Proceedings of Statistics Canada Symposium 2001
Achieving Data Quality in a Statistical Agency: A Methodological Perspective

PROCEDURES TO ACCOUNT FOR ENTRIES IN BUSINESS SURVEYS[1]

Carol S. King and Robert E. Struble[2]

ABSTRACT

The Economic Directorate of the United States Bureau of the Census conducts a number of surveys designed to provide estimates for several sectors of the economy, including Mining, Manufacturing, Wholesale and Retail Trade, and Services. Most of these programs rely on probability samples to represent the specific target populations. The true populations are constantly changing; there are entries, exits, organizational changes, and classification changes. To assure that our samples remain representative of the changing populations, sample maintenance procedures have been developed and implemented. One significant part of sample maintenance is accounting for births and entries. This paper presents and contrasts the sample maintenance procedures for births and entries currently in place for the various sectors.

KEY WORDS: Sample maintenance; births; entries.

1. INTRODUCTION

The identification of entries in the Retail, Wholesale, Service, and Manufacturing sectors relies heavily on administrative data received from other agencies of the United States government, specifically the Internal Revenue Service (IRS), the Social Security Administration (SSA), and the Bureau of Labor Statistics (BLS). The confidentiality and authorized use of this information is strictly regulated by law (United States Code). The administrative data received from these agencies includes business name and address, industry classification, quarterly and annual payroll, number of employees, annual sales or receipts, and company affiliation. These data are the basis for construction and maintenance of the Census Bureau's Business Register (BR). The BR contains all known employer business establishments in the United States.

New businesses that have paid employees must file an application for a Federal Employer Identification Number (EIN) with the IRS. The EIN assigned by IRS is used by the employer businesses as their taxpayer identification number when reporting payroll and employment data. According to legal provisions, the IRS regularly transfers selected data from tax returns to the Census Bureau for the purpose of producing official statistics. The BR uses the EIN as the primary identifier for single establishment businesses, and as the secondary identifier for multiestablishment businesses. This information is used to identify births in the Retail, Wholesale, and Services sectors and entries in the Manufacturing sector.

The term 'entries', as used in this paper, includes both new businesses (births) in the business sector as well as existing businesses that enter the sector through changes in industry classification. The procedures described
for the Retail, Wholesale, and Services sectors (section 2) deal exclusively with births while the procedures

---

[1] This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a Census Bureau review more limited in scope than that given to official Census Bureau publications. This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress.

[2] United States Census Bureau, Washington, D.C., USA, 20233

described for the Manufacturing sector (section 3) cover both births and selected changes in classification.

2. PROCEDURES TO ACCOUNT FOR BIRTHS IN THE RETAIL, WHOLESALE, AND, SERVICE SECTORS

2.1 The Initial Samples

About every five years, as soon as possible after the most recent Economic Census, new samples are selected for surveys of the retail, wholesale, and services sectors. These surveys collect data used to compute estimates of totals and trends at the U. S. level for various kinds of businesses as defined by the 1999 North American Industry Classification System (NAICS). The data collected for the retail and wholesale sector include sales, end-of-year inventory, and value of purchases. The services survey collects receipts data, as well as data specific to particular NAICS industries.

The initial employer frame for these surveys is made up of two types of sampling units, companies and EINs. Companies are multiestablishment businesses with annual sales exceeding specific sales size cutoffs for their industries while the EINs are either EINs of multiestablishment businesses not sampled as companies or the EINs of single establishment businesses. All EINs selected in the initial samples, including EINs associated with certainty companies, are marked on the BR with a code denoting each EIN's sampling status: selected in retail, wholesale or services; subjected and not selected or out-of-scope either because of kind-of-business classification (e.g., manufacturing) or because of geographic location (e.g., Canada).

2.2 Identification of Births

An integral part of maintaining the quality of the estimates produced from these surveys is assuring that the surveys' samples, as much as possible, represent the target population for the time period that the estimates are made. Therefore it is necessary to update the initial samples in a timely manner with new businesses that start operation subsequent to the initial sample selection. For the retail, wholesale and services monthly and annual surveys, all new businesses added to the initial sample are added as EIN sampling units. These EINs associated with these new businesses are called birth EINs, or births.

Our goal is to identify and represent births in the samples as soon as possible after the new businesses start operation. We attempt to meet this goal by extracting births from the BR and subjecting them to sampling on a quarterly basis. Note that we identify new EINs. Also, births may not actually be new businesses, but established businesses that for some reason (e.g., reorganization) acquired a new EIN. Even though a new EIN can be associated with a multiestablishment business, most of these EINs are associated with single establishment businesses. In any case, we traditionally refer to the EIN as the specific business unit it is associated with. That is how the term "birth" will be referred to for the remainder of the discussion.

We extract births that satisfy the conditions noted below:

1) has not been identified as subjected to sampling or identified as out-of-scope
2) is classified inscope to the retail, wholesale or services sector
3) if not classified, the birth has:
   a) more than 3 employees or
   b) nonzero payroll in at least one of the latest four quarters
4) is Business Master File (BMF) active, i.e., has an IRS Form 941, "Employer's Quarterly Federal Tax Return", active filing requirement. IRS mails a Form 941 to each business that has employees and the business uses the EIN to file the quarterly report.
5) has a valid mailing address

2.3 First Phase Sampling Process

The extracted births are input into a first phase sampling process. This process begins with the identification of births associated with units already in our samples. Once identified, these births are removed from further processing. The remaining births are assigned two weights. One weight is based on the birth's quarterly payroll and the other on its employment. The weight ultimately assigned is the one that gives the birth the largest probability of selection. In most instances this is the payroll weight because payroll is available more often than employment. Equal probability systematic sampling is done within NAICS industry by size stratum. Births with no classification are sampled in an unclassified industry by size stratum. About 125,000 - 170,000 births are identified each quarter. Approximately 10% of the births are selected in the first phase.

We mail the selected births a form called the SQ-CLASS, Business and Professional Classification Report. The forms are mailed from and returned to Jeffersonville, Indiana for clerical processing and keying. The canvass of the selected births in this first phase results in the collection of two months of sales (receipts), type of operation, company organization information, taxable or tax-exempt status, wholesale inventories, and electronic commerce information. The survey is not mandatory and the response rate is about 70%.

2.4 Second Phase Sampling Process

Births mailed a SQ-CLASS form are input into the second phase sampling process. Part of this process is to update the administrative data of the births since their first phase mailing. This allows us to use the most recent data for kind-of-business classification and size determination. Just as some births are excluded from the first phase sampling, there are also some births excluded from second phase sampling. These births include those identified as having an association with units already in our samples, having been mailed as unclassified and determined to be out of scope based on the business' responses to the SQ-CLASS form, units associated with central administrative offices, or units associated with government entities.

Based on information collected on the SQ-CLASS form, each birth is assigned a new or more detailed NAICS code and a measure of size (MOS). The MOS is used to determine the second phase sampling weight assigned to the birth and is computed using the two month of sales (receipts), if reported. For the births that do not provide the monthly data, the MOS is computed from payroll or employment.

If the birth has not yet started business, or has not provided enough information to be sampled, it is retained on what is called the unsampled birth register. If administrative data updates from the subsequent quarter provides sufficient data, the birth on the unsampled birth register can be subjected to sampling in that subsequent quarter. Otherwise, the business associated with the birth is mailed a SQ-CLASS form the following quarter, i.e., six months after the birth's initial mailing. The business must still be in operation and have recently reported payroll. This every-six-months mailing will continue until a response is received or sufficient administrative data becomes available. For any given quarter there are about 15,000 births on the unsampled birth register, most of which are there because of insufficient NAICS classification for sampling.

The second phase sampling is done three months after first phase sampling. About 20% of the first phase selects are selected in second phase sampling. Sampling for this phase is probability proportional to size systematic sampling conducted from quarter to quarter.

Selected second phase births are added to the monthly and annual samples. The average amount of lag for a birth to be added is 6-9 months. The lag is influenced by a number of factors including the time between the birth starting operation and filing of the IRS Form SS-4 (Application for Employer Identification Number) and the time needed to conduct the mailings for the first phase selected EINs. As done for the EINs in the initial samples, a status code is assigned to each birth reflecting the results of the sampling operation.

2.5 Quality Issues Associated with the Birth Processing

It is agreed that reducing the effects of nonsampling error is important to the goal of producing quality estimates. The quarterly birth process attempts to address coverage issues affecting our economic programs by including in the quarterly birth process procedures to identify inscope births, to sample in two phases to obtain accurate kind-of-business classification and size information, and to check for duplicate coverage. However, two of our major concerns are the timeliness of our process and the accuracy of our kind-of-business codes.

2.5.1 Timeliness of the Birth Process

Ideally we would like to represent a birth in our samples as soon as it has business activity. As previously noted the lag of introducing births into the surveys is about 6 to 9 months. We introduce these births into the monthly canvass every three months. Realistically, the quickest that we could possibly introduce births into our surveys would be monthly. However, there are a number of factors that come into play if the attempt is made to introduce births sooner than on a quarterly basis. These factors are:

1) Receipt of administrative records. We do not receive all administrative data used to identify inscope births in a uniform manner during the year. This results in the number of births identified on a quarterly bases being more uniform than the number that would be identified on a monthly basis. This uniformity would affect the number of births that would be introduced on a monthly bases and could very well affect our month-to-month change estimates.

2) Processing. We currently give respondents 30 days to respond to the SQ-CLASS form. We could require a response to a collection in fewer than 30 days if we (1) have good cause, and (2) explain that good cause to the Office of Management and Budget (OMB). OMB requires that agencies explain in the clearance package any circumstances requiring response to a collection of information in fewer than 30 days after receipt.

3) Effect of introducing of births into the monthly estimates. We would need an effective method for introducing births so that they would not artificially effect the month-to-month change estimates of sales and inventories. We do have a method in place with our quarterly processing.

2.5.2 Business Classification

The business classification, if available at the birth identification process, can come from a variety of sources, in particular, the SSA, IRS, or BLS.

- The basis for SSA's industry coding is the IRS SS-4 which SSA receives a copy in order to provide a business classification code. This form includes information on the principal business activity, number of expected employees, type of organization, and class of customer.

- On a quarterly basis, the Census Bureau provides BLS with single establishment businesses from the BR that are unclassified on the BR, partially classified, or have randomly assigned classification codes. BLS matches these establishments to their files and provides to the Census Bureau with any business classification codes they may have.

- A principal business activity code may also be supplied from IRS on the BMF records we receive.

Even with these sources about 41% of the births extracted in any particular quarter are unclassified. Many of the births with classification are partially coded, i.e., have fewer than the full six-digits in the assigned NAICS code. Additionally, about 20% of the units mailed out with a particular NAICS code are recoded to an entirely different NAICS code (which may or may not be inscope to retail, wholesale and services surveys). About 30% of the codes match at the five-digit level, which is the minimum number of digits needed for sampling.

2.6 Future Work

We are currently evaluating the feasibility of a one phase sampling process. This would require an accurate kind-of-business classification as well as MOS estimates determined solely from administrative data. Additionally, we are considering the introduction of births into our monthly and annual samples annually instead of quarterly.

There has also been a suggestion for the birth processing to expand to:

- include births currently considered out-of-scope to the retail, wholesale and services surveys
- increase the first phase sample in selected sectors/subsectors to emphasize industries that are known to have or are suspected of having classification problems (e.g., wholesale versus retail trade)
- increase the sampling rate for smaller units
- lower the payroll cutoff that determines whether a birth is taken as certainty.

The advantages of this expansion are the following:

- Eliminating the need for retaining out-of-scope EINs on reserve scope. In the identification phase of our quarterly birth processing, we miss some retail, wholesale and service EINs that have been misclassified as out-of-scope. To compensate, we keep in our surveys out-of-scope EINs which were classified as inscope at the first phase mailing but are later coded as out-of-scope based in information from the birth form.
- improving in birth kind-of-business classification
- providing information on the NAICS distributions of births
- using NAICS codes to evaluate the quality of administrative industry codes
- collecting ancillary information such as company affiliation and locations of operations both of which would improve the BR multiestablishment business coverage.

3. ENTRIES IN THE MANUFACTURING SECTOR

3.1 Survey Characteristics in the Manufacturing Sector

The Census Bureau utilizes a variety of surveys to canvass the manufacturing sector. Most of these surveys use the individual establishment (rather than the company) as the sampling and data collection unit. In other words, if a company operates at several different locations (one establishment for each location), then each location would be considered separately for sampling and data collection purposes. This means that procedures to identify entries in the manufacturing sector must also be focused on the establishment level.

The survey frequency characteristics of the manufacturing surveys are another key factor in molding the process for identifying manufacturing entries. The Census of Manufactures is the keystone survey for data collection in the manufacturing sector. This census is conducted every five years and serves as the frame for nearly all other manufacturing surveys. The Manufacturing Energy Consumption Survey is conducted every three or four years and uses the census, updated for entries, exits, and other changes, as the sampling frame and source for measure of size information. The Annual Survey of Manufactures (ASM) collects data annually. A new sample is drawn every five years using information from the most recent census. Again, the

census is used to construct a sample frame and provide measure of size data. The ASM sample is updated annually to reflect entries and other changes identified from the sample maintenance operations. Several other manufacturing surveys (Survey of Plant Capacity, Pollution Abatement Costs and Expenditures Survey, and many of the surveys in the Current Industrial Reports program) are also conducted on an annual basis and utilize information from the most recent Census of Manufactures. These surveys also incorporate entries on an annual basis.

Since most of the surveys in the manufacturing sector are based on information from the five year census yet are conducted at more frequent intervals, they rely on sample maintenance updates to remain representative of the changing population. Also, since the majority of these surveys are conducted annually, these updates must be identified annually so that the survey panels can be updated prior to mailout of the next survey cycle. Procedures to identify entries in the manufacturing sector are designed to generate lists of manufacturing entries that need to be incorporated into the frames of the annual manufacturing surveys.

3.2 Sources and Types of Entries

Entries in the manufacturing sector may originate from several different sources. Some entries are identified from existing surveys within the Census Bureau while other entries are a result of administrative data received from the IRS, the SSA, and the BLS. Since the BR is the repository for the administrative data received from these external sources as well as information from the Census Bureau's economic programs, the BR becomes the direct source for identifying entries in the manufacturing sector. The BR is maintained and updated on an annual cycle that coincides with the processing cycle of the Bureau's Company Organization Survey. Each year, when the BR is reinitiated for the new processing year, the prior year version of the BR is saved in an archive. The existence of multiyear versions of the BR is useful for identifying entries in the manufacturing sector.

As stated in the introduction, manufacturing entries include both new manufacturing establishments (births) and existing establishments that enter the manufacturing sector through selected classification changes. Specifically, the three types of manufacturing entries that will be discussed in this paper are the following: new manufacturing establishments of multiestablishment companies; single establishment manufacturing companies that became active in the current year (births); and, single establishment manufacturing companies that were active in the prior year but had no prior year industry classification. Separate procedures are utilized for identifying each of these types of entry. Establishments that switch industry classification from nonmanufacturing (prior year) to manufacturing (current year) will not be covered in detail in this paper.

3.2.1 Births from Multiestablishment Companies

Most births that occur within multiestablishment companies are identified through the Company Organization Survey (COS). This survey, administered annually by the Census Bureau, is designed to obtain an accurate and up-to-date list of the establishments that comprise the selected multiestablishment companies. The COS is mailed to all large and medium-sized multiestablishment manufacturing companies (companies with 50 employees or more) and a sample of the smaller companies. Approximately 90 percent of all manufacturing establishments of multiestablishment companies are canvassed annually by the COS. Companies receiving the COS are instructed to write-in any establishments that are not pre-listed on the COS form, to describe the primary activity of these establishments, and to indicate whether they are 'new' or acquired establishments. Based on the reported activity descriptions, the 'new' establishments are assigned an industry classification code. All 'new' establishments that are classified as manufacturing are added to the manufacturing universe and automatically mailed a report form for the ASM. Using this method, multiestablishment manufacturing births are incorporated into the ASM in the year that they are identified. Roughly 3,000 multiestablishment births are identified annually. A list of these births is also generated so they can be added to the other manufacturing surveys.

3.2.2 Births of Single Establishment Companies

Births of single establishment manufacturing companies are identified annually using the current BR (containing IRS and SSA updates) and the prior year BR. All single establishment manufacturing companies are identified on the current BR (based on industry classification and activity indicators) and these cases are matched to a file of single establishment manufacturing companies from the prior year BR. The unmatched current year records are then reviewed to remove cases that are new to manufacturing due to changes in industry classification. The remaining establishments are referred to as births of single establishment manufacturing companies.

At this point these births are split into three files based on establishment size. Payroll data from the IRS are used to identify and separate 'very large' births and 'very small' births from the remaining births. The cases classified as 'very large' are removed from birth processing because they are generally found to be simply new EINs for existing multiestablishment companies rather than true births of single establishment companies. The files received from the IRS and the SSA do not distinguish between new EINs of multiestablishment companies and births of single establishment companies. The Census Bureau does attempt to link new EINs with existing multiestablishment companies, however, many times these linkages are not immediately identified.

At the other end of the size spectrum, 'very small' births are identified using payroll size cutoffs that vary by industry. Small single establishment companies are normally excluded from the mail portion of the manufacturing survey panels. In the ASM, for example, this nonmail portion of the panel consists of 166,000 establishments out of 366,000 total manufacturing establishments. Although this seems like a high percentage of the population on a count basis, the nonmails account for only two percent of total manufacturing shipments. Data for the nonmails are estimated based on information obtained from the IRS and the SSA. These estimates are included in the published ASM estimates. The births of single establishment companies that are below the payroll size cutoff (approximately 4,000 per year) are included in the nonmail portion of the ASM estimates. For other manufacturing surveys (Survey of Plant Capacity, Pollution Abatement Costs and Expenditures Survey, and the Current Industrial Reports surveys) these small births are not added to the panels since these surveys cover only manufacturing establishments above the nonmail cutoff.

The remaining births of single establishment companies (neither 'very large' nor 'very small') are added to the ASM and the other manufacturing surveys each year. This amounts to about 6,000 births. For the ASM, an attempt was made in the mid-1990s to compensate for births of single establishment companies that were misclassified as nonmanufacturing when they were added to the BR. Weight adjustment factors were utilized to address this situation. When single establishment companies are added to the BR, their initial industry classification codes often do not represent their true activity. These incorrect codes will generally not be corrected until the establishments are mailed during the following economic census. In the manufacturing sector, a significant influx of these reclassified recent births was noted during the census year. The weight adjustment for births of single establishment companies is the strategy that was implemented for the 1994-1998 ASM cycle to spread out the effect of these cases over the years between the censuses. This strategy was suspended for the current ASM sample panel, however, due to uncertainty over the effects of recent coding system changes on the quality and consistency of the classification codes.

3.2.3 Entries of Previously Unclassified Single Establishment Companies

When industry classification codes are initially assigned by the SSA there is a significant percentage of establishments that are unable to be classified due to insufficient or nebulous information reported for the description of business activity. As stated earlier, additional assistance in classifying these establishments is obtained through an agreement with the BLS. As updated classification codes are received from the BLS, the codes are carried to the BR. A separate operation is utilized to identify and incorporate the previously unclassified records that are coded into the manufacturing sector.

The first step in this operation is to identify the establishments that were unclassified in the prior year version of the BR. These establishments are matched to the current version of the BR to determine their current classification status. Establishments that are now classified as manufacturing are separated for inclusion in the manufacturing programs. These previously unclassified manufacturing entries total roughly 5,000 cases per year (including 4,000 below the nonmail payroll size cutoffs). A file is created of the manufacturing establishments which are larger than the nonmail cutoffs. This file is used for adding previously unclassified entries into the manufacturing surveys other than the ASM. For the ASM, a different procedure is followed.

The ASM takes the complete file of previously unclassified records that have been coded into manufacturing (including those records below the nonmail cutoffs) and inflates these records so that they represent the unclassified records on the current BR that will eventually be classified into the manufacturing sector. To develop this inflation (weight adjustment) factor, counts are maintained of the number of previously unclassified establishments that are now classified as manufacturing as well as the number now classified as nonmanufacturing. A count is also generated of the establishments on the current BR that are unclassified. Using these counts, the relative percentage of establishments classified into manufacturing is calculated and this percentage is applied to the total number of unclassified establishments on the current BR. This yields an estimate of the expected number of currently unclassified establishments that will eventually be classified into manufacturing. These cases are represented in the current year ASM estimates by applying weight adjustment factors to the previously unclassified manufacturing entries.

3.3 Quality Issues Associated with Manufacturing Entries

Since entry processing for the manufacturing sector is handled on an annual basis, timeliness is not the primary concern that it is for the retail, wholesale, and service sectors. For manufacturing, the main quality issues are the misclassification of births of single establishment companies and the large fluctuating volume of unclassified records on the BR. As noted earlier, a significant percentage of births receive an initial classification code that does not reflect the true establishment activity. If these codes do not get updated until the next economic census then the manufacturing universe is deficient during the intervening years. We addressed this issue for the 1994-1998 ASM cycle by estimating the amount of the deficiency (based on the previous census) and spreading it over the intervening years. However, with the conversion to NAICS and the subsequent changes to the classification systems, we have suspended this procedure. Once these changes have been implemented we would like to see a study of the quality/reliability of the new classification techniques.

The number of unclassified records on the BR varies from under 100,000 to over a quarter of a million records. Since industry classification for new EINs is highly dependent on information from other government agencies, any disruption to this flow of information may have a profound effect on the number of unclassified records. Although the percentage of unclassified records that are eventually coded as manufacturing is relatively small (typically 4-5%) this still adds a measure of uncertainty to the manufacturing universe each year.

4. WRAP-UP

Both the retail, wholesale and services surveys and the manufacturing surveys have put in place procedures to keep the initial samples updated with new businesses. The timing of the introduction of the births and entries into each of the samples is directly linked to the frequency of the surveys. However, it has been noted that we will investigate the impact of adding births annually rather than quarterly for the retail, wholesale and services surveys.

The quarterly process currently provides some service to the manufacturing programs through the NAICS codes assigned as a result of the first phase mailing. Some births initially coded as unclassified as well as some

initially classified as retail, wholesale, and service get assigned codes in the manufacturing sector as part of the process. The birth processing should be able to provide a greater service to the entire Economic Directorate in the area of kind-of-business coding if the proposal to expand the quarterly birth processing is put into place.

REFERENCES

BSR-2K Action Memo 2N1, "First and Second Stage Birth Sampling," February 2001, Internal Census Document.

Konschnik, C. (1988), "Coverage Error in Establishment Surveys,"paper presented at the Annual Meeting of the American Statistical Association, New Orleans, LA..

Konschnik, C., Monsour, N., Detlefsen, R. (1985), "Constructing and Maintaining Frames and Samples For Business Surveys," *Proceedings of the Section on Survey Research Methods, American Statistical Association*, pp.113-122.