The U.S. Bureau of Labor Statistics Longitudinal Establishment Database

Michael Searson, Kenneth Robertson, Richard L. Clayton; all of the U.S. Bureau of Labor Statistics

ABSTRACT

In this paper we describe the key processes used in the construction of the Bureau's Longitudinal Database (LEDB). This database serves several functions: as a business register; to conduct longitudinal studies of businesses and employment decline and growth; and as a rich and comprehensive source of information on employment and wages. In addition to a review of the overall program, data sources and outputs, we describe the following topics: the Annual Refiling Survey; the processes required to implement an industrial classification change; recent improvements in the record linkage methodology utilized to identify and maintain the continuity of establishments over time; and, the implementation of Permanent Random Numbers for sampling.

Key Words : Business Register, Classification Systems, Record Linkage, Permanent Random Numbers

1. Background

The Longitudinal Establishment Database (LEDB) is one of the outputs of the Bureau of Labor Statistics' (BLS) Covered Employment and Wages (ES-202) program. The ES-202 program is a Federal/State statistical program managed by the BLS and the State Employment Security Agencies (SESAs) in the 50 States, the District of Columbia, Puerto Rico, and the Virgin Islands. The LEDB is used as the sampling frame for most of the BLS' programs. Nationwide, in 1999, Federal and State Unemployment Insurance (UI) programs covered 127.0 million full- and part-time workers who received over $4.23 trillion in pay. ES-202 program data were collected for approximately 7.8 million employer establishments in 1999. The LEDB is the most complete and timely source of business establishment information used as a sampling frame in the United States.

The SESAs conduct four principal ES-202 program activities: data collection; assigning initial industrial and geographical codes for new employers; periodic review of these codes including review of the business identification information (names, addresses, etc.); and, micro/macro data editing. The SESAs submit micro-level data (Enhanced Quarterly Unemployment Insurance File-EQUI) to the BLS each quarter. The EQUI file is the primary input to the LEDB.

The EQUI file includes the following information for each active employer subject to Unemployment Insurance (UI) coverage during the reported quarter: State UI Account Number (UIN), Establishment Reporting Unit Number (RUN), Federal Employer Identification Number (EIN), four-digit Standard Industrial Classification (SIC) code, six-digit North American Industrial Classification System (NAICS) code, county/township codes, employment for each month during the quarter, total quarterly wages, and the establishment's name(s), addresses and telephone number. Known predecessor and successor relationships are also identified by UI Account Number and establishment Reporting Unit Number (UI/RUN). These numbers are used as administrative codes for matching records from one quarter to the next. The State code, UIN and RUN allow each establishment to be uniquely identified. Imputed employment and wage data are assigned specific codes to distinguish them from reported data. Codes are placed on the records to identify the type of address (i.e., physical location, mailing address, corporate headquarters, address on UI tax file, or "unknown"). The database can store up to four different addresses for each establishment.

2. Overview of the program : Data sources, editing, outputs and uses

2.1 LEDB Data Sources

2.1.1 The EQUI file

As stated above, the EQUI file is a product sent by the SESAs to BLS each quarter which contains data from the administrative files of the State UI programs. These files are supplemented by data from the Multiple Worksite Report and the Report of Federal Employment and Wages, both are described in section 2.1.5.C and 2.1.5.D.

BLS, a user of UI program data, does not control the procedures and processes SESAs use for collecting and maintaining the UI administrative and accounting data. BLS contracts with each State's LMI unit to provide selected UI data for statistical purposes. The contract, called the LMI Cooperative Agreement, specifies the quality standards and the types and timing of data submitted for the five BLS Federal/State statistical programs.

All of the ES-202 program data originates from employers. Each State's UI program collects and maintains quarterly administrative record data and business identification information on employers who are subject to State UI laws. The UI program provides these micro, or employer-level, data to the LMI unit of the respective SESAs. The LMI unit then supplements, edits, and processes these data for use in the ES-202 program. Included in these administrative data are the registration information that is obtained from every new employer.

All new employers are required to file a "Status Determination Form" (see section 2.1.5.A) to obtain a UI account number and UI tax rate. Information from this form is also used to assign to each establishment a set of classification codes (industrial, geographical, ownership, and auxiliary). All covered employers are also required to file a "Quarterly Contribution Report (QCR)" (see section 2.1.5.B). The employers must report the total wages paid to all workers during the calendar quarter; the wages that are subject to UI taxes; and, the UI taxes that are due to the State.

The LMI unit's activities are centered on four primary areas from the UI administrative files. First, the LMI staff uses information collected on the Status Determination Form to assign the classification codes previously mentioned. Second, the LMI staff supplements the data provided by UI to obtain a greater level of geographic and industry detail for selected employers' classification codes. Specifically, they collect employment, wages, and business identification information for the individual establishments of most multi-establishment employers, which are not separately reported on UI administrative records. Third, the LMI staff directly collects data from Federal government agencies each quarter to add the employment, wages and business identification information of workers covered under the UCFE program, as they are not included in the State UI administrative. Fourth, the LMI staff is responsible for conducting the Annual Refiling Survey (ARS). The purpose of the ARS is to contact one-third of all the establishments each year to verify, and update if necessary, their industrial activity, location (county code), auxiliary status, type of ownership, and both physical location and mailing address information. The updated classification information from the survey is incorporated in the ES-202 program with the data submitted for the first quarter of each year. In addition to the ARS, the BLS regional staff conduct an annual quality assurance review in each SESA for SIC coding for two purposes: 1) to improve the quality and consistency of assigned SIC codes; and, 2) to provide information regarding quality measures of these codes at the national level. The accurate assignment of SIC codes is crucial to the published employment and wages data and since most BLS data collection programs derive samples based upon these characteristics.

2.1.2 Data Element Definitions

All of the data elements listed below have the potential to be updated each quarter, since the main source of the data for the LEDB is the administrative tax records that the employers file with the SESAs for tax purposes each quarter. Generally, the SIC (or NAICS), geographical and ownership codes are only updated with the data being submitted for the first quarter of each year. Thus, any break in the series caused by an update to one of these data elements will occur between calendar years.

BLS follows the *1987 SIC Manual*'s definition of **establishment** (including auxiliary establishments) in its LEDB. An **auxiliary establishment** is primarily engaged in performing management or support services for other establishments of the same enterprise and is separately identified by a one-digit auxiliary code based on the primary activity performed by the auxiliary establishment.

The LEDB also contains a **Multi-Establishment Employment Indicator (MEEI)** code which designates whether the employer is one of the following: a) single establishment employer; b) multi-establishment employer reporting their sub-units on the MWR or RFEW (i.e., master record or parent record); c) establishment of a multi-establishment employer; d) sub-unit of a multi-establishment employer that includes more than one establishment for reporting purposes; e) multi-establishment employer reporting as a single unit because it does not meet the minimum criteria for filing the MWR or RFEW; and, f) multi-establishment employer reporting as a single unit because the employer refuses to complete the MWR or RFEW.

Also collected on the LEDB are two types of **business names**. The first is the **legal or corporate name** whereas the second is the **trade name**, which is essentially the name that the employer is using to conduct its business. In addition to these names, the LEDB also retains a **Reporting Unit Description (RUD)** for each establishment of a multi-establishment employer. The purpose of this field is to further delineate these units by using terminology (store or unit number, plant name, etc.,) provided by the employer, and thus familiar to them, to separately identify each of the establishments in the event that only one of their establishments is selected for a survey. The LEDB has recently begun to collect information on the **type of legal entity** (individual, partnership, corporation, and other-limited partnership, S-corporation, household, etc.) that the business or employer represents.
A number of **addresses** for each establishment are also stored on the LEDB- 1) a **tax mailing address** that is carried on the State's UI employer master file 2) a **mailing address** for statistical reporting purposes and 3) a **physical location address** for each establishment.

2.1.3 Data Coverage
UI coverage is quite broad and comprehensive covering approximately 98 percent of all non-farm wage and salary workers. All employers who have one or more workers employed on one day in 20 weeks or more in a calendar year or who have a payroll of more than $ 1,500 in a calendar quarter are required to have their workers covered by UI. Under the Federal Unemployment Tax Act (FUTA), nonprofit organizations who employ more than 4 workers during the same time period just cited or who meet the same payroll requirement must also cover their workers. Many States, however, have extended this coverage to the "one or more worker criteria" and thus draw no distinction between profit and nonprofit businesses.

Agricultural coverage also has different coverage requirements from nonagricultural employers. The agricultural employers are required to be covered if they employ ten or more workers on one day in 20 weeks in a calendar year or have more than a $20,000 payroll in a quarter. Some of the more densely populated States have also chosen to extend coverage beyond the Federal minimum, resulting in an agricultural coverage rate of 77 percent. Private households, social clubs, and college fraternities and sororities which employ domestic help and pay wages less than $1,000 in a quarter are excluded from UI laws. State and local government employers are also required to cover their workers with only a few minor exceptions. These exceptions are principally students who work part-time at the colleges or universities that they attend; student nurses and interns; elected officials and temporary emergency employees hired to handle disaster relief activities.

The major exclusions to coverage are workers in selected industries, such as nonprofit organizations that employ less than 4 workers; Railroad workers; workers in school systems that are owned and operated by religious institutions; and, selected classes of workers (straight commission life insurance agents, self-employed, unpaid family workers and proprietors).

2.1.5 Data Collection
A. Status Determination Form (SDF)
All new employers which become subject to UI coverage are required to file a SDF with the UI unit of the SESA. This form, which varies in structure but generally not content from State to State, is used to determine an employer's tax liability under the State's UI laws and to collect administrative information such as the employer's EIN. The SDF also requests information on which the classifications for industrial activity, county (township in the New England area), auxiliary (i.e., warehouse, central administrative office, research and development office, etc.), and ownership (private sector, or Federal, State or local government) codes are based. These codes are assigned by the LMI staff. The assignment of the SIC code is based on the establishment's primary economic activity, that is determined by its principal product or group of products produced or distributed, or services rendered. If there is insufficient information on the SDF, the employer will either be contacted by telephone or mailed the Industry Classification Statement to obtain the necessary information.

After employer liability is determined, the SESA UI unit assigns a UI account number. Most new employers are aware of their UI liability and request the SESA to supply a Status Determination Form when they begin their business operations. Some liable employers, usually small ones, fail to file a Status Determination Form. These employers may be discovered through information on new firms applying for EINs supplied by the Internal Revenue Service (IRS) to the SESA each quarter. Other means of discovering liable employers are through the UI claims process, the UI field auditor investigations, and an initial filing of the QCR without an UI account number. From

sampling and business birth perspectives, this state-specific registration process allows BLS to capture employers whose business operations begin in one state and at a later date, begin operations in other states.

B. Quarterly Contribution Report (QCR)

All liable employers are required to file a QCR with the SESAs for their UI accounts. These reports, like the SDF, are administered by the UI program and also differ slightly in design for each State. All of the QCR forms, however, request employment values for each month of the quarter and total wages, taxable wages, and UI taxes due for the quarter. This information and the taxes that are due are necessary for the operation of the UI tax system but they are also important for statistical purposes for the ES-202 program. Employers are asked to report, among other items, the total number of covered workers (full and part-time) who earned wages (subject to UI taxes) during the pay period(s) which includes the 12th of each month in the quarter and the total payroll for the quarter. This report is mandatory for employers with a single location as well as employers with multiple locations in the State. The latter group of employers report a summary of these data for all of their establishments covered under the same State UI account on the QCR. Therefore, establishment-level data for these employers do not exist in most State UI administrative files.

C. Multiple Worksite Report (MWR)

Multi-establishment employers with 10 or more employees in the sum of their secondary physical locations and/or industrial activities covered under one UI account, are requested to provide establishment level data using the MWR. Thus, BLS has a database (the LEDB) which is almost entirely at the establishment level. Data collection procedures for multi-establishment employers differ from those for single units. For multi-establishment employers, the State LMI unit is responsible for the mail-out, processing, and review of the MWR forms each quarter. As part of this process, multi-establishment employers are asked to verify the business identifying information (trade name, worksite description, and physical location address) for each establishment (worksite) that is pre-printed on the MWR. In addition, the employer is requested to provide the employment for each month (using the proper reference period) and total wages for each worksite for the given quarter. New worksites are manually added to the MWR by employers. The State LMI unit then adds these worksites to their database. When that employer receives the next quarter's MWR, these new worksites will be pre-printed on the MWR along with their other worksites. The situation for deaths or business transfers is handled in the same manner. The employer provides the information on the affected worksite and the LMI staff deletes the worksite from that employer's file. Thus, the MWR captures business births and deaths for these multi-establishment employers on an on-going quarterly basis. The MWR is the only source of quarterly data for these multi-establishment employers and is thus an excellent tool for business cycle and policy analysis.

D. Report of Federal Employment and Wages (RFEW)

Federal agencies, whose civilian employees are covered under the separate but comparable UCFE program, do not file QCR forms with State UI programs but instead report employment and wages data directly to the State LMI unit. Since 1993, all States have been using a standardized form, the RFEW, that was developed by BLS to collect these data each quarter. The RFEW was modeled after the MWR to facilitate its use in the State processing systems. Likewise, the Federal agencies providing the RFEW data have been requested to use the electronic transmittal procedure, with approximately 45 percent being collected in this manner in 1999.

E. The Annual Refiling Survey (ARS)

The purpose of the ARS is to review and update, if necessary, the classification codes (industrial, geographical, ownership and auxiliary) currently assigned to the establishments stored on the LEDB. The survey, conducted by the State LMI units, is initiated in October of each year with approximately one-third of the establishments being reviewed annually. The establishments are selected on the basis of the 7th and 8th digit of their Federal EIN, which are essentially random digits within the nine digit number. This selection process ensures that the industrial distribution of the survey respondents is proportional to the establishments in the economy. In other words, no industrial sector is specifically targeted in any one year. For an employer currently coded as a single establishment, the ARS questionnaire requests that the respondent review an industrial classification statement. This statement is a general description of the economic activities for that 4 digit SIC, or 6-digit NAICS code in the future, followed by some specific economic index items that comprise the industry. If the statement reflects the establishment's previous 12 month economic activity, then the respondent simply checks the "Yes" box and underlines the relevant wording that applies to their economic activity. If the employer thinks that the description is not correct or is unsure, then they are requested to check the "No" box and provide a description of the economic activities of their

business along with an approximate percentage of the sales or revenue for each activity listed. The State LMI staff then review this information and determine whether the SIC/NAICS code needs to be updated. The current SIC/NAICS code may have been assigned from the Status Determination Form or it may have been updated from a previous ARS questionnaire. In addition to the industrial classification review, the respondent is also requested to review and update the following, or provide the information if it is not preprinted: 1) physical location address; 2) mailing address; and, 3) county in which the establishment is located.

The respondent is also requested to answer a question concerning whether they provide support services for other units of the same enterprise or to the general business community or the public. This information is used to update the auxiliary code, if necessary. The ARS questionnaire also helps identify new multi-establishment employers. Employers are asked if the establishment whose listed physical location address is the only establishment in that State under that UI account number. If no, then the employer is requested to complete the back of the ARS form which provides space to list the physical location address, economic activity and employment for each of the previously unidentified establishments. This information is then reviewed by State staff to determine if the employer should file a MWR form each quarter.

The ARS procedures for known multi-establishment employers is quite similar to that for single establishment employers. The main difference is that the former group will receive a separate 4 digit industry description for each SIC code currently assigned to that employer in that State under that UI account number. The questions on the survey form are identical to those for a single establishment employer described previously. The States are required to achieve a 75 percent usable response rate to the ARS. Most States conduct at least one, and sometimes two, follow-ups to achieve the required response rates.

2.2 Editing of Micro Data

2.2.1 Current Editing Procedures

Micro data collected on the QCR, MWR, and RFEW are edited by the State LMI staff and corrected, as necessary. The micro data, including imputed values, are then aggregated to the appropriate ES-202 macro-level cells, including the size class break-outs for first quarter. The State LMI unit also reviews the macro level data. When necessary, one of the standard data editing systems is used to edit and update the appropriate micro level records, based on the macro level review. Both the micro and macro edits include checks for invalid and inconsistent data as well as checks for large and unusual employment and wages fluctuations between and within quarters.

Every quarter, a relatively small number of employers fail to submit either a QCR, MWR, or RFEW. Others may submit incomplete reports, typically QCRs with missing employment data. Delinquent and missing data notices are sent to these employers.. Usually the SESA unit that initially mailed the form is responsible for this follow-up. Therefore, the UI unit generally contacts employers who do not complete the QCR, while the LMI unit pursues delinquent MWRs and RFEWs. The follow-up procedures for delinquent and incomplete reports vary slightly in each State, but every attempt is made to minimize the amount of missing data on the ES-202 file. For those employers who fail to respond to follow-up requests, the data are imputed, generally by employing methods that use historical data for the establishment. This imputation procedure is automatic for one quarter. If the report remains delinquent for a second quarter, the record will be flagged for further review. A State LMI analyst will determine the status of the establishment (i.e., active or inactive). If it is determined that the establishment is still active, the data will be imputed. A new imputation methodology that emphasizes a current quarter industry's trend has been tested with data from several States. This new methodology will be included in a future revision of the States' standardized systems.

After making corrections and adding comments to the micro level file, States submit the data to the BLS national office, where it is due approximately four months and three weeks after the end of the reference quarter. A long-term goal is to ultimately have a four month lag in the receipt date translating into accelerated publication of quarterly and annual data.

Over the past 5 years, BLS has invested in the development of standardized ES-202 computer systems to standardize processing, editing, and imputation methodology in the SESAs; improve the data quality of the LEDB; and, control program maintenance costs. Furthermore, this approach permits a more timely introduction of program changes in a more cost-effective manner. These 2 standardized ES-202 State processing systems were developed by Utah and Maine staff. These original system requirements were developed by the State staff and approved by BLS.

Improvements to the systems are implemented with new versions. The system requirements for these versions are developed jointly by BLS and the State developers.

2.2 Data Outputs and Uses

The employment and wages data produced by the ES-202 program represent the universe of workers covered under State UI laws (this includes the private sector, State and local governments) as well as civilian workers covered by the program of Unemployment Compensation for Federal Employees (UCFE). Since coverage is so broad (approximately 98 percent of all non-farm wage and salary employment), the ES-202 program provides a virtual census of these employees and their wages. It is the most complete and timely source of monthly employment and quarterly wages information by detailed industry and county. Consequently, ES-202 data are used extensively in many economic and statistical applications. These include UI program administration, macro-economic research, survey employment benchmarking, and micro-economic analysis. The Bureau of Economic Analysis uses the macro level ES-202 data for about 54% of the quarterly Personal Income component of the Gross Domestic Product for state Personal Income estimates. BLS programs and surveys use the micro level ES-202 data for sampling purposes. These BLS programs and surveys include: 1) Current Employment Statistics 2) Occupational Employment Statistics 3) Producer Price Index 4) Occupational Safety and Health 5) National Compensation Survey (including the Employer Benefits Survey and the Employment Cost Index Survey) 6) Productivity programs and 7) Job Openings and Labor Turnover survey.

BLS also uses the micro level ES-202 data as an input to the Longitudinal Database (LEDB). The LEDB is also used to produce a data series to analyze the job creation and job destruction process. BLS recently completed the development of its establishment LEDB that can be used to analyze the job creation and job destruction process. Data from the first quarter of 1990 forward are now available to researchers for qualified projects. .

3. Conversion from SIC to NAICS

3.1 Basic Principles

Periodically, the system used to classify industrial activities is updated to reflect changes in the structure of the economy. In this manner, new and emerging technologies and services are introduced into the system. To implement a major revision to a classification system requires an enormous amount of planning and coordination among the various federal statistical agencies, the employers who provide these data, and the various users of the economic time series. At the present time, BLS is in the middle of a multi-year project to change the industry codes of its 8.2 million establishments from the 1987 Standard Industrial Classification (SIC) to the North American Industry Classification System (NAICS). The new system was jointly developed by the various statistical agencies of the United States, Canada and Mexico to insure that the economic data developed by these agencies will allow industry comparisons among the countries. The initial work to determine the composition and philosophy of the new system began in 1995 with the final agreements being signed in 1998. (See Stamas and Morisi ICES II paper on this subject).

In any major revision to an industrial classification system, the emphasis at BLS is on introducing the new system as soon as possible while trying to maintain the continuity of as many economic time series as possible. As noted earlier, BLS conducts a periodic review of the industrial and geographical codes of all establishments (one-third of the establishments are reviewed each year). Consequently, the number of establishments with out-of-date codes is greatly reduced when the need to implement a new classification system is proposed. To implement the conversion in a cost-effective and timely manner, BLS developed a detailed, comprehensive multi-year plan that included an adequate amount of lead time to perform the following activities: 1) request multi-year funding from Congress for the States and BLS for planning and implementation purposes, 2) identify and implement the necessary modifications to the State and BLS processing systems, 3) develop and prepare training materials and provide training to the States and BLS staff, 4) develop a data collection strategy, design questionnaires, and implement procedures for non-response, 5) develop conversion tables that show the relationships between SIC and NAICS codes, 6) develop an implementation plan and a timeline for all BLS programs, 7) prepare and maintain comparisons of data on the old and new classification systems to assist data users to assess the impact of the new classification system on the time series

As part of the conversion process, BLS and the States review the current SIC codes assigned to each establishment and update them, if necessary, before assigning the NAICS code. This procedure allows data users to assess the

impact of the change caused directly by the classification systems as opposed to the change being masked by moving from the old SIC code (which was incorrect) to NAICS If one were to look only at the old 1987 SIC code and then at the NAICS code, the relationships between the 2 systems as identified in the conversion tables would not be accurate.

3.2 Current Status

The States, in cooperation with BLS, are currently recoding all reporting units in the ES-202 program from an SIC to NAICS basis. BLS is also working with other U.S. statistical agencies, Canada, and Mexico to complete work for NAICS 2002, which is an update to the original NAICS agreed to by the partners in 1997.

The BLS implementation plan to recode all ES-202 reporting units to NAICS began in late 1998, when all active units with a direct relationship between the SIC and NAICS were automatically assigned NAICS codes. This affected approximately one-half of establishments. In October 1998, units with employment > 50 and ½ of the units with employment < 50 and whose SIC split between more than one NAICS code were targeted to be surveyed. The States sent these units survey forms in order to classify the units in the correct NAICS code. In October, 1999, the remainder of the units with no valid assigned NAICS were selected for the survey. In addition, BLS is conducting a survey this year to verify the auxiliary status of reporting units, because auxiliary units will be treated differently under NAICS. Auxiliaries are worksites within a company that primarily serve other establishments within the same company (examples are warehouses or corporate offices). Under NAICS, auxiliary units will carry the NAICS code for their primary activity, while under SIC, auxiliaries were classified according to the primary activity of the company they served. In October 2000, BLS plans to survey those units that will impacted by the changes introduced with the implementation of NAICS 2002.

3.3 NAICS 2002

The three countries did reach agreement on detailed structures for the Construction and Wholesale Trade sectors during NAICS 1997. Therefore, plans are for these two sectors to be revised for NAICS 2002. The major changes in the construction sector include a new national industry for residential remodeling, and the return of operative residential builders. In addition, BLS will use the 6-digit national industry level to separate the special trades subsector between residential and nonresidential construction and will publish data at this detail via the ES-202 program. The wholesale trade sector will be changed to separate merchant wholesalers (those that take title to goods and play the role of principal in transactions) from agents and brokers that act on behalf of sellers or facilitate transactions.

Other changes in NAICS 2002 include more detail in the Information sector in order to better capture Internet-related activities, separation of the Department Stores industry between discount department stores and traditional department stores, and more detail in the Electronic Shopping and Mail-Order industry.
All of the above changes for NAICS 2002 are still preliminary pending a review of the comments received from data users and interested parties following publication of these plans in the Federal Register.

3.4 Publication Plans

Current plans are for BLS to introduce NAICS 2002 in the ES-202 program with data for the first quarter of 2001. The decision to forego the introduction of the NAICS 1997 was based on the need to keep breaks in these time series (and the other BLS programs that use these data as a sampling frame and/or an employment base or benchmark) to a minimum. It was felt that publishing data for 1999 on the 1987 SICM followed by data for the year 2000 on NAICS 1997 with data for calendar year 2001 on NAICS 2002 would have been too many revisions in a short time span. Thus, the decision to move from the 1987 SICM to NAICS 2002 was both pragmatic and cost-effective.

4. Improvements in Record Linkage methodology

The LEDB is composed of establishment records linked across time. The linkage process used to create and update the database is based upon files that have the same structure across time; these files, therefore, are linked to a new iteration of themselves each quarter. The linkage identifies business establishments which may have gone out of business; establishments that remain in business for both periods; and, new establishments. The quality of the administrative codes is very good; most records, therefore, are correctly linked using these codes. A probability-based linkage process is used to identify the small percentage of links that are missing the appropriate administrative codes.

The LEDB has been in existence for only a short period of time. Its precursor was the Universe Database, or the UDB. The UDB contained four quarters of QUI data, and was linked in a way that minimized the number of incorrect linkages. One goal of the LEDB is to identify and tabulate statistics on business establishment creation and destruction. To meet this goal, the linkage philosophy had to be revised. The linkage process needed to offset to the degree possible, the potential joint errors of invalid linkages and missed linkages. To understand the changes that were introduced in the LEDB process, it would be easier to explain the previous system.

4.1 Universe Database (UDB) Record Linkage

This matching system was composed of four main components. The first component identified the most obvious continuous establishments - those with the same State code-UI/RUN combination. These are establishments that from one quarter to the next did not change their UI reporting - no change of ownership, reorganization, etc. The second component matched units that States submitted with codes identifying predecessor/successor relationships.

The third component matched units based upon certain shared characteristics, e.g. identical trade name, phone numbers, physical location address, etc. Pre-specified weights were assigned based on common data element values. To reduce the number of false matches, a blocking process was used to limit the potential matches to those that had specific shared characteristics. Two matched units were considered a valid match when they exceeded a cutoff level. The fourth component of the matching routine attempted to capture changes that occurred within a quarter as opposed to those that occurred between quarters.

4.2 Reasons for Modifying the UDB Record Linkage Process

The UDB record linkage process effectively linked over 96 percent of all the records received each quarter. Nevertheless, because its methodology was designed to limit the number of false matches, the original linkage system may not have been the most effective at identifying all valid relationships that existed in the remaining four percent of establishments. The result was a potential under-counting of continuous businesses and over-counting of business births and deaths. For that reason a research project was undertaken to improve the record linkage processes. Furthermore, experience with the previous matching process had highlighted specific areas of the process that needed improvement or enhancement. Although these areas affect only the four percent of the records mentioned above, the net effect on the number of births and deaths identified is significant.

4.3 New Approach

The historical linkage procedure described above used four processes to identify continuous units within each state. The modified matching process identifies continuous units using the five procedures described below:

A. UI/RUN Linkages

The files are first linked by UIN/RUN. These administrative codes link most of the records.

B. Imputed Records process

The next step in the new linkage process is to identify records for establishments that are assumed to remain in business, that did not report data for the current quarter (fully imputed records). The corresponding record in the preceding quarter of the match is then flagged and the fully imputed current quarter records are temporarily removed from the file. Rather than assume that these units are delinquent, an attempt is made to identify the units that actually may have been reported under new ownership. At the end of all of the other match processes, the remaining unmatched flagged records on the past quarter file are identified. These records have their matching imputed record restored to the current quarter file, and the link between them is restored. If the previous quarter record finds a better match, the fully imputed current quarter record is deleted.

C. Predecessor/Successor Code Linkages

The files are then linked by Predecessor and Successor codes. In general, the linkage by UIN/RUN and these other administrative codes link over 96 percent of the current quarter file, depending on economic conditions.

D. Probability-based Linkage

The probability-based linkage process involves only the records that are unlinked to this point. In this process, BLS matches approximately one-tenth of one percent of the current quarter records. While this is not a large portion of

the total number of records, it is still an important part of the overall process. The more accurate the linkage process is, the more useful the database will be in identifying economic occurrences.

As described for the historical probability based match, this process links records based upon selected shared characteristics. Linkage weights for each establishment pair are developed based on selected data element values. Two matched units are considered a valid match when they exceed a cutoff weight. To reduce the number of false matches, a blocking process limits the potential matches to those that have specific shared characteristics. The UDB Record Linkage system utilized three basic blocks to identify probability- based linkages, while the new system utilizes 21 blocks. The additional blocks provide more control over the error structure for this part of the linkage system. Within these 21 blocks, there are three groups which block on certain data elements. The first group contains blocks that include either exact name or exact street address. The second group blocks on phone number, and the third group blocks on various other data elements, such as ZIP code and EIN.

E. Within-quarter Matches, Breakouts and Consolidations

Establishments that experience a within-quarter reporting change are generally assigned either a predecessor code or successor code pointing to another record within the same quarter. Many of these within-quarter links are legitimate, so a process to identify them was included in the linkage system.

After the within-quarter linkages are identified, situations where multi-establishment reporters changed the basis of their reporting to the state, e.g. as filing only a QCR to filing a QCR and MWR or stopping the filing of the MWR. States encourage these reporters to supply data for each worksite, using the MWR. When a reporter changes from reporting all worksites consolidated into one report (the QCR) to reporting disaggregated data for each establishment using the MWR, there is a possibility of failing to capture this as a non-economic event. If the objective were to merely count records, it would appear that more establishments were present in the current quarter than in the past quarter. The reverse situation is also possible.

BLS is interested in identifying these links in order to exclude them in the counts as business openings or closings. The limited number of situations found are sent to a data editing routine, where the employment values are checked for reasonableness. If the match fails the edits, it is not counted as a breakout or a consolidation unless the match is based on predecessor or successor code information.

4.4 Future Linkage Research

The LEDB record linkage system is an improvement over its precursor. The system has been designed to balance the potential joint errors of invalid linkages and missed linkages. There are still, however, improvements that can be made with additional research. Given the timing of the program, the next quarter's preliminary files are available shortly after the time the final linkage is being completed for the current quarter. Improvements might be made by taking advantage of the information in the next quarter's preliminary files. Improvements might also be made by developing a methodology to handle deleted imputed records and intra-quarter matched records that return in subsequent quarters. There are also several improvements that might be made in the probability based linkage process. Among these are 1) designing state-specific linkage frequencies and cutoffs, as opposed to using one standard for all states, 2) exploring alternative methods for developing the frequencies which underlie the probability weights, and 3) developing a within-quarter probability based linkage

5. Implementation of Permanent Random Numbers (PRNs)

5.1 PRNS

In the near future, Permanent Random Numbers (PRNs) will be developed and retained on the Bureau's LEDB. These numbers are currently developed and retained on a system external to the LEDB. These numbers are used to reduce the burden that the Bureau imposes upon small businesses by reducing the probability that they will be included in more than one Bureau survey. The reduction in burden is accomplished by assigning each major survey a start point within a stratum, and then having the survey select sample units sequentially from that point. Randomness is retained as a survey property by randomly assigning the PRNs. The PRNs are collocated within defined strata. The collocation process spreads the PRNs evenly within each stratum. Having the quarters birth units evenly distributed within a stratum allows periodic surveys to supplement their sample periodically with a representative sample of birth units.

5.2 PRN Collocation Procedure.

PRNs are developed and collocated using the following procedure. First, we generate a uniform random number, in the range from zero to one, for each birth record (a birth record is an establishment record with no link to the previous quarter). For collocation purposes a cell is defined by MSA, industry, and employment size class. The random numbers are collocated within each cell. To do this, the birth records in the cell are sorted by the uniform random number. We then determine R for each of the sorted records, where R is a sequential number from 1 to N, and N is the number of birth records in the cell. We then generate e , a uniform random number between 0 and 1. Finally, we assign a Permanent Random Number to each birth record i in the cell according to the following formula:

$$PRN_i = \frac{R_i - e}{N}$$

Several Bureau surveys have already begun to use the PRNs to limit their overlap with other Bureau surveys. The primary surveys which are utilizing these numbers are the Current Employment Statistics survey, the Occupational Employment Statistics survey, and the new Job Openings and Labor Turnover survey. Other Bureau programs are evaluating the PRN implementation to determine if a sampling strategy which utilizes these numbers meets the goals of their individual sample designs.

Note: Any opinions expressed in this paper are those of the authors and do not constitute policy of the Bureau of Labor Statistics.

6. REFERENCES

EXECUTIVE OFFICE OF THE PRESIDENT AND OFFICE OF MANAGEMENT AND BUDGET (1987), Standard Industrial Classification Manual 1987, Springfield, Virginia,: National Technical Information Service.

Farmer, Tracy E. and Michael A. Searson (1995), "Use of Administrative Records in the Bureau of Labor Statistics' Covered Employment and Wages (ES-202) Program," 1995 Bureau of the Census Annual Research Conference, Washington, DC, March 1995.

Searson, Michael A. and John Pinkos (1990), "The Bureau of Labor Statistics' Business Establishment List Improvement Project," 1990 Bureau of the Census Annual Research Conference, Washington, DC, March 1990.

U.S. DEPARTMENT OF LABOR, BUREAU OF LABOR STATISTICS (1992), "ES-202 Operating Manual," Employment Security Manual, Washington, DC: U.S. Department of Labor. (Includes revisions from BLS ES-202 Program Technical Memorandums from 1993 to 1996).

U.S. DEPARTMENT OF LABOR ETA (1997), Comparison of State Unemployment Insurance Laws, Washington, DC: U.S. Department of Labor.

U.S. DEPARTMENT OF LABOR ETA (1995), UCFE Instructions for Federal Agencies, Washington, DC: U.S. Government Printing Office.

U.S. DEPARTMENT OF LABOR, BUREAU OF LABOR STATISTICS (1997), "Improvements in Record Linkage Processes for the Bureau of Labor Statistics' Business Establishment List", Kenneth Robertson, Larry Huff, Gordon Mikkelson, Timothy Pivetz, and Alice Winkler

U.S. DEPARTMENT OF LABOR, BUREAU OF LABOR STATISTICS (1995), "Implementing a Standard Industrial Classification (SIC) System Revision ", Brian MacDonald, published by Wiley, Chapter 7 in '*Business Research Methods*'.