Development of Business Data: Tracking Firm Counts, Growth, and Turnover by Size of Firms

by

Catherine Armington
Washington, DC

for

under contract number SBAHQ-03-MO-0563

Release Date: December 2004

SMALL BUSINESS
RESEARCH SUMMARY

SBA Office of Advocacy

No. 245
December 2004

Development of Business Data: Tracking Firm Counts, Growth, and Turnover by Size of Firms

by Catherine Armington, Washington, DC
under contract no. SBAHQ-03-M-0563 (2004, 42 pages)

Purpose

Just over a few decades ago, data on U.S. industries were mainly focused on agriculture and manufacturing; thus, comprehensive data on small business were not available. Tremendous efforts were undertaken by individuals and organizations to obtain the extremely valuable data by firm size that are taken for granted today. Documentation was needed to chronicle the creation of these data sources so that users could understand the strengths and weaknesses of the data and so that later efforts to produce new data sources could benefit from the previous experience.

Overall Findings

Creating firm size data on the dynamic U.S. economy has been difficult. Administrative data involve delays for the purposes of capturing births, closures, corporate restructuring, mergers and spin-offs. Within the last quarter century, great progress has been made by government data agencies in overcoming the challenges to develop dynamic firm size data that will allow policymakers and academics to understand the role of new and small businesses in the U.S. economy.

Highlights

• In the mid-1970s, the Small Business Administration's Office of Advocacy leased data from Dun & Bradstreet and contracted with the Brookings Institution to edit, link, and produce firm size tables from the data. The biennial data were referred to as USEEM/USELM. The program was discontinued in the late 1980s because of cost and quality concerns.

• In the late 1980s, the U.S. Department of Commerce, Bureau of the Census, with partial funding from the Office of Advocacy, began creating firm links among its established County Business Pattern data to produce annual firm size data under the Statistics of U.S. Businesses (SUSB) program. Longitudinal linkages followed later, allowing for the preparation of figures on births and closures of businesses, along with data on job creation/destruction by firm size. SUSB is an example of the cooperative efforts to produce data undertaken by many government agencies.

• In the early 1990s, the Bureau of Labor Statistics (BLS) began working on a project to link the BLS internal establishment list longitudinally. This work contributed to the still evolving Business Employment Dynamics program, which produces establishment birth and death figures along with job creation/destruction figures. A business size database is in process, but has not yet been published.

• The experience of the Federal Reserve Board in collecting data for the Survey of Small Business Finances reflects issues that arise when a private list of businesses is used. Extensive time and effort are needed to screen the data for accuracy.

• Common to all of the longitudinal datasources are issues related to linking firm data through time and determining whether a record is closed or is simply not recorded for the period.

• Because most of the data sources contain very large confidential datasets, the author suggests that synthetic data be created to allow researchers to better understand the dynamics of the U.S. economy.

Scope and Methodology

The researcher relied upon firsthand knowledge for the report, as she was involved in the technical aspects of creating and evaluating many of the data sources. Older tabulations and reports were used as source documents.

This report was peer-reviewed consistent with Advocacy's data quality guidelines. More information on this process can be obtained by contacting the Director of Economic Research at advocacy@sba.gov or (202) 205-6533.

Ordering Information

The full text of this report and summaries of other studies performed under contract to the U.S. Small Business Administration's Office of Advocacy are available at *www.sba.gov/advo/research.* Copies are also available from:

National Technical Information Service
U.S. Department of Commerce
5285 Port Royal Road
Springfield, VA 22161
(800) 553-6847 or (703) 605-6000
(703) 487-4639 (TDD)
www.ntis.gov
NTIS order number: PB2005-101237
Pricing information:
Paper copy, A04 ($31.50)
Microfiche, A01 ($14.00)
CD-ROM, A00 ($18.95)
Electronic download, A00 ($ 8.95)

To receive email notices of new Advocacy research, press releases, regulatory communications, and publications, including the latest issue of *The Small Business Advocate* newsletter, visit *http://web.sba.gov/list* and subscribe to the appropriate Listserv.

Table of Contents

Executive summary

In 1979 Congress instructed the SBA Office of Advocacy began to develop data for measuring the state of small business in the U.S. economy. Existing data by size of firm were incomplete and infrequent. Dun & Bradstreet had archived data files covering most U.S. businesses, so these were leased and the SBA contracted with Brookings Institution to edit these data, link them over time, and to design and produce tables for analysis of the distribution and growth of large and small businesses over time. This project developed into the 1976-1988 U.S. Establishment and Enterprise Microdata (USEEM) file, covering most U.S. private sector businesses, and providing basic descriptive data for over 5 million establishments and the 4 million firms that owned them, for each even year during this period. The large subset of these that had robust longitudinal data (USELM) was weighted to represent the universe for accurate measurement of changes over time. Updating these data became increasingly complex, and it was found that D&B's was very slow to acquire data on new firms.

The SBA then initiated a cooperative project with the U.S. Bureau of the Census to link the microdata in their business register (formerly known as SSEL) to construct longitudinal microdata for tracking all private sector establishments with employees, while identifying the firms that own them, essentially creating an enterprise version of their County Business Patterns. The resulting Statistics of U.S. Businesses (SUSB) now provide annual observations on each business from 1988 through 2002 with the Longitudinal Establishment and Enterprise Microdata (LEEM, also known as BITS) providing data from 1989 though 2001. The Office of Advocacy has an extensive library of aggregate data prepared from these for analysis of business distributions and dynamics. They are also exploring alternative aggregate databases, and linkages between the older USEEM aggregate data and specially tabulated BITS data to facilitate more detailed analysis of patterns of change since 1976.

The Bureau of Labor Statistics is also considering systematically aggregating data from their new longitudinal Quarterly Census of Employment and Wages (QCEW) to the legal entity (EIN, or pseudo-firm) level, which would provide more timely and more frequent measures of the dynamics of U.S. businesses, at a level very close to enterprise. It might also be feasible to construct limited public use samples from these, and an approach to this is proposed, to stimulate discussion of that alternative.

The dynamic nature of the U.S. business population makes it very difficult to capture comprehensive statistics covering it. All data collection systems suffer from delays in capturing and classifying new firms, discontinuities due to corporate restructuring and mergers and acquisitions, and ambiguities in identifying businesses that have closed. Nevertheless, the last quarter century of work on this challenge has made great progress, and has supported the development of a much clearer view of the contributions of new and small businesses to the U.S. economy.

I. Small Business data needs of the Office of Advocacy of the U.S. Small Business Administration

A. Introduction

Although traditional economic theory generally focuses on firms as being homogeneous, in practice they are very heterogeneous. They differ by their lifestages, industries, legal form, internal organization, and production processes. All of which leads to rapidly changing populations of firms of various sizes. Academics and policymakers need data to understand and address the roles and status of different types of firms for innovation, competition, regulatory, finance, employment and social reasons.[1]

In the past, the U.S. statistical agencies focused almost exclusively on agricultural and manufacturing businesses, so even data on the total number of firms was difficult to come by, let alone data classified by size of firms. Two sources of data on the number of firms did exist in the early -1900's, one private and one governmental. D&B's Dun's concerns go back to 1870, but expansion of their scope over time make trend analysis of numbers of 'concerns' difficult. And the precursor to the Bureau of Economic Analysis, the Office of Business Economics, produced data on the number of firms by major industry annually from 1929 to 1963. However this data program was discontinued because of doubts about the quality of the data. "The last substantial revision was made in January 1963 and revealed errors in the earlier estimates for absolute number and rate of growth; these errors were due partly to the cumulative effect of imperfect estimates for discontinued businesses." (U.S. Department of Commerce 1975, p.909)

In the mid-1900s, the U.S. Census Bureau had two data programs that helped fill the data gaps, the Economic Census and County Business Patterns. By 1954 the Economic Census, a quinquenial program, was established enough to allow firm size tabulations to be created for the Enterprise Statistics Program. But the slow steady expansion over time of the economic census to include almost all industries makes creating a comparable time series dataset for that an extremely difficult task. The other program, County Business Patterns (CBP), began in 1946 but became available annually with the data year 1964. CBP's unit of analysis is the business establishment, or location, not the entirety of establishments owned by a parent company, which is often referred to as a firm or enterprise. With sporadic firm size data from the Enterprise Statistics and annual establishment size data from CBP, firm size data was far from complete.[2]

[1] The author would like to gratefully acknowledge the helpful comments of Richard Clayton, U.S. Bureau of Labor Statistics, Trey Cole, U.S. Census Bureau and Brian Headd, U.S. Small Business Administration.

[2] Other data on small businesses have existed historically, but they have not been classified by firm size. The U.S. Census Bureau has produced owner demographic data as part of the Economic Census, under the older names, Surveys of Minority- and Women-Owned Business Enterprises and Characteristics of Business Owners, and under the newer name of Survey of Business Owners (see http://www.census.gov/csd/mwb/ and http://www.census.gov/csd/sbo/ for details). A large trade association for small business, the National Federation for Independent Business, has also produced data with a monthly survey of the trends and views of their members. These datasources are very useful to academic and policymakers, but out of scope for this paper on firm size data. A further discussion of many of these data can be found in USSBA 1994.

The various data development projects discussed in this paper should serve two purposes. First, they provide a guide for individuals trying to understand the strength and weaknesses of various firm size data sources. Secondly, they should help organizations in both the U.S. and other countries that are planning development of new data for study of business dynamics to avoid some of the pitfalls experienced in the U.S. efforts over the past 25 years.

B. Brief history of small business data development by the SBA

In 1976 Congress (PL94-305, sec. 202) mandated that the U.S. Small Business Administration (SBA) research the role of small business in the U.S. economy, but without directly collecting significant new data. In 1979 the SBA Office of Advocacy was further instructed by Congress (P.L. 96-302) to develop data for analysis of the impact of public policy on American small businesses, publish a wide variety of small business economic indices on a regular basis, and provide data for an annual presidential Report on Small Business and Competition.

At that time, the only reasonably comprehensive business data that classified by size of firm were the Enterprise Statistics data for selected industries, which the Census Bureau constructed every five years from data collected in its Economic Census and its Company Organization Survey. However, the creation of Enterprise Statistics was so complex and time-consuming that they generally were not released until four or five years later (e.g. U.S. Dept. of Commerce, Bureau of the Census, 1991 for 1987 Enterprise Statistics). Nevertheless, they did provide snapshots of the distribution of many types of businesses by firm-size at five-year intervals, but these cross-sectional data (snapshots) offered no clue about what was happening to businesses in the interim. The businesses in any size-class might be new firms that formed since the last survey, smaller businesses that recently grew into that size-class, the same ones that had been there previously, or larger businesses that had recently shrunk into that size-class. One could only conclude from these data that the number, or share, of businesses that were small had increased or decreased, and similarly for their employment, payroll, and sales.

Census was experimenting with construction of longitudinal data, but their Longitudinal Establishment Database (LED) was limited to manufacturing plants with at least 250 employees. They did not have the LED linked to any enterprise data, so the overall size and industry of the enterprises that owned these fairly large establishments were unknown.

About this time a great deal of excitement was generated by the findings of David Birch , who had used four Dun & Bradstreet (D&B) files to analyze the gross employment changes (those from business births, deaths, expansions, and contractions) of establishments that underlie the net annual changes in employment, for the period from 1969 to 1976. In a study for the Economic Development Administration of the Department of Commerce, Birch determined that during this period 66 percent of net new jobs were generated by enterprises with not more than 20 employees, and only 13 percent of net new jobs were in enterprises with over 500 employees (1979a, p. 30). He further identified establishments less than 5 years old as the source of 80 percent of all net job growth (ibid, p. 32). The SBA Office of Advocacy contracted with him for additional analysis of his 1969-1976 database, which he reported in Birch (1981), but his data

were not available for SBA use, nor were they sufficiently documented (Birch 1979b) to replicate.

After study of the alternative approaches (see Popkin 1980), the SBA Office of Advocacy began by leasing from D&B archived data files of data covering most U.S. businesses, and contracting with Brookings Institution to edit these data, link them over time, and to design and produce tables for analysis of the distribution and growth of large and small businesses over time. It was expected that these D&B data would be used to establish the size and characteristics of the entire universe of small and large enterprises and all their establishments. These would then be used to properly weight the file of selected income and balance sheet data from D&B, which would in turn be statistically matched to appropriate business tax returns from a sample of unidentified corporate, partnership, and Schedule C returns prepared by the Internal Revenue Service (IRS). Brookings staff had previous pioneered the development of a similar statistically matched database of U.S. Individual Income Tax-paying units, using Census data, the Current Population survey, and an IRS sample file supplemented with Social Security data (Armington and Odle, 1975).

Starting with the Duns Market Identifier (DMI) file for 1978, this project developed the 1976-1988 U.S. Establishment and Enterprise Microdata (USEEM) file, covering most U.S. private sector businesses, and providing basic descriptive data for around 5 million establishments and around 4 million firms that owned them, for each even year during this period. The large subset of these that have complete longitudinal data (USELM) were weighted to represent the universe for accurate measurement of changes over time. However, the D&B FINSTAT file of income and balance sheet data was found to have high rates of missing data for crucial variables, in addition to having weak representation of many industries and firm-sizes (Hirschberg and Phillips, 1982), and the IRS ceased its earlier practice of making sample files available to qualified researchers, so this plan for greatly expanding the range of variables for a weighted sample of businesses was abandoned. Furthermore, eventually this D&B-based data system was found to be both costly and increasingly complex to maintain, so it was discontinued after completion of the 1988 linkage and construction of growth tables.

Meanwhile, the Bureau of Labor Statistics (BLS) had completed its Business Establishment List (BEL) improvement project, which they began in 1988 to convert their quarterly database of businesses covered by the Unemployment Insurance system (ES-202) to establishment-basis, rather than the reporting unit they formerly used. In the early 1990s they had constructed a prototype Longitudinal Establishment and Firm (LEF) database, using newly developed statistical matching techniques (MacDonald and Armington, 1991) to link BEL microdata for consecutive years from 1989 through 1994, and to link Dun & Bradstreet Corporate Affiliation File (CAF) data to BEL data to better identify components of multi-location enterprises (Armington, 1995b). In 1991 the SBA requested data from the LEF (then called Firm-linked UDB, and note that the UDB is now referred to as the LDB), and BLS responded with an estimate of costs for preparation of relevant tables, because the microdata were confidential to BLS. However the SBA Office of Advocacy decided to focus its data efforts on the Bureau of the Census, which had a longer tradition of collecting and organizing economic data to the enterprise level.

Although the Bureau of the Census had previously not followed up on SBA Office of Advocacy initatives to cooperate on development of useful data for measurement of small business, in the early 1990s an agreement was reached to have Census prepare longitudinal establishment-level microdata from the edited microdata underlying their annual County Business Patterns (CBP) publication, supplemented with enterprise information for each establishment from their Company Organization Survey. This approach provides more complete longitudinal microdata for tracking all (non-farm) private sector establishments with employees and their associated firm information. With continued SBA support, this developed into the Census Bureau's Statistics of U.S. Businesses (SUSB) and Longitudinal Establishment and Enterprise Microdata (LEEM), which are referred to by the SBA's Office of Advocacy as the Business Information Tracking Series (BITS) data.

The BITS data now provide annual observations on each employer business from 1989 through 2001 (Acs and Armington 1998, Robb 1999, and USSBA 2000). Census offers annual tabulations by size of firm from the underlying Statistics of U.S. Businesses (SUSB), and they have discontinued their previous quinquenial Enterprise Statistics publications, which had been based on the Economic Census. Because Census data are confidential, the (BITS/LEEM) microdata can be used only for approved research at one of Census' Centers for Economic Research. However, the Office of Advocacy has developed an extensive library of tables of aggregated data prepared from them and approved for public disclosure, and these are widely used for analysis of business distribution and dynamics for various sizes of businesses. These tabulated data are located on the SBA Office of Advocacy web site (www.sba.gov/advo/stats).

More recently, BLS renamed its improved ES-202 data as the Quarterly Census of Employment and Wages (QCEW), and built on methods developed for their LEF to construct a new dynamic data series called the Business Employment Dynamics (BED) (Spletzer et al, 2004). This now has longitudinal establishment data covering the period from September 1992 through 2003. If the Employer Identification Numbers (EIN) from the ES-202 raw data were used to aggregate across establishments to measure the total size of multi-location business legal entities, the dynamic data classified by EIN-size would provide a very close approximation to firm-size data. These data would be both more frequent (quarterly) and more timely (less than one year lag) than any previous data available for evaluating the state of small business in America.

C. Enterprise data, not establishment data, needed to identify small businesses.

Most business activities are based in a specific location where goods or services are produced, and these locations are called 'establishments'. Business establishments are the basic unit used for organization of data describing the location and industry of business activities and their associated employment and payroll. The majority of business establishments are also legal entities (including sole proprietors) that are not owned or controlled by any other legal entity. These are called single-location firms or single-unit enterprises. However, about 23% of establishments are components of larger enterprises – a cluster (or multi-level tree of multiple legal entities) of establishments that are owned or controlled by a legal entity, which together make up a multi-location firm or enterprise. While their firm-size is clearly bigger than their establishment-size, their firm-level industry and location classification may also be quite

different from their own establishment classifications, and their business decisions may be made locally or by some distant headquarters.

Small business policy is clearly directed toward small enterprises (or firms, or companies), not toward small establishments that are majority owned or otherwise controlled by a firm that is large. Sometimes labor policy or safety regulations may be targeted on small or large establishments, but the overall financial and management responsibility for small establishments belonging to large businesses is, or could be, with the highest-level legal entity that controls it. Thus, however the distinction between 'small' and 'large' businesses is specified -- using employment, assets, or revenue -- and whatever quantitative boundary is chosen, it must be applied to an enterprise or firm. In most analyses where establishment data were used to represent enterprises the substitution seriously distorted the intended analysis of small business performance. Many of the secondary locations (plants, branches, et cetera) of large firms are small establishments, and their performance should be considered in the context of the total size of the firm that controls them.

D. Longitudinal establishment data, rather than cross-sectional data, to analyze establishment changes.

The performance of small businesses can be measured only by tracking small businesses themselves, not by tabulating changes in the numbers of small businesses. A growing small business might become large, or merge with another business with the resulting merged enterprise being large, while a shrinking small business may disappear (death, or failure), and a shrinking large business may be reclassified as small. Therefore a comparison of the number of small businesses (or their employment or revenue) in consecutive time periods cannot be used as an indicator of the activity or success of small businesses. Comparing the static shares of small businesses in an area or industry does show how the role of small business has changed in that area or industry.

Similarly, comparisons of the size of an industry sector, or even of the businesses in a specific location, do not accurately indicate the fate of the population of businesses at the beginning of the period, because a substantial portion of businesses change sufficiently over time so that their size classification, their industry, their ownership, or their location may be different at the end of a period. Using data for individual business establishments, we can track how they have grown or shrunk, and whether they have changed ownership, changed firm-size classification, changed industry classification, or changed location. The analysis plan for each research project must include determining the most suitable way of handling each of these kinds of changes. The economic research literature is littered with reports that either failed to get significant results, or found peculiar results, primarily because they used aggregate data which was defined without allowing for these types of changes.

Furthermore, each data source has built-in assumptions about these types of changes that may trigger changes of establishment identification numbers, generating challenges for the data matchers who are trying to set up systems to track the performance of business establishments. D&B generally cannot track establishments that change ownership, because their focus is on financial responsibility. BLS has difficulty tracking establishments that move across State lines

(even just across the river or across the road, if the state boundaries are there), because their data come from State Employment Security Agencies. Census often loses track of a business establishment if it changes two basic characteristics during the same year (two among ownership, legal form, industry, or address). It also has to give up on tracking businesses that change the degree of consolidation of their reporting—e.g. If a multi-unit enterprise changed its reporting of employment in a given state from 20 employees consolidated in one establishment to 10 establishments with 2 employees in each newly reported establishment, then no data would be available to track what happened to the original consolidated pseudo-establishment, nor when each of the newly reported establishments started, nor how their employment changed from their start-up level.

It is generally not feasible to track large enterprises across time, because their options for change are too varied to allow meaningful summary data. Large firms may acquire other firms or establishments, resulting in radical changes in size, industry, and headquarters location (Armington and Robb, 1998). They may also split themselves into multiple firms, create new legal entities to hold ownership of themselves, legally move off-shore, sell or assign away all their assets (including their workforce), sell themselves, lay off all their employees and then lease them back from a worker leasing firm, close or sell all their operating establishments and go into another sort of business, or disappear temporarily into bankruptcy. If an enterprise is large relative to the industry and locations under study, the effect of any of these changes may swamp the rest of the aggregate data so the performance of the rest of the population is lost behind the large enterprise change. This is a real world problem for analysis, not just a data problem, but negligent analysts often generate data problems by failing to allow appropriately for such real world changes.

E. Merits of commercial data versus government data sources

Collection and maintenance of a database covering all U.S. businesses is a very expensive project. Commercial business data suppliers expect to make a profit supplying such data to their customers, while government data agencies are responding either to Congressional mandates to collect such information, or are reorganizing administrative data to assist with their own goals.

D&B undertook their Duns Marketing Indicators project in order to profit from commercial leasing of the data for credit rating and market research applications. Most of their customers were interested in current data about recent performance, not in time series data, nor aggregates of data for classes of businesses. This market niche led to D&B decisions about editing, updating, and expansion of scope that were less than ideal for longitudinal analysis. Changes in the data were based on D&B's analysis of the market potential in their large commercial customer base, and the SBA Office of Advocacy and other government clients were left to suffer, or benefit, as the case might be. Furthermore, since D&B had a virtual monopoly on the supply of such large-scale business databases, neither Advocacy, nor any other government agency, had the leverage to influence D&B's processing to make the data more accurate, useful or consistent for economic analysis, nor any basis for getting a reduction in price per record or a lease with more generous use restrictions.

When the SBA Office of Advocacy first explored its options for data on small businesses it encountered seemingly intractable agency rigidities and confidentiality restrictions. As all agencies have suffered from budgetary restrictions and pressure to make their data more relevant to government policy needs, the propensity of statistical agencies to find ways to cooperate has greatly increased. However, statistical agencies are similar to commercial data producers in having entrenched bureaucracies that resist changes in procedures, and established customers that oppose changes in data products. Therefore it has been a very long run process to establish cooperative relationships with both the Census Bureau and the Bureau of Labor Statistics, and to learn enough about their existing business data to participate in an effective dialog about how to modify them so that they would be useful for analysis of small business.

While many of the benefits of greater inter-agency cooperation are obvious, the limitations imposed by federal policies on data confidentiality and the vagaries of the federal budget impose both constraints and uncertainty on the future of cooperative data projects. Just as the SBA Office of Advocacy lost access to its expensively developed USEEM/USELM data when its budget for leasing D&B data was reduced, the BLS Division of Occupational and Administrative Statistics lost its prototype Longitudinal Establishment and Firm (LEF) database when its budget for enterprise data was eliminated. Access to economic microdata such as the BITS/LEEM data files at the Center for Economic Research was sharply reduced as privacy concerns increased.

On the optimistic side, there is increasing awareness of the potential benefits from greater inter-agency data-sharing legislation, and therefore increasing political support for such legislation. In addition, agencies are figuring out ways of adding value by merging different types of data within each agency – adding employee wage and employment history data to establishment data, or merging economic census data to create longitudinal series, or merging special economic survey data with other establishment data at Census. As the sophistication and effectiveness of software for data merging increases, it increases the potential for such combined data products.

On the other hand, this increased capacity for statistical merging also increases the potential for illegal disclose of confidential data. This possible loss of data privacy arises because many different databases with some common coverage, but without common identification numbers, can now be efficiently merged, allowing unrestrained users to disclose identities that were supposed to be confidential. This is particularly likely if unidentified business microdata from government sources were to be merged with microdata from commercial sources that included the business identities.

II. Issues common to all longitudinal enterprise databases.

A. Defining the target business population.

A useful starting point for defining the target population is 'all private-sector business establishments with any employees.' However, many elements in this definition should be questioned, both for their desirability and their feasibility. First is the brutal fact that no one knows how many active businesses there are in the United States, because we have no uniform registration system for businesses that indicates whether they are actually 'in business'. The number of entities registering Employer Identification Numbers (EINs) or Taxpayer Identification Numbers (TINs) annually is far in excess of the number of new employer-firm formations, and we have no system to deregister when businesses close. Tabulations of new incorporations also include many entities that never become active, and many that are legal entities for purposes other than employer businesses. Therefore we have only an upper limit on 'all.'

Based on business tax returns filed with the IRS in 1976, there were 2.1 million corporations, 1.1 million partnerships, and 11.4 million sole proprietorships, making a total of 14.6 million possible businesses, but it is not clear how many of each of these types were active businesses, or how many had employees. Our best guess about the number of private sector enterprises with employees in 1976 is around 4.5 million, or just under a third of those filing income tax returns. We attempted to refine those numbers using Statistics of Income data from the IRS and looking at the numbers of various types of businesses by revenue classes, or by payroll, but this did not lead to any clear conclusions about how many of these businesses were within the scope of the desired database.

In fact, we know that a large fraction of corporations are financial shells, rather than operating businesses. However, because even owner-operators are legally treated as employees of corporations, we cannot distinguish how many have non-owning employees.

Partnerships are similar to corporations in frequently being a vehicle for ownership of property, rather than an operating business. However, their partners are not treated as employees under federal law, so they might actually be substantial operating businesses where all workers are partners and none are employees. In practice though, it seems that most partnerships that are active businesses have at least some employees to complement the work of the owning partners. Neither Census not BLS would include the partners as employees, but D&B counts working partners as employees, so this might cause the total employment reported by D&B to be several million higher than that reported for the same partnerships by the government-administrative-sourced data.

Among the tax returns for sole proprietorships (Schedule C of the 1040 Individual Income Tax) are many privately owned and operated businesses, as well as many hobbies that are reported as businesses so that their expenses are deductible, and retired folks and employees who work full-time in other businesses who get paid small fees for minor work that they report as Schedule C income. Again, the federal government does not count any of these individuals as employees, nor do they count any unpaid family members who work for them. Thus those without payroll

could include a considerable number of 'mom and pop' stores or service companies with substantial revenue and profits. D&B count all workers in their employment figures, so both the owner-operator and any unpaid family members who worked for them are included in DMI 'employment', but not counted in the CBP or QCEW (see USSBA 1988, p18).

Although our focus on private sector seems non-controversial, even the Census CBP data waive this rule for government-owned liquor stores and wholesalers, depository institutions and credit unions, and hospitals, because they are often in direct competition with private sector businesses. However, D&B's DMI file included large numbers of public school systems, public universities, military commissaries, and many other government-owned utility systems and local transit systems that had to be identified and removed in the attempt to limit the scope to private sector. There is another class of quasi-governmental or public/private ventures whose inclusion is uncertain, such organizations as government research laboratories, research institutes in public universities, post offices in general stores, city-owned sports arenas, and so on. The non-profit sector offers further scope for confusion, as their for-profit ventures should logically be included, and many of their not-for-profit businesses are competing with for-profit ones, but they often have difficulty in distinguishing employees from volunteers who have some employment benefits, so their data may not be comparable. BLS' BED includes the non-profit organizations that are required to join the Unemployment Insurance system, and those regulations differ somewhat from State to State.

While our primary focus is small businesses (by whatever definition of 'small' is chosen), data on non-small businesses are obviously needed for comparison. Therefore we need similar data for both large and small business enterprises. In fact, every year many businesses cross the boundaries between large and small, in both directions, so it is important that there be flexibility to refine, or re-time, the assignment of size class definitions to be appropriate to different research needs.

Most very large businesses are publicly-owned corporations, and a great deal of current and historical financial and employment data is commercially available on them. However, because many of these very large corporations undergo frequent corporate restructuring, with multiple mergers, acquisitions and divestitures, comparing their data over time requires expert restatement for comparability. Therefore use of commercial databases such as Standard and Poor's Compustat for evaluating the performance of these businesses over time is extremely difficult, and subject to huge errors, especially from double-counting or missing large parts of firms in transitions.

The few very large privately-owned companies (which may include corporations, partnerships, limited liability companies, and trusts of various types) tend to guard their data closely. D&B admitted that for companies like Mars (candy manufacturers, headquartered in McLean, Virginia) they resort to counting cars in the parking lots during various shifts to estimate their number of employees, and multiplying that by an average revenue-per-employee figure derived from similar public corporations in order to estimate their revenue. Obviously D&B's ability to detail all of the establishments owned or controlled by a private non-cooperating company like Mars in the U.S. is even more limited. We presume that such privately-owned corporations comply with regulations and file administrative data from which they will be reasonably

represented both in the CBP and the resulting BITS/LEEM files, and in the QCEW and the resulting BED files.

Finally we have the issue of variability in employment. Not only do employment levels in continuing businesses sometime move such that firms are classified as small in some years and large in others, but a large number of businesses have unstable employment that drops to zero seasonally or irregularly. The DMI tends to report average employment ranges, and companies probably try to hide any temporary periods without employees from D&B, which is fairly easy because the questionnaire does not specify a point in time for which employment should be reported. Anyway, if there is anyone there to report it, that person counts as one. Thus many annual changes will not be detectable, unless they cause a change in the reported range.

CBP includes a large number of businesses that report no employees, even though they have positive payroll (around 600,000 in 1990, increasing to 700,000 in 1995, or over 10% of all establishment records). Their positive payroll indicates that the business had at least one paid employee at some time during the year, but the reported number of employees is supposed to cover only the March 12 pay period each year. Therefore newly formed businesses that hired their first employee after March 12 will report zero for that year, and we will not have any employment data for those establishments until the following March, if they survive that long. Seasonal summer, fall and winter businesses might have no employees in March, regardless of how big they may be in August or December. Distressed businesses may lay off all their employees while the owners regroup, and then rehire their staff the following year. This appears to happen to about 90,000 businesses in the BITS/LEEM each year. Finally, businesses may close down permanently by March, but have employees and/or payroll earlier in the year.

The BED, being based on quarterly data from the QCEW, has both employment and payroll data for the third month of each quarter. Therefore most patterns of seasonal employment would be evident, and new formations and closures should be more quickly evident. However, State Employment Security Agencies have traditionally filled in estimated employment when they find one or two quarterly reports with zero employment, on the assumption that it represents a late report. If these records with zero employment actually represent closures, that will not be noted until the third quarter of zeros in the records for that establishment, and by that time the estimated data might have been extensively used, under-representing the closures.

B. Classifying industry of establishments and enterprises.

All of the databases under discussion used the Standard Industrial Classification (SIC) system as the basis for their industry coding. Since most current data are being classified only according to the newer NAICS system there will be serious problems for industry analyses extending over the period of classification change.

There are substantial differences in establishment and employment distributions by industry among the business databases using SIC, because of differences in how the SIC system is applied. D&B do not classify auxiliary establishments (those providing support services only to other parts of their own enterprise) separately. The biggest impact of this difference is that large corporate headquarters, especially of manufacturing firms, are coded in their primary enterprise

industry. MacDonald (1985) used the USEEM to analyze only firms active in food and tobacco processing, and found huge employment in New York City as a result of the inclusion of many corporate headquarters there. Census 1987 Enterprise Statistics included 2.8 million employees in auxiliary establishments, of which 1.2 served manufacturing firms and 0.8 served retail firms.

Firms that engage in petroleum extraction, refinement, and sales are a particular problem for classification, because these activities are in completely different sectors, so 'oil companies' classifications tend to shift dramatically as they acquire or grow establishments in different categories in this very large and vertically integrated business. Census 1987 Enterprise Statistics noted that only 44 percent of the sales of firms classified in Petroleum and Coal Products were in manufacturing, while 28 percent were in central administrative and sales offices, and 20 percent in mining. Since the 'oil companies' have been subject to a great deal of corporate restructuring over the last several decades the reclassification of their enterprise industry classification may have substantially affected the data for Mining and for Manufacturing.

Census has traditionally used procedures to make industry classifications 'sticky' (slow to change over time) to prevent frequent reclassification of establishments whose activities are almost evenly divided across several sectors so that the strict application of classification rules would cause flip-flopping. They tend to correct industry classifications in the year preceding the quinquenial Economic Censuses, because the Census forms they send out are specific to industries. And of course many of the replies received during the Census indicate further changes in classification of primary activity, so the greatest number of changes occur in Economic Census years (those ending in 2 or 7).

BLS has focused a lot of money and time on establishing rigorous procedures for doing their industry classifications, but the results may be biased because the Unemployment Insurance tax rate of a business often depends on its industry classification. Therefore, a new business may tend to describe itself as working in a lower-taxed industry whenever there is room for ambiguity (but many states base the tax rate on experience rating after a business reaches a certain age). BLS has a memorandum of understanding with Census to provide BLS industry codes for around one million new establishment records annually for the Census SSEL, now referred to as the Business Register.

D&B industry coding may also have some biases resulting from their interest in selling data for marketing, but none were identified during the work on the USEEM development.

C. Specifying continuity of a business -- tracking establishments.

In theory, each data source assigned establishments unique identification numbers, and if they reappear in the next time period with that same number we can assume that those businesses are still active. D&B used DUNS numbers, Census uses CFNs, and BLS used UDB (or LDB) numbers. Generally, the appearance of a new identification number denoted the formation of a new establishment, and the disappearance of an identification number denoted closure of a business. As secondary users of business data we do not generally make the decisions about standards for how much a business can change and still be considered to be the same business. BLS followed a policy that allowed an establishment to change its location, its primary industry,

or its name (or EIN), and still be considered the same business, but if more than one of these characteristics changed in a specified time period it would be considered a new business. However, as a practical matter, most business databases are based on administrative data that assigns identification numbers to the businesses, so both tax regulations and common practice generally determine the continuity standards. The formal requirements for defining continuity are relevant primarily when setting up new match systems for identifying continuing businesses that have changed their identification numbers.

Unfortunately, there were inevitable complications to the simple scheme of depending on the continuity of establishment identification numbers. In the mid- 1980's D&B ran out of DUNS numbers and started reusing DUNS numbers from businesses that had closed at least five years earlier, so it was necessary to check that a DUNS number had been unused for 5 years, and then add a flag to indicate that it was a different business than the previous one that had used that number. D&B eventually modified their numbering scheme to solve that problem. Furthermore, whenever a business changed ownership it usually was assigned a new DUNS number, so it would appear to be a closure and a new formation, even if it was the same set of employees doing the same thing in the same location. This inflated turnover rates, and we had no way to estimate how much without embarking on a statistical match exercise, which would have been too expensive at that time, because appropriate software packages were not yet available to facilitate matching.

Census CFNs are of two types, those for single-unit enterprises (based on their EINs) and those for establishments that belong to multi-unit enterprises. Therefore, when a single unit establishment opens an additional location (or at least when Census discovers that they have), its Census identification number is changed but the EIN is still useful in tracking establishments from year to year and another field, called its Permanent Plant Number, or a predecessor or successor number is also used to create matches. If it returns to single-unit status it usually retains its multi type of CFN, because of the likelihood that it will convert to multi again.

CFNs (and EINs) are also changed whenever an establishment changes its ownership or legal form. In order to track businesses through such transitions Census developed a complex matching system to look for strong similarities between establishments with discontinued CFNs and those with new CFNs, and to link the strongly matched CFNs through a directory called the Longitudinal Pointer File. For each new year of data, such links were sought between it and the prior year, it and two years earlier for some records, and within it (for midyear reorganizations).

On an annual basis, the percentage of surviving establishments with any type of CFN change ranged from a low of 1.9 percent between 1992 and 1993, to 3.1 percent from 1991 to 1992. The percentage of employment with CFN changes is higher, ranging from 2.5 percent and 5.5 percent, suggesting that those with CFN changes have more employees, on average. The most common type of CFN change was a change from one single unit to another single unit. Note that the highest change rate was in the period of the Economic Census in 1992, when all single units were asked if they had any additional locations, resulting in a dramatic increase in changes from single unit firms to multi-unit firms. Part of this increase is due to actual changes that year, but another large part results from delayed reporting of the opening of secondary establishments since the prior Census.

In the early 1990s BLS changed its reporting standards so that all but the tiniest secondary locations belonging to multi-units enterprises (within a county and industry, with a total of less than 10 employees) were required to report all of their secondary locations, whereas previously they could use consolidated 'reporting units' for multiple units within a county and industry. As a result, a large number of UDB numbers belonging to previously consolidated reports of multiple units were discontinued, because it was impossible to assign continuity from the old consolidated record to any specific newly broken out unit. While the transitions could be tracked with predecessor and successor numbers, the establishments themselves could not be tracked, since they had no prior specific data.

When analyzing business over time, the assumption is often made that the industry of a business stays constant. Analysis of data in the LEEM file (SBA 2000, p. 9) showed that single units are much more likely to have a SIC code change than are establishments of multi-unit firms (although this is partially a result of their being fewer initial sources of information for industry coding of single-unit establishments). During the 1989-1996 interval, almost 25 percent of surviving single unit establishments with an industry code (excluding those with 9999, for unclassified) experienced a change in industry code. These changes were almost evenly distributed across the levels of SIC code changes (1 digit, 2 digit, 3 digit, and 4 digit). Over that same time period, about 14 percent of establishments from multi-unit firms experienced a SIC code change. The percentage of employment experiencing SIC code changes closely mirrored the percentage of establishments.

There was a very high incidence of industry changes during the Census year, as well as the years immediately before and after that year. The changes during the 1992 Economic Census year, were probably a combination of corrections to codes that were initially wrong, additional refinement of more general industry codes, and actual changes in primary activity. Census also puts extra effort into updates of SIC codes before the Census in order to send the correct industry specific Census form to each business. The annual changes for the other years are probably an understatement of actual changes that are occurring to establishments in the natural course of business.

In processing the BEL microdata that preceded the QCEW files, BLS postponed accepting most reports of changes of industry coding until the first quarter updating for each year, in order to simplify use of the file for generating comparable change data by industry during each year. It is likely that there are still similar provisions in the procedures for the QCEW, so most industry changes in continuing establishments in the BED would appear in the first quarter of each year, and any calculation of 'enterprise industry' should probably be based on first quarter data for all enterprises that had first quarter data.

D. Identifying new business formations and closures.

With every source of business data there are questions about how quickly, how accurately, and how completely they identify new businesses and incorporate their data into the file. Since the process of starting a new business is often long, the point at which it should be added to the database is sensitive to the precise definition of businesses to be included – first paid employee,

first revenue, first tax registration? Closures also may involve a long process and considerable uncertainty, leading to questions discussed above about continuity.

However, even if we assume that the reporting is complete, there are many further issues to consider. As with identifying continuing establishments, one can establish some simple rules for identifying new formations and closures, but there are many cautions and exceptions that must be dealt with. Generally, a new employer establishment formation is identified when an establishment record appears with a new identification number and with at least one employee. Of course, if the database systems matching program has already identified the apparent formation as a continuation of an establishment that changed its identification number, then it cannot be considered a new formation. Similarly, if the apparently new establishment record has data on 'year business started', then it should also be used as a filter, requiring it to be within a year or two of the year of first appearance of employment. An additional filter that might be used is a limit on the number of employees reported for the initial year – new establishments with more than 500 employees often appear, on closer inspection, to be transfers of employees to employee leasing firms, reorganizations or joint ventures of existing firms that are restructuring their establishments, or simply data errors that have slipped past all edit checks.

Because D&B were working constantly to increase their coverage of U.S. businesses, successive DMI files frequently represented vastly different scopes from the universe of the U.S. business population. Editing new DMI files involved a constant struggle to identify and weed out new records that represented newly covered businesses, rather than new business formations. In June 1986, for instance, D&B proudly announced that they had added 50,000 records for additional secondary locations of the 500 largest firms, and were getting ready to add another 50,000 secondary locations of the next 1000 largest firms. The employment in these new records would already have been represented in the USEEM by imputed branch records, but these could not be linked to the newly reported data. In other cases D&B added in thousands of new records representing establishments in industries that had not previously been well covered. In such cases, if we could identify the new records introduced we had to choose between imputing earlier data for these establishments, or excluding them permanently from the longitudinal file.

The identification of closed establishments is somewhat less certain than new formations, because some establishments suspend operations for years, and then are re-activated in the same legal form, with the same identification number. Each data source has its own standards for the period of time it maintains records in the active file after their establishment has ceased activity, and how long these records then stay in an archived file, before being considered definitively out of business. As a practical matter, many analysts feel that they need to assume an establishment is closed after just a single period of zero employment. The reporting of two sequential years without employees greatly reduces the probability of subsequent re-activation, and therefore only a small overstatement of the closure rate would be expected if closure were defined as 2 years of zero employment, following positive employment.

Another problem that BLS had with UDB numbers was referred to above – that States routinely generated data to replace missing reports, on the assumption that the business was continuing and its report was late. If the report were actually missing because the establishment had closed, this would cause an appearance of continuity, even several quarters after it closed.

E. Tracking enterprises -- avoiding distortions resulting from business restructuring – mergers, acquisitions, divestitures.

As long as enterprises comprise only a single establishment, we can use the same rules to track them through time as we use for tracking establishments. But as soon as the owner or manager of a single-location enterprise opens another establishments, tracking the enterprise becomes a potentially complex challenge, and it can only be attempted with a set of fairly arbitrary rules devised to facilitate research on a particular analytical plan.

To give a simple illustration of how quickly enterprise-tracking gets complex, suppose that in our example above, the enterprise with 2 establishments then divested one of them. The next period then would show those same two establishments as 2 different single-location enterprises. Is this an enterprise death and two enterprise births, even though all establishments continue their operations? If the microdata source did not include data showing that one of the establishments still had the same owner as previously, then would you want to assume that the larger part of the split enterprise stayed with the original owner? This relatively simple transaction might also cause changes in the enterprise industry classification, and the firm-size classification. You might find that both establishments grew, but the firm-size of each might shift from large to small when one was divested.

Most analysis addressing the behavior of enterprises can be better investigated by redefinition in terms of tracking certain of their establishments. Armington and Robb (1998 and 1999), for example, investigated mergers and acquisitions from 1990 to 1994, using the LEEM data to track a limited set of establishments that were acquired during the period and then evaluating how employment in these changed after their change of ownership.

III. USEEM and USELM constructed from Dun & Bradstreet's DMI (1978-1988).

D&Bs DMI was an extension of their traditional credit database (which would have covered only business legal entities using credit or insurance) to create a marketing tool, covering all business locations that might purchase goods or services in any given location or industry. Many large businesses used D&B's unique identification numbers for establishments and firms, the Duns number, as the basis of their accounting system for suppliers and customers, so most businesses wanted to be covered by D&B. The information the DMI provided on each establishment included not only its Duns number, location, industry, and employment, but also an estimate of its annual sales (unless it was a branch location that did not keep separate books from its headquarters), and the Duns number and total employment of the firm to which it belonged. Additional data included the startup year of the establishment, when the data were last updated, codes for its position in any multi-location firm, and flags for various types of estimated data.

Most annual DMI files were copies of the December archive of the most current data for active businesses that D&B covered – the result of a continuous updating process. Data for the most active firms were updated every few months, while those for firms with little activity were expected to be updated at least within 13 months, but a few were ignored for years. Thus the employment data do not represent a point in time, but an average over the year, with larger figures generally closer to the end of the year. Partners, proprietors, and unpaid family workers were included in the employee counts, unlike the government's employee counts, which are limited to those receiving salaries or wages.

A. Editing and imputing missing and old data

The DMI data were not designed for statistical use either cross-sectionally or longitudinally, and because these data were continuously updated, rather than being refined into an edited database representing a point in time. Therefore D&B had not established procedures to edit the employment and sales data that were crucial to the SBA's interests. Furthermore, D&B specifically did not include sales branches of manufacturing firms, and they had never attempted to get complete representation of all domestic establishments belonging to multi-location firms. In addition, their updating procedures involved systematic updating of data for large firms, and those for which it had credit inquiries, but many small-firm records were neglected to the extent that they could not be used for analysis of two-year changes, and in extreme cases some old records were presumed to represent closed businesses.

Potential problems with the employment and sales data included missing data, simple clerical errors, use of estimates that were inconsistent with subsequently reported data, and lack of updating of records, especially for smaller firms. Missing data that had no relevant reference data were replaced with industry averages for each state, using CBP. Clerical errors could be detected when they resulted in extreme changes over time, and these were corrected, but of course less extreme errors remained undetected. Inconsistent estimates were corrected to be consistent with subsequently reported data. Data that were over 2 years old were flagged so that they would not be used in calculations of change rates, and records over five years old were deleted, as experience eventually demonstrated that nearly all such records represented closed

establishments. For 1984, for instance, 550,000 old records were deleted, and less than 10 percent of them were later restored because updated reports were found for them in 1986.

The DMI data included details of the corporate structure of most multi-location businesses, with establishments identified as headquarters, branches, and subsidiaries (which might also be headquarters). The total employment figure provided for the enterprise represented the sum of the local employment in all locations, domestic and foreign. However, in practice, differences in timing, missing establishments, employment in foreign affiliates, double-counting of employment of some subsidiaries, and random errors all contributed to frequently substantial differences between the reported total for the enterprise and the actual sum of its establishments' employment. After extensive analysis of the patterns of errors, techniques were developed to reconcile the employment for each corporate 'tree' of affiliated establishments. Any differences between the total employment reported for a multi-location firm and the sum of all its reported locations was reconciled either by adjusting the total employment, or by adding imputed, or proxy, establishment to represent missing branches. Then many of these proxy locations were assigned to represent foreign affiliates, based on distributions provided by the Bureau of Economic Analysis.

B. Defining consistent scope of coverage over time for the LEEM.

As discussed above, D&B was continuously increasing their coverage of U.S. businesses, adding more records representing secondary establishments of multi-location firms, increasing their coverage of small businesses in industry sectors that had previous been skimpily covered (primarily personal services and local government businesses, including public schools), and adding many more non-employer businesses (but counting the proprietors, partners, and unpaid family member workers as employees).

In converting the biennial DMI archived data to USEEM data it was necessary to identify and delete these records representing new coverage of existing businesses, as well as those that we could identify as beyond the intended scope. Having begun with a DMI file representing U.S. businesses in 1978, all subsequent files were edited to represent approximately the same scope. This became an increasingly complex problem as time passed, and the raw DMI swelled from 5 million records to over 8 million by 1988.

The additional records representing previously omitted branches of multi-location enterprises provided valuable information for cross-sectional analysis, so they were retained in the cross-sectional USEEM files, but they lacked historical data back to 1976, so they could not be included in the USELM file for longitudinal analysis. Because employment in missing branch records was previously represented by proxy branch records, their addition did not involve any change in scope of coverage.

C. Processing of cross-sectional data to produce longitudinal data, and weighting to compensate for missing information.

The majority of establishment records in each year of LEEM data represented continuing business establishments that were linked over time simply by merging on their DUNS

identification numbers. However, if no match was found for an establishment record, it might indicate either that the establishment did not exist that year, or that the data were merely temporarily missing (some had been deleted because they were over 5 years old). Therefore, after merging the files of establishment data for all available years, an analysis of the patterns of missing data was undertaken in order to formulate decision rules for filling gaps in data for continuing establishments, and flagging each merged record for establishments that apparently were newly formed or closed during the period. Most of the extreme gaps in reporting were found to be the result of D&B's temporary reuse of retired DUNS numbers for new establishments, and such records were split to correctly represent the two different establishments covered.

By the time data for 1976 through 1986 had been merged together to create a Linked USEEM file, nearly half of the longitudinal records had some employment data that had been estimated, or had not been updated since the prior reported data. These records were not suitable for longitudinal research, because their employment changes were not accurate if they were frequently based on actual versus estimated, or all estimated, or simply stable because the information had not been updated. Of course all the proxy records imputed to account for missing domestic branch locations were included with these records flagged as unsuited for longitudinal analysis because of reporting problems. The USELM file includes only the subset of longitudinal establishment records without serious reporting problems.

The proportions of records with reporting problems varied with the size and type of business, its industry, and the periods for which it reported data, but within strata defined by these factors the reporting problems appeared to be random. Therefore the USELM subset of records could be weighted to represent the entire USEEM population, by considering the USELM to be a random stratified sample from the Linked USEEM, and calculating the appropriate weight for each strata. Weights were calculated for each year of data for 15,435 cells, and averaged across years to calculate a weight for each USELM record.

See USSBA (1984) for further discussion of the USEEM and USELM data development, and Applied Systems Institute (1987) for details on the weight calculations.

The 1988 DMI file had even more challenges than prior DMIs as a result of D&B improvements in coverage, and a new contractor was chosen to incorporate it into the Linked USEEM and USELM. Some of the resulting problem proved to be unresolvable, so they recommended that the 1976-1988 USELM be used only for analysis of four-year changes, rather than the two-year change data that had been provided by earlier files.

D. Strengths and limitations of resulting USEEM/USELM database.

Because the microdata covering nearly all private sector businesses for 1976 through 1988 were already edited and available in a choice of two files – the USEEM for cross-sectional comparisons and the USELM for longitudinal analysis – these data could support a wide variety of analyses based on business and employment distributions or changes. Extracts were prepared for narrow industry studies, for small area analysis, for broad studies of particular types of change in employment or in company structure, and for basic research on growth in both large

and small businesses. Because so much data was available, aggregate data could be conveniently prepared according to specific definitions to satisfy the needs of researchers and policy decision-makers.

Many investigators undertook studies to validate these data, either by comparing them to other data from the government or from commercial sources such as business telephone listing, and all found problems. However, although these data have a large number of known differences from Census and BLS-based data, those data also have weaknesses, and were subject to change during the seventies and eighties as those agencies addressed their own shortcomings. Many of the published critiques of the DMI and USEEM data suffered from problems with specification of the data – having asked D&B of the SBA for microdata that did not include all of the businesses they were looking for, by failing to define their industry, region, establishment types, or time periods adequately. Some of these studies helped to identify problems with the DMI or USEEM that were then minimized by better handling in subsequent versions of the data. Other problems were inherent, either to DMI or to all business data. Many of these are reviewed in Armington and Odle (1988a) and USSBA (1988).

Because the SBA has recently been using primarily CBP-based data, some detail on how the USEEM compares with it might be useful. As mentioned earlier, the USEEM employment data include proprietors, partners and unpaid family workers, so aggregate USEEM data would be expected to show higher employment, and the difference should be roughly equal to the number of establishments, since all but corporations would have at least one additional worker, and some would have more than one extra included.

CBP annual employment represents workers on the payroll on March 12 of each year, while DMI employment is probably an average over the previous year, or even the maximum. DMI records for smaller establishments (less than 50 employees) tended to be much older than those for large, so probably represent employment two years earlier in many cases.

Until 1983 CBP included only the establishments that were active in the fourth quarter, along with their employment at the end of the first quarter. But in 1983 they changed to include all establishments with any positive payroll in a year, thus including all of the newly formed, and all of the establishments that closed during the year n their annual counts. This raised the CBP establishment counts since 1982 by about 15 percent.

Comparing the numbers of establishments by industry that were recorded by CBP and USEEM in 1986, one must first take into consideration that CBP separates out auxiliary establishments such as headquarters, research labs, and sales offices that serve only their own company, while the DMI generally classifies these by the primary activity of the enterprise. In addition, CBP had nearly half a million establishments in 1986 without any industry classification yet – presumably mostly new businesses. On the other hand, the USEEM includes some sole proprietorships and partnerships that technically have no employees, which should increase its establishment count.

CBP was considerably lower in Agriculture, because it does not include Agricultural production in its scope. It was also much lower than USEEM in mining, probably primarily because Census tends to classify all petroleum-related activities under manufacturing, rather than under mining.

However, the USEEM counts also exceeded the CBP counts in Construction, Manufacturing, Transport, communications and public utilities, and Wholesale trade, for no special reason. USEEM was slightly lower in Retail trade, 10 percent lower in Finance, insurance and real estate, and just over a third lower in Services. Looking in more detail at Services showed that although the USEEM establishment counts were very low in medical, legal, educational, and social services, the employment data were generally higher than CBPs. Indeed employment from the USEEM exceeded that from CBP for every industry except Retail trade, as expected because of their different definitions of 'employment'.

The DMI data are generally agreed to be fairly slow at picking up newly formed establishments, and at identifying establishments that have closed. About 45 percent of the newly appearing firms on the DMI already are over two years old, according to their Start year data. It was originally assumed that D&B would be prompt in dropping records for closed firms, and indeed, they are prompt for those for which they get credit inquires. However, if D&B gets no credit inquiry it often retains old records on the DMI until they eventually have a special project to purge records for closed businesses. This is partially dealt with by dropping all records more than five years old from the USEEM.

Calculations of employment change in branch establishments are imprecise, because of the frequent use of estimates for breaking out consolidated employment reports into multiple locations in both the CBP and the USEEM. Both these databases also have difficulties with firms sometimes changing the extent to which they break out employment of various secondary locations. Substantial numbers of firms do not report their March employment annually for all their secondary locations, so Census estimates it from the reported payroll data, using average payroll-per-employee ratios from similar firms that so report employment. Furthermore, multi-location firms with less than 250 employees are not surveyed every year to determine whether they have changed structure or ownership, but are on a survey schedule that promises coverage at least twice every five years. The DMI generally provided current data on the total employment of each firm and the employment at the headquarters and a number of locations, but not all, so the USEEM has proxy branches imputed with average employment for similar firms, and these cannot be used for analysis of change.

IV. BLS' prototype Longitudinal Establishment and Firm (LEF) data as pilot for BLS' Business Employment Dynamics (BED).

The Quarterly Census of Employment and Wages (QCEW) program at the Bureau of Labor Statistics (BLS), formerly known as the Business Establishment List (BEL) or the ES-202 program, obtains information on all employers that have employees covered by the State UI laws. Each quarter, the State Workforce Agencies edit and process the data and send the information to BLS in Washington, DC. These data items include the name, address, telephone number, monthly employment and quarterly wages, the date of initial UI liability, federal Employer Identification Number (EIN), and other information. The result is a quarterly universe list of UI-covered businesses that serve as the sampling frame for most BLS business statistics surveys. QCEW total employment counts also serve as the annual employment benchmark for the CES establishment survey and as the population base in business birth-death estimation (Spletzer et al, 2004).

Much of the design for the BED was developed from a series of special projects in BLS' Division of Occupational and Administrative Statistics investigating methods for linking their newly expanded BEL data over time and across States to construct longitudinal data at the establishment and firm level. Some further details on this work are reported in Armington 1995a and 1995b.

BED is currently producing establishment openings and closings counts along with associated employment changes (see http://www.bls.gov/bdm/home.htm). They have plans to produce business size data in the near future (Okolie, 2004).

A. Linking annual BLS Business Establishment List (BEL or ES-202) data.

The UDB numbers identifying establishments in the BEL are assigned during their processing at BLS. They are primarily based on the State-assigned UI account numbers, then on any State-reported unique predecessor account number, and finally on statistical matching of entering and exiting (apparently discontinuous) units. The statistical matching essentially applies rules that attribute continuity to an economic unit if it changes only one of the following three dimensions: its name or EIN, its location within a county, or its industry classification.

For most establishments the UDB number assignment system sufficed to provide the basis for matching data for each establishment for a range of years. New business formations are identified as the first appearance of a given UDB number, and a UDB number disappearance strongly suggests a closed business. Thus the details of defining a continuing economic unit are subsumed under the specifications for the system for assignment of UDB numbers.

The exception, which may lead to ambiguous changes, involves the State UI accounts that cover than one establishment (sub-units). Certain types of changes in the number of locations that are reported for an account make it impossible to determine whether the appearance of a new UDB number and the disappearance or substantial change in an old location are the result of administrative reporting changes, or the result of real economic changes – a formation or closure of a business location. In these cases the UDB number assignment system assumes discontinuity,

and therefore tends to inflate the turnover rate of these multi-location businesses. Records with these ambiguous discontinuities should be flagged so that the analyst can easily recognize them as possible breakouts or consolidations of continuing units.

B. Linking BEL data into enterprises using D&B data on ownership of multi-establishment enterprises to identify common ownership of EINs.

D&B's Corporate Affiliations File (CAF) contains descriptive data for all U.S. establishments recognized by D&B as being components of multi-establishment firms. Each establishment record also includes much information on the ownership structure and an identification number for the Ultimate owner, which serves as an enterprise identification number. By statistically matching a large proportion of the CAF records to records with matching descriptive data on the UDB, the EINs identifying legal entities in multi-location firms can be linked with the D&B firm identification numbers.

The AUTOMATCH software system was used to match each year of CAF and UDB data to construct linkages of the EINs denoting legal entities into enterprises (firms) that reflect the ownership structure of each time period. Using a combination of the reported federal EINs and the D&B-based EIN linkage to firms for each year, all of the affiliated UDB establishments were assigned Firm Identification numbers (FID). The characteristics of each multi-establishment firm are then determined by aggregation of the UDB establishment level data for their affiliated establishments. The resulting Firm Attribute Files for each year were used to append the appropriate firm attributes to each establishment record for affiliates of multi-establishments firms, thus creating a Longitudinal Establishment and Firm (LEF) file for each year.

In the 1992 prototype LEF file there were 4,510,000 single-location enterprises/establishments, making up 93 percent of the total number of firms. The other 321,700 firms included 424,400 legal entities (EINs), with 1,550,000 establishments and 55,284,000 employees. Thus the multi-location firms made up 7 percent of the firms, 9 percent of the legal entities, 26 percent of the establishments, and 63 percent of the employment of all private U.I.-covered businesses that year. Since the number of legal entities per multi-location enterprise was only around 1.3, it seems pretty safe to assume that most multi-location enterprises have only a single EIN, so they would be accurately represented by data aggregated to the EIN level. The remaining fraction of multi-location enterprises with multiple EINs are highly unlikely to include many legal entities that would be classified as small, when the total size of the enterprise is actually large. Thus it appears that legal entity size would make an effective proxy for firm-size for the purpose of classification into small and large firm categories.

As part of an exercise to validate some of the matching algorithms, BLS had D&B supply a sample of their DMI establishment and enterprise data for components of multi-location firms in Kansas that included the EINs that had been reported to D&B. These were duly matched to BEL establishment records, and it was found that the matches were generally credible, but that over a third of the EINs reported to D&B were for different legal entities than the ones filing U.I. tax returns in that State.

This phenomenon of businesses apparently using different EINs for different administrative agencies had been observed earlier, in the cooperative project with Census that matched new establishments entering the SSEL with the corresponding records in the BEL in order to improve Census' industry classification data. There BLS found that less than two-thirds matched on EIN. There are probably more recent summary data on this continuing project available from Census or BLS.

C. Evolution into BLS' Business Employment Dynamics (BED) data.

Many lessons learned during the construction of the prototype LEF were incorporated into the design for the QCEW and the BED. The QCEW incorporates many improvements in matching into its version of the prior Continuous Unit Match system, to improve the ability to track establishments by more consistent assignment of UDB numbers to establishment records that have changed State identification number. The issue of systematically delaying the incorporation of many annual changes (in industry classification especially) until the first quarter appears to have been resolved, probably by spreading out the changes over time, but flagging them for purposes of calculating quarter-to-quarter changes that might be affected by such reclassifications. The seasonal adjustment of gross employment change data accommodates not only seasonal employment patterns, but also any residual seasonal patterns in data processing by BLS or any of the State agencies that provide the quarterly data.

The problem of not being able to define business deaths on a contemporaneous basis has been cleverly resolved by the use of closings and openings, based on the finding of a single quarter of employment changing to, or from, zero. Spletzer et al (2004 p.41) explain:

> Businesses in the UI system are allowed to, and often do, report zero employment for several quarters after they have effectively closed. This undoubtedly occurs when a business owner temporarily shuts down, but anticipates starting up the business again when economic conditions improve. By reporting zero employment and wages on the quarterly contributions form, the business owner can keep his or her UI account active. Reporting zero employment results in many observed business closings, but which of these closings will start up again and which will die is not observed for several more quarters. …we can define both births and deaths in the historical data… A business birth is defined as an opening establishment that did not have any positive employment during the previous four quarters (thus differentiating seasonal openings from business births). Hence, births are a subset of openings. Likewise, a business death is defined as a closing establishment that does not have positive employment during the subsequent four quarters. Deaths, then, are a subset of closings.

They must have also found a way to solve the prior problem with some State agencies assuming that zero employment often implied only a late report, and therefore replacing the zeros with estimated employment base on prior employment.

D. Probable strengths and limitations of resulting extended BED database.

The BED's capacity to provide detailed, edited, comprehensive data on employment dynamics within 8 months of the end of each quarter is incomparable, and extremely valuable.

Aggregating by EIN to construct pseudo-firm size classification codes would make them even more valuable.

The data on openings and closings clearly overstate actual formations (births) and closures (deaths) of establishments by including in openings and closings the departures and re-entries of both seasonal businesses and other temporarily suspended businesses. Presumably, the seasonal adjustment process should mitigate the impact of the large number of seasonal businesses. In any case, most of such seasonal businesses will never even appear in the LEEM/BITS dynamic data, because Census has only their March employment, and assumes they are inactive if that is zero.

When BLS later provides birth and death data based on 4 quarters of zero employment, that will still overstate establishment births and deaths by about 10 percent, based on what was learned about the duration of temporary zero employment in the LEEM file. There it was found that the proportion of apparent deaths reviving was close to zero only if two years of no employment was required to define a death. About 10 percent of the deaths defined by one year of zero employment revived the following year.

Most of the changes from consolidated reporting units to specific locations of multi-unit firms took place before 1992, so the prior problem with tracking businesses across those transitions should be minor in subsequent years.

V. Census County Business Patterns (CBP) files used to create their SUSB and the Business Information Tracking System (BITS) for the SBA.

With support from the SBA Office of Advocacy, Census has edited its microdata from the CBP program and supplemented them with enterprise data from the Company Organization Survey (technically now part of their Business Register). Census now offers tabulations from these data, under the title, Statistics of U.S. Businesses (SUSB), and they have discontinued their Enterprise Statistics program. Most of these annual SUSB files have been linked over time to facilitate analysis of change at the establishment level with enterprise information. A subset of variables from these linked files for 1989 through 2001 is known as the Longitudinal Establishment and Enterprise Microdata (LEEM) file. The SBA Office of Advocacy usually refers to these data as the Business Information Tracking Series, or BITS. These are also used to provide standardized annual tables for the SBA, showing the distribution and changes in establishments and employment, classified by the size of the enterprise they belong to, for a variety of industries and geographic areas.

A. Enterprise microdata constructed from County Business Patterns (CBP) annual microdata supplemented by Company Organization Survey.

The Census Bureau's annual County Business Patterns (CBP) tabulations of private sector business establishments in the U.S. are based on specially edited microdata selected from Census Business Register (BR). The BR is a data system that systematically incorporates information on new businesses from the IRS Master Business File (for names and addresses of business tax filers) and IRS Form 941 (for payroll and employment reported with Social Security tax payments) for existing businesses. Refer to U.S. Department of Commerce, Bureau of the Census, 1979 for further details on the BR, which was based on their earlier SSEL. The Census File Number (CFN) used in the BR to uniquely identify each establishment also incorporates the identity of its owning firm if the establishment is part of a multi-location firm.

The CBP program selects data on establishments and their owning enterprises, including all private sector establishments (excluding railroads and farms) that had any payroll payments during the year. Some competitive government organizations, such as liquor stores and wholesalers, depository institutions and credit unions, and hospitals, are included. The key data are then edited to ensure consistency with the previous year's data for the same establishment, and supplemented with firm-level data calculated by aggregating all of the establishments belonging to each firm.

Establishment employment in the CBP microdata is derived from IRS form 941 filings. Employment should include full and part-time employees, salaried personnel, and persons on sick leave or vacation in the pay period of March 12. In the case of sole proprietorships and partnerships, this figure does not include proprietors or partners of the business. This figure also excludes all contractors and volunteers, but does include temporary employees unless they work for an employee leasing company. While reporting of both payroll and employment is mandatory on their 941 reports, the IRS does not put a lot of emphasis on the reporting of employment, and it is missing for 15 to 18 percent of the establishments. A higher proportion of

the larger EIN entities fail to report their employment, but employment data for most multi-unit establishments are collected by the Company Organization Survey (although not annually for enterprises with less than 250 employees). Other data must be imputed from payroll changes. This is either derived from the prior year's reported employment and payroll, or from the ratio of employment to payroll reported from similar businesses. Usually less than a quarter of the multi-location firms in the Business Register are surveyed in a given year so the CBP data for the remainder are estimated on the assumption that they are still active.

Values for the employment, payroll, receipts (available in Economic Census years), 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. The resulting SUSB file is tabulated to produce the annual SUSB static tables describing the business population. When a table cell covers very few businesses, the number of establishments may be provided, but any further information about those establishments will be suppressed to prevent disclosure of confidential information. See Armington (1998) for further details.

B. Linking consecutive years of SUSB establishment and enterprise data.

Most establishment records in an annual SUSB file can be linked to their corresponding record in the following year simply by matching on their CFN identification number. However, CFNs are changed whenever a business is sold, when it changes its legal form, or when it adds a secondary location (in the case of a single unit firm).

Census has used Permanent Plant Numbers and a complex assortment of alternative match techniques to construct a Longitudinal Pointer file to link establishment records for each establishment from the various years of SUSB files, so that surviving establishments can be identified even when a business changes its identification number (Moore and Trager, 1995). When linking two files, for records that do not have matching CFNs in the 2 files, matches are sought among their PPNs and then among their EINs, then using combinations of other attributes, and finally further matches are sought within each file in order to identify cases of mid-year reorganization of establishments of enterprises that may have caused changes in CFNs of continuing establishments. The Longitudinal Pointer file for each year contains that year's CFNs and up to 2 alternate CFNs for each, allowing for both midyear reorganizations and between-year changes.

C. Annual tables on the distribution of, and change in, businesses and their employment, by firm size, industry, and location.

A Composite SUSB file is constructed with two years of merged SUSB data, using the Longitudinal Pointer file to guide the merge. This Composite SUSB is then used to construct the annual BITS standard tables for the SBA Office of Advocacy (found on their web site, www.sba.gov/advo/stats). These include cross-sectional tables that show the distributions of establishments, firms, employment, annual payroll, and receipts (in Economic Census years) by size of firm and by either State, MSA, or industry class. The number of firms in each cell of these static tables is the number of qualifying single-location firms plus the number of different

multi-location firms that have any qualifying locations. Therefore if tabular firm data for two states or two industries were added together, the sum would double-count every firm that operated in both states or both industries. The firm-size is always determined by the firm's total nationwide employment in all covered industries.

The standard tables of dynamic data focus on businesses that are have employees in the first quarter of each year – they actually measure March-to-March changes in employment. Thus new businesses that hire their first employees after the first calendar quarter will not be recognized until the following year, when they also have better data for classification purposes.

Each dynamic table includes static data describing the initial distribution of the covered businesses, combined with dynamic data on how those establishments changed (expanded, contracted, or closed) before the next year, and on new business formations. For each type of change, the net and gross changes in both employee and establishment numbers are provided. Establishments are classified by their characteristics at the beginning of the period, except for new firm births, for which data exist only for the end of the period. In addition to firm-size, these dynamic change tables also classify establishments by their industry or geographic location, and by their enterprise type (single- or multi-location).

D. Specification of BITS/LEEM files as user-friendly subsets of SUSB, for analysis at Census CES.

To facilitate longitudinal analysis of the microdata from the SUSB files, the Office of Advocacy of the SBA has been contracting with Census to construct multi-year composite files for use by researchers at Census' Center for Economic Research (CES). These Longitudinal Enterprise and Establishment Microdata (LEEM) files, which are also called Business Information Tracking System (BITS) files, were designed to include all of the relevant available information about each establishment that was likely to be needed for a wide variety of research projects.

Each LEEM record represents an establishment, and includes the Start year of the establishment plus data from the Longitudinal Pointer file and from each annual SUSB file. The annual information includes the establishment's Census File Number, Standard Industrial Classification, state, county, MSA, enterprise employment, establishment employment, and annual payroll in thousands. The LEEM for 1989-1998, for instance, included 12,377,530 records, with 99 variables in each, requiring 3.9 billion bytes for storage.

Most researchers would want to start by selecting only the subset of records needed for their research design, with only the variables needed for either the research itself or the later disclosure processing necessary to determine that any results needed for publication, or even for academic discussion, are not confidential under the rules of the Census Bureau and the IRS. Data from other Census programs may be linked to the LEEM data by CFN to enrich the database for special analyses.

Alternatively, if the desired subset of data is already well-defined, and the research can be done with aggregate data, the construction of the required tables from the LEEM may be specified by the researcher and programmed and produced under a contract arrangement with Census. This

latter arrangement is far less costly than access to the LEEM microdata at CES. It also produces a reusable aggregate database that is not further subject to confidentiality restrictions, and therefore does not require the investigator to comply with the regulations on work at CES.

VI. Additional approaches to providing further data by size of firm

A. Statistical options for extending time series of aggregate data across different databases.

In spite of the many underlying weaknesses and eccentricities with each of the data sources that have been explored, the pictures that they provide of the distribution, growth, and turnover of small and large businesses in the U.S. is remarkably similar. Table 1 compares data on the numbers of firms in various size-classes from Census' SUSB and D&B-based USEEM, along with the number of establishments in Census' CBP over the period from 1976 to 2002. Comparable data for the numbers of employees are shown in Table 2, both of which were assembled by the Office of Advocacy of the SBA. Although these data sources differ in many details, their similarities have supported similar conclusions during the last quarter of a century of research on small business dynamics.

Research using firm-size data would benefit greatly from time series data covering longer periods, especially for analysis of policy impacts and business cycle effects. This could be achieved by combining the older USEEM and USELM aggregate data with either the data from the existing SUSB tables, or with specially tabulated BITS data, to facilitate more detailed analysis of patterns of change since 1976. While the data from the Census Bureau's SUSB have become the gold standard for firm size data, they are only available from 1988 onwards. If the historic bi-annual firm-size data from the USEEM and USELM were adjusted to better match the annual establishment size data from Census' County Business Patterns and/or Census' Enterprise Statistics, then these new estimated data could be concatenated with SUSB data to create annual firm, establishment, and employment data from 1976 to 2002. With completion of this project, at least 25 years of aggregate data on the U.S. business population, classified by firm size, would be available for more detailed analysis of industrial organization, entrepreneurship, and labor studies.

Two relatively simple approaches to the use of the combined data for statistical analyses of the 25-year time series data are suggested here, followed by a brief discussion of two more ambitions approaches.

If research questions can be structured so that only data on short-term changes (one or two year changes) are needed to test the hypotheses, then USELM-based changes for 1976 through 1988 can easily be combined with SUSB changes from 1988-current year (now 2002). If each change is measured relative to the stock during the same period (beginning, end or mean), then most inconsistencies between the two data sources will not be relevant. The underlying differences between the SUSB and USELM datasets are unlikely to affect most measures of business and employment distributions or rates of change, such as business birth and death rates, employment growth and shrinkage rates, employment shares in firms of various sizes or industries, and so forth. Therefore many research questions can be investigated by analysis of such measures of short-term change in business populations or employment, or changes in the distribution of firms, establishments, or jobs.

An example of this approach is the following analysis of differences in the impact of cyclical factors on short-term rates of growth in employment in large and small businesses. There was

much discussion in the 1980's both in the popular press and among researchers about 'the small business share of job growth.' It was popularly assumed that this share would be relatively constant over time, so that any finding of differences in the share of total job growth attributed to small businesses was taken as an indication of measurement errors. However, by 1987, when the SBA Office of Advocacy had constructed USELM data for measurement of employment dynamics by firm-size for 1976 through 1986, it seemed pretty clear that small and large businesses reacted differently to cyclical variations in the U.S. economy, and that these differences caused great volatility in the relative growth rates of small and large businesses. But the five data points available from these ten years of data for even years were not sufficient to support a robust statistical analysis of these differences.

Since the job-growth rate for each period of available data can be calculated completely within each of the separate databases, and these growth rates will not be substantially affected by the relatively small differences in coverage of the separate databases, we can simply concatenate the respective growth rates from each to create time series of growth rates from 1976 to 2000 for large and small firms. One must carefully transform the explanatory variables to correspond appropriately to the one or two-year time periods over which the employment growth was measured, and equal care is needed to correctly specify any lagged variables, to correspond to the various unequally timed observations. Regression analysis can then be used to estimate the impact of lagged macroeconomic variables such as GNP growth, increases in the productivity of private labor, and changes in the size of the labor force on employment growth rates by firm size. We would expect that increases in GNP or in labor force would lead to greater employment, and increases in labor productivity would lead to less employment.

Table 3 shows the results of this regression analysis in terms of standardized betas, so that we can directly compare the sizes of the parameters estimated for different explanatory variables for the various sizes of firms. The expected signs were found only for growth of businesses with at least 500 employees, and the strongest of that group's relationships by far was that of employment growth with GDP growth. For smaller businesses, the relationship with GDP growth is small and negative, and labor productivity exerts a positive effect on growth rates of small firms during the last quarter of the twentieth century. By far the strongest relationship for small firm growth rates was that with growth in the labor force in the prior period, suggesting that a surge in the labor force may subsequently promote both small firm hiring and faster rates of new firm formation. As also expected, the correlations among these variables show that there is considerable collinearity, but the regressions were replicated with multiple subsets of the variables, and with a D&B dummy, and in both linear and logarithmic form, and the results were quite consistent. This analysis could easily be extended to focus on whether new firm formation is substantially responsible for this relationship, and to see whether similar relationships hold within most industrial sectors during this period.

A second approach is to merge, or append, the two databases, and use one or more dummy variables in a regression analysis framework to estimate the impact (on levels or slopes) of the differences in the two databases, and to isolate those differences from the relationships that you are testing for.

For primarily descriptive data for comparison over the longer term, there are two other approaches to choose from. The first of these is to adjust one series to be more similar to the other. To the extent that data can be found or estimated to measure known differences between the USEEM and the SUSB-based data, these can be used to adjust particular data series from the USEEM data to match the comparable series from the BITS data for the purpose of investigating a particular research question. Thus we could adjust the USEEM establishment counts (from D&B) by subtracting out the numbers and employment of establishments coded as farms, so that the remaining agricultural sector data would more closely match that of the BITS. However, unless we also had specific data on the numbers of farm births and deaths, and job creation and destruction from farms from the USELM data we could not make parallel adjustments to the employment dynamics data. Another important difference in coverage of these two databases is the inclusion of more government-owned enterprises in the D&B-based USEEM. No data are available to quantify how much employment these add to the USEEM.

The final alternative is the most complex, and the SBA Office of Advocacy doubts that it would be cost-effective. Some or all of the available aggregate D&B-based data from the USEEM and/or USELM could be used in combination with historical data from CBP and Enterprise Statistics to construct a simulated SUSB historical aggregated database. This is the most labor-intensive approach to the combination of the two databases – construction of a complete new database composed of detailed estimated annual or biannual aggregate data for the years from 1976 through 1988, combined with either the existing aggregate SUSB data for 1988 through 2002, or respecified newly aggregated data from the SUSB from 1988 to the present. This might also be though of as constructing simulated aggregate 1976-1988 data to concatenate with later SUSB data, with the simulated data based on actual SUSB data for 1988-1990, Enterprise Statistics data for 1972-1992, Country Business Patterns data for 1972-1992, and USEEM/USELM data for 1976-1988.

The appropriate choice of methods for rebasing and annualizing the USEEM and USELM aggregate data for the years before SUSB data are available is very sensitive to the intended research purpose of the estimated data, since any estimation method is built on specific assumptions, which must be consistent with the planned research and not affect the outcome significantly.

B. Utility of a Public Use Sample of unidentified microdata, or a synthetic microdata file.

Over the years of frustration with data that were imperfect, late, incomplete, and/or inaccessible, it has often been suggested that the state of knowledge about small and new business would be greatly advanced much sooner if some sort of representative microdata could be made generally available. This sample could be used for flexible tabulation of timely data for policy analysis, for sophisticated research projects, and for training more graduate students to deal with the concepts of business dynamics and the technicalities of microdata analysis.

Aside from the usual challenges of funding and approvals for such a project, there are a number of other challenges that have discouraged prospects for progress in this direction.

1) The sample would have to be updated at least annually in order to effectively track the progress (or demise) of small and new businesses.

2) The sample could not include specific data on both industry detail and geographic detail while avoiding disclosure problems.
3) The sample could not include much specific data on very large firms, or on publicly owned firms, without risking disclosure by matching to commercial databases.
4) Details of the sample design and content will pre-determine the range of research projects that it is suitable for.

As a basis for discussion, a general design approach that might be feasible and useful within this context could include the following features:

a) Sample S would provide detail on industry sectors, while Sample G would provide geographic detail, perhaps at the Labor Market Area level, which has proven generally useful for a wide variety of analyses.
b) Each sample would be selected on the basis of chosen digits in EINs from a comprehensive microdata file such as BLS' QCEW, with employment and enterprise data updated annually, and new establishments added to the sample.
c) Selection by EIN would allow recalculation of pseudo-firm size within the sample.
d) Actual data, or minimally randomly disturbed data, would be used for establishments with less than X employees, where X might be around 20. More vigorously disturbed data might be required for establishments with employment from X to Y, where Y might be 500. All establishments with more than Y employees would be represented by actual averages of all qualifying similar establishments, where the definition of similar includes a breakout of growing, shrinking and stable (within a small percentage) employment, so that gross employment changes can still be calculated from this small population with no sampling error.
e) While the population of tiny businesses is huge, their behavior is quite diverse in many respects, so a 1 percent sample of all establishments with employment under Y would probably be necessary to adequately represent them with either industry or geographic detail.

After the original setup of the sample design and approval, and software development and testing costs, such a pair of samples from an existing administrative dababase should serve as a relatively inexpensive permanent resource of data on U.S. businesses and their dynamics.

C. Other Specialized data on Small Businesses

The Federal Reserve Board's Survey of Small Business Finances (SSBF) development offers a sample of cross-sectional data on small businesses.[3] The universe for the SSBF is nonfarm, nonfinancial firms with fewer than 500 employees and data are now available for the years 1998, 1993, and 1987. SSBF collects financing and firm characteristic information for different demographic types of business owners.

[3] See http://www.federalreserve.gov/pubs/oss/oss3/nssbftoc.htm for details on the SSBF program and note that the SSBF was originally titled the National Survey of Small Business Finances.

The program can be viewed more as a microdata source for researchers and less as a source of tables for the casual observer. As such, the data has been used in numerous small business financing research studies, such as the impact of bank consolidation, loan discrimination, financial constraints, trade credit and equity financing.[4]

The Fed utilized D&B's Dun's Market Identifier (DMI) to select a sampling frame and contracted out telephone data collection for the survey. (Note that the first survey included a sample of 400 firms with SBA-guaranteed loans and was partially funded by the SBA Office of Advocacy). The Fed did recognize that the DMI under-represents very new firms and sole proprietors, who are often nonemployers (Cox et al, 1989). Stratifying groups, oversampling and reweighting the data were all needed to construct a useful dataset from this sample. In addition, they used focus groups, pretested the survey instrument and mailed an information package prior to telephone contact.

Before they could conduct telephone surveys, they found that they needed to pre-screen the companies selected for the sample to clean up the raw information from the DMI, as it was often incomplete or out of date. Call screening within firms made contact with the actual business owners difficult, the sensitive nature of business finances made getting responses to questions difficult and numerous calls for each business were often needed to get useable responses.

Time comparisons among the different SSBF periods show that collecting data is an increasing difficult task. Even after increased collection efforts, the response rate dropped from 52 percent for the 1992 survey to 33 percent for the 1998 survey (Hagerty et al, 2001). Since the SSBF is not longitudinal in nature, this did not create issues associated with tracking firms through time.

Overall the SSBF data collection experience re-enforces the earlier lessons that massive efforts that must be undertaken to obtain an accurate business dataset. Data derived from commercial business lists must be confirmed and a laborious multistage data collection process must be undertaken to obtain even minimal response rates.

D. New Data on Firm Formations and Lifecycles

With the quality and availability of firm size data vastly improved over the last quarter century, we are beginning to realize how rapidly the population of American businesses is changing, and has probably always been changing throughout history. Although the data sources mentioned in this paper put great effort into finding new firms promptly, determining when new firms close, dealing with unreported data, and identifying mergers and spin-offs, they still imperfectly represent the universe of firms that is their target. In particular, researchers have increasingly come to realize that much of the distinctive behavior of dynamic small firms is more closely associated with the youth of many of them than with their size. It appears that the weakest aspect of most of the existing data is its inability to adequately track the beginning of the life

[4] See http://www.federalreserve.gov/pubs/oss/oss3/abstract.html for a bibliography of studies.

cycle of new firms. Thus the next horizon is earlier capture of new firms, and better data on firm age. This missing link will help researchers and policymakers better understand the business lifecycle (firm growth and industry evolution) for innovation, competition, regulatory, finance, employment and social reasons.

Two encouraging projects have recently been undertaken to provide better data for analysis of business lifecycle issues. One is creating a longitudinal database from the U.S. Census Bureau's nonemployer data, and the other is Kauffman Foundation's efforts to create a longitudinal database covering private sector businesses, that would be widely available for microdata research.

Researchers Rick Boden, University of Toledo and Al Nucci, U.S. Census Bureau, are striving to add to our knowledge of the smallest firm size, nonemployers. When complete, the U.S. Census Bureau will be able to annually produce entry and exit figures for individuals involved in nonemployer businesses, as well as figures for nonemployers growing into employers. Their research paper offering details for this data effort was not yet complete, but when finished will be available at www.ces.census.gov/ces.php/papers.

The Kauffman Foundation has recently discussed creating a survey of new firms using listings from D&B sources and conducting follow-up surveys over time to track their progress. In addition to the standard variables on firms that are already tracked in current government longitudinal data, this new project would gather expanded information on the characteristics of each surveyed firm. This database would be available to researchers for conducting microdata research on firms' life cycles.

The problems of capturing a representative group of firms, and of maintaining an adequate sample by continuously adding new firms at their formation, will be a big challenge. Even if only large firms were to be represented in a sample, the dispersion of growth among firms listed in the stock market over the last century, and the turnover in the list of active firms, indicate that this would be a very difficult universe to sample adequately. Let us hope that the lessons learned in past projects to provide longitudinal firm data will provide a basis for improvements in the future.

References

Acs, Z. and C. Armington, 1998, "Longitudinal Establishment and Enterprise Microdata (LEEM) Documentation," Center for Economic Studies, U.S. Bureau of the Census, CES Discussion Paper 98-9.

Acs, Z., C. Armington, and A. Robb, 1999, "Measures of Job Flow Dynamics in the U.S. Economy," Center for Economic Studies, U.S. Bureau of the Census, CES Discussion Paper 99-1.

Applied Systems Institute, 1987, 'The 1986 USEEM and the Weighted Linked 1976-1986 U.S. Establishment Longitudinal Microdata (USELM): Phase 1 Final Report, Contract No. SBA-2037-OA-87,' Office of Advocacy, U.S. Small Business Administration, Washington DC.

Armington, C, 1998, "Statistics of U.S. Business- Microdata and Tables of SBA/Census Data on Establishment Size," Office of Advocacy, U.S. Small Business Administration, Washington DC.

Armington, C., 1995a, "Deriving Establishment Births from Unemployment Insurance Data," Proceedings of the 1995 Annual Research Conference, Bureau of the Census, U.S. Dept. of Commerce.

Armington, C., 1995b, 'Development of a Longitudinal Establishment and Firm Database at the U.S. Bureau of Labor Statistics,' Techniques and Uses of Enterprise Panels, Office for Official Publications of the European Communities, Luxembourg, pp. 74-78.

Armington, C., 1989, Differences in Business and Employments Data: Alternative Measures of the U.S. Wholesale Trade Sector in 1982, Office of Advocacy, U.S. Small Business Administration.

Armington, C., 1988a, Uses and Limitations of USEEM/USELM Data, Office of Advocacy, U.S. Small Business Administration.

Armington, C., 1988b, Users Guide to the 1976-1986 Linked USEEM, Office of Advocacy, U.S. Small Business Administration.

Armington, C., 1987,"Removing Business Restructuring from Data on Startups and Closures," 1986 Proceedings of the Business and Economics Statistics Section, American Statistical Association, pp. 265-269.

Armington, C and M. Odle, 1988a, Uses and Limitations of USEEM / USELM Data, Office of Advocacy, U.S. Small Business Administration, Washington DC.

Armington, C and M. Odle, 1988b, 'Users' Guide to the Linked 1976-1986 USEEM, Contract No. SBA-2037-OA-87,' Office of Advocacy, U.S. Small Business Administration, Washington DC.

Armington, C and M. Odle, 1981, "Associating Establishments in Enterprises for a Microdata File of the U.S. Business Population," Statistics of Income and Related Administrative Record Research, Internal Revenue Service, U.S. Dept. of the Treasury, GPO, pp. 71-77.

Armington, C and M. Odle, 1975, Research on Microdata Files Based on Field Surveys and Tax Returns: Minimizing a Distance Function to Create a Large Micro-Economic Data Sample, , Working Paper No. 1, Brookings Institution Tax Project.

Armington, C., and A. Robb, 1999, "An Investigation into the Volume and Impact of Business Acquisition Activity in the United States: 1990-1994," Office of Advocacy, U.S. Small Business Administration, Washington DC.

Armington, C. and A. Robb, 1998, "Mergers and Acquisitions in the United States: 1990-1994," Center for Economic Studies, Bureau of the Census, CES Discussion Paper 98-15.

Berney, R. E., 1982, Developing a Microdata Base to Support Research on Small Business Issues," Review of Public Data Use 10:167-178.

Berney, R. E., 1978, "Report on the Small Business Data Needs Workshop," Office of Advocacy, U.S. Small Business Administration, Washington, D.C.

Birch, D. L., 1979a, "The Job Generation Process," M.I.T. Program on Neighborhood and Regional Change, Cambridge MA.

Birch, D. L., 1979b, "Using the Dun and Bradstreet Data for Micro Analysis of Regional and Local Economies," M.I.T. Program on Neighborhood and Regional Change, Cambridge MA.

Birch, D. L., 1981, "Corporate Evolution: a Micro-based Analysis," Cambridge, Mass.: MIT Program on Neighborhood and Regional Change.

Birch, D. L. and S. MacCracken, 1982, "The Small Business Share of Job Creation: Lessons Learned from the Use of a Longitudinal File," Cambridge, Mass.: MIT Program on Neighborhood and Regional Change.

Brown, S.H. and B. Phillips, 1989, "Comparison Between Small Business Databases (USEEM) and Bureau of Labor Statistics (BLS) Employment Data: 1978-1986," *Small Business Economics*, 1, (4), 273-284.

Cox, Elliehausen, and Wolken, 1989, "The National Survey of Small Business Finances: Description and Preliminary Evaluation," Finance and Economics Discussion Series, Federal Reserve Board, November.

Haggerty, Grigorian, Harter, and Stewart, 2001, "The 1998 Survey of Small Business Finance: Methodology Report," Federal Reserve Board of Governors, April.

Hirschberg, D.A., and B.D. Phillips, 1982., 'Using Financial Statement Data to Evaluate the Status of Small Business,' Statistics of Income and Related Administrative Record Research: 1982, Internal Revenue Service, U.S. Dept. of the Treasury, GPO, pp. 71-73.

Kirchhoff, B.A. and D.A. Hirschberg, 1980, "Small Business Data Base," Office of Advocacy, U.S. Small Business Administration, Washington, D.C.

MacDonald, B. and C. Armington, 1991, 'Exact Matching for Constructing BLS' Enterprise Register,' Proceedings of Fifth International Roundtable on Business Survey Frames, October 1990, U.S. Bureau of the Census.

MacDonald, J.M., 1985, 'Dun & Bradstreet Business Microdata Research Applications and the Detection and Correction of Errors,' Journal of Economic and Social Measurement, 13:173-185.

Moore, R. and M. Trager, 1995, "Development of a Longitudinally-Linked Establishment Based Register: March, 1993 Through April, 1995". Presented at the Joint Statistical meetings of the American Statistical Association.

Okolie, C., 2004, "Why size class methodology matters in analyses of net and gross job flows," Monthly Labor Review, July, pp. 3-12.

Popkin, J., 1980, "Developing a Micro and Macrodata Base for Small Business Research," Interim Final Report, SBA 2624-OA-79.

Robb, A., 1999, "New Data for Dynamic Analysis: The Longitudinal Establishment and Enterprise Microdata (LEEM) File," Center for Economic Studies, U.S. Bureau of the Census, CES Discussion Paper 99-18.

Spletzer, J.R., R.J. Faberman, A Sadeghi, D.M. Talan, and R.L. Clayton, 2004, "Business employment dynamics: new data on gross job gains and losses," Monthly Labor Review, April, pp. 29-42.

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

U.S. Department of Commerce, 1975, Historical Statistics of the United States, Colonial Times to 1970, Part 2, Washington, DC

U.S. Small Business Administration Office of Advocacy, 2000, "New Data for Dynamic Analysis: The Business Information Tracking Series."

U.S. Small Business Administration Office of Advocacy, 1998, "Mergers and Acquisition in the United States, 1990-1994," October.

U.S. Small Business Administration Office of Advocacy, 1994, Handbook of Small Business Data, U.S. Gov't Printing Office, D.C.

U.S. Small Business Administration Office of Advocacy, 1988, "Uses and Limitations of USEEM / USELM Data," Revised November.

U.S. Small Business Administration Office of Advocacy, 1986, "The Small Business Data Base: A User's Guide," June.

U.S. Small Business Administration Office of Advocacy, 1985, "The Development of the Small Business Data Base of the U.S. Small Business Administration: A Working Bibliography," March.

U.S. Small Business Administration Office of Advocacy, 1984, "Constructing a Business Microdata Base for the Analysis of Small Business Activity," November.

Table 1: Number of Businesses by Business Size, 1976 - 2002

	SUSB Firms			D&B-based USEEM Firms			CBP Establishments		
	Total	<20 empl.	<500 empl.	Total	<20 empl.	<500 empl.	Total	<20 empl.	<500 empl.
2002	NA	NA	NA	NA	NA	NA	7,200,770	6,198,512	7,182,661
2001	5,657,774	5,036,845	5,640,407	NA	NA	NA	7,095,302	6,083,043	7,076,000
2000	5,652,544	5,035,029	5,635,391	NA	NA	NA	7,070,048	6,068,523	7,050,971
1999	5,607,743	5,007,808	5,591,003	NA	NA	NA	7,008,444	6,035,988	6,990,146
1998	5,579,177	4,988,367	5,562,799	NA	NA	NA	6,941,822	5,990,738	6,924,085
1997	5,541,918	4,958,641	5,525,839	NA	NA	NA	6,894,869	5,968,233	6,877,784
1996	5,478,047	4,909,983	5,462,431	NA	NA	NA	6,738,541	5,842,606	6,721,844
1995	5,369,068	4,807,533	5,353,624	NA	NA	NA	6,613,218	5,732,778	6,596,856
1994	5,276,964	4,736,317	5,261,967	NA	NA	NA	6,509,276	5,661,525	6,493,700
1993	5,193,642	4,661,601	5,179,013	NA	NA	NA	6,403,367	5,577,433	6,388,129
1992	5,095,356	4,572,994	5,081,234	NA	NA	NA	6,317,690	5,506,581	6,302,868
1991	5,051,025	4,528,899	5,037,048	NA	NA	NA	6,200,650	5,392,152	6,185,764
1990	5,073,795	4,535,575	5,059,772	NA	NA	NA	6,175,563	5,354,342	6,160,389
1989	5,021,315	4,493,875	5,007,442	NA	NA	NA	6,107,413	5,305,307	6,092,487
1988	4,954,645	4,444,473	4,941,821	4,004,743	3,484,812	3,988,595	6,018,600	5,246,381	6,004,638
1987	NA	NA	NA	NA	NA	NA	5,937,060	5,186,099	5,923,538
1986	NA	NA	NA	3,868,553	3,377,315	3,853,860	5,806,973	5,081,627	5,793,680
1985	NA	NA	NA	NA	NA	NA	5,701,485	5,000,492	5,688,315
1984	NA	NA	NA	3,809,502	3,346,766	3,795,669	5,517,715	4,849,103	5,504,822
1983	NA	NA	NA	NA	NA	NA	5,306,787	4,686,051	5,294,633
1982	NA	NA	NA	3,663,203	3,223,368	3,649,850	4,633,960	4,013,272	4,621,372
1981	NA	NA	NA	NA	NA	NA	4,586,510	3,971,979	4,573,824
1980	NA	NA	NA	3,584,598	3,151,270	3,571,328	4,543,167	3,931,892	4,530,222
1979	NA	NA	NA	NA	NA	NA	4,535,653	3,929,109	4,522,481
1978	NA	NA	NA	3,610,689	3,209,967	3,598,784	4,409,223	3,839,164	4,396,787
1977	NA	NA	NA	NA	NA	NA	4,352,295	3,832,610	4,340,689
1976	NA	NA	NA	3,333,777	2,971,549	3,322,300	4,142,809	3,648,513	4,131,439

Notes: SUSB and CBP figures exclude farms. D&B figures include some non-employers.

Sources: U.S. Census Bureau / U.S. Small Business Administration joint project, and USEEM.v8 for 1976-84, v9 for 1986-88.

Table 2: Employment by Business Size, 1976 - 2002

	SUSB Firms			D&B-based USEEM Firms			CBP Establishments		
	Total	<20 empl.	<500 empl.	Total	<20 empl.	<500 empl.	Total	<20 empl.	<500 empl.
2002	NA	NA	NA	NA	NA	NA	112,400,654	28,116,153	89,552,667
2001	115,061,184	20,602,635	57,383,449	NA	NA	NA	115,061,184	27,680,970	90,928,175
2000	114,064,976	20,587,385	57,124,044	NA	NA	NA	114,064,976	27,569,325	90,451,895
1999	110,705,661	20,388,287	55,729,092	NA	NA	NA	110,705,661	27,288,714	88,188,738
1998	108,117,731	20,275,405	55,064,409	NA	NA	NA	108,117,731	27,130,586	86,437,139
1997	105,299,123	20,118,816	54,545,370	NA	NA	NA	105,299,123	26,883,041	84,506,779
1996	102,187,297	19,881,502	53,174,502	NA	NA	NA	102,198,864	26,115,321	81,898,388
1995	100,314,946	19,569,861	52,652,510	NA	NA	NA	100,334,745	25,784,823	80,351,496
1994	96,721,594	19,195,318	51,007,688	NA	NA	NA	96,733,300	25,372,516	77,558,013
1993	94,773,913	19,070,191	50,316,063	NA	NA	NA	94,789,444	25,233,111	75,871,011
1992	92,825,797	18,772,644	49,200,841	NA	NA	NA	92,800,870	24,999,753	74,256,007
1991	92,307,559	18,712,812	49,002,613	NA	NA	NA	92,301,543	24,482,212	73,756,481
1990	93,469,275	18,911,906	50,166,797	NA	NA	NA	93,476,087	24,373,021	74,712,170
1989	91,626,094	18,626,776	49,353,860	NA	NA	NA	91,631,203	23,992,301	73,207,699
1988	87,844,303	18,319,642	47,914,723	98,887,128	17,625,130	48,454,276	87,881,632	23,582,752	70,819,691
1987	NA	NA	NA	NA	NA	NA	85,483,378	23,068,645	68,904,748
1986	NA	NA	NA	91,678,252	16,897,707	45,729,012	83,380,465	22,295,991	56,866,736
1985	NA	NA	NA	NA	NA	NA	81,119,257	21,809,879	54,759,126
1984	NA	NA	NA	85,367,675	16,427,726	43,598,391	77,995,566	21,170,742	61,968,283
1983	NA	NA	NA	NA	NA	NA	72,971,318	20,135,382	57,735,109
1982	NA	NA	NA	82,470,054	15,624,912	41,409,476	74,297,252	19,898,017	58,485,667
1981	NA	NA	NA	NA	NA	NA	74,850,402	19,514,851	58,723,517
1980	NA	NA	NA	81,835,455	15,122,499	40,621,686	74,835,525	19,423,092	58,430,322
1979	NA	NA	NA	NA	NA	NA	74,681,388	19,405,548	57,925,086
1978	NA	NA	NA	74,823,166	14,841,000	38,006,377	70,289,236	18,723,358	54,662,640
1977	NA	NA	NA	NA	NA	NA	64,975,580	17,605,043	50,375,627
1976	NA	NA	NA	68,802,494	13,680,289	34,581,991	62,647,846	--	--

Notes: County Business Pattern data is available annually back to 1964. SUSB and CBP figures exclude farms.

Sources: U.S. Census Bureau / U.S. Small Business Administration joint project, and USEEM.v8 for 1976-84, v9 for 1986-88.

Table 3: Firm size differences in accounting for employment growth variation, 1976-2000

Estimated coefficients from least squares logarithmic regressions
All explanatory variables represent lagged annual (or 2-year) changes
(standardized betas with t-ratios below, significant at .05 if bold)

Firm size classes:	Under 20 employees		Under 500 employees		500+ employees	
Adj. R squared	0.578	0.493	0.547	0.578	0.796	0.780
Labor force	**0.929**	**0.726**	**0.693**	**0.620**	0.290	**0.398**
	3.36	3.95	2.42	3.62	1.51	2.18
Labor productivity	**0.485**		0.436	**0.389**	-0.218	
	2.18		1.90	2.27	-1.41	
Gross domestic product	-0.403		-0.110		**0.767**	**0.562**
	-1.23		-0.32		3.36	3.08

n=16

Data for employment growth by size of firm tabulated for the SBA's Office of Advocacy from the USELM database for 1976-1988 and the SUSB database at Census for 1989-2000.