25th Meeting of the Wiesbaden Group on Business Registers
- International Roundtable on Business Survey Frames

Tokyo, 8 – 11 November 2016

*Erica Marquette*
*U.S. Census Bureau*
*Session 4*

*Quality and Coverage*

# Editing and Imputing Measure of Size Variables on the U.S. Census Bureau's Business Register

**Disclaimer:** Any views expressed are those of the author(s) and not necessarily those of the U.S. Census Bureau

## I. Introduction

The U.S. Census Bureau's Business Register (BR) has many roles, but its primary purpose is to provide a high-quality source for creating survey sampling frames that serve the needs of the U.S. Census Bureau's business statistics programs. The BR is constructed and maintained through a combination of both administrative records and survey response data. Administrative records include quarterly payroll tax filings and various data from annual business income tax returns. These data are received and verified on a near-continuous basis and are applied to the BR every month. The tax data are supplied under identifiers that are assigned by the Internal Revenue Service (IRS) and one of the most important of these is the Employer Identification Number (EIN). Sources of survey response data that update the BR directly include: the annual Report of Organization (also known as the Company Organization Survey or COS), the Annual Survey of Manufactures (ASM), and the quinquennial Economic Census. Among other things, all of these data are used to assign a *measure of size* to BR business entities. Measures of size are important in business survey sample design, but their accuracy has implications in many other aspects of the survey life cycle as well. Typically, the variables used for determining a measure of size include: March 12th employment, first quarter payroll, annual payroll, and annual receipts[1]. The primary objective of this paper is to describe the methods and processes that are used to ensure that these BR variables are as accurate as possible.

## II. BR Business Entities

There are essentially two types of BR business entities: single-units and multi-units[2]. A single-unit is a business operating at a single physical location (establishment) while a multi-

[1] Receipts is generally defined as gross business sales or receipts minus returns and allowances.
[2] The scope of this paper only includes BR employer units. The editing and imputation of nonemployers is not covered.

unit enterprise[3] consists of more than one establishment under common ownership or control. For a single-unit, the administrative data and response data are directly comparable because there is a one-to-one relationship between the payroll tax-paying entity (EIN) and the statistical unit. For a multi-unit enterprise, this is not the case. A payroll tax-paying EIN of a multi-unit company is known as a *submaster* in the BR business unit model. A given multi-unit company can have more than one submaster, each of which may be linked to one or more establishments. Ideally, the aggregated payroll tax data of all submasters in an enterprise would be comparable to the corresponding survey response data when summed for all of the establishments of an enterprise.

## III. Processes

There are four processes in the BR system that are used to edit and, if necessary, impute the variables that are used for determining a measure of size: *EIN Imputation; General Data Prep; Multi-unit Imputation; and Receipts Edits*. Each of these processes has a specific purpose and uses different methods to accomplish its goals. EIN Imputation and Receipts Edits operate on administrative units, while General Data Prep and Multi-unit Imputation focus on statistical units.

### A. EIN Imputation

The scope of the EIN Imputation process covers the editing and imputation of payroll and employment for administrative data derived from payroll tax records. The process corrects administrative payroll and employment data that may be incorrectly reported or not available at all. These data are provided to the Census Bureau on an EIN basis and are stored on the BR as attributes of single-units and submasters. The EIN Imputation process is run when the payroll tax data are loaded to the BR each month and a quality assurance check is done in order to verify that it was run correctly. Most companies are required to file payroll and employment tax information with the IRS at the end of each quarter (i.e. first quarter taxes are filed in April for the period ending in March).

#### 1. Payroll

Payroll data for a given EIN are evaluated by comparing values for a particular quarter against the values of the previous and subsequent quarters. In this way, the payroll data are always edited at least one quarter behind whatever is currently being applied to the BR for the EIN. For example, when Quarter 3 payroll is received, Quarter 2 will then be compared to both Quarter 1 and Quarter 3. Therefore, Quarter 2 receives the edit and Quarter 3 (the most recently received quarter) remains unedited until Quarter 4 is introduced. When the reported payroll of the edited quarter is significantly higher or lower than that of the adjacent quarters, the process determines whether the EIN has any cyclical or seasonal patterns[4]. When a cyclical or seasonal pattern is not identified, the procedure imputes a payroll value whenever it is missing or determined

[3] The terms “enterprise” and “company” are used interchangeably.

[4] Payroll is not generally imputed or referred to an analyst for review within the first six quarters of business for an EIN, allowing time for cyclical trends to emerge.

to be extremely different compared to the surrounding quarters. Alternatively, the process may flag or refer the EIN to an analyst for review without taking any automated action. When imputation is required, the original payroll value is replaced with the average of the adjacent quarters. The newly created value is also flagged so that BR users can readily see that imputation has occurred.

2. Employment

Employment imputation may occur once the corresponding quarterly payroll value has been edited. This can happen when the ratio of payroll to employment is out of tolerance or when the employment provided is zero or missing[5]. For these cases, a three-level hierarchy is used to impute new employment values.

The first imputation method in the hierarchy is to utilize the average employment of adjacent quarters. This method is used if the adjacent quarters are reported, greater than zero, and relatively close in value. If data for the adjacent quarters is not available or is unsuitable, then the next method uses data from the prior year to calculate the current year employment value. This is done by taking the ratio of the current year payroll to the prior year payroll for a given quarter and multiplying it by the corresponding prior year employment. This result is then adjusted by an inflation factor that is created based on EINs with similar characteristics.

Otherwise, if prior year employment and payroll are not available then the last method in the hierarchy starts with the current year payroll value for the quarter and divides it by an “average wage factor” which is the average ratio of quarterly payroll to employment that is derived from EINs with similar characteristics.

As with payroll, employment values that are imputed get flagged so that BR users are aware of the source of the data.

3. Annual to Quarterly Payroll Allocation

Very small companies are only required to file payroll taxes annually and no employment information is collected. For such companies, if they reported at least $5,000 in annual payroll, the data are evenly distributed among the four quarters and the employment value is calculated using one of the imputation methods described above. For companies that reported less than $5,000 in annual payroll, the value is randomly assigned to one of the four quarters and employment is imputed for that same quarter.

[5] Additionally, if the provided employment value is greater than zero, employment is automatically set to zero if the corresponding quarterly payroll is zero.

### 4. Annual Wrap-up Process

At the conclusion of a particular BR reference year, all EIN payroll tax data are subject to an Annual Wrap-up Process. This process occurs after all the data for a given tax year, as well as the first quarter of the subsequent tax year, have been collected. This process takes all nine quarters of data (four prior year, four current year, and one subsequent year) into account. This process serves as a final check before closing out the BR reference year to ensure all four current year quarters are inline with each other. Additionally, the process corrects for anomalies that may arise due to the timing and processing of the payroll tax data. For instance, sometimes payroll quarters for certain EINs may be received out of sequence—e.g., Quarter 2 arrives before Quarter 1. If there are any inconsistencies detected during the Annual Wrap-up, data may be further edited and imputed using the methods described above.

## B. General Data Prep

General Data Prep is one module of the overall system that is used to edit and cleanse survey response data before it is applied to the BR. As mentioned earlier, the primary sources of survey data that can directly update the BR include: the COS, the ASM, and the Economic Census. In addition to its use in editing survey response data, the General Data Prep module is also called when data are corrected by analysts via the BR interactive application. The module operates on establishment-level data and edits or imputes the following items: March 12th employment[6], first quarter payroll, and annual payroll, leased March 12th employment, leased first quarter payroll, and leased annual payroll[7]. General Data Prep first determines the path a record should follow based on the type of establishment being edited and the data that has been reported. There are three main paths a record can take within the module: multi-unit establishment data, single-unit establishment data, and leased employment data. Each path is comprised of five main parts: (1) an edit for rounding errors, (2) an edit for payroll switch errors, (3) the imputation of missing data items, (4) a check for consistency between current year items, and (5) a comparison of current year items with prior year survey data or administrative data.

### 1. Rounding Edits

The rounding edit checks to see if any payroll data items have been reported in whole dollars and cents. On most Census Bureau surveys, payroll items are requested in thousands of dollars. Unfortunately, many times respondents do not report as instructed. In the edit, the current year values are first compared to either prior year data (for multi-units) or administrative data (for single-units). If the magnitudes of the data being compared are not consistent enough, the reported value is rounded

[6] For the remainder for this section, March 12th employment will simply be referred to as "employment".

[7] The "leased" versions of these data elements exist on the BR as separate and distinct variables from their "traditional" counterparts. They are intended to accommodate those companies that have an employee-leasing arrangement or that are engaged in a co-employment relationship with a Professional Employer Organization (PEO). Separation between "leased" and "traditional" variables on the BR is useful for statistical program purposes and for verifying the coverage of multi-unit companies against payroll tax records.

accordingly (for example, by a factor of 1,000). Depending on the size and degree of difference in the data, there can be multiple attempts at rounding made, using different factors, in order to get a consistent value. As is the case with all items that get changed by General Data Prep, a rounded value would be flagged appropriately.

2. Switch Edits

   Switch edits correct the data where first quarter payroll is greater than annual payroll. This can happen when the respondent inadvertently transposes or “switches” the reporting of these two items on the questionnaire or in the electronic instrument. This can be particularly problematic if a large company reports electronically by “mapping” their data incorrectly and doing a mass import operation. For single-units, in the event that a switch is detected, administrative data may be used to replace the erroneously reported data. If for some reason administrative data are not available, then the reported data are evaluated. If the values are relatively small, then the edit switches them automatically. In some scenarios, the reported annual payroll value will be kept and the first quarter value will be blanked out and imputed later. If the payroll is relatively large, then the single-unit is referred to an analyst for corrective action. For multi-units, the reported data are compared to the prior year values. If the values are comparable, then the current year items are automatically switched in the edit. In some situations, the edit determines which of the two payroll items is more reliable based on the prior year data and then imputes the other one accordingly. If the reported values are not comparable to the prior year data, then a decision is made based on their magnitudes. If the payroll data are relatively small, then they are switched automatically. Some scenarios call for the reported annual payroll value to be kept and for the first quarter value to be blanked out for imputation. As with single-units, if the payroll is relatively large, then the multi-unit establishment is referred to an analyst for review and corrective action.

3. Imputation of Missing Data

   After the edit process, all establishments where any of the data items are missing go through the imputation process. For both single-units and multi-units, other reported data items are first used to try to impute missing values. For single-units, administrative data are used as a basis for imputation. For example, the ratio of payroll to employment as taken from the payroll tax data may be used to determine an imputed employment value if this item is not reported on a survey but annual payroll does happen to be provided. If for some reason administrative data are not available for single units, then prior year values may be used. For multi-unit establishments, prior year data are the only available source for imputation since there is not a one-to-one correspondence with the EIN. Whether a single-unit or a multi-unit, if no other data are available for imputation, the establishment is referred to an analyst for review and possible correction.

4. Consistency Checks

After editing and imputation, all cases then go through some additional data quality tests. These tests are designed to flag or refer cases to BR analysts, but no further data changes via editing or imputation are done. One set of tests is done for inter-item consistency. For multi-unit establishments, a consistency check compares all the ratios among the current year data items (i.e. the payroll to employment ratio). If any of the ratios fall outside of predetermined tolerances, then the case is sent to an analyst for review and possible corrective action.

5. Comparison Checks

Another data quality check is made by comparing the current year data against another available source. For single-units, current year reported data are compared to both current year and prior year administrative data. For multi-units, current year data are compared to prior year data. If any item is not comparable based on predetermined tolerances, the record is sent to an analyst for review and possible corrective action.

C. Multi-unit Imputation

Multi-unit Imputation consists of imputing payroll and employment for: (1) *non-mail companies*-- those that were not mailed in one of the surveys that directly updates the BR, (2) *fully delinquent companies*-- those that were mailed but that did not respond in any way, and (3) *partially delinquent companies*—those that responded in part but that did not fulfill their total reporting obligation. The primary purpose of Multi-unit Imputation is to ensure that all active multi-unit establishments on the BR have updated payroll and employment measures. This process runs annually after the BR is considered complete for a given reference year, typically after the surveys have closed out collection activities and the payroll tax data are reasonably complete.

1. Complete Company Impute

Non-mail companies and fully delinquent companies will be treated as "complete company imputes" since no current year payroll and employment values are available on the BR for any of their establishments. The basic approach is to allocate the current year payroll tax data of the EIN submasters to the individual establishments. This is done by creating distribution percentages or "factors" that are developed based on either prior year data or, in the case of new companies or establishments, industry averages. For most companies, the result is that aggregated establishment data agrees with the current year EIN submaster payroll tax data and the relative values allocated to the individual establishments matches the prior year proportions.

2. Delinquent Establishment Impute

   Partially delinquent companies are handled differently from those subject to the complete company impute process described above. With these companies, some but not all of the establishments are missing current year values. For these cases, prior year data are effectively substituted for the missing current year values which are then adjusted via inflation factors. If no prior year data are available because the establishment is newly opened (i.e., a "birth"), then industry-based imputation factors are used to determine the current value. If for some reason, no prior year data are available and the establishment is not actually a birth, then it will be flagged as "unable to impute" and no values will be assigned.

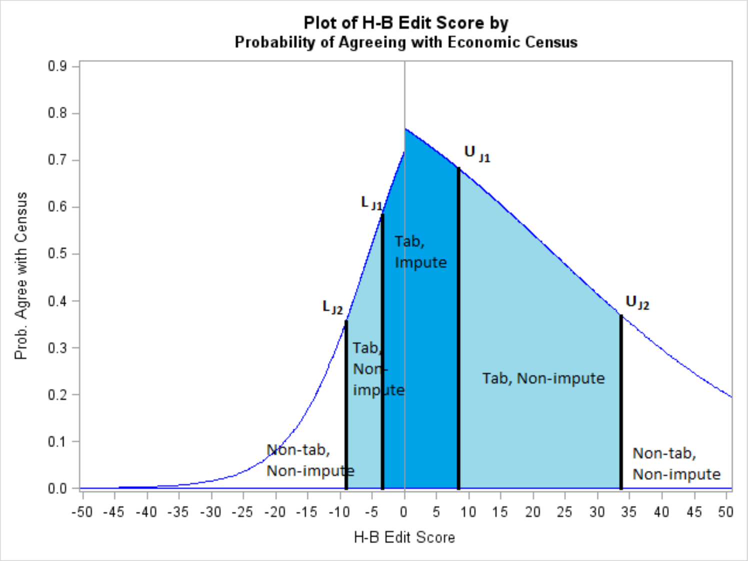
D. Receipts Edits

Currently, no automated editing or imputing of administrative records receipts data are done on the BR. However, some statistical programs are evaluating the possibility of publishing receipts values at a sub-national level. As such, the quality of these data and the feasibility of developing an edit and imputation process for receipts are currently being researched. As an initial step, starting in the Fall of 2016, extreme receipts values for single-units only will be flagged appropriately using the Hidiroglou-Berthelot (H-B)[8] edit method. The flag will inform users about the quality of the receipts values on the BR. In this first implementation of the edit, no values will be changed or imputed. The Receipts Edits process determines if an administrative receipts value for a single-unit is acceptable by calculating a normalized ratio of receipts to payroll value and then comparing it to score cutoffs defined using the H-B Method. The edit will occur whenever administrative records containing income or payroll tax data are loaded to the BR. Only single-units that have an annual payroll or administrative receipts value of at least $100,000 will be subject to the edit.

Two sets of H-B score cutoffs are used in the process. The first set of cutoffs define an inner range to identify receipts that would be acceptable to use as imputes in a survey program. The second set of cutoffs defines an outer range identifying what is acceptable for tabulating administrative receipts. Figure 1 shows how the score cutoffs are applied. If the score is outside the outer limits ($L_{j2}$ and $U_{j2}$), then receipts are flagged as "Non-tab, Non-impute", implying that the receipts value should not be tabulated in an estimate or used as an impute. If the score is inside the outer limits, but outside the inner limits ($L_{j1}$ and $U_{j1}$), then receipts are flagged as "Tab, Non-impute," implying that the receipts value can be tabulated in an estimate, but are not recommended for imputing receipts. If the score is inside the inner limits, then receipts are flagged as "Tab, Impute," implying that the receipts value can be tabulated in an estimate and used as an imputation for sales or receipts.

[8] Hidiroglou, Michael A. and J.M. Berthelot, 1986, "Statistical Editing and Imputation for Periodic Business Surveys," Survey Methodology, Vol. 12 (No. 1), pp 73-83
(There is no known Internet source of this paper.)

**Figure 1: Receipts Edits H-B Score Cutoffs**



## IV. Conclusion

The processes that are used to edit and impute the measure of size variables are critical in ensuring that the BR database is of high quality and meets the needs of the Census Bureau's business survey programs.  Analysts and statisticians are continually monitoring these processes for accuracy and constantly seeking ways to enhance their capabilities.  The immediate future will focus on the practicality of taking the Receipts Edit beyond just flagging extreme values into possibly imputing receipts values. It may also be feasible to expand the scope of this edit to encompass multi-unit companies.  Further, research into "Big Data" and the practicality of leveraging other non-survey sources of information may lead to further enhancements to these processes.