, for instance, shows 71,366 unclassified establishments with 45,568 employees. CBP that year showed 86,614 with 51,167 employees. As a general rule, classifications are found for about 15 thousand of the unclassified establishments from CBP each year. However, this has occasionally picked up some inadequately edited SICs.

There are occasional further differences in SIC coding because the CBP editing changes a few of the original SSEL SIC codes to group tiny industries for publication, and CBP handles coding of auxiliary establishments differently . SIC codes for auxiliary establishments are re-coded (from CBP values) to their original (SSEL) values, which indicate the industry of the enterprise for which they perform services, but only to the 2-

digit level. The SUSB codes are generally based on the original SIC codes from the SSEL, not the special CBP publication codes. See the appendix of an annual CBP publication for further details on this.

4.3 Calculation of enterprise data

Enterprise (or firm) data are constructed from data in all of the establishment records affiliated with each enterprise. Establishments with single-unit status represent single-location firms, so their firm-level data can be copied directly from their establishment-level data. For multi-location firms, the data for all affiliated establishments must be aggregated to construct enterprise-level data, which are stored in a Multi-unit enterprise file. From there, they are copied onto the records for each of their affiliated establishments.

To construct the Multi-unit enterprise file, the records for all establishments with multi-unit status (excluding establishments in Puerto Rico) are extracted from the preliminary SUSB Tabulation file. The employment, payroll and receipts of all the establishments affiliated with each enterprise are aggregated to determine the enterprise's total employment, payroll and receipts. Primary state, primary industry division, and primary (3-digit SIC) industry within the primary division for the entire multi-unit enterprise are defined as those with the largest share of annual payroll.

4.4 Differences between LEEM and SUSB annual data

Other than elimination of double counting of single unit businesses that undergo reorganizations during a year, the population of the LEEM is the same as the population of the SUSB for each year. An average of 44,000 single-establishment firms undergo midyear reorganizations such that their administrative data represent both their new for

and their prior form during that year. Between 1989 and 1995, the number of such midyear reorganizations identified in the SUSB has varied between 40 thousand and 49.5 thousand. These businesses appear in both CBP and the SUSB Tabulation file in both their old and their new forms, under two different CFN's, causing double counting of the number of these firms and establishments. Using the Longitudinal Pointer File (see Section 7) to identify these alternate forms of the same establishment, the LEEM avoids this double counting.

Most of these 'reorganizations' are actually just changes in ownership or legal form, with little economic impact. Frequently such reorganizations show no first quarter payroll for the new form, so that March employment will not be double-counted. However, in about 22 percent of the cases there is March employment reported for the second form, and that is double-counted in SUSB annual files, as in CBP aggregates.

4.5 Distributions of establishments by firm type and presence of employment

Table 4-1 shows the numbers of LEEM establishments of each type for each year. Looking first at the establishments with positive employment, about 77 percent of the total are single-unit establishments, and their average employment is around 8. The other 23 percent are affiliates of multi-unit firms, and the average employment of those establishments is about 39. However, there are a few very large single-unit firms, and many very small multi-unit establishments.

Table 4-1 also indicates that a relatively large proportion (about 13 percent) of the single-unit establishments do not report having any employees. It is important to remember that, following the practice of County Business Patterns and the Statistics of U.S. Businesses, the universe of "active establishments with employees" is defined as

those establishments with positive annual payrolls in a given year. The employment, on the other hand, should represent only the number employed in the pay period including March 12 of that year. Thus, the establishments reporting positive annual payroll, but no employees, represent primarily the businesses that had employees at some other time during the year, and secondarily those that still owed pay to employees from a prior year. In practice, the majority of these establishments with no employees are new businesses which had not yet hired employees by March of their first year. Only 3 percent of the establishments in multi-unit firms had no employees.

4.6 Addition of County Codes

The SUSB Tabulation files from which the LEEM was constructed did not include county codes. Therefore, to facilitate analysis of smaller geographic areas than states, count codes have been added to the LEEM for 1990 and 1995.

The county codes were acquired by matching each year of the LEEM with an extract from the appropriate SSEL, matching on both the CFN which was used to provide employment data for that year of the LEEM and on the State code, and filling in the county code for records which matched on both fields.

A partial analysis of these codes showed that the county data included some unspecified codes – 999, which probably indicates state-wide operations, and a few in the range between 960 and 998, which may indicate other aggregations, foreign operations, or errors.

5.0 LEEM Comparisons with County Business Patterns (CBP)

5.1 Aggregate establishment, employment and payroll comparisons

Because the LEEM data are derived from SUSB data, and the SUSB data, in turn, are derived from the CBP data files, the coverage of the LEEM and the CBP data is virtually identical. In each year, the SUSB preparation filters the CBP data to eliminate a few special types of records which might occur -- those without State codes, those representing large pension and other funds, and certain extra records representing duplicates or predecessors.

Table 5.1, and all other tables in this section, show aggregate data from the LEEM along with the percentage of that aggregate by which it exceeds the corresponding CBP aggregate. Thus, a -0.1 percent difference, for instance, indicates that the LEEM data are a tenth of one percent lower than CBP data for the same aggregate.

Table 5.1 compares the overall number of establishments, March employment and annual payroll of the LEEM with the CBP for the years 1990, 1994 and 1995. In 1990 the aggregate LEEM employment of 93,425,129 was 0.05% less than CBP. This difference was virtually identical in each of the three years. The difference was slightly larger for the number of establishments, suggesting that many of the establishments which were filtered out of the SUSB and LEEM had little or no employment. The somewhat smaller difference for payroll is associated with the handling of establishments with mid-year reorganizations. When the LEEM was constructed, the duplicate record resulting from each such reorganization was dropped, but the data on annual payroll from both part-year records were added together to represent the annual payroll of the continuing establishment.

5.2 LEEM establishments by industry, and differences from CBP

Table 5.2 compares the number of establishments in LEEM and CBP for each major industry sector. The industry classification of LEEM data differs from that of CBP primarily because the first step of SUSB processing after extracting the data from the CBP system is to search for missing industry classifications, and to fill them in with more current data from the following year's SSEL. In most years, this procedure finds classifications for an average of about 15 thousand of the unclassified establishments from CBP each year. Additional industry classifications may be filled in from other years of data for the same establishment.

In 1990 there were 258,646 uncoded establishments in CBP, in comparison with only 62,659 in LEEM. This proportion had been typical of CBP until then, after which a variety of new programs to improve the speed and completeness of industry classification in both the SSEL and subsequently in the CBP program. Because of the extraordinarily large number of records with uncoded industry in the CBP in 1990, the more complete industry classification in the LEEM resulted in each of the coded industry divisions showing more LEEM establishments in 1990 than CBP shows.

In 1994 the LEEM showed 47,311 uncoded records, only 19 percent less than CBP. In 1995 the difference between LEEM and CBP was -32 percent. In these two years the shift from uncoded to coded is not great enough to offset the lower numbers of LEEM establishments resulting from the elimination of double counting of reorganized establishments. For these years all industry counts for the LEEM are less than the number of CBP establishments, with the differences ranging from –0.86 to -0.22 percent in 1994, and from -0.58 to -0.10 percent in 1995.

5.3 LEEM employment by industry, and differences from CBP

Table 5.3 compares employment by major industry sector between LEEM and CBP. The general pattern of differences is similar to that in the establishment comparisons, but the size of each difference is smaller. This smaller scale follows primarily from the relative scale of the overall differences, about 0.05 percent for employment, compared to 0.70 percent for establishments. In addition, a large proportion of the records with uncoded industry do not have any employment. Many of these are new businesses which have not yet been industry-coded, and had not yet hired any employees by March, but had some payroll for the year.

In the redistribution of the large number of 1990 CBP records without specific industry codes, the amount of newly identified employment in manufacturing was not great enough to cause the employment comparison with CBP to be positive, as it was in all other industries. It appears therefore that the manufacturing establishments with employment in CBP were nearly all properly classified, in contrast to those in other industry divisions.

5.4 LEEM establishments by state, and differences from CBP

Table 5.4 shows the numbers of LEEM establishments in each state for each year, and indicates the percentage by which the LEEM exceeds the CBP numbers. The overall differences each year appear to be nearly evenly distributed across states. For example in 1990 the LEEM had 0.81 percent fewer establishments overall than CBP, and this difference by state ranged from a high of 1.07 percent in Alaska to a low of 0.63 in New Jersey.

The pattern was similar for the smaller differences in 1994 and 1995. In 1994 the greatest difference was 1.04 percent for Montana, and the least was 0.50 percent for the District of Columbia. For 1995 the range was from 0.92 for Idaho to 0.38 for Hawaii.

5.5 LEEM employment by state. and differences from CBP

Table 5.5 compares employment in each state as reported by LEEM and by CBP. Again the differences between LEEM and CBP each year center on the overall employment difference, which is only –0.05 percent. Most differences were within the range from -0.02 to -0.10 percent. The specific states that were negative outliers from this range (with differences ranging from -0.17 to -0.30) differed across the years. This would be expected, because the primary reason for the differences is mid-year reorganizations of establishments with signficant employment, which should be random with respect to location.

Two notable exceptions to the general pattern occured in 1990, when two states showed an unexplained tiny surplus of LEEM employment over CBP employment, with Mississippi showing 0.17 percent more employment in LEEM than in CBP, and New Hampshire 0.37 percent more.

5.6 LEEM establishments by establishment size, and differences from CBP

Table 5.6 compares the numbers of establishments within establishment size classes, as measured by the LEEM and by CBP. The published tables from CBP include establishments with no employees in the March 12 pay period along with the 1-4 employee size class. The LEEM permits separate examination of the counts of establishments with zero employees. The LEEM has 619,153 zero-employee establishments in 1990. The

difference that year between LEEM and CBP for the aggregated 0-4 firm size class was -1.43 percent. This fell to -1.20 percent in 1994 and -1.04 percent in 1995.

The differences between LEEM and CBP for all other establishment size classes are much smaller, and generally decrease with increasing size. In 1990 these differences range from - 0.12 percent for the 5-9 firm size class, down to - 0.02 percent for establishments in firms with 1,000 or more employees. The ranges were smaller in the later years.

5.7 LEEM employment by establishment size, and differences from CBP

Table 5.7 shows the comparison of employment classified by establishment size class as reported in the LEEM with that in CBP. Employment in both files is based on the pay period including March 12th. Obviously, there is no establishment employment in the zero firm size class. In the 1-4 employee firm size class in 1990 the difference between LEEM and CBP was - 0.15 percent, which is probably roughly the same as the difference in numbers of establishments with 1-4 employees. As with the numbers of establishments, the differences decreased with increasing firm size, and remained virtually constant in 1994 and 1995.

6.0 LEEM Establishments and Employment Distributions by Firm-size

6.1 Establishments classified by firm-size and by establishment-size

Because of the scarcity of data on businesses classified by size of firm, researchers have frequently used size of establishment as a proxy for size of firm. Of course, at the upper end of the range this has some validity – a large establishment must be in a large firm. However, a large number of small establishments are also parts of large firms, and

many large firms contain only small and medium-sized establishments. Table 6.1 does not show the distributions of establishments cross-classified by both their firm size and their establishment size, but it is sufficient to demonstrate the order of magnitude of the errors introduced when establishment size is used as a proxy for firm size.

Looking first at the largest size class, with 1000 or more employees, there are generally about 700,000 establishments in the firms in that size class, and only 6,000 of these establishments have 1000 or more employees themselves. At the other extreme, there are generally about 300,000 more establishments with 1-4 employees than the number of establishments in firms with 1-4 employees. These 300,000 establishments must belong to larger firms, and are probably distributed across the larger firm size classes roughly in proportion to the distribution shown, with a bias upwards, where more of the firms have multiple establishments. Many of these tiny establishments are affiliated with firms in the largest size class. Clearly classification by establishment size class is normally a poor proxy for classification by firm size, even for the smallest employment size class.

6.2 Employment classified by firm-size and by establishment size

Table 6.2 shows employment classified by firm size and by establishment size for 1990, 1994, and 1995. For all but the largest employment size class, employment by firm size class is somewhat less than employment in the corresponding establishment size-class. This rough equivalency has led many analysts to assume that, except in the largest size class, there are not substantial differences in these two classifications. However, even in these employment-weighted distributions of establishments, the gross differences are much greater than appears. Taking, for example, the numbers of employees in the 10-19 size-class in 1990, there were 7.5 million in the firm-size class, but 10.3 million in the

establishment-size class. The difference appears to be only 2.8 million employees. However the 7.5 in that firm-size include many in establishments with less than 10 employees, and many of the 10.3 in establishments with 10-19 employees are in firms with 20 or more employees. Thus the 2.8 million difference represents only the net difference between the number of employees in smaller establishments which belong to firms of this size, and the number of employees in larger firms in establishments of this size.

6.3 Establishment distribution by firm-size and establishment industry

Tables 6.3a, 6.3b, and 6.3c show the distributions of establishments by firm-size and establishment industry, for 1990, 1994 and 1995, respectively. In all industries, the number of establishments decreases as firm-size increases, except for larger numbers in the open-ended class with 1000 or more employees. This largest firm-size class accounted for less than 1 percent of the establishments in Agricultural services, Construction, and Uncoded, and around 5 percent in Services. In Mining, Manufacturing, and Wholesale trade around 10 percent of the establishments were in firms with at least 1000 employees. Around 20 percent of the establishments in Transportation, communication and public utilities, in Retail Trade, and in Finance, insurance, and real estate were in these largest firms.

Most of the uncoded establishments are in the zero and 1-4 firm-size class. Other than the uncoded establishments, other establishments in firms with no employees are fairly evenly distributed across all the industry sectors. Services accounted for the largest number of establishments in zero-employee firms, with around 200,000 each year. These patterns are similar for all three years.

6.4 Employment distribution by firm-size and establishment industry

Tables 6.4a, 6.4b, and 6.4c show the distribution of employment by firm-size and establishment industry for 1990, 1994 and 1995, respectively. The differences among industries in the shares of employment in the largest firm-size are magnified versions of those for establishments shares. Thus, the industries with less that 1 percent of their establishments in firms with at least 1000 employees have about 10 percent of their employment in these firms. Those with over 50 percent of their employment in such large firms included Mining, Manufacturing, Transportation, communication and public utilities, and Finance, insurance and real estate.

Again, the distribution of employment by firm-size appears to be consistent over all three years, except for that of establishments with uncoded industry. Apparently the changes in handling of industry coding over these years has taken firm or establishments size into consideration, so it has substantially changed the firm-size distributions of employment of uncoded establishments.

7.0 Longitudinal Linking of Establishment Records

7.1 Construction of Longitudinal Pointer File for establishments

The Longitudinal Pointer file is a directory for tracking each continuing establishment. It lists up to two Census File Numbers (CFN) for each year, allowing for a maximum of one midyear reorganization during each year, as well as a possible change in identity between each year's SSEL file.

The CFN is the basic Census identification number, which is assigned to each new establishment, and it is generally retained consistently over time. However, a change in ownership or legal form, or a change in status between multi-unit and single-unit, will

cause a change in CFN. A complex system of computerized matching of records for establishments which might have changed CFN's is used to identify continuing establishments in the SUSB and to update the longitudinal pointer file each year.

The annual updating of the Longitudinal Pointer file uses a wide variety of information to track continuing establishments that have changed CFN's. These include matching Permanent Plant Numbers (PPN's), matching on EIN's to track changes from single-unit to multi-units businesses, and statistical matching of records for single units, based on their attributes – such as name, address, zip code, and industry.

Each establishment location has a Permanent Plant Number (PPN) which identifies its physical location and industry and this PPN should not change with ownership changes. The PPN's were revised in 1988 because problems with their processing had led to some duplication, so it is difficult to consistently track establishments prior to that revision.

An Employer Identification Number (EIN) is also associated with each establishment, identifying the legal entity to which it belongs. The EIN may be unique to an establishment or may be shared by many establishments belonging to the same legal entity. Multi-unit enterprises may be composed of one or many legal entities. An establishment's EIN will change when it has a change of ownership or legal form, but not when it changes from single-unit to multi-unit. Therefore EIN matching can be especially helpful in identifying those continuing establishments if the PPN fails. When there is more than one potentially matching establishment record with the same EIN the longitudinal match system picks the establishment with matching EIN and the same 5-digit zip code. If there are still multiple eligible match locations the system picks the one with the greatest employment within the matching zip code.

When no matches are found on any of the above bases, the remaining single-unit establishment records are further processed in search of matches within each 5-digit zip code, based on either name matching or 3-digit industry and street number matching. This identifies many of the remaining independent businesses that have changed EIN (and therefore changed CFN) and have not been already tracked with a PPN.

Using the production of the 1992-93 update to the Longitudinal Pointer file as a typical example, the linkage process began with about 6.3 million establishment records in each year's SUSB Tabulation file. A corresponding name and address file was constructed for each, containing the attributes used for statistical matching. The residual from each match step was passed to the next match step. Here are the results:

5,564,000 record pairs matched on CFN,

32,000 of the remainder matched on PPN,

2,500 additional matched on EIN.

The remaining unmatched multi-unit records represent 110,000 deaths in the 1992 file and 86,000 births in the 1993 file.

The remaining single-unit records were passed on to further match processes, grouped by their 5-digit zip codes. Records showing no quarterly payroll in the last quarter were considered as potential matches with other records (usually limited to those showing no quarterly payroll in the first quarter of the same year). The results were as follows:

- matching on business name

19,100 matched across years

25,300 matched within 1992 (midyear reorganization)

24,300 matched within 1993 (midyear reorganization)

- matching on industry (3-digit SIC) and street number

10,600 matched across years

11,400 matched within 1992 (midyear reorganization)

12,500 matched within 1993 (midyear reorganization)

Many of the remaining records were not eligible to match because they lacked the crucial data for determining a match. These all became births and deaths by default, although with more complete information some might have been matched:

- not eligible to match because of missing or invalid zip code or industry

129,000 single unit records in 1992

95,000 single unit records in 1993

- not eligible to match because of non-unique zip/SIC/Street #

10,500 single unit records in 1992

23,100 single unit records in 1993

These and the other remaining unmatched records represent 402,000 single unit deaths in the 1992 file and 531,000 single unit births in the 1993 file.

The results of this sequence of matches were used to extend the Longitudinal Pointer file to 1993, adding the CFN's for all qualifying newly discovered 1992 mid-year reorganizations, the 1993 CFN values, and all qualifying 1993 mid-year reorganizations. Within-year matches that indicate a possible mid-year reorganization of a single unit (establishment) are subjected to an additional condition before they qualify to be added to the Longitudinal Pointer file – the unit must have existed for at least two years under the same EIN before reorganizing.

Unfortunately, each additional year of longitudinal data provides evidence in some cases that previously matched CFNs indicating mid-year reorganizations were errors, so those previously linked CFNs must be unlinked when updating the Longitudinal Pointer file. Specifically, in our example of updating the 1989-92 Longitudinal Pointer file with matches from the 1992-93 linkage process, some of the single establishments that were previously identified as having reorganized (changing to a new CFN) during 1992 will be found continuing under their original CFN in 1993. In each of these cases the updating process must eliminate the previous mid-year change, and figure out the appropriate alternative handling of the new establishment.

The volume of such revisions may be estimated roughly by noting that around 12,000 CFNs are changed during each update of the Longitudinal Pointer file. However, many of these would be counted twice, so perhaps 7-8,000 mid-year reorganizations from the last year of each version of the file are later found to be inconsistent with the following year's data.

The Longitudinal Pointer file for each set of years is fixed in format, with each establishment record containing space for two CFN's for each year covered by the Pointer File. The second value for each year is blank if no mid-year reorganization was detected, but the first value for each year is filled in for every year that the establishment exists. Thus, if an establishment continued under the same CFN throughout the period, its CFN would appear as the first value for each year, and blanks would appear in the space for the second value.

7.2 Construction of a 3-year composite LEEM file

The Longitudinal Establishment and Enterprise Microdata (LEEM) file is a simple composite file with each establishment record including data for three time periods – 1990, 1994, and 1995. This file was designed primarily for analysis of long and short term changes in the establishment employment and payroll for establishments of different ages and industries, belonging to firms of various sizes. Studying such changes for the single-year1994-95 period, and the five-year 1990-95 period facilitates investigation of the broader validity of many of the recent job generation research results based on more limited data.

The data source for each year of economic data is the SBA Tabulation files produced by Census' Economic Planning and Coordination Division under contract to the Office of Economic Research of the Office of Advocacy, Small Business Administration. The Longitudinal Pointer file which Census constructed to track individual establishments, even when they reorganize (changing identification numbers), is used to properly link the separate years of data.

The preliminary LEEM was a fixed format file with data including the Census File Numbers (CFN's) for each year (and possible mid-year reorganization), the first year in the Longitudinal Pointer file for 1989 to 1995, and selected variables from each of the relevant annual SBA Tabulation file records. This file was supplemented at CES by adding Source year for establishments that existed in 1989, and county for 1990 and 1995.

Because of the possibility of mid-year reorganizations, a complex procedure was needed when merging data from each annual SUSB file to the LEEM. In these cases, the annual file has two records for a single establishment which changed its identification sometime during the year. To consolidate the two records into one which best represents

the establishment for the year, one must decide first which record has the better (March 12) employment numbers. The employment data from the second CFN was whenever it had positive employment, or if neither CFN had positive employment. The classification data, attributes such as location, industry, and enterprise employment, were taken from the same record that supplied the employment data, except that if industry was missing (SIC =9999 or SIC =0000) in that record, then it was taken from the other record. The annual payroll numbers from both part-year records were added together to represent the entire year.

For each covered year, the CFNs were copied from the Longitudinal Pointer file, and if there were a mid-year reorganization, the CFN flag specifies which CFN (1 or 2) for that year supplied the employment data (and most other data). This information is stored in fields of the following form, where x denotes the last digit of the year:

CFN9x1	ID of establishment before any midyear reorganization
CFN9x2	ID of establishment after midyear reorganization
CFNFLG9x	Flag = 0 if no CFN9x2
	1 if employment from CFN9x1
	2 if employment from CFN9x2

This preliminary LEEM (LEEM952) was transferred to the Center for Economic Studies (CES), and then modified by substituting the Source Year from the 1990 Standard Statistical Establishments List (SSEL) when the first year in the LPF was 1989. County locations for each establishment for 1990 and 1995 were also merged into the preliminary LEEM, to allow more detailed geographic analysis.

The complete list of fields in the basic LEEM is provided in Appendix A, along with their attributes in the archived SAS data file at CES. Within each record data are recorded as missing for any year during which the establishment was not active (positive annual payroll). In SAS data files missing numeric fields are represented with a dot, while missing character fields contain a space.

This basic 3-year LEEM has 9,122,982 records, and uses about 1.1 gigabytes in compressed SAS format. It is stored sorted by CFN901, CFN941 and CFN951. Thus the records for single-unit establishments which existed only in 1995 appear first, followed by multi-unit births in 1995, then a similar arrangement of 1994 births, followed by establishments that have data for 1990.

7.3 Construction of Start Year

The preliminary LEEM included an implied starting year which was simply the firs year that the establishment appeared in the 1989 to 1995 Longitudinal Pointer file. Thus all businesses which already existed in 1989 would be coded as 1989. In order provide more complete data to represent start years, additional data were merged in from the 1990 SSEL. The SYR field on the SSEL indicates the first year that the establishment (as identified by CFN or PPN) existed on the SSEL. This might be a year before or after the establishment commenced hiring, or selling its product. However, since the LPF provides dating of establishments starting after 1989, the rough indicator for earlier years is useful for dating older establishments.

All LEEM records with CFN90-1 present and First year in LPF =89 were matched to the 1990 SSEL, matching on CFN90-1. When matches were found, if the SYR field in the SSEL was less that 89 but greater than 73, the SYR field was used to replace the First

year field on the LEEM. If the SYR field value was less than 74 and greater than 02, the truncated value 73 was used to replace the First year field on the LEEM. When the SYR field had a value of 00 or 01 it was judged to be a probable error, and was not used. The resulting composite LEEM field was called STRTYEAR.

Table 7.1 shows these start year counts for all establishments in the LEEM file, classified by their firm type in their first year of data. The value 73 was used for any date earlier than or equal to 1973. This distribution must be interpreted with care, because it represents two type of concepts. The counts of Startyears for the years covered by the LEEM file – 1990, 1994 and 1995 – represent the total number of businessses starting up in those years. For all other years, the counts represent the numbers of surviving businesses starting up in those years.

Examining the counts for single units, it appears that the series is credibly smooth except for surges in some of the Economic Census years – 1982 and 1987 -- but not in 1977 or 1992. If business age were calculated with intervals that allowed for the delayed processing represented by these surges, these data are probably fairly representative.

However, the similar series for multi-unit establishments is more eccentric, and clearly is not representative prior to 1978. The surge in 1987 and the low number for 1993 suggest that further inquiry into inconsistencies in procedures that year would be appropriate before making use of the earlier portions of these data.

8.0 Tables tracking changes in establishments over time

8.1 Changes between years in establishment identities (CFNs)

To reiterate, the basic unit of LEEM data is a business establishment, which is usually a physical location where business is conducted or services are performed. Each establishment is owned by a legal entity, which is typically a corporation, partnership, or sole proprietor (or individual). A firm (or enterprise, or company) is the largest aggregation of business legal entities under common ownership or control. Most firms are composed of only a single legal entity that operates a single establishment, so their establishment data and their firm data are identical. These are referred to as "single units" in most Census business data. Firms that operate at more than one location are referred to as "multi-unit" firms, and their affiliated establishments are also referred to as multi-units. In many respects, the sources and processing of data on U.S. businesses are quite different for these two types of firms and establishments.

The Census File Number (CFN), which is the primary identifier of an establishment in most Census files, is changed when an establishment changes from single-unit to multi-unit status, and sometimes when it changes the other way. It also changes whenever an establishment changes ownership or legal form. Because the LEEM construction incorporates all available information to help track the identity of continuing business operations, the LEEM provides an opportunity to quantify the rates at which these CFNs change. The bottom line of Table 8.1 shows that, of the establishments that survived from 1990 to 1995, 10.8 percent had some change in ownership, legal form or firm type which caused a change in CFN. The 2.3 percent change over the one-year period from 1994 to 1995 suggests that the average rate of change is fairly consistent (rather than having a large number of changes centered on the Economic Census around 1992).

Looking further at the establishments that survived from 1990 to 1995, 78 percent were single units in 1990. Just under a tenth of these had different identities by 1995, with less than a third of those shifting to multi-unit status (either being acquired or opening other locations themselves). The largest single category of change was the 5.3 percent that were, and remained, single units, while changing owners or legal form. Of the nearly 22 percent that were multi-units in 1990, almost a sixth changed identities by 1995. Most of these remained multi-units, but were affiliated with different multi-unit firms.

8.2 Employment-weighted changes in CFNs between years

When the distribution of changes in CFNs are weighted by employment, the single units and multi-units have identical proportions changing during the 1990 to 1995 interval -- 16.7 percent. This suggests that such changes are strongly concentrated in the largest of the single units and the smallest of the multi-units.

The shares of various types of change also shift considerably. Half of all employment-weighted changes result from change in ownership of multi-units. Another third is accounted for by changes from single unit to multi-unit status. Recall that this might be due to either acquisition of the single unit, or to the single unit actually growing into multi-unit status (by opening or acquiring another location). Unfortunately, this type of change appears to be high during the 1990 to 1994 period, and quite low during the 1994 to 1995 interval. This suggests that much of the change may have taken place during the Economic Census of 1992, when many single units are asked if they have any additional locations. Reclassification at this time represents both actual changes and delayed reporting of additional locations since the prior census. This delayed reporting of the existence of secondary establishments does not affect the accuracy of the aggregate

firm employment or payroll reporting, but may distort the establishment size classification and the geographic and industrial classification of the secondary establishments.

8.3 CFN changes within years – mid-year reorganizations

Although it requires enormous effort to identify and correctly process information associated with mid-year reorganizations, their gross impact on the aggregate data is quite small. An average of 44,000 mid-year reorganizations of single-unit establishments are identified in the LEEM, and they have an average of 7 employees. If these reorganizations were not properly identified, they would each appear in the LEEM as an establishment death and an establishment birth, which would significantly raise the apparent rate of gross job changes due to births and deaths. (In the CBP aggregate annual data each of these is counted in both its old form and its new form.)

Table 8.3 provides more detail about the numbers of establishments and the employment involved in such mid-year reorganizations. Much of the process for identifying these imposes a rule that the new form cannot have any first quarter payroll, and thus that those establishments cannot have any March employment. However, it has recently been discovered that the matches found among establishment records with some positive first quarter payroll appear to have a lower error rate than those with the new form constrained not to have any first quarter payroll. This suggests that many more mid-year reorganizations might be identified were the funds available to redo these matches without this constraint. The constraint will probably be removed for matches of data from 1996 onward.

8.4 SIC code changes in surviving establishments with coded industries

Most longitudinal analyses of businesses are limited to a single industry or set of industries, and the analyst assumes that the businesses under study remain in the same industry classification. Table 8.4 provides some detailed data on the frequency with which this assumption is false. Establishments whose industry was not coded (SIC = 9999) were not included in this tabulation. Examining first the aggregate rates of change in each of the three intervals, it is clear that most of the changes took place in the 1990 to 1994 interval, and they were probably concentrated in 1992, resulting from the Economic Census. The changes associated with this census represent a combination of corrections to codes which were originally in error, or incomplete, and the accumulated actual changes in primary industry since the previous census (1987). The LEEM can provide additional evidence to identify changes from incomplete to complete 4-digit coding, but it cannot identify other corrections versus actual changes in primary industry.

Nearly 24 percent of the single units had a change in their SIC code during the 4 or 5-year intervals. The multi-units rate of change was only 13 percent, or just over half as frequent. But both types of units had roughly the same distribution across levels of SIC code change.

Looking at single units during the 1990 to 1995 interval, about 14 percent of the surviving units with coded industry change their coding within their 2-digit industry class. But 10 percent of them had changes which shifted them into other 2-digit classes. Over half of these changes across 2-digit classes involved changes of industry division. To be more precise, 5.4 percent of the single-unit establishments with coded industry in 1990 that survived to 1995 were in a different industry division by 1995. If not handled

carefully, this type of coding change might have a significant impact on analyses of other types of changes within industry divisions.

8.5 Employment-weighted rates of SIC code changes

Table 8.5 shows the rates of different levels of SIC code changes with employment weighting, and the rates appear virtually identical. Thus, it seems that the probability of SIC code changes, due either to actual industry changes or to corrections, is independent of the employment size of the establishment.

8.6 Establishment turnover by base-year firm-size and establishment industry

Dynamic employment change tables are often called "job generation" tables. They provide measures of the business population at a point in time, associated with measures of various types of changes in those businesses that took place before the next measurement point. Thus, for instance, if comparison of the static data in Table 5.2 for 1990 and 1995 showed a small increase in the number of establishments in services, the dynamic tables for 1990-95 could identify more about the sources of this increase, showing the gross deaths of 1990 establishments in services, as well as the births into services by 1995. Any remaining differences between the net dynamic change and the static end period total must be attributed to the net transfer of establishments into or out of the industry during that year.

Tables 8.5, 8.6, and 8.7 data cover only businesses that have employees on March 12 of the beginning or ending period. This has the disadvantage that certain seasonal businesses are never included, and businesses with temporary lack of employees in March will appear to go out of business and may later start up anew. On the other hand, this

definition of active business has a number of important advantages. It is clean and simple – the universe is businesses with employees, and all businesses in it have employees. The employment data necessary for enterprise size calculations are always available for the years in which he establishment is active. In addition, the establishment counts are more closely associated with the employment counts (both cover first quarter only), so that calculations of average employment and payroll are more representative.

Thus, births are not recognized until they have employees, and on average they will have had some employment for six months before the March 12 reporting period. Deaths are recognized equivalently -- the first March after they lose all employees, which also averages six months after closing. Using the data from the year after a birth and the year before a death allows the impact of these events to be better measured.

The base period number of establishments in Tables 8.6 and 8.7 therefore represents the population of businesses at a point in time, in contrast to many static table establishment counts (as in CBP), which include every business that existed (had a positive payroll) at any time during the year. The establishment counts in such static tables include all establishments with employees on March 12, plus all those that had employees earlier that year and died by March 12, plus all those that were born later that year, plus those that are seasonal which were inactive in March. Thus, a higher business turnover rate will result in a larger difference between the static count and the dynamic count of establishments.

Table 8.6 shows the number of establishments in 1990 and 1994 and the birth rate and death rate for new establishments. All size and industry classifications are determined at the beginning of the period, except those of new establishments, which are classified by

their ending period characteristics. Note that the birth rate for new establishments is higher for the one year interval than for the five year interval. Over a five year period some births and deaths are not recorded because establishments enter and exit the file undetected between the end points. Also, the birth and death rate varies across firm size and by industry sector. For example, while the birth and death rate are almost identical for manufacturing, in the service sector the birth rate is significantly higher then the death rate. The birth rate in services is considerably higher than that in manufacturing.

8.7 Employment changes by base-year firm-size and establishment industry

Table 8.7 shows establishment employment in 1990 and the employment changes from births, deaths and surviving establishments, all classified by the employment-size of the firm. All size, industry, and geographic classifications are determined at the beginning of the period, except those of new establishments, which are classified by their ending period characteristics. The net employment change is a decreasing function of firm size in all sectors, with the <20 firm employment size showing the highest rate of employment change. The only exception is in the distributive industries.

8.8 Employment changes by mean firm-size and establishment industry

Table 8.8, like Table 8.7, shows the employment of establishments in 1990 and their changes from births, deaths and surviving establishments, all classified by the employment size of the firm. However in Table 8.8 the mean firm employment size is used for classifying surviving establishments. The net change for the <20 firm employment size class is no longer as prominent as above, primarily because of the increased employment loss from deaths of establishments from the next larger firm-size

class that get averaged with zero (size after dying). Overall the distribution of net growth across mean firm-size classes is more proportional to the employment in those classes.

9.0 Strengths and Weaknesses of the LEEM for Future Research

Strengths of the LEEM

As a convenient source of basic cross-sectional data on the population of U.S. non-farm businesses with employment, the LEEM has a number of advantages over its alternatives – the SSEL and the CBP microdata.

- Reduced number of missing values in industry codes
- More precision in coding auxiliaries to industry firm
- MSA calculated
- Firm employment data is calculated

As a source of longitudinal data the LEEM has no competition, since it is the only U.S.-wide longitudinal data base covering all industries.

- Covers all establishments in all industries with any annual payroll (only farms, railroads, and government are excluded).
- Better start year data (based on Longitudinal Pointer File back to 1989, the SSEL)

Establishments are tracked across years as well as across CFNs (ownership and legal changes, even allowing for mid-year changes).

Weaknesses in the LEEM

Several characteristics of the data might introduce errors into a cross sectional analysis of the LEEM if they were not handled carefully.

- Exclusion of payroll and employment of partners and owners results in a probable understatement of average employment and wages for non-incorporated firms.
- Employment only for the March 12th pay period so cannot account for most part year businesses.

For analysis of longitudinal aspects of the LEEM, the investigator must keep in mind several limitations of the LEEM data.

- Imperfections in tracking establishments changing ownership and legal form (missed linkages), especially between single units and multi-units.
- A few false linkages exist between establishments that are not the same.
- Employment is frequently estimated, especially for multi-unit establishments, so year to year employment comparisons are imperfect (depend on the estimation method).

The LEEM file will next be used for a preliminary investigation of job generation by major industry between 1990-1995, with special attention to comparisons of annual gross flows in manufacturing during 1994-1995 with the average annual rates found by Davis, Haltiwanger and Schuh for earlier periods using the LED. Then the patterns of creation and destruction for manufacturing will be contrasted with those found for other sectors of the U.S. economy.

In another project, the LEEM will be used to measure the impact of mergers and acquisitions on the distribution of employment and on changes in employment and payroll during the 1990 through 1995 period.

This research is part of a multi-year cooperative project between the Office of Economic Research of the Office of Advocacy in the U. S. Small Business Administration,

the Bureau of the Census and the Center for Economic Studies. It is anticipated that a similarly defined LEEM file with annual data for 1989 through 1996 will be prepared for use during the fall of 1998. This will facilitate much more detailed analysis of patterns of job generation, persistence of changes, growth rates of new establishments, and survival rates. It would also support a variety of research projects investigating patterns of corporate dynamics.

In order to better understand the evolution of the rapidly growing service sector during the 1990s, the authors will use the extended LEEM data to analyze annual employment changes of establishments in the service sector. This will begin with study of the net employment changes, which will be analyzed in terms of differences by age, wage levels, and employment size of the establishments, and by firm size and type. We will then analyze differences in job creation, job destruction, job reallocation, excess reallocation in the service sector. The survival rates of both new and existing jobs will be calculated for different types of establishments and firms, and correlated with their net growth rates. The distribution of job changes and of change rates will also be investigated.

REFERENCES

Armington, C., "Statistics of U. S. Business - Microdata and Tables of SBA/Census Data on Establishment Size," Office of Advocacy, U. S. Small Business Administration, December 31, 1997.

Davis, S.J, Haltiwanger, J.C., and Schuh, S, Job Creation and Destruction, The MIT Press, Cambridge, 1996.

U.S. Dept. of Commerce, Bureau of the Census, Technical Paper 44, "The Standard Statistical Establishment List Program," January 1979.

U.S. Dept. of Commerce, Bureau of the Census, County Business Patterns, 1992 " General Explanation" and "Appendices".

U.S. Dept. of Commerce, Bureau of the Census, Economic Planning and Coordination Division, (internal) Transmittals of Specification/Procedures:

"Program Modifications for the SSEL Multiunit Complete Company Imputation Model," 4/30/92.

"Sampling Specifications for the Company Organization Survey Probability Sample," 10/13/94.

"Processing Requirements to Create a Longitudinal Data File," 3/23/95.

"Processing Requirement for the Creation of Longitudinal Composite Records," 7/12/96.

"Data Base Load Requirements for the Small Business Administration and International Business Machines Tabulations," 10/10/96.

U.S. Dept. of Commerce, Bureau of the Census, Economic Surveys Division, (internal):

"County Business Patterns: Specification for Tabulating the Cell Edit Universe and Identifying Cells for Analyst Review," 4/7/88.

U.S. Dept. of Commerce, Bureau of the Census, Enterprise Statistics: Company Summary -1987. General Explanation" and "Appendices".