IMPLEMENTING THE NAICS FOR BUSINESS SURVEYS AT BLS

Gordon Mikkelson, Teresa L. Morisi, George Stamas, U.S. Bureau of Labor Statistics
George Stamas, Bureau of Labor Statistics, Suite 4985, 2 Massachusetts Ave NE, Washington, DC 20212
Stamas_g@bls.gov

ABSTRACT

To implement the North American Industry Classification System (NAICS), the Bureau of Labor Statistics and State partners are assigning NAICS codes to the approximately 8.2 million employers covered by State unemployment insurance (UI) laws. Employer UI reports are the basis of the Longitudinal Data Base (LDB), which serves as the frame for BLS establishment surveys. The NAICS conversion includes a multi-year process of gathering information from employers in order to assign NAICS codes. The collection procedure allows for interim assessment of the effect of the NAICS conversion on industry classification and BLS products. When employers do not provide adequate information for industry classification, BLS will assign NAICS codes based on the distribution of those codes across other establishments with the same Standard Industrial Classification (SIC) and other characteristics. These procedures will be applied to current and, to the extent feasible, historic data on the LDB including establishments that are out of business. This provides a frame for surveys requiring stratification by NAICS and aids in the conversion from SIC to NAICS for ongoing surveys. In addition, the availability of a continuous history with NAICS codes will permit seasonal adjustment and other time-series analysis of the data.

Key Words: Industry classification, Sampling frame, Nonresponse

* All opinions expressed in this paper are those of the authors and do not constitute policy of the Bureau of Labor Statistics.

1. INTRODUCTION

The Covered Employment and Wages Program, commonly referred to as the ES-202 program, is a cooperative program between the Bureau of Labor Statistics (BLS) of the U.S. Department of Labor and the State Employment Security Agencies (SESAs). The ES-202 program produces a comprehensive tabulation of employment and wage information for workers covered by State unemployment insurance (UI) laws. Employer UI reports also are the basis of the Longitudinal Data Base (LDB), which serves as the sampling frame for BLS establishment surveys. For more information on the ES-202 program, see the *BLS Handbook of Methods*, Bulletin 2490 (Bureau of Labor Statistics, April 1997).

In order to implement the North American Industry Classification System (NAICS) for the ES-202 program, the BLS and its State partners are assigning NAICS industry codes to the approximately 8.2 million employers covered by UI laws. The NAICS conversion includes a multi-year process of gathering information from employers in order to assign NAICS codes. Whenever employers do not provide adequate information for industry classification, BLS assigns NAICS codes based on the distribution of those codes across other establishments with the same Standard Industrial Classification (SIC) and other characteristics. These procedures will be applied to current and, to the extent feasible, historic data on the LDB including establishments that are out of business. This provides a frame for surveys requiring stratification by NAICS and aids in the conversion from SIC to NAICS for ongoing surveys. In addition, the availability of a continuous history with NAICS codes will permit seasonal adjustment and other time-series analysis of the data.

2. WHAT IS NAICS?

NAICS was established in 1997 through a cooperative effort among the United States, Mexico, and Canada. The Bureau of Labor Statistics worked closely with the Bureau of the Census, the Bureau of Economic Analysis, and

other U.S. statistical agencies to achieve the goal of developing NAICS. NAICS replaces the SIC (Standard Industrial Classification) system that has been in place since the 1930s and was last revised in 1987 (NAICS, 1998).

NAICS was developed based on the economic concept that establishments should be grouped together according to similar production processes. This coding system focuses on the identification of new and emerging industries and high technology industries, and provides increased detail in the services sector over what was available under the SIC system. It uses a six-digit classification system that generally provides three-country comparability at the five-digit level. Under NAICS, the highest level of aggregation is the sector, of which there are 21. This compares to the 10 divisions available under the SIC system. NAICS includes nine new service sector aggregations that were not found under the SIC system. For additional information concerning the NAICS coding system, see Ambler (1998) and Murphy (1998).

3. OTHER CHANGES WITH NAICS

The treatment of auxiliaries will change 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, auxiliary units were classified according to the primary activity of the company they served. BLS is conducting a special survey in fiscal year (FY) 2000, in order to verify auxiliary status and assign NAICS codes to auxiliary units reflecting that status. Non-auxiliary units will be automatically assigned a code that matches their NAICS code. The success of the survey on auxiliaries will be especially important, because ES-202 data under NAICS will be tabulated and published using the NAICS treatment of auxiliaries.

A revision to NAICS 1997 is on the way -- NAICS 2002. The three countries are currently working on proposed changes to the construction and wholesale trade sectors, because agreements were not reached on these two sectors during NAICS 1997. In addition, changes are planned within the Information and Retail Trade sectors in order to better capture Internet-related activities.

The task for BLS and its State partners is to assign NAICS codes to the 8.2 million business establishments in the ES-202 program. At the same time, the ES-202 program will verify SIC codes (and assign SIC codes to new units) in order to create linkages between the two industry classification systems. BLS plans to implement NAICS over a four-year period. By the end of FY 2001, all establishments in the ES-202 program will be assigned NAICS 2002 codes. The first step in assigning NAICS codes will be to contact the employer directly, as described in the next section. Those units that do not receive NAICS codes through this process (i.e., nonrespondents) will be assigned one by an imputation process developed by BLS (described later in this paper). BLS will incorporate NAICS 2002 changes, using the NAICS treatment of auxiliaries, with the first publication of ES-202 data under NAICS. This will be published in 2002, for reference year 2001. This schedule will ease the burden on data users by providing a single change in coding structure from the 1987 SIC to NAICS 2002.

4. ASSIGNING NAICS CODES BY CONTACTING THE EMPLOYER

4.1 The Refiling Process

The ES-202 program updates classification codes using a process known as "refiling," in which the employer receives a form from their State Employment Security Agency (SESA). The employer will verify or update the information contained there, including the primary business activity of the establishment. The form will ask the employer to select an appropriate NAICS-based industry description for the establishment. The SESA will then assign a NAICS code based on this response. Some SIC and NAICS code combinations will be direct matches, that is, the SIC code is associated with only one NAICS code. Split combinations, or non-directs, occur when the SIC code maps to more than one NAICS code. During the refiling process, BLS targeted directs and non-directs, as well as records that had no NAICS code, an unclassified NAICS code, or an invalid one. During the last year of implementation, BLS will refile establishments affected by changes in NAICS 2002. Details are as follows:

FY 1998 Establishments in direct industries were automatically assigned a NAICS code by a computer program. This affected approximately one-half of establishments in the ES-202 program. The direct match program is run periodically to assign codes to any records that have direct match SICs but no NAICS code.

FY 1999 All units with employment greater than or equal to 50 (including directs recoded in FY 1998) were selected to receive a refiling form as well as units that had SIC 9999 (Unclassified), or SIC 9621, (Regulation and Administration of Transportation Programs). Records collected by the BLS EDI (Electronic Data Interchange) center were also refiled. Finally, a random sample of the UI accounts with less than 50 employees and worksites with SICs that could not be directly matched to one NAICS code were refiled.

FY 2000 Selected during this fiscal year were units that lacked a NAICS code, had an unclassified NAICS (NAICS 999999) or had an invalid NAICS code. A survey is being done in FY 2000 in order to verify the auxiliary status of auxiliary units and to assign a corresponding NAICS code.

FY 2001 Included in this year's refiling will be those SICs impacted by the NAICS 2002 revision.

4.2 Response rates

The success of the revision from SIC to NAICS requires that BLS and its State partners work diligently to ensure accuracy and completeness in the conversion to NAICS codes. To meet this objective, States pursue a goal of achieving usable response rates of at least 90 percent, in both units and employment, during each year's refiling cycle. Usable responses are those that receive a NAICS code through the refiling process. Establishments receive up to three non-response follow-up mailings. By December 1999, 72 percent of records in the ES-202 program had received NAICS codes from the refiling process; in terms of employment, 84 percent had been assigned NAICS codes. See industry details in the adjacent table. For the remainder of units that do not receive NAICS codes from the refiling process, BLS will assign NAICS codes through an imputation process as described later in this paper.

Division	Percent of Records Coded	Percent of Employment Coded
Agriculture, forestry, fishing	86.8	85.5
Mining	94.6	98.0
Construction	88.8	90.1
Manufacturing, durable	71.2	86.9
Manufacturing, nondurable	67.7	86.7
Transportation, public utilities	59.9	79.4
Wholesale trade	63.0	73.7
Retail trade	69.0	80.3
Finance, insurance, real estate	62.1	73.7
Services	72.7	85.5
Government	97.7	98.0
Total with NAICS	71.9	84.0

5. ESTIMATION WITH FIRST QUARTER 1999

Most establishment surveys that BLS conducts use historical time series data in order to evaluate current economic activity. The implementation of a new industry coding system has a significant impact on the continuity and value of these time series. Because the assignment of NAICS codes is phased in over a four year period, BLS programs that maintain time series need to be able to estimate the movement of economic activities between the SIC and NAICS codes before all of the establishments have been assigned NAICS codes.

For purposes of estimation, the records are divided into three types: direct matches, certainty records, and sample records. Each of these is handled separately during the estimation process. For the purpose of calculating weights used for estimation, UI accounts were stratified by state, 4-digit SIC, and employment size class.

5.1 Direct Matches

The direct matches are records that have only one NAICS code associated with the SIC code for that record. Included are single worksite accounts with an average monthly employment (AME) of 50 or less, and multiple

worksite accounts with all worksites in direct SICs and a total AME of 50 or less. In the estimation process, these records receive a final weight of 1.000, and a non-response adjustment is not needed.

There is no accounting for out-of-business UI accounts although some proportion would fall into this category. Since these are small UI accounts, out-of-business units could be a substantial part of this category. This may lead to an overestimation of the number of units and employment.

There is also no accounting for movement of these direct units into non-direct NAICS codes or into other direct NAICS codes among these small employers. For example, all units in SIC 0112 would be coded to NAICS code 111160. They cannot be classified into NAICS codes like 111140 or any of the other non-direct NAICS codes like 111150. Therefore, the number of units and employment for direct NAICS codes would be overstated while for non-direct NAICS codes it would be understated.

5.2 Certainty Records and Sampled Records

We designated UI accounts with AME of at least 50 as "Certainty" and selected 100 percent. We randomly selected about one-half of the UI accounts consisting of single records in split SICs with an AME of less than 50, and multi-unit accounts that have at least one sub-unit in a split SIC, and an AME of less than 50. We called these "Sampled." Each of these groups of UI accounts, Certainty and Sampled, was stratified across 4-digit SIC and size class.

Within each stratum, we calculated sampling weights, *N/s*, where *N* is the number of UI accounts in a stratum and *s* is the number of accounts selected for refiling. The weight was generally 1.000 for certainty strata and about 2.000 for sampled strata. The non-response adjustment factor is *s/r*, where *s* is the number of UI accounts that were selected and *r* is the number of UI accounts that responded including out-of-business UI accounts. In the absence of any other information, the assumption is made that the distribution of non-respondents is the same as that of respondents.

For multi-establishment UI accounts, a partial response is considered a respondent. For these UI accounts, a weight adjustment is done to account for the non-responding sub-units. This adjustment, *p*, is the ratio of the sum of employment across all reporting units in the account divided by the sum of employment across all of those with NAICS codes. In addition, for multi-establishment accounts, all sub-units have to be out-of-business for the account to be classified as out-of-business. The final weight is equal to the sampling weight times non-response adjustment times the partial adjustment, *(N/r) * p*, and is assigned to each sub-unit of a UI account. Estimates were calculated by summing data of appropriate establishments to aggregated levels. Essentially, the formulas in the box were used. fwt_i are the final weights and the summation is across all reporting units in any group of interest. Ratio tables that show the distribution of units, employment, and wages from each SIC across the NAICS codes associated with the SICs were also produced.

$$\hat{N} = \sum_i fwt_i$$

$$\hat{EMP} = \sum_i (fwt_i)(EMP_i)$$

$$\hat{Wages} = \sum_i (fwt_i)(Wages_i)$$

6. IMPUTING NAICS CODES WHERE THEY ARE MISSING

We need NAICS codes assigned to every record in the database for sampling on a NAICS code basis and for aggregating records to publish summaries and other statistics. We will apply an imputation procedure, state by state, to assign NAICS, NAICS corresponding to auxiliary status, and NAICS 2002 codes where they are missing on the 2000 and 2001 files.[1] For an overview of imputation procedures, see Kalton and Kasprzyk (1982).

[1] In order to approximate the distribution of NAICS codes across records with reported SIC and NAICS codes, the Bureau of the Census used a random assignment process. This process used digits from the Employer Identification Number (EIN) from each record missing a NAICS as random numbers. They established ranges for

Our imputation of NAICS 1997 in the summer of 2000 and NAICS 2002 in the summer of 2001 will use a nearest neighbor procedure. First, we will assign NAICS codes automatically to records with direct match SICs. Then we will apply an imputation procedure to assign NAICS codes to any records that remain without them. This nearest neighbor procedure will choose a donor record with the closest average employment from among those records with the same SIC and a state-assigned NAICS code. Ties among donors will be broken with a random assignment process. The process is based on the assumption that among records with the same 4-digit SIC, employment is a significant explanatory variable when determining NAICS assignment. The algorithm will be applied first to records from UI accounts with multiple worksites reported, and then to any remaining records without NAICS codes. Before imputing codes, the files will be edited for invalid SIC/NAICS conditions. Records that do not pass this edit will not be used in the imputation process and will be forwarded to the states for correction.

Occasionally, none of the records for a given SIC will have a state-assigned NAICS. In these cases, the procedure will go to a national summary file, with records of observed SIC and NAICS combinations, and will choose a donor record with the closest average employment from among those with the same SIC, and assign that NAICS code.

The first type of record requiring imputation comes from UI accounts with multiple worksites, where some records in that account for a given SIC have NAICS codes reported but others do not. The imputation will be carried out using only records reported with that UI account. For each SIC assigned to records in any such UI account, we will determine whether any records have a NAICS code assigned. If none of the records has a NAICS code, then we will calculate the average employment across the records, search the national summary file and choose from among those with the same SIC the NAICS code with the closest average employment. If some of the records with the same SIC in the UI account have a NAICS code assigned, we will determine whether there is only one NAICS code or more than one. If there is only one NAICS code, then we will assign that code to every record with that SIC. Otherwise, we apply the nearest neighbor method to assign codes.

We will impute for any single-site UI account without a NAICS code using the same algorithm. First, we will attempt nearest neighbor imputation using responses from records with the same SIC. If no such record are available, we will search the national summary file and choose the NAICS code with the closest average employment from those with the same SIC. In the event that a particular SIC has no responses for NAICS in the national summary file, we will determine the possible NAICS codes for that SIC and assign those codes randomly across the UI account having that SIC.

The process of imputing NAICS codes that reflect auxiliary status where those codes are missing will treat records the same whether they come from a multiple-worksite UI account or a single-site account. Because these records will be so limited in number, the procedure will go directly to one of two national summary files. One file has all reported NAICS, and auxiliary code combinations with reported NAICS codes reflecting auxiliary status and average employment. The other file has all reported SIC and auxiliary code combinations with reported NAICS codes reflecting auxiliary status and average employment. First, any record without an auxiliary code or with a code indicating that the record is an operating facility will have the NAICS code assigned to the field for the NAICS code corresponding to auxiliary status. Records with an auxiliary code indicating that the record is a headquarters or regional office, will be assigned the NAICS code “551114” indicating it is a headquarters. For any other record with the NAICS and auxiliary codes reported, we will search the national file for the records with that NAICS and auxiliary code, and choose the auxiliary treated NAICS code with the closest average employment. If the record missing an auxiliary treated NAICS code is also missing the NAICS code, we will search records with the same SIC and auxiliary codes and assign the auxiliary treated NAICS with the closest average employment. If

---

random assignment of NAICS based on the proportion of records assigned each NAICS in a given SIC/NAICS group (Census, unpublished internal memo). Statistics Canada did not have to deal with missing NAICS codes because they handle industry coding centrally.

there is neither a NAICS nor an SIC code, we will search records with the same auxiliary code and assign the auxiliary treated NAICS with the closest average employment.

7. CREATING A TIME SERIES

The BLS maintains a longitudinal data base (LDB) that links UI reports from businesses through ownership changes, to the extent possible. Each quarterly record on the LDB has an LDB number that links the records for an establishment back through time. This makes historical data for the establishment easily available for analysis. Over time the industry classification of some establishments in the data base have changed. The ES-202 program is considering how to assign NAICS codes to historical records, in light of these classification changes. One approach would be to assign the code imputed for the most recent quarter to earlier quarters as well, regardless of changes in classification. An alternative approach would be to assign the NAICS codes independently each time the classification changes, applying the same algorithms used to assign codes to First Quarter 2001 records. That NAICS code would then be carried back in time as long as the classification remains unchanged.

Establishments that have gone out of business prior to First Quarter 2001 are also part of the LDB and historic time series, therefore they must be assigned a NAICS code based on SIC. These records will be assigned codes by applying the same algorithms within each quarter.

8. CONCLUSION

Implementing the NAICS at BLS will involve a multi-year process that is nearly complete. It is imperative that the implementation be done in both an accurate and timely manner. Other BLS programs will begin publication based on NAICS effective with data year 2001. (See the box.) Therefore, by data year 2001, all 8.2 million employers in the ES-202 program must have NAICS codes. BLS has produced estimates of the effect of NAICS on employment and wages reported under the ES-202 program. Normal nonresponse will require BLS to assign NAICS codes to those employers that remain without one after the refiling process ends. BLS has designed an imputation process to assign NAICS codes to those records that do not have them.

BLS IMPLEMENTATION SCHEDULE

Reference Year	Program
2001 -	Covered Employment and Wages--ES-202
	Job Openings and Labor Turnover Survey
2002 -	Occupational Employment Statistics
	Mass Layoff Statistics
2003 -	Current Employment Statistics
	Productivity measures for selected industries
	Foreign Labor Force Statistics
	Occupational Safety and Health Statistics
	Current Population Survey
2004 -	Employment Projections
	National Compensation Survey
	Producer Prices Indexes

9. REFERENCES

Ambler, Carole A. (1998), "NAICS and U.S. Statistics," *Proceedings of the Section on Government Statistics and the Section on Social Statistics, American Statistical Association, pp. 21-30.*

Executive Office of the President, Office of Management and Budget (1998), *North American Industrial Classification System - United States, 1997*.

Kalton, G. and Kasprzyk, D. (1982), "Imputing for Missing Survey Responses," *Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 22-31*.

Murphy, John B. (1998), "Introducing the North American Industry Classification System," *Monthly Labor Review*, July 1988, Vol. 121, No. 7, pp.43-47.

U.S. Department of Labor, Bureau of Labor Statistics, Bulletin 2490 (1997), *BLS Handbook of Methods, pp. 42-47*.