# Using Hot Deck Donor Imputation Methodology in the Service Annual Survey

Carol S. King, Rebecca D. Bogle, United States Census Bureau
Carol S. King, U.S. Census Bureau, SSSD, Rm. 2651-3, Washington, DC 20233-6500

Key words: imputation, reporting patterns

## Introduction

The Census Bureau's Service Annual Survey (SAS) is a large, complex survey conducted to provide national estimates of annual revenue and expenses for selected service industries. The SAS is divided into several components including finance, computer, health, transportation, information, and general services. This paper addresses the subject of accounting for missing data in the finance portion of the SAS (SAS-F).

The target population for the SAS-F is all business establishments in the United States that provide financial services. The population is stratified by kind of business (KB) and within KB by annual revenue with the largest units selected with a probability of one (certainty). Sampling units are either companies or Employer Identification Numbers (EINs) used for filing payroll with the Internal Revenue Service (IRS), both of which are groups of one or more employer establishments under common ownership. The kinds of business for SAS-F are shown below. Each kind of business is preceded by its five-digit North American Industry Classification System (NAICS) code:

52311 Investment banking and securities
52312 Securities brokerage
52313 Commodity contracts dealing
52314 Commodity contracts brokerage
52392 Portfolio management
52393 Investment advice

In addition to total revenue, the survey collects revenue for twelve components of the total, as well as exported services revenue and electronic commerce revenue.

[1]T*his 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.*

A description of each revenue item preceded by its item number is shown below:

1. Commissions from the sale of securities and commodities

Net gains (losses) in trading accounts in securities and commodities:

2. Net interest income from trading accounts insecurities
3. Net gains (losses) from trading accounts in securities
4. Net gains (losses) in trading accounts in Commodities

Net gains (losses) from underwriting and selling groups of securities:

5. Mortgage Backed Securities (MBS, Collateralized Mortgage Obligations (CMO), and Real Estate Mortgage Investment Conduits (REMIC) transactions
6. All other securities transactions
7. Net gains (losses) on investment accounts
8. Dividend income
9. Margin interest and other interest income, including Repurchased Agreement (REPO)
10. Other investment income (net)
11. Asset/Portfolio management fees
12. Other revenue

The principal statistics estimated by the survey are the annual totals for each of the above items. The estimates are Horvitz-Thompson linear estimators computed as the sum of the weighted data (reported and imputed) for all selected sampling units. The weight for a given sampling unit is the reciprocal of its probability of selection into the sample.

This paper focuses on the estimation of items 1 through 12. The distribution of these items is semi-continuous, a mixture of zeroes and continuously varying dollar values. The sum of these items, total annual revenue of the unit, has a lognormal distribution.

## Reporting Patterns in the 2000 SAS-F

In the SAS-F survey conducted to collect data for the year 2000, there were 1535 total units, comprised of 784 respondents, 42 partial respondents, and 709 total non-respondents. Unit respondents had all revenue items reported. Unit non-respondents had no revenue reported. Partial respondents had some revenue items reported and

some not reported. The breakout by four-digit NAICS code, size of firm, and type of response is provided below:

**Table 1: Reporting Patterns**

| Size | Response | NAICS | | TOTAL |
|---|---|---|---|---|
| | | *5231* | | *5239* |
| Certainty (selected with probability=1) | *Respondents* | 106 | 163 | 57 |
| | *Partial respondents* | 6 | 12 | 6 |
| | *Total non-respondents* | 171 | 267 | 96 |
| Non-Certainty (selected with probability <1) | *Respondents* | 328 | 621 | 293 |
| | *Partial respondents* | 21 | 30 | 9 |
| | *Total non-respondents* | 233 | 442 | 209 |
| **T O T A L** | | **865** | **670** | **1535** |

Some of the total nonresponse for certainty units results from the fact that the survey form is not sent to units having relatively small unit size within the certainty strata. It was decided that in order to reduce respondent burden, large companies with small receipts for particular NAICS codes would have the annual receipts for these NAICS codes imputed using administrative data.

Other reasons why an item can be missing in the SAS-F include nonresponse to the form or the item fails an edit. A particular item can be left blank by a respondent because the business does not have the item or has it but does not report it. Note that edits are done to zero fill items left blank if the reported detail data sums to the total revenue.

Of the 621 reporting EIN units, the most commonly reported item was item 1, Commissions from the Sale of Securities and Commodities. The least reported item was item 5, MBS, CMO, and REMIC Transactions. On average, EIN reporters gave estimates for 2.45 items out of 12.

Of the 163 reporting company units, the most commonly reported items were item 1, Commissions from the Sale of Securities and Commodities and item 12, Other Revenue. The least reported item was item 2, Net gains (losses) in Trading Accounts in Commodities. On average, alpha reporters gave reported values for 4.81 items out of 12, about twice as many as for EIN units.

Current Imputation Methodology

Missing current year (CY) values for the component revenue items are currently imputed using one of two methods. The first method is to multiply the unit's prior year (PY) item by the CY-to-PY revenue ratio of the particular unit. The second method is to multiply a ratio of identicals, which is the weighted value of the item to be imputed to the weighted revenue for all units having both the item and revenue reported, to the revenue of the unit.

$$(1) \quad PY\ item * \frac{CY\ revenue}{PY\ revenue},$$

if the PY item is present. (PY data is not available for new businesses added to the sample in the survey year and in the year a new sample is introduced)

$$(2) \quad \left(\frac{\Sigma CY\ item}{\Sigma CY\ revenue}\right)_I * revenue\ of\ unit,$$

otherwise

where *I(identicals)* denotes only units with both the numerator and denominator reported are included in the ratio.

The result of this second methodology is that if there is at least one unit in a particular NAICS code that has reported positive data for an item, data will be imputed for that item for any nonresponse unit that does not have PY data. That estimate, though it may be small if the item is not reported by many units, will be carried forward into the next year using Equation (1) unless a response is obtained. The rational behind this method is that for any individual reporting unit the imputed value may not represent what the unit would have reported but the total estimate across all units would be accurate.

The question arose as to whether we could preserve the reporting patterns of respondents and produce an estimate that is at least as accurate as the current method. We decided that donor imputation, in particular, multiple donor imputation (MDI) was the method that might achieve these two goals. In the MDI method, a missing value is matched with M donor values according to a distance metric. For our purposes, the ratio of each donor value to the donor's total revenue is multiplied by the recipient's total revenue to create a "donated" value.

Determination of the Distance Metric

Using the 784 good respondents, data values for the component items were blanked out for 347 of these cases. Total revenue was not blanked out since this variable would continue to be imputed using the current imputation methodology. Several different metrics were used to determine the donor record, and results were compared with actual values.

To determine which cases to blank out, all 1535 of the cases were sorted by total revenue then divided into quartiles. Quartiles were determined by number of units. The percent of cases having unit non-response in each quartile was determined. In each of these quartiles, the percent of non-response that occurred in the actual data was used to determine the number of units that should be blanked out in quartiles created with the 784 respondents. The specific units to be blanked out were selected randomly within each quartile. This process was done separately for company units and EIN units. The rate of unit non-response in each quartile is shown in the following table:

**Table 2: Rate of Nonresponse**

| | EIN Unit | Company Unit |
|---|---|---|
| 1st Quartile | 52% | 84% |
| 2nd Quartile | 46% | 78% |
| 3rd Quartile | 35% | 52% |
| 4th Quartile | 29% | 28% |

Note that the nonresponse rates are higher for the company units because we do not mail parts of companies that have small revenue.

Once the cases were divided into donors and recipients, the cases were grouped by four-digit NAICS code (5231 or 5239) and type of unit (company or EIN). The grouping was done by four-digit NAICS code because the grouping by six-digit NAICS code did not provide a large enough donor pool. Recipients were matched with a donor in the same group using one of four different matching methods. In each of the following methods, the donor with the smallest distance was selected:

*(3)* $D1 = |P_d - P_r|$ (Payroll Distance)

*(4)* $D2 = |R_d - R_r|$ (*Revenue Distance)*

*(5)* $D3 = \sqrt{D1^2 + D2^2}$ *(Euclidean Distance)*

*(6)* $D4 = MAX(D1, D2)$ (*Minimax Distance)*

*Where* $P_d$ = *payroll of donor*
$P_r$ = *payroll of recipient*
$R_d$ = *receipts of donor*
$R_r$ = *receipts of recipient*

We determined the "best" distance method by blanking data five times with a different set of units blanked each time. The differences between imputed and reported estimates were averaged and compared. The Euclidean distance method provided the overall smallest difference across all the items.

The 2000 SAS-F and MDI

The MDI method was tested using the Euclidean distance matching by performing a study using the 784 respondents. Recipients were created by blanking out the data in the component revenue items for 347 of the respondents. To determine which cases to blank out, the universe of cases was sorted by total revenue in increasing order then divided into ten groups with approximately the same number of units in each group. We changed from quartiles to deciles to determine the nonresponse rate because we thought we would get a better pattern of nonresponse with more groups. The percent of cases having unit nonresponse in each group was determined. In each of these groups, the percent of nonresponse that occurred in the actual data was used to determine the number of units that should be blanked out in that group. The units to be blanked out were selected randomly within each group. This process was done separately for certainty and noncertainty units.

Once the cases were divided into donors and recipients, imputation cells were defined by four-digit NAICS code (5231 or 5239) and size of unit (certainty or non-certainty). Recipients were matched with the five closest donors in the same group using the Euclidean distance matching function (See Equation (5).) For each of the 5 selected donors the item revenue/revenue of each donor was multiplied by the revenue of the recipient. For each of the five imputations, data was summed across all units (reported data for the donors and imputed data for the recipients) for each item. The five resulting estimates were summed and averaged to produce the final estimates for each item.

Estimates were calculated after multiple donor imputation was completed. Model-based imputation was also run on the same group of recipients to compare the current imputation methodology with multiple donor imputation. In addition, the totals for each item were calculated using the reported data so that the results of both methods could be compared with actual values. Comparisons showed that multiple donor imputation based on one sample without regards to sampling error produced estimates that were generally closer (i.e., most of percent differences were smaller and/or the range of percent differences were smaller when compared to the reported data than the current imputation methodology). Table 3 shows the percent difference of each item estimate from the actual total for the two methods.

**Table 3: Percent Difference by Item**

| Item | Current Imputation | Multiple Donor Imputation |
|---|---|---|
| 1 | 0.0045 | 0.0673 |
| 2 | 0.1250 | 0.1028 |
| 3 | 0.0598 | -0.0108 |
| 4 | -0.3066 | -0.1213 |
| 5 | 0.0609 | 0.0238 |
| 6 | -0.0283 | -0.1190 |
| 7 | 0.0607 | -0.1176 |
| 8 | 0.3169 | 0.0917 |
| 9 | -0.0415 | -0.0304 |
| 10 | 0.2428 | 0.1272 |
| 11 | -0.0311 | 0.0098 |
| 12 | -0.0648 | -0.1035 |
| **Absolute Average** | **0.1119** | **0.0771** |
| **Min** | **-0.3066** | **-0.1213** |
| **Max** | **0.3169** | **0.1272** |

Advantages of MDI

One of the main advantages of multiple donor imputation is that it appears to preserve the distribution of responses. Imputed and reported values were plotted against the total revenue for each item. Even though multiple imputation does not provide for such estimates, we took the liberty to compute item data within each recipient by averaging the imputed item data within each recipient across the five imputations. Logarithms of the plotted values were used to facilitate more symmetrically dispersed values. These plots appear to demonstrate that multiple donor imputation, unlike model-based imputation, maintains the distribution of the actual values.

Another advantage of multiple donor imputation is that it lends itself to variance estimates that adjust for the fact that values have been imputed. Currently, a random group method is used to estimate the variance of each estimate in SAS-F. This technique does not take into account the additional variance caused by imputation which could understate the variance. At the same time, it tends to ignore the finite population correction which could overstate the variance. The net result of these weaknesses is not known. Using multiple donor imputation, it is possible that improved variance estimates could be produced using methods first proposed by Rubin (1987) for multiple imputation.

Variance Computations

Before discussing an alternative technique of calculating variance estimates, it is important to note that inference using Rubin's methods has an underlying assumption that data is missing at random. One of the reasons why this assumption may not be entirely satisfied for the certainty units is the practice of not mailing smaller sized units a survey questionnaire. Further study should be conducted to determine how departure from the missing at random assumption affects inference in this case.

Let $Y_i$ be the $i^{th}$ item estimate with variance $v_i$, i=1,..,M, where M imputations have been provided for each missing value. The final item estimate Y was given by $Y=\Sigma Y_i / M$. The variance of this final estimate can be approximated with the following equation (Little, 1988):

(7) $V = s_w^2 + ( 1 + 1/M ) s_b^2$ where $s_w^2 = \Sigma v_i / M$

(the average variance within imputed data sets. These variances were computed using the random group method.)

and $s_b^2 = \Sigma (Y_i - Y)^2 / (M - 1)$
(the between imputation variance)

The variance for each item estimate produced by multiple donor imputation was calculated using this method with M=5 imputations. The standard error was divided by the item estimate to determine the coefficient of variation (CV). Table 5 at the bottom of the next page displays the CVs from multiple donor imputation calculated using Rubin's method. We took the 347 cases and imputed them using the current method and the donor method and computed CVs accordingly. It also shows the CVs from the current imputation methodology as well as the CVs from the reported data, both produced using the random group technique:

The efficiency of the final estimate based on M imputations is given by (Rubin, 1987):

(8) $Eff = (1 + R/M)^{-1}$
where $R = [(r + 2) / (df + 3)] / (r + 1)$
(the fraction of missing information for the quantity being estimated)

and $r = ( 1 + 1/M ) s_b^2 / s_w^2$
(the relative increase in variance due to nonresponse)

and $df = (M - 1)[1 + M s_w^2 / (M+1) s_b^2]^2$

The degrees of freedom, the fraction of missing information (R), and the efficiency for each item estimate is shown in the following table. The fraction

quantifies how precise the item estimate is relative to there being no missing data.

**Table 4: Efficiency**

| Item | df | R | Efficiency |
|---|---|---|---|
| 1 | 4.1802 | 0.1450 | 0.9718 |
| 2 | 4.0009 | 0.1430 | 0.9722 |
| 3 | 4.4549 | 0.1467 | 0.9715 |
| 4 | 4.5885 | 0.1417 | 0.9724 |
| 5 | 4.1425 | 0.1447 | 0.9719 |
| 6 | 4.0594 | 0.1437 | 0.9721 |
| 7 | 8.0219 | 0.1255 | 0.9755 |
| 8 | 4.2737 | 0.1458 | 0.9717 |
| 9 | 4.0722 | 0.1439 | 0.9720 |
| 10 | 7.0833 | 0.1327 | 0.9741 |
| 11 | 4.3760 | 0.1464 | 0.9716 |
| 12 | 4.8497 | 0.1468 | 0.9715 |

Finally, 95% confidence intervals may be determined for the estimates using the approximation $Y \pm t_{df} \sqrt{V}$, where Y is the item estimate, V is the variance estimate as given in equation 7 and $t_{df}$ denotes a Student's t-distribution with degrees of freedom as given in the above table (Schafer and Olsen, 1998).

Conclusions

Multiple donor imputation should be considered as a possible alternative to model based imputation in the SAS. This study suggests that this method produces more accurate estimates than model-based imputation for the component revenue items in SAS-F. The method has several advantages over model-based imputation including the tendency to preserve the distribution of responses and the ability to conveniently calculate variance estimates that adjust for imputed values. Further research may show that this method performs well on other components of the survey besides SAS-F. Note that these conclusions were based on one sample. A simulation study should be performed. Also this methodology needs to be applied to the complete sample of survey units to get a complete picture of how the hot deck donor imputation compares to the current methodology.

## REFERENCES

Kinyon, D., D. Glassbrenner, J. Black, R. Detlefsen (2000). "Designing Business Samples Used for Surveys Conducted by the United States Bureau of the Census," paper presented at the second International Conference on Establishment Surveys, Buffalo, NY.

Little, R. J. A. (1988). "Missing Data Adjustments in Large Surveys," *Journal of Business and Economic Statistics*, Vol. 6, pp. 287-296.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*, New York: J. Wiley & Sons.

Schafer, J. L., M. K. Olsen (1998). "Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst's Perspective," The Pennsylvania State University.

U.S. Office of Management and Budget (2002), *North American Industry Classification System: United States, 2002*, Lanham, MD: Bernan Press.

Wolter, K. (1985). *Introduction to Variance Estimation,* New York: Springer-Verlag.

**Table 5: Estimated Coefficients of Variation**

| Item | Current Imputation (using random group method) | Multiple donor imputation (using random group method) | | | | | Multiple Donor Imputation (Little estimate) | Reported Data (using random group method) |
|---|---|---|---|---|---|---|---|---|
| | | M=1 | M=2 | M=3 | M=4 | M=5 | | |
| 1 | 0.0223 | 0.0228 | 0.0188 | 0.0387 | 0.0223 | 0.0234 | 0.1235 | 0.0234 |
| 2 | 0.0135 | 0.0130 | 0.0135 | 0.0081 | 0.0107 | 0.0164 | 0.2928 | 0.0134 |
| 3 | 0.0396 | 0.0417 | 0.0431 | 0.0445 | 0.0454 | 0.0633 | 0.1531 | 0.0531 |
| 4 | 0.0178 | 0.0966 | 0.2331 | 0.1516 | 0.1793 | 0.2605 | 0.3703 | 0.1381 |
| 5 | 0.0030 | 0.0019 | 0.0020 | 0.0020 | 0.0024 | 0.0092 | 0.0242 | 0.0020 |
| 6 | 0.0185 | 0.0198 | 0.0652 | 0.0240 | 0.0215 | 0.0255 | 0.3088 | 0.0379 |
| 7 | 0.1999 | 0.1680 | 0.2834 | 0.3053 | 0.2223 | 0.3017 | 0.3926 | 0.2135 |
| 8 | 0.0311 | 0.0390 | 0.0312 | 0.0429 | 0.0327 | 0.0404 | 0.1497 | 0.0405 |
| 9 | 0.0328 | 0.0456 | 0.0394 | 0.0280 | 0.0396 | 0.0278 | 0.2621 | 0.0337 |
| 10 | 0.0476 | 0.0644 | 0.0507 | 0.0671 | 0.0490 | 0.0812 | 0.1111 | 0.0547 |
| 11 | 0.0206 | 0.0251 | 0.0209 | 0.0260 | 0.0236 | 0.0266 | 0.0870 | 0.0284 |
| 12 | 0.0250 | 0.0356 | 0.0501 | 0.0437 | 0.0355 | 0.0332 | 0.1052 | 0.0251 |