QUALITY INITIATIVES FOR THE ANNUAL REFILING SURVEY

Aaron Leatherwood and Gordon Mikkelson, Bureau of Labor Statistics
Gordon Mikkelson, 441 G St. NW, Room 2821, Washington, D.C. 20212

KEY WORDS: Quality Assurance, Chi-square test

1. Introduction

The Covered Employment and Wages, or ES-202, program of the Bureau of Labor Statistics (BLS) summarizes quarterly data for workers covered by state Unemployment Insurance (UI) laws and the Unemployment Compensation for Federal Employees (UCFE) program. The data are submitted by the state employment security agencies of the 50 states, the District of Columbia, Puerto Rico, and the Virgin Islands. This program provides an almost complete census of nonagricultural employers and thus represents a comprehensive and accurate source of employment and wage data, by industry, at the national, state, and county levels. As such, the ES-202 series has broad economic significance in evaluating labor trends and major industry developments.

As a major Federal statistical agency in the United States, the BLS conducts dozens of establishment surveys each year involving hundreds of samples. For these surveys, the BLS maintains a file of the names, addresses, and other information for over 6.0 million U.S. business units. This universe file is used by survey offices as the business frame from which most establishment samples are drawn. In addition, it is used during the estimation process for ratio adjusting the survey estimates to the known target population values (benchmarking figures). These benchmark values are summaries of the employment and wage data of units on the universe file at the industry and occasionally at the industry size class level. These summaries also serve as a source for general economic uses and programmatic uses.

Covered Employment and Wages data are used by the Bureau of Economic Analysis of the Department of Commerce to estimate a large portion of the wage and salary component of the national personal income and gross national product. The Social Security Administration also uses ES-202 data to update economic assumptions and forecast trends in the taxable wage base.[1]

In this paper, we discuss some of the initiatives the BLS has undertaken to improve the quality of the ES-202 data and to assure consistency of procedures among the states. In Section 2 we describe the methodology used to update Standard Industrial Classification (SIC) codes, as well as some work that the BLS has undertaken to enhance the methodology. Section 3 describes a method that has been devised to detect potential keypunch errors during data entry. The results of a study that examined the relationship between SIC code changes and the age of the establishment are reported in Section 4. Section 5 summarizes the material presented in the paper, and in Section 6 we offer suggestions for future improvements.

2 Updating SIC Codes - The Annual Refiling Survey (ARS)

Since data collected by the ES-202 program are usually tabulated using a combination of geographic and industry codes, assigning accurate SIC codes to units in the ES-202 program is of fundamental importance. To this end, the BLS, in cooperation with the state employment security agencies, updates the SIC codes of all employers covered under state UI laws over a three year rotational cycle for most industries. SIC codes are updated and verified through the Annual Refiling Survey (ARS).[2]

Industry codes are assigned according to the principles outlined in the Standard Industrial Classification (SIC) Manual.[3] Included in the SIC manual is a description of the activities for each industry code. In the ARS, each unit is assigned a 4-digit industry code based on its primary activity, as determined by the principal product or service rendered.

In 1987 the SIC Manual was revised in order to add new industries, merge some industries, and subdivide some existing industries. Following the 1987 revision, the BLS reviewed and updated the SIC codes of all establishments affected by the changes to the SIC manual. Since then, the BLS has completed another three-year cycle. Concurrently, the BLS has developed new methods to improve the quality of these data.

Annual Refiling Control System (ARCS)
A control file is maintained by each state for the purpose of identifying nonresponses and maintaining records of reporting units for which SIC codes should be updated as a result of the ARS. A carryover file is also produced that enables the states to continue the refiling activities into the next cycle by capturing late responses and carrying over non-respondents into the next year's ARS.

In order to promote uniformity among state operations, the BLS has developed and exported to the states the Annual Refiling Control System (ARCS) which manages all aspects of the ARS. The ARCS creates the control file for the yearly refiling, updates the control file as the refiling cycle progre sses, selects employers for mailing, prints forms, and creates a file that contains all code changes that were made during the ARS.

The ARCS has the means to record survey responses, track individual records throughout the refiling process, and generate summary information about mailings and survey responses. Each record has a response code that is either assigned by the system or the user. When a response is received, a response code is assigned based on whether the response has been reviewed and whether it is a usable response. The system automatically includes in the mailings any mailable record that has not yet responded to the survey.

The ARCS is currently being enhanced to incorporate many suggestions made by the states. The enhancements will include the capability to print newly designed forms with separate industry descriptions for each industry of multi-establishment employers. The capability to print bar codes on forms will be implemented to save processing time and improve the accuracy of data entry during the check-in of returned questionnaires.

The new system will generate on-demand reports that summarize the progress of the survey, tabulate information for different types of records, and review individual records using a variety of selection and sort options. As a result, users will be able to list separate response rates for single-establishment employers, multi-establishment employers, and non-respondents from the previous survey. This will enable the BLS and the state agencies to monitor the survey as it progresses and identify problems as they arise. Quality assurance checks are also being incorporated into the new ARCS. It will identify records that were selected for a quality assurance review, and also record the results of the review. It will also identify records whose employment and/or wage data were changed based on employment and wage edits.

SIC Coding Software
The SIC Manual contains text defining the scope of each industry and line items providing examples of the products or services in that industry. These line items are also listed in alphabetic order in the index of the SIC Manual.

Recently, the BLS has developed an SIC coding system for use by the states in conjunction with ARS activities. The software enables users to search for line items in the SIC manual using key words corresponding to text in the line item for each industry. Additionally, users are able to easily locate key words other than the specific key words listed in the index of the SIC Manual. This system contains approximately four key words for each of the nearly 18,000 line items in the SIC manual. Users can then determine which of the alternative industries presented by the system best characterizes the activities of the firm being coded.

This system will also provide the BLS with the ability to add index items and key words to the SIC Manual. [4] As a result, the BLS will further standardize its coding procedures and thoroughly integrate these standard coding procedures into on-going operations. The system will benefit all users, and will be particularly helpful for inexperienced coders.

3. Detection and Correction of Data Entry Errors

When a large number of SIC code changes occur from one SIC to another, a relationship often exists between the old code and the new code. For example, an establishment classified in SIC 2421 (Sawmills and Planing Mills, General) is more likely to change to SIC 2411 (Logging) than to SIC 5411 (Grocery Stores).

For the SIC code changes that occurred during the past two years, frequency counts were generated indicating the number of establishments changing from each old SIC to each new SIC. Whenever this frequency was greater than twenty [5], the change was defined to be a common change. In the above example, over 100 changes from SIC 2421 to SIC 2411 were recorded. For each common code change,

a list of SIC codes which were numerically similar to the new SIC was generated. Numerically similar SIC codes were defined as any two SIC codes that differ by one digit, or have two digits transposed. If the numerically similar new SIC was sufficiently uncommon (generally, the change occurred either zero or one time), then the numerically similar new SIC was added to a cross-reference table of potential keypunch errors. These potential errors were classed into different potential error types that were defined as follows:

Error Type	Definition and Example
N1:	One digit is different and is horizontally adjacent on the numeric keypad. (2411 - 2511)
N2:	One digit is different and is not horizontally adjacent on the numeric keypad, although the difference between the 2 digits is one. (6513 - 7513)
N3:	One digit is different and is vertically adjacent on the numeric keypad. (2411 - 5411)
N9:	One digit is different, and the difference is not described by N1, N2, or N3. (1542 - 7542)
T:	Two digits are transposed. (7841 - 8741)

Each of the 'N' error types was further differentiated by identifying the position of the digit which differs.

Within each of the general error types (i.e., N1, N2, N3, N9, T), it is assumed that the probability of an SIC code being miskeyed will not vary on the basis of the digit in error. Thus, miskeying 2411 as 2511 should be as likely as miskeying 2411 as 2421, since both of these instances of miskeying 2411 are of general type N1. However, an analysis of the data revealed that approximately twice as many SIC codes were flagged as potentially miskeyed in the third and fourth digits of the SIC code as in the first two digits of the SIC code.

Manual examination of the records flagged as potentially miskeyed in the third and fourth digits indicated that many of these records were not miskeyed records, but were instead valid changes. Further examination of the data also revealed that some general types of keypunch errors (as defined above) were more likely than others. This is consistent with the expectation that keys which are closer together on the keypad are more likely to be accidently pressed than keys which are further apart on the keyboard.

Using this information, records were excluded from the potential keypunch error cross-reference table if the frequency of the common code change was less than an error-specific parameter and they belonged to less likely error categories (general errors N2 or N9; or specific errors involving the third or fourth digit). To screen these less likely general error categories and errors in the last two digits, potential errors were eliminated if the number of changes from the old industry to the new industry was not sufficiently large. The following table provides the minimum number of times a common change must occur for a given error type to remain in the potential keypunch error cross-reference table:

Error Type/ Digit	Min. freq.	Error Type/ Digit	Min. freq.
N1/1	20	N3/1	20
N1/2	20	N3/2	20
N1/3	35	N3/3	35
N1/4	50	N3/4	50
N2/1	35	N9/1	35
N2/2	35	N9/2	50
N2/3	35	N9/3	70
N2/4	50	N9/4	85
T1	20		
T2	20		
T3	35		

For example, if only 40 code changes had been observed from SIC 2421 (Sawmills and Planing Mills, General) to SIC 2411 (Logging), the following types of potential keying errors would not be flagged: N1, N2 or N3 in the fourth digit and N9 in second, third, or fourth digit.

In the new version of the Annual Refiling Control System (ARCS), the system will look for changes from SIC 2421 (Sawmills and Planing Mills, General) to SIC codes that are numerically similar to 2411. When such a change is detected, the system will ask the user if a change to SIC 2411 (Logging) was intended, since the likelihood of a change from SIC 2421 to SIC 2411 is much greater than the likelihood of a change to the codes listed in the numerically similar look-up table for SIC 2411. The

cross-reference table that will be incorporated into this system will include approximately 10,000 of these potential errors.

4. Relationship Between SIC Code Changes and Age and Size of Unit

When a new establishment opens for business for the first time it is assigned an SIC code based on the description of the work activity that is provided on a status determination form that the employer completes. The SIC code is reviewed whenever the industry is included in the Annual Refiling Survey (ARS). This can occur up to three years after the initial SIC code is assigned. The relationship between the age of a unit and the probability that its SIC code changes as a result of the ARS was examined in order to measure the extent to which newer firms change industries more often than older firms and to measure the quality of the initial coding.

Data were used from the 1990 ARS that included wholesale and retail trade. The age of a unit was calculated based on the "Date of Initial Liability" that is stored on the Quarterly Unemployment Insurance Name and Address file. A date of initial liability is assigned by the Unemployment Insurance section of the state employment security agency when the status determination form is filed.

Under certain circumstances the field on the ARS control file that contains the date of initial liability is changed. This occurs whenever there is a change in ownership or reporting configuration. As the Federal-State ES-202 program has moved from county-level reporting of smaller firms to establishment-level reporting, many changes in the reporting configuration of multi-establishment employers have occurred. For this reason, the study excluded multi-establishment employers.

Not all states have fully implemented reporting the date of initial liability on their Quarterly Unemployment Insurance Name and Address files. For the purposes of this study only five of the states which have initial liability dates assigned to all records were selected. Future studies will included an expanded number of states. The five states are Connecticut, Nebraska, South Carolina, Texas, and Virginia.

The size of a unit (i.e., number of employees) is also a factor that was considered when analyzing the relationship between the age of a unit and SIC code changes. Smaller units historically are more likely to change SIC classification. Table 1 illustrates this relationship for the five states in the study.

Table 1. SIC Changes by Size of Unit

Number of Employees	Percent With SIC Changes	Number of Observations
0 to 4	8.71	97,161
5 to 9	6.86	40,723
10 to 19	6.75	25,546
20 to 49	5.97	15,859
50 or More	5.70	6,679
Average	7.70	185,968

The data in Table 1 show that the smaller units tend to change SIC codes more frequently. The null hypothesis that the same percentage of SIC changes occurred in each size class was tested against the alternative hypothesis that the percentage of SIC changes was not the same for all size classes. This was accomplished by performing a χ^2-test to compare the number of observed changes with the number of expected changes for each size category. The result was significant at the $\alpha = .001$ level.

Table 2 shows the relationship that exists between the age of a unit and the likelihood that its SIC code would change during the ARS.

Table 2. SIC Changes by Age of Unit

Date of Birth	Percent With SIC Changes	Number of Observations
88.2 - 89.1	10.70	22,569
87.1 - 88.1	9.03	18,798
84.3 - 86.4	8.51	35,438
81.3 - 84.2	7.64	28,934
75.1 - 81.2	6.47	38,150
Before 75.1	5.95	42,079
Average	7.70	185,968

Note: The entries in the date of birth column refer to the date of initial U.I. liability, and are expressed as the year and quarter (e.g., 88.2 refers to the second quarter of 1988). The last column refers to the number of records in the size category and not the number of records with SIC changes in the size category.

The data in Table 2 show that there is a relationship between the age of a unit and the probability that the SIC code will change in the ARS. Again, a χ^2 one-sample test was performed to compare the number of observed changes and non-changes for each age category. This result was also significant at the $\alpha = .001$ level. The results confirmed what was believed to be true - that newer units have a greater tendency to change business activities.

A remaining question is whether the newer units change SIC codes more frequently because smaller firms tend to be newer. Table 3.1 combines the data from Tables 1 and 2 to show the percent of firms with SIC changes based on age and size.

Table 3.1 SIC Changes by Age and Size

	SIZE				
Date of Birth	< 5	5-9	10-19	20-49	50+
88.2 - 89.1	12.26	8.43	7.22	5.38	5.62
87.1 - 88.1	9.81	8.05	8.49	5.31	8.01
84.3 - 86.4	8.95	8.14	8.23	7.52	5.22
81.3 - 84.2	8.00	7.71	7.01	6.41	6.90
75.1 - 81.2	6.95	5.79	6.57	6.12	5.21
Before 75.1	6.93	5.22	5.40	5.27	5.36

Table 3.2 Number of Records Refiled by Age and Size

	SIZE				
Date of Birth	< 5	5-9	10-19	20-49	50+
88.2 - 89.1	15195	3867	1965	1115	427
87.1 - 88.1	11712	3651	1906	1092	437
84.3 - 86.4	20511	7626	4117	2379	805
81.3 - 84.2	15184	6720	3868	2278	884
75.1 - 81.2	17844	9147	6138	3659	1362
Before 75.1	16715	9712	7552	5336	2764

A χ^2-test was performed on the data in Table 3.1. The result was significant at the $\alpha = .001$ level. Within each age category (i.e., row in Table 3.1) the highest percentage of units with SIC changes occurs in the smallest size class. This illustrates that smaller units have a higher probability of changing SIC codes than larger units of the same age.

One final point is worth noting concerning the data in Table 3.1. The units whose date of initial liability was between the second quarter of 1988 and the first quarter of 1989 had become active after the previous refiling of wholesale and retail trade. SIC changes for these units, therefore, are changes from the SIC codes that were initially assigned to the firm rather than SIC codes that were verified or changed from a previous cycle of the ARS. The data in Table 3.1 show that for establishments with less than 10 employees the percentage of cases with SIC changes is greatest for the cases that have not been through a previous ARS. Larger firms that have not been through a previous refiling did not tend to change more often than older establishments; however, the estimates for these larger establishments are based on a much smaller number of observations, as shown in Table 3.2.

5. SUMMARY

As the capabilities of the ARCS expand, the Annual Refiling Survey will continue to become more effective. Additional systems, such as the computer-assisted SIC coding software the BLS has developed, will complement the capabilities of the Annual Refiling Survey. Within the context of the ES-202 program, other projects such as the keypunch error detection methodology and our studies of SIC changes by age and size of unit have increased our ability to collect accurate data and led to a greater understanding of the nature of this data.

The methodology used by the BLS to detect keypunch errors during data entry of SIC codes can be generalized to improve data integrity for other data collection processes when the following conditions are met:

1. Data are not being double-keyed. If data are entered twice, error rates are likely to be very low and this methodology will provide little benefit.

2. Some types of data changes within a cohort are more common than others and these common data changes can be identified.

3. Unlikely data changes within the cohort exist and are numerically similar to common data changes.

Under these conditions, this methodology provides a relatively inexpensive alternative to double-keying data.

The BLS has also observed that small businesses tend to change industries more often than larger businesses. This trend is not dependent on how long an establishment has been in business: within any

given age group, small businesses are more likely to change industry. One possible explanation is that small businesses are more capable of adapting to changing market conditions and adjusting their mix of products and services.

6. FUTURE IMPROVEMENTS

As the Covered Employment and Wages program has expanded, the importance of improving data quality in the SIC coding process has increased. In the future, the BLS will analyze survey response rates for each mailing of single-establishment employers, multi-establishment employers, nonclassifiable employers, and non-respondents to the previous year's refiling. The BLS will also develop a feedback process which will utilize data obtained through other surveys. These data will be analyzed to examine if some industries are impacted more than others as a result of SIC changes. Finally, the BLS is planning to conduct a response analysis survey as time and resources permit to measure the magnitude of response errors that are caused by misinterpretation of questions and definitional difficulties on the questionnaire.

REFERENCES

[1] BLS Handbook of Methods, pages 35-40

[2] Bureau of Labor Statistics. "ES-202, Operating Manual," *Employment Security Manual*, Part III, Section 0424, revised on a regular basis.

[3] Executive Office of the President, Office of Management and Budget. "*Standard Industrial Classification Manual*, 1987".

[4] The source of these index items and key words is separately identified by the coding software. They are then incorporated in the electronic copy of the SIC manual.

[5] Our research indicated that: 1) using code changes that occurred less than 20 times caused a disproportionate number of records to be incorrectly flagged as potentially miskeyed; and 2) the shift in business activity represented by changes that occurred less than 20 times did not represent the types of changes that were likely to have actually occurred, and were more likely to be keypunch errors.