EMPLOYER REPORTING UNIT MATCH STUDY (ERUMS) -- A FINAL REPORT!

Warren L. Buckler, Social Security Administration
Thomas B. Jabine, Committee on National Statistics
Warren L. Buckler, 1105 King Arthur Ct., Sykesville, MD 21784

Introduction

The Employer Reporting Unit Match Study (ERUMS) was a pilot record linkage study carried out under the auspices of the Federal Committee on Statistical Methodology (FCSM), Office of Management and Budget. The study linked records of employers and their reporting units from three agencies: the Bureau of Labor Statistics (BLS), the Social Security Administration (SSA) and the Internal Revenue Service (IRS). The primary linkages involved samples of the agencies' records for employers in the State of Texas, covering their activities in 1982.

The ERUMS project was planned and carried out by an interagency workgroup under the general guidance of the Federal Committee on Statistical Methodology. Planning began in 1983 and the project operations were completed in 1989. The motivation for ERUMS came from earlier work of the FCSM Subcommittee on Statistical Uses of Administrative Records, which had determined that effective and efficient statistical uses of administrative records were being hampered by the existence of noncompatible systems for reporting employer information at the establishment level.

The goal of ERUMS was to demonstrate the feasibility of matching employer and reporting unit data from different agency record systems as a means of obtaining more precise information about differences in the coverage and content of the data in those systems. The study focussed on the BLS and SSA record systems, with employer-level data from IRS being used primarily to reconcile and explain BLS-SSA differences. It was expected that ERUMS, as a demonstration study, would provide valuable experience with the technical aspects of data linkage and the administrative requirements for gaining access to the data and carrying out the matching operations.

This paper summarizes the ERUMS project. A more detailed report can be found in Statistical Policy Working Paper 16, " A Comparative Study of Reporting Units in Selected Employer Data Systems", Statistical Policy Office, Office of Information and Regulatory Affairs, Office of Management and Budget, 1990

Data sources

The primary source of data for ERUMS from BLS was the first quarter 1982 Unemployment Insurance (UI) Address File. For each State, the UI Address File contains data for individual employers and their reporting units, which are often but not always equivalent to establishments. The data for this file are submitted annually (more recently quarterly) to BLS by the State employment security agencies that operate the Federal-State UI Program. The BLS uses the data submitted by the States as a basis for periodic statistical reports on employment and wages and uses the UI Address File as a national sampling frame for its establishment surveys.

The principal SSA files used for ERUMS were files developed for statistical uses within SSA. They included an edited file of Form W-3 annual wage reports for 1982 and the Single Unit and Multi Unit Code Files. The Form W-3 file provided wage data for individual employers and, in some cases, for each of their reporting units, which are frequently but not always equivalent to establishments. The Single Unit Code File, which is updated annually, contains a record for every entity that has filed an application for an Employer Identification Number (EIN), excluding non-employing entities and household employers. The Multi Unit Code File contains a record for each reporting unit of multi unit employers who are participating in the Establishment Reporting Plan, a voluntary program under which employers report their annual wage information on Form W-3 separately for each of their reporting units.

The main source of IRS data used for ERUMS was a Census-edited file based on Forms 941 and 943 for Tax Years 1981-83. These forms are used by employers to report each quarter (annually for Form 943) to IRS on income taxes withheld from wages and other payments to employees and on taxes under the Federal Insurance Contributions Act (Social Security taxes). Extracts of data from these forms are provided annually by IRS to the Census Bureau for use in the latter's County Business Patterns Program and other statistical purposes. The Census Bureau edits the files to use the best available industry code for each employer and impute certain missing data. A copy of the edited file has been made available to the IRS Statistics of Income Division for use in its statistical programs. Data from this Census-edited file were obtained for most of the employers in the Phase II ERUMS sample (see below). In addition, copies of Form 940, Federal Unemployment Tax Return, for 1982 or 1983 were obtained for a substantial proportion of the Phase II sample cases.

The study design

Because of the ERUMS Workgroup's limited resources, the study was restricted to a single State, Texas, and a small sample of employers and their reporting units from that State. The sampling unit was the employer, identified by a unique EIN. A probability sample of all EINs active in the State of Texas in 1982 was selected from the BLS and SSA files described above. Employers were considered to be active in the BLS system if they had one or more records in the 1982 UI Address File and in the SSA system if they had filed a W-2/W-3 wage report for 1982.

The sample was selected in two phases. The sampling fraction for Phase I was 6 in 100, and the selection was based on the 7th and 8th digits of the EIN. The BLS sample, which was selected first, contained 16,336 distinct EINs. The BLS sample was compared to the SSA files and an

additional sample was selected (using the same pairs of digits) of 3,628 EINs which had at least one Texas reporting unit, had wage reports for 1982 and did not appear in the 1982 UI Address File. The Phase I sample EINs were stratified by match status (match, SSA only, BLS only) and single/multi unit status. A Phase II sample of 401 EINs was selected from the Phase I sample, using disproportionate stratified sampling, with equal probability systematic selection within each stratum. Nonmatch and multi unit EINs were oversampled in Phase II because of their greater interest for the purposes of ERUMS.

The Phase II sample provided the basis for the detailed analyses presented in this report. For matched cases, BLS and SSA geographic and industry codes were compared. The industry codes from both sources were compared with those in the IRS/Census-edited Form 941 file. The status of unmatched EINs was clarified by reviewing additional data sources in the agency for which the EIN did not show up in the initial match. Several of the EINs not located initially in the SSA edited 1982 W-3 file were found among groups of delinquent reporters or cases for which the W-2/W-3 wage report and IRS Form 941 data were being reconciled. In addition, several of the Phase II sample employers originally classified as SSA multi unit were reclassified as single unit because it could not be established that they reported 1982 wages for two or more reporting units in Texas. As a result of these reviews and changes, the final distribution of the sample EINs by match status and single/multi unit classification differed substantially from the preliminary distribution of the Phase II sample.

Administrative arrangements

For the ERUMS Workgroup to gain access to the data sets needed for the study, it was necessary to develop working arrangements that complied with the provisions of confidentiality statutes, regulations and policies of the Federal and State agencies that controlled these data sets. After protracted negotiations, this was accomplished primarily through the development of two bilateral agreements.

In one of these agreements, the IRS contracted with BLS for the performance of those parts of the ERUMS project that required access to tax data, including the wage report information that was to be provided by SSA. Under this agreement, SSA staff could be designated as special agents of BLS to carry out their part of the linkage and analysis operations. By law, the purposes of IRS participation in the project and its service contract with BLS had to be related to IRS administration of the tax laws.

The second agreement was a conditions of use agreement between SSA and BLS which allowed SSA to release relevant data from its employer files to BLS and authorized BLS to link data from these files with data from the UI Address File and certain data to be furnished by IRS, and prohibited any other linkage. Both agreements incorporated several safeguards, with emphasis on limiting access at each stage of the project to those persons who needed to use identifiable data, keeping the number of such persons to a minimum and having them sign non-disclosure affidavits.

To meet the statutory confidentiality requirements of the State of Texas, BLS obtained the permission of the Texas State Employment Commission to use the 1982 Texas UI Address File microdata for the ERUMS study.

Results

All results based on the ERUMS sample are estimates weighted to account for the disproportionate sampling used in the selection of the Phase II sample, unless otherwise noted. The main quantitative results are shown in Tables A-1 through 8 of the Appendix.

Of the Texas EINs that were active in 1982 in the BLS or SSA systems, 67.1 percent were active in both systems, 27.6 percent were active only in the SSA system and 5.3 percent were active only in the BLS system (Table A-1). Only about 1.0 percent of all active EINs were classified as multi unit in one or both systems, and most of these were classified as multi unit only in the BLS system (Table A-4).

For the matched single unit EINs, i.e., those that were active in both systems, an estimated 81.6 percent had the same State and county codes in both systems. The remaining cases were about equally distributed in three categories: same State, different county; same State with no county code in the SSA file; and different State (Table A-5). An estimated 70.2 percent of the matched single unit cases had the same two-digit industry codes. About half of the remaining cases were not classified by industry in the SSA system (Table A-5). When matched against the IRS/Census-edited Form 941/943 file, about three-fourths of the matched single units from both the BLS and SSA files had two-digit industry codes that agreed with those in the IRS/Census file. However, when the SSA unclassified cases were excluded from this comparison, the proportion of SSA cases that agreed with the IRS/Census two-digit code was somewhat greater than the corresponding proportion for the BLS matched single unit cases (Table A-8).

Only a few EINs (nine sample cases) were classified as multi unit in both the BLS and SSA systems. Matching individual reporting units for these cases proved to be difficult. Overall, the nine sample employers had 105 Texas reporting units in the BLS system and 60 in the SSA system for 1982.

Of the active SSA EINs not found in BLS's first quarter 1982 UI Address File, it was estimated that 69.2 percent had reported no first quarter employment to IRS on Form 941 and therefore would not normally be expected to appear in the BLS system (Table A-6). For another 10 percent of these employers, the analysis suggested that they may not have met requirements for UI coverage in Texas either because they had no operations in Texas, because of nonprofit status or because their payrolls were too small. For the remaining 20 percent, the reasons for their absence are not always clear, but it may have resulted in part from lags

in incorporating new employers in the UI State agency and BLS files.

Most of the employers who were included in the 1982 UI Address File but did not file 1982 W-2/W-3 wage reports (22 sample cases) appeared to have ceased hiring employees, gone out of business, or gone through other changes that altered their reporting to IRS and SSA. Half of the employers in this group reported no employment in the 1982 UI Address File. Many of the remainder had filed their final Form 941 with IRS (at least for the period 1981-1983) for a quarter in 1981.

An analysis of the sample EINs that appeared in SSA's Multi Unit Code File provided some indication of the extent to which multi unit employers were participating in SSA's Establishment Reporting Plan (ERP) in 1982 (Table A-7). An estimated 35.9 percent of these EINs had been incorrectly added to the Multi Unit Code File as the result of a processing error that has since been corrected. Most of the remaining employers had initially agreed to participate in the ERP, but more that half of this group did not provide separate data for each reporting unit in their W-3 wage reports for 1982.

Limitations of the study

Several factors limit the broad applicability of the ERUMS findings. The results reflect the reporting requirements and operating procedures associated with the agency record systems in 1982. There have been significant changes since then. In particular, BLS has taken several steps to improve the timeliness and the completeness and accuracy of data in its UI Address File.

The study was based on data for a single State, Texas, and on a small sample of employers and reporting units. The UI system gives the States some latitude in their record-keeping practices, so indications of the coverage of employers in the record systems of the Texas State Employment Agency in 1982 should not be assumed to apply fully to the UI systems of other States at that time. The small sample size means that estimates based on the Phase II sample are subject to relatively large sampling errors. Because of limited resources and the complexity of the Phase II sample design, we were able to compute sampling errors only for a few key estimates (see Table A-4).

The analysis of the results was complicated by differences in concepts and coverage in the record systems used in the study. These differences occurred in the basic filing requirements for the UI and SSA/IRS systems, the time reference of the basic BLS and SSA files used for matching, the definition of reporting units in the BLS and the SSA/ERP systems, and the structures of the BLS and SSA industry classification systems. In addition, certain file deficiencies and operational problems made the analyses more difficult. About 1.3 percent of the records in the 1982 UI Address File for Texas did not have EINs and therefore were not included in the Phase I sample of EINs from that file. In the SSA files, a significant proportion of employers lacked county and industry codes. The most serious problem was that a high proportion of multi unit employers were not reporting separately in 1982 for each reporting unit, so that we were unable to do a thorough comparison of reporting units for multi unit employers active in both the BLS and SSA systems.

Although these differences and file deficiencies made the analyses more difficult, the fact that we succeeded in identifying and documenting them is an indication that the ERUMS project succeeded in its main goal, which was to demonstrate the feasibility of doing matching studies as a means of evaluating the suitability of administrative record systems for statistical uses.

Findings

The detailed analyses of the ERUMS data did not suggest that large numbers of employers who report wages in one of the payroll tax systems were failing to report in the other system when they should have been. They do, however, suggest that late reports and different procedures for processing the reports in the two systems created potential problems for using both of the systems' data files for statistical purposes.

Perhaps the clearest finding was that it is not possible to maintain a usable establishment reporting unit plan for multi unit employers in the absence of systematic procedures for monitoring employer reporting and updating files for changes in the number, location and industry of each employer's reporting units. SSA's Establishment Reporting Plan clearly lacked the necessary resources to do this in 1982 and there is no reason to think that the situation has improved since then.

There was a moderately high but by no means perfect correspondence between county and two-digit industry codes for employers included in both the BLS and SSA systems. A substantial proportion of the differences arose from the absence of county or industry codes in the SSA system. Comparisons of industry codes at the three and four-digit level were not attempted because of the differences in the industry classification systems used by the two agencies.

With some qualifications, we were successful in matching the records of employers, as defined by their EINs, in different systems. However, we were not successful in matching BLS and SSA records for reporting units, the main reason being the incompleteness of SSA's data for reporting units provided under the voluntary ERP. Other reasons were the lack of a common identifier, analogous to the EIN at the employer level, for reporting units and the slight differences in the reporting unit definitions used by BLS and SSA.

We learned what we believe are some important lessons for others who may wish to match business records from different agency sources, whether for research or operational purposes. First, the plans and the necessary interagency agreements should be developed well ahead of the earliest date at which the files to be linked are expected to be available. In particular, the development of interagency agreements for the exchange of identifiable records is a painstaking process and

considerable time may be needed for their completion and approval.

Second, successful matching requires in-depth knowledge of all of the record systems involved and of the specific files that exist within those systems. An interagency team approach, with full exchange of information, is essential because there is unlikely to be a single individual who has all of the necessary information, even for the files of a single agency.

Finally, whenever possible, it is essential to pretest matching procedures before embarking on large-scale operational applications.

Recommendations

ERUMS was designed primarily as a demonstration project and was therefore limited in its coverage and scope. Nevertheless, the Workgroup believes that the study results, along with other information acquired in the course of the study, justified the inclusion in its report of five formal recommendations addressed specifically to the BLS and SSA record systems for employers and reporting units. These recommendations, along with relevant discussion, are as follows:

Recommendation #1 - SSA should undertake a full review of the current status and uses of the Establishment Reporting Plan and decide either to continue it with adequate resources for maintenance and improvement of quality or to discontinue it entirely.

The level of compliance with the ERP is so low that it is clearly of little value for its intended uses. If continued at this level, it would represent an unjustifiable burden on those employers who continue to participate. Discontinuance of the ERP would affect the level of detail available for coding individuals by industry and geography in SSA's Continuous Work History Sample (CWHS). Industry could continue to be coded, but in a single unit context. County codes based on ERP reporting unit locations could be replaced by county codes based either on W-2 addresses or on taxpayer addresses in the IRS individual master file, provided the necessary arrangements could be worked out with the IRS.

Concurrent with the the latter stages of the ERUMS project the Workgroup learned that a full evaluation of the ERP was being undertaken by the Office of Research and Statistics (ORS) at SSA, which we strongly supported. Subsequently, that evaluation has been completed with a resultant conclusion that employer participation in the ERP has declined to the extent that it no longer provides usable information to the statistical systems for which it was intended. With little prospect for adequate resources being available to improve and maintain the system properly, ORS is recommending discontinuing the Establishment Reporting Plan. The recommendation includes alternatives for obtaining geograghic and industry data needed for the statistical records.

Recommendation #2 - BLS should review the State Employment Security Agencies' procedures for identifying employer births (including those resulting from mergers and changes of organization) and seek ways of reducing the apparent lag between filing of applications for EINs and inclusion of new employers on State Agency and BLS lists used as frames for statistical surveys and reports.

It should be noted that the new requirement that states submit UI Address Files to BLS for each quarter is one step in this direction. Delays in deleting deaths from the UI Address File were apparently due in part to the States' practice of imputing employment and payroll for employers who appear to be late filing their quarterly reports.

Recommendation #3 - Data in the UI Address File on employment and wages paid should be labelled to distinguish imputed data from data reported by employers.

The Workgroup have been informed that as of the first quarter of 1989, 40 states had adopted this practice. A related issue which needs to be considered is whether the actual data for these employers, when available to the States, should be submitted to BLS to replace the imputed data in its files. We also noted that slightly more than one percent of the records in the 1982 UI Address File for Texas did not have EINs. The absence of EINs could cause problems for linkages of data for the same employer between states within the UI system or for any linkages with other systems that might be undertaken.

Recommendation #4 - The EIN should be identified as a key item in the UI Address File and efforts should be made to achieve 100 percent reporting initially and current reporting of changes in EINs.

The Workgroup has been informed that BLS has put increased emphasis on complete reporting of current EINs. Also note that the reporting unit definitions used by BLS and SSA are similar, but not identical. Under its new Business Establishment List project, the BLS will be moving toward the collection of establishment-level data, using the OMB definition of an establishment. We have also noted that BLS and SSA use somewhat different adaptations of OMB's Standard Industrial Classification for their own classification of employers and reporting units by industry.

Recommendation #5 - If SSA concludes that it wishes to continue the ERP, BLS and SSA should adopt common definitions of the units for which data are to be reported by employers and identical industry coding structures, based on the SIC. Whether or not the ERP is continued, identical industry coding structures should be used by SSA for coding new employers identified on Form SS-4 and by BLS for coding employers and their reporting units or establishments.

Implementation of this recommendation would be an initial step in following the broad recommendation contained in Statistical Policy Working Paper 6 for agencies to follow consistent procedures in coding reporting unit characteristics (Subcommittee on Statistical Uses of Administrative Records, 1980, Recommendation 3).

In a broader context, the ERUMS Workgroup concluded that current efforts to collect economic data at the establishment level are dispersed among Federal and State agencies, are poorly coordinated, and place unnecessary burden on employers. The Workgroup believes that further, more intensive and extensive interagency matching studies have an important role to play in resolving these problems and in determining the possible effects on statistical programs of prospective major changes in administrative reporting systems for employers. Therefore the Workgroup further recommended that:

Recommendation #6 - Further matching studies should be directed at acquiring information that will support the eventual development of a mandatory reporting system to meet the needs of all Federal and State statistical programs for establishment lists, including SIC codes. An interim goal should be that all agencies requiring or requesting employers to provide data at the establishment or reporting unit level adopt common definitions of units and data items to be submitted for these units.

To the extent possible, such a reporting system should derive most of its information from the major administrative reporting systems. All supplemental information required for statistical purposes should be collected as part of a fully-integrated program, using concepts and definitions agreed on by all users. Three agencies -- the BLS, the Census Bureau and the National Agricultural Statistics Service -- play a dominant role in the direct collection of establishment-level economic data. Recent initiatives of these agencies, under the general guidance of OMB's Statistical Policy Office, have been directed at greater coordination of their respective list-building and maintenance activities. Further integration of business lists will require fuller understanding of the similarities and differences of the three systems, based on matching of individual establishments and reporting units in the different systems.

REFERENCES

American Statistical Association

1980 "Business Directories: Findings and Recommendations of the ASA Committee on Privacy and Confidentiality". The American Statistician, 34:8-10.

Buckler, W.L.

1985 "Employer Reporting Unit Match Study (ERUMS): A Progress Report". Proceedings of the Survey Research Methods Section, American Statistical Association: 434-437.

1988 "Employer Reporting Unit Match Study (ERUMS) -- What have we learned?" Presented at the annual meeting of the American Statistical Association, New Orleans, LA.

Bureau of the Budget

1961 "Brief History of the Movement in the Federal Government for a Central Directory and of Related Efforts Aimed at Improving Quality and Comparability of Economic Statistics". Unpublished report, Office of Statistical Standards. Washington, DC: Bureau of the Budget.

Bureau of the Census

1965 "Final Results of BES-Census Retail Payroll Reconciliation for the State of Delaware". Memorandum from Peter Ohs and Ralph Woodruff to Harvey Kailin and William Hurwitz, July 22. Washington, DC: U.S. Department of Commerce.

Bureau of Economic Analysis

1972 An Evaluation of the Usefulness of the Social Security Administration's Continuous Work History Sample. Report prepared for the Manpower Administration, U.S. Department of Commerce. Washington, DC: Department of Commerce.

Cartwright, D., Levine, B. and Buckler, W.

1983 "An Update on Establishment Reporting Issues: Practical Considerations". Proceedings of the Survey Research Methods Section, American Statistical Association: 481-486.

Grzesiak, T. and Lent, J.

1988 "Estimating Business Birth Employment in the Current Statistics Program". Paper presented at the Annual Meeting of the American Statistical Association, New Orleans, August 21-25.

Harte, J.

1986 "Some Mathematical and Statistical Aspects of the Transformed Taxpayer Identification Number: A Sample Selection Tool Used at IRS". Proceedings of the Survey Research Methods Section, American Statistical Association: 603-608.

Jabine, T.

1984 The Comparability and Accuracy of Industry Codes in Different Data Systems. Committee on National Statistics, National Research Council. Washington, DC: National Academy Press.

MacDonald, B.

1989 "Progress Report, U.S. Bureau of Labor Statistics". Paper prepared for the Fourth International Roundtable on Business Survey Frames, Newport, Gwent, United Kingdom.

Montana Department of Labor and Industry

1987 Montana Business Birth-Death Study: 1984 to 1986. Research and Analysis Bureau, Employment Policy Division.

Office of Federal Statistical Policy and Standards

1980 Report on Statistical Uses of Administrative Records: Statistical Policy Working Paper 6. Washington, DC: Department of Commerce.

Office of Management and Budget

1983 Establishment Reporting in Major Administrative Record Systems. Establishment Reporting Working Group, Administrative Records Subcommittee, Federal Committee on Statistical Methodology. Unpublished report, October 17. Washington, DC: Office of Statistical Policy.

1984 A Review of Industry Coding Systems: Statistical Policy Working Paper 11. Washington, DC: Office of Management and Budget.

1990 A Comparative Study of Reporting Units in Selected Employer Data Systems: Statistical Policy Working Paper 16. Washington, DC: Statistical Policy Office.

Social Security Administration

1988 2000: A Strategic Plan. Washington, DC: Department of Health and Human Services.

NOTE: Because of space limitations, the tables referenced in this paper have not been included. They can be obtained from the contact author upon request.