

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, Bureau of Labor Statistics*

Abstract

The Bureau of Labor Statistics has historically maintained a Universe Database file that contains quarterly employment and wage information for all covered employees under the Unemployment Insurance Tax system. It is used as a sampling frame for establishment surveys, and also as a research database. Each quarter approximately seven million records are collected by the States and processed for inclusion on the file. There are many data items of interest associated with this database, such as an establishment's industry, county, employment information, and total wages. Historically, this database has contained five quarters of data. These data have been linked across the five quarters by both administrative codes and through a weighted match process. Recently, a project has been undertaken to expand this database so that it will include multiple years of data. Once several years of data have been linked, the database will expand as new data are obtained. This will create a new "longitudinal" establishment information database, which will be of prime interest to economic researchers of establishment creation, growth, decline, and destruction.

As one step in the creation of this new resource, research was initiated to refine the existing record linkage process. This paper will provide details of the processes used to link these data. First, we will briefly cover the processes in place on the current system. Then we will provide details of the refinements made to these processes to improve the administrative code match. These processes link nearly 95 percent of the file records. The remaining records are processed via a revised weighted match process. Information on the current state of the revised weighted match will be provided, as well as the details of work still in progress in this area.

Introduction

In preparation for the building of a new longitudinally-linked establishment database, the Bureau of Labor Statistics decided to review its current system for linking business establishments across time. Because the new database will be used to produce statistics on business births and deaths and job creation and destruction, we had to ensure that the linkage procedures used in building the database would yield the most accurate results possible. Since the current linkage system was built for different purposes than the new system, there were areas where we could potentially improve the process. This paper provides an explanation of the current linkage procedures, details of the work completed to date, and areas of research that need to be explored in the future.

Background

Quarterly Unemployment Insurance Address File

The Bureau of Labor Statistics oversees the Covered Employment and Wages, or ES-202 program, that provides a quarterly census of information on employers covered under the State Unemployment Insurance (UI) laws. These data are compiled into a data file, the Quarterly Unemployment Insurance (QUI) Address File.

The QUI file includes the following information for each active employer subject to UI coverage during the reported quarter: State UI Account Number, Establishment Reporting Unit Number (RUN), federal Employer Identification Number (EIN), four-digit Standard Industrial Classification (SIC) code, county/township codes, monthly employment during the quarter, total quarterly wages, and the establishment's name(s), address 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, EIN, and UI/RUN allow establishments 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 Universe Database

The State QUI files are loaded to a database, the Universe Database (UDB), for access by users for survey sampling and research purposes. The UDB is composed primarily of data elements drawn from the QUI files. In addition, there are a few system-assigned and derived data elements, as well as information on SIC code changes merged from other sources. An important system-assigned field is the UDB Number, a unique number identifying continuous business establishments.

UDB Record Linkage

When considering the linkage of these records, the reader should understand that we are linking files which have the same structure across time. These files are linked to a new iteration of themselves each quarter. This linkage allows us to identify business establishments which may have gone out of business; establishments which remain in business for both periods; and, new establishments. The quality of the administrative codes are very good, so we expect that we correctly link most records which should be linked. We follow the administrative code match with a probability-based match. This procedure is followed to identify the small percentage of links which are missing the appropriate administrative codes.

Each quarter prior to loading the QUI files to the UDB, a matching procedure is performed to link businesses. By default, all units that do not link are identified as either new establishments or closed establishments. In order to have accurate data on business births and deaths, it is critical that the matching system accurately link establishments. The intent of the original linkage system was to minimize the number of invalid matches. Unfortunately, however, this causes some good matches to be missed. Because statistics on business births and deaths were not being produced from these linked data and only a small percentage of the total number of records was affected, this situation was acceptable.

The match 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. Given that State personnel have access to the information needed to determine these relationships, they are assumed to be correct.

The third component matched units based upon certain shared characteristics. Prespecified weights were assigned based on data element values that the units had in common. This weighted match routine processed the data in three steps (or blocks). All three blocks limited potential matches to those units coded in the same 4 digit SIC code and county. (The New England States also use township codes.) The first block included all units that also matched on a key constructed from the Trade Name field. This "Name Search Key" was composed of the first seven consonants of the Trade Name. The second block included all units that also matched on the first 15 positions of the Street Address field. The third block included all units that also contained identical phone numbers. Two matched units were considered a valid match when they exceeded a cutoff weight. A limitation stemming from this three-block structure is that units that had a valid relationship but had different 4 digit SIC or county codes are missed by the linkage system.

The fourth component of the matching routine attempted to capture changes that occurred within a quarter. It first linked units that had State-identified predecessor relationships already coded. It next performed a within-quarter weighted match to capture relationships not previously identified by the States. A significant restriction placed upon both parts of the within-quarter matching was that the potential predecessors had to contain zero employment in the third month of the quarter while the potential successor units had to contain zero employment in the first month of the quarter. Because of inconsistencies in reporting by some employers, valid relationships could exist that did not meet this criteria, and were not matched.

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 between the remaining four percent of establishments. The result was a potential under-counting of continuous businesses and over-counting of business births and deaths. It is for that reason that the research described in this paper was undertaken.

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 could be significant.**

New Approach

The matching process consists of the two major procedures described below -- an administrative code match and a probability-based weighted match.

Administrative Code Match

Imputed Records

The first step in our new linkage process is to identify the imputed records (i.e., non-reporting records that are assumed to remain in business), and flag the corresponding record in the preceding quarter of the match. We then temporarily remove the imputed records from the current quarter file. Rather than assume that these units are delinquent, we attempt to identify the units that actually may have been reported under new ownership. At the end of all of the other match processes, we identify the unmatched flagged records on the past quarter file. These records have their matching imputed record restored to the current quarter file, and the link is made between them.

Within-Quarter Matches

Establishments that experience a reporting change within a quarter are generally assigned either a predecessor code or successor code pointing to another record within the same quarter. We determined that these within-quarter links were legitimate, so we included a process to find them.

Remove Breakouts and Consolidation

After the within-quarter matches were identified, we examined situations where multi-establishment reporters changed the way they reported. States encourage these reporters to supply data for each worksite. When a reporter changes from reporting all worksites on one report to supplying multiple reports, there is a possibility of failing to capture this as a non-economic event. If we were just counting records, it would look like we have a lot more establishments in the current quarter than we did in the past quarter. The reverse situation is also possible.

We were interested in identifying these links in order to exclude them in the counts as business openings or closings. The limited number of situations found were sent to a data editing routine, where the employment values were checked for reasonableness. If the match failed the edits, it was not counted as a breakout or a consolidation, and not included in counts of establishments increasing or decreasing in employment. Those cases failing the edits were still linked. However, since there is some type of economic change occurring along with the reporting change, the units failing the edits are included in counts of establishments increasing or decreasing in employment.

All Other Administrative Code Matches

The files were then linked by UI/RUN. These administrative codes linked most of the records. Additional links were identified using Predecessor and Successor codes. In general these administrative code match processes link over 96 percent of the current quarter file, depending on economic conditions.

Probability-Based Match

The probability-based weighted match process involves only the unmatched records from the administrative code match process. In this process we generally expect to match less than one-half of one percent of the current quarter records. This can also be expressed as linking less than ten percent of the current quarter residuals. While this is not a large portion of the overall records, it is still an important part of the overall process. The more accurate we can make the overall linkage process, the more useful the database will be in identifying economic occurrences.

Theoretical Basis for Weighted Matching

The weighted match process is accomplished using the software packages AutoStan and Automatch, from Matchware Technologies Incorporated. The first is a software package used to standardize names and addresses for linking. The second package uses a record linkage methodology based on the work of Ivan P. Fellegi and Alan B. Sunter. Automatch uses the frequency of occurrence of selected variable values to calculate the probability that a variable's values agree at random within a given block. The probability that the variable's values agree given that the record is a match can also be calculated by the software. These match and nonmatch probabilities form the basis of the weight assigned to the variable in the match process. The sum of these variable weights are assigned as the overall weight for a given record pair. The distribution of these summed weights, along with a manual review of selected cases, allows us to determine an appropriate region where we find mostly matches. The lower bound of this region is set as the match cutoff value. We expect that above this cutoff will be mostly good matches, and that below this cutoff will be mostly bad matches.

These theoretical constructs are the foundation of probability-based record linkage. However, the nature of the data, in combination with software, hardware, and resource limitations, sometimes requires that additional steps be taken to fine-tune this process. Fortunately, Automatch provides some capabilities in this direction. The weights assigned to a matched or nonmatched variable can be overwritten or augmented as needed. This allows the user to augment the weight of important variables, as well as to penalize certain combinations of variable values, so that a record pair will not match.

Weighted Matches

Blockings

While the UDB Record Linkage system only utilized three basic blocks (trade name, address, and phone number), the new system, using Automatch, provides the option to use as many blockings as needed to match records. Based on empirical studies using California data, we constructed 21 blocks for the new system. All blocks match on two to four data elements. 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.

Adjustments to Blockings

After the first few runs of Automatch, we adjusted the blockings and their probability weights to enhance their matching potential. One weight adjustment we made was to records with similar street addresses. If the street addresses contained different suite numbers, we reduced the weight. Similarly, we reduced the weight if primary names contained different unit numbers. If one data element was unknown or blank, we increased the weight because these data elements did not necessarily disagree. However, if both data elements were unknown or blank, we deducted weight because there was a greater possibility that they would disagree. Finally, we deducted weight if both records were part of a multi-establishment employer.

We also made adjustments based on the address types. Some accountants submit data for many companies. Therefore, more than one record could have the same accountant's address and telephone number. If two records contained the same physical location address, they were considered a good match and we gave them more weight. If one record contained a physical location address and the other record

contained an unknown or tax address, it is possible that it would be a good match, so we gave it slightly more weight.

Subjective Results and Cutoffs

Although these records contained some common data elements, frequently it was difficult for us to decide whether the records were good matches. We subjectively identified matches as being “good,” “bad,” or “questionable.” We reviewed these data to determine the quality of each matched pair. Then we set the cutoff weights for each of the 21 blocks, approximately in the middle of the questionable records.

Results

California data files were linked forward from the first quarter of 1994 (1/94) through the first quarter of 1995 (1/95). We evaluated the matches resulting from the final two quarters (4/94 to 1/95). These results are shown in Tables 1 through 4. Additionally, two quarters of data were matched for three other States -- West Virginia, Georgia, and Florida -- using preliminary match parameters developed for California. The results were evaluated by the four analysts using the same rules used in evaluating the California results. Although the results are not tabulated, they are approximately equivalent to those obtained for California. This finding is significant since there are insufficient resources to manually review cases which fall close to the match cutoff parameter. It is, therefore, important that we find match cutoff parameters for each block which produce satisfactory results in all States.

Number of Units Matched

Table 1 provides a summary of the matches in California which were obtained from the current matching procedures and the new procedures as tested. Both procedures produce the same number of matches on administrative number identifiers (79.58% of the file matched on UI/RUN). The first improvement in the matching process appears among those “delinquent” reporters which are assumed to remain in business. In the new procedures being tested, these imputed records are not generated until after all other matching processes are completed. The rationale for this change is that these non-reporting records may represent administrative business changes such as a change in ownership and may be reporting with a new UI/RUN. These units were matched with new UI/RUNs in 342 cases in California (0.04 percent of the file). These cases represent Type II errors (erroneous matches) for the previous matching process. The remaining 154,143 delinquent reporters were later matched to a new imputed record, as in current procedures.

The second improvement in the matching process appears in the within-quarter administrative match. These within-quarter matches represent units which have undergone some administrative change such as a change in ownership in a quarter and appear twice in the quarter with different UI/RUNs. It has become apparent during study of these files over several years that these units do not always cease reporting in one month during the quarter and begin reporting as a new entity in the next month of the quarter. The previous match procedures restricted these within-quarter matches to those reporters which report in a very precise manner. The new procedures allow for some reporting discrepancies in the monthly employment in matching these cases. The new procedures obtained approximately 1,110 (0.12%) additional matches for these situations.

Table 1. -- Results: Match Comparison for CA 95 Qtr. 1

	Current	New

Match Type	Method		Method	
	Count	%	Count	%
UI/RUN to UI/RUN	739,442	79.58	739,442	79.58
Correct “Delinquent” Matches	154,143	16.59	154,143	16.59
Incorrect “Delinquent” Matches (Type II Errors)	342	0.04	0	0
Pred/Succ. Codes/Non-Economic Reporting Changes /Within-Quarter Administrative Matches	821	0.09	1,978	0.21
Weighted	686	0.07	1,513	0.16
Births	33,723	3.63	32,081	3.45
Total (Records = 929,157)	100.0		100.0	

Note that the new system identifies 1,642 more links than the old system.

Finally, the third improvement in the matching process is in the weighted matching for all units in both quarters which do not match during any of the administrative matching procedures. The new procedures make use of many additional block structures which make possible incremental increases in the number of matches without significantly increasing the number of Type II errors. This is accomplished by tailoring the match cutoff parameter for each block so that most of the good matches fall above the match cutoff parameter without including a large number of Type II errors. The good matches falling below the parameter in one block are captured as matches in other blocks without picking up significant numbers of additional errors. The number of weighted matches went from 686 to 1,513 for an increase of 827 (0.09%). The total number of additional matches from the new procedures over the current procedures is 1,642. This reduces the number of business births (and business deaths) by 1,642 per quarter.

Although the results of the new linkage procedures do not appear dramatically different from the results of the current linkage procedures, the marginal improvements are significant in terms of the uses of the linkages. As stated earlier, one of the principal uses of the linked data files is to estimate the number and characteristics of business births and deaths and to track business births over time to determine when they increase or decrease their employment and how long they continue in business. It is easy to see that even though a large portion of the units match through the administrative codes, it is the remainder of the units which are considered business births and business deaths. Marginal improvements in matching these other units can have a relatively large impact on the number of business births and deaths and the ability to track them over time.

Quality of Matched Units

Tables 2 and 3 compare the quality of the weighted matches resulting from each procedure. There are two conclusions of interest from these tables. First, there are many more good matches resulting from the

new procedures and fewer Type II matching errors. Also, there are approximately 150 to 300 good or questionable weighted matches obtained from the current matching procedures which are not being identified during the new weighted matching procedures. There are two possible explanations. The first is that we are missing these matches with the new procedures and we must find methods which will identify the good matches. The second is that although we are not identifying these matches during the weighted match, they may be identified in the enhanced administrative matching procedures which would preclude them from the weighted matching process. The truth may lie somewhere between these possibilities and will be one focus of our future research efforts.

Table 2. -- A Comparison of Weighted Match Counts and Quality

Match Quality	Current Method		New Method	
	Count	%	Count	%
Good	262	49.1	1,317	87.0
Questionable	198	37.1	173	11.4
Bad	74	13.9	23	1.5
Total	534		1,513	

Note that all weighted match results are based on a manual review of linked records, and are based on the subjective opinions of several reviewers.

Table 3 continues the comparison of the quality of matches obtained from the current and new match procedures. It is obvious from this table that, although, the new match procedures apparently miss some good and questionable matches at or near the cutoff parameters for a match, the new procedures identify many additional good matches which are missed by the current weighted match procedures. This is accomplished by the new procedures while picking up fewer questionable and bad matches than the current procedures.

Table 3. -- Weighted Matches

Match Quality	Current Method Only		New Method Only		Both Methods	
	Count	%	Count	%	Count	%
Good	156	38.4	1,211	87.4	76	91.6
Questionable	178	43.8	153	11.0	7	8.4
Bad	72	17.7	21	1.5	0	0
Total	406		1,385		83	

Finally, Table 4 provides an analysis of the overall quality of the weighted matches obtained from the new procedures. Those units above the match cutoff parameter are identified as matches while the units below the match cutoff parameter are not identified as matches. There are at least 23 Type II errors while there are at least 51 Type I errors. This rough balance in these error Types seems a reasonable one for the purposes for which we are matching the files. Since there are only 142 good or questionable matches which fall below the match cutoff parameter, it seems that a substantial portion of the weighted matches identified only by the current weighted match procedures are identified during the enhanced new administrative match procedures.

Table 4. -- New Weighted Match Distribution and Quality

Match Quality	Group			
	Above Cutoff	%	Below Cutoff	%
Good	1,317	12.2	51	0.5
Questionable	173	1.6	91	0.8
Bad	23	0.2	9,098	84.6
Total	10,753			

Nonmatches are only counted within the twenty-one designated blocks, and with a match weight greater than or equal to zero.

Future Areas of Research

The results shown in Tables 1 through 4 are based on the research completed to date. As we are now aware from this preliminary effort, the matching procedures used here can be improved and there are more areas of study which may yield further improvement. In addition, there is additional testing which will be necessary to complete an initial assessment of the quality of the matching process.

- Since the files which make up the UDB are the product of each of the State Employment Security Agencies, it is important that the new match procedures be tested on data files from each of the States. This is the only way to insure that anomalies in any of the State files will not adversely affect the match results. The short time available for completing each of the quarterly matches and the size of the files does not allow for a manual review of the quarterly results. This initial review of the matching process using the final parameter values will provide some measure of the quality of matches obtained. It may also be advantageous to tailor the match cutoff parameters independently for each State.
- It is apparent from our initial analysis that additional analysis of the results of the current and new match procedures is necessary to determine how many good matches are being missed by the new procedures and how many of these are being identified by the new enhanced administrative match procedures. Once it is determined how many of these matches are being missed by the new procedures and their characteristics, the new match procedures must be modified to identify these matches.
- Intra-quarter weighted matching procedures should be tested to determine if such a procedure should be added to the new match procedures and its impact on overall results.
- Once the new procedures are enacted, an ongoing review of selected States may be recommended to insure that the match results do not deteriorate over time.

Acknowledgments

The authors would like to acknowledge the contributions of Larry Lie and James Spletzer of the Bureau of Labor Statistics and Catherine Armington of Westat, Inc.

Disclaimer

Any opinions expressed in this paper are those of the authors and are not to be considered the policy of the Bureau of Labor Statistics.