COMPARISON OF METHODS FOR IMPUTING MISSING RESPONSES IN AN ESTABLISHMENT SURVEY

Jill M. Montaquila and Chester H. Ponikowski, Bureau of Labor Statistics
Jill M. Montaquila, BLS, Postal Square Building Suite 3160, 2 Massachusetts Ave. NE, Washington, DC 20212

KEY WORDS: Item Nonresponse, Imputation, Hot Deck, Nearest Neighbor

I. INTRODUCTION

The problem of missing responses to data items is one of the most common problems to surveys. Missing responses occur because some respondents refuse or are unable to provide data for a particular item or items, the interviewers sometimes fail to ask for or record the data items, data entry clerks may omit keying the data item, or an editing process deletes an inconsistent data. A common procedure for dealing with this problem is to use some form of imputation method to assign values for the missing responses.

Various methods have been proposed for imputing missing item responses (Kalton and Kasprzyk, 1982; Rubin, 1978; Sedransk, 1985). Kalton and Kasprzyk (1982) describe a variety of imputation methods being used and their properties. They point out that imputation has three desirable features: "First ... it aims to reduce biases in survey estimates arising from missing data .... Second, by assigning values at the microlevel and thus allowing analyses to be conducted as if the data set were complete, imputation makes analyses easier to conduct and results easier to present. Complex algorithms to estimate population parameters in the presence of missing data ... are not required. Third, the results obtained from different analyses are bound to be consistent, a feature which not need to apply with an incomplete data set."

This paper describes the establishment survey used to compare the performance of imputation methods (Section II), describes imputation methods studied (Section III), presents empirical analysis and results (Section IV), and proposes issues for further research (Section VI).

II. DESCRIPTION OF EMPLOYEE BENEFITS SURVEY

The Employee Benefits Survey (EBS) is an establishment survey conducted annually by the Bureau of Labor Statistics (BLS). The goal of the survey is to produce estimates of the incidence and characteristics of benefits provided to employees by their employers. All State and local governments and private sector industries, except for farms and private households, are covered in the survey during a two-year cycle. All employees are covered except the self-employed. Data for small private establishments (under 100 workers) and State and local governments are collected in even-numbered years, and data for medium and large private establishments (100 workers or more) are collected in odd-numbered years.

State Unemployment Insurance (UI) files of establishments serve as the sampling frame for the EBS. The EBS sample is selected using a 2-stage stratified design with probability proportional to employment sampling at each stage. The first stage of sample selection is a probability sample of establishments and the second stage of sample selection is a probability sample of occupations within the sampled establishments.

The sample of establishments is drawn by first stratifying the sampling frame by industry group and establishment employment. The number of sample establishments allocated to each stratum is approximately proportional to the stratum employment. Each sampled establishment is selected within a stratum with a probability proportional to its employment.

After the sample of establishments is drawn, occupations are selected in each establishment. The probability of an occupation being selected is proportionate to its employment within the establishment. For a more detailed description of the EBS sample design, refer to the *BLS Handbook of Methods* (Bulletin 2414, September 1992).

The EBS collects information on provisions of benefits as well as incidence of benefits. Occasionally, responding establishments refuse to provide or are unable to provide data corresponding to the provisions and/or the number of employees within a given occupation(s) that participate in an offered benefit plan. Thus, item nonresponse results. Ignoring the item nonresponse and using only complete data records could result in substantial bias in estimates and incorrect variance

estimates. Frequently the distribution of the characteristic of interest is different for units that provide data versus units that do not provide data. In EBS, an adjustment for participation item nonresponse is made by imputing participation ratios (i.e., the ratio of number of occupational group participants in a given plan to occupational group employment) into the record with "unusable" participants, and then multiplying the imputed participation ratio by the occupational group employment in order to obtain an imputed participant figure.

In our study, we used the data from the 1991 EBS. The 1991 EBS had a sample of 2,144 establishments which consisted of 11,274 sampled occupational observations. The dataset included auxiliary data from the frame as well as reported data obtained during collection. We did not wish to artificially induce item nonresponse among the complete respondents. To perform the evaluation of the methods being considered, we left the dataset in its original form, imputing for the originally missing items.

III. IMPUTATION METHODS

The imputation methods studied are nearest neighbor within-cell hot-deck, random within-cell hot-deck, and cell mean imputation. These were chosen for our study because they appear to be most commonly used in establishment surveys.

A. Random Within-Cell Hot-Deck. Imputation classes ("cells") are formed, based on auxiliary data that is known for all units. Within each cell, a unit that is missing the characteristic of interest (i.e., an "unusable") takes the value of the characteristic of a "usable" unit ("donor") that is selected at random within the same cell. In our application,

$y_{ij}^* = y_{ik}$, $j \neq k$,
where y_{ij}^* = imputed participant ratio for plan j in cell i,
y_{ik} = actual participant ratio for plan k in cell i,
and plan k is chosen at random from among all usable plans in cell i.

An advantage of the random within-cell hot-deck method is that, unlike the cell mean method, it retains the respondent distribution of the characteristic (Kalton and Kasprzyk, 1982); within cell k,

$E(s^2_{k_{RHD}}) \cong s^2_{k_r}$,

where $s^2_{k_{RHD}}$ = variance of characteristic within cell k when random within-cell hot-deck method is used,
and $s^2_{k_r}$ = respondent variance of characteristic within cell k.

B. Nearest Neighbor Within-Cell Hot-Deck. The "unusable" takes the value of the characteristic of the "usable" unit within the same cell that is "nearest" to the unusable, where "nearness" is defined by a pre-specified distance function. Currently, the EBS uses the nearest neighbor within-cell hot-deck method to impute missing participation ratios. In our application,

$y_{ij}^* = y_{ik}$, $j \neq k$,
where y_{ij}^* = imputed participant ratio for plan j in cell i,
y_{ik} = actual participant ratio for plan k in cell i,
and plan k is chosen from among all usable plans in cell i such that $|e_{ij} - e_{ik}|$ is minimized, where
e_{ij} = establishment employment for establishment corresponding to plan j in cell i.

As with the random within-cell hot-deck, the nearest neighbor within-cell hot-deck preserves the respondent distribution of the characteristic of interest conditional on the cell-defining auxiliary variables. However, this method allows for the use of additional auxiliary information that may be highly correlated with the characteristic of interest in choosing the donor.

C. Cell Mean. The "unusable" takes the mean of the characteristic among all "usables" within the same cell. In our application,

$$y_{ij}^* = \frac{\sum_{k \in R_i} w_{ik} y_{ik}}{n_{R_i}}, \qquad j \neq k,$$

where y_{ij}^* = imputed participant ratio for plan j in cell i,
w_{ik} = weight applied to plan k in cell i,
y_{ik} = actual participant ratio for plan k in cell i,
R_i = set of all usable plans in cell i,
and n_{R_i} = number of usable plans in cell i.

Imputing the cell mean results in a spike in the conditional distribution of the characteristic, conditional on the cell-defining auxiliary variables, at the cell mean. That is, the distribution of the characteristic is distorted in that the variance of the characteristic is attenuated (Kalton and Kasprzyk, 1986).

The cell mean method is deterministic, while the random hot-deck and nearest neighbor hot-deck methods are stochastic. In general, stochastic imputation methods preserve the variance and covariance structures in the data better than do deterministic methods.

Each of the imputation methods considered is based on the formation of disjoint imputation cells, and the subsequent collapsing of cells when necessary. We assume that the missing responses are missing at random (MAR) within cells. That is, we assume that the conditional distribution of the characteristic of interest for unobserved units (which may or may not have been included in the sample) given the cell-defining auxiliary variables and the observed values is independent of the sampling and response mechanisms.

To maintain comparability between methods, the cells and collapse patterns used in this study are the same for each of the three methods under consideration. Imputation cells are constructed based on characteristics that include industry (SIC), major occupational group, region, and union status. Analysis of variance results showed that, among plans with usable participant data, each of these variables have highly significant main effects on participation ratio (p-values were less than or equal to 0.001 for each test of significance of main effects). Thus, the predictive distribution of participant ratio given these observed variables should have small variance (Rubin, 1978).

Using donors multiple times may result in a nonnegligible loss in precision of the estimators due to an increase in imputation variance (Kalton and Kasprzyk, 1986). Thus, for the random within-cell hot-deck and nearest neighbor within-cell hot-deck, we required that, whenever possible, usables be used at most once in imputing missing participation for unusables. If a cell had one or more unusables but no available usables, the cell was collapsed with other similar cells according to the predetermined collapsing pattern until a usable donor was found.

D. Other Methods. There are several other commonly used imputation methods that we chose not to consider at this time. Regression methods would involve regressing the participation ratios of usable plans on other known auxiliary variables and using the estimated regression equation to "predict" values for unusables. Another variation of the regression method involves adding a residual (selected either randomly from among the "observed" residuals or otherwise) to each predicted value. West *et al* (1989) considered several regression models in an evaluation of imputation methods for employment data using data from the Current Employment Statistics (CES) Survey of establishments. For their purposes, they found that a regression method appeared superior to other methods considered. In our case, most of the auxiliary variables are categorical; several have many categories. Thus, in our case, regression imputation with this set of auxiliary variables is impractical, if not impossible.

Multiple imputation methods (Rubin, 1978) involve independently imputing $J>1$ values for each missing value. That is, for each missing participation ratio, J participation ratios would be drawn with replacement from the predictive distribution of the participation ratios, given the observed values of the participation ratios. This method enables the analyst to obtain valid variance estimates by incorporating into the variance estimate an estimate of the imputation variance.

IV. EMPIRICAL ANALYSIS AND RESULTS

We focused on estimates from three different benefit areas: Benefit areas 01 (Health Care), 04 (Long Term Disability), and 05 (Defined Benefit and Defined Contribution Plans). Our reason for choosing these benefit areas in this study was that these benefit areas have high participation item nonresponse rates relative to other benefit areas, as indicated in Table 1. The figures in Table 1 represent, for each benefit area, the proportion of plans having missing participation data.

Each estimate is a ratio estimate of the following form:

$$\hat{R} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m_i} w_{ij}\, y_{ij}}{\sum_{i=1}^{n} \sum_{j=1}^{m_i} w_{ij}\, x_{ij}},$$

where w_{ij} = weight for occupation j in establishment i,
y_{ij} = number of participants in occupation j in establishment i,
x_{ij} = occupational group employment for occupation j in establishment i,
n = number of establishments in sample,
and m_i = number of occupations selected in establishment i.

Table 1. Participation Item Nonresponse by Benefit Area

Benefit Area	Participation Item Nonresponse Rate
01 (Health Care)	0.2219
02 (Life Insurance)	0.0652
04 (Long-Term Disability Insurance)	0.0820
05 (Def. Benefit & Def. Contribution Plans)	0.1111
12 (Personal Leave)	0.0136
20 (Short-Term Disability)	0.0259

NOTE: All other benefit areas had participation item nonresponse rate of less than one percent.

The estimates and variance estimates were calculated using software for survey data analysis (SUDAAN Release 6.0) for multistage sample designs.

Regardless of the imputation method used, the usual variance estimator will underestimate the variance of $\hat{R}$, since it does not account for additional variability due to imputation, i.e., "imputation variance." Underestimation of true variance can be a very serious problem when the proportion of missing values for a particular characteristic of interest is high (Rao and Shao, 1992). Multiple imputation methods have been proposed in order to account for imputation variance. Rao and Shao (1992) have recently proposed a jackknife variance estimation method that accounts for between-imputation variability.

Table 2 presents the results for each of the three imputation methods we considered. The absolute differences among the estimates based on the three imputation methods are in the range of 0.0003817 to 0.0076791 for health care benefit estimates, 0.0007497 to 0.0017416 for long-term disability estimates, and 0.0000687 to 0.0083537 for defined benefit and defined contribution estimates. These differences across imputation methods are not significant at α=0.05 level. The similarity in these estimates is due in part to the cell definitions. Several auxiliary variables, many having several levels, were used in constructing the cells. This was done in an attempt to ensure homogeneity of participant (item) response propensity within cells, and thus reduce participant (item) nonresponse bias. Given our current imputation cell definitions, the analysis of these data is robust to the imputation method used.

In general, one would expect the cell mean method to yield smaller variance estimates than the random within-cell hot-deck and nearest neighbor within-cell hot-deck, since the cell mean method distorts the distribution of the characteristic of interest by inducing a "spike" at the cell mean. However, for several of the estimates in Table 2, the cell mean method did not result in the smallest standard errors. There may be several reasons for this "counter-intuitive" result:

1. **Differences in the extent of collapsing of cells.** In general, the cell mean method required less collapsing than the other two methods.
2. **Variability in the variance estimates.**
3. **Post-Imputation Edits.** A series of edit constraints was imposed upon the data. For example, in health care, for a given occupation within a given establishment, the total number of participants in medical plans could not exceed occupational group employment. As a result, imputed participant ratios were sometimes modified in order to satisfy the edit constraints.
4. **Rounding.** Although participation ratios are imputed, integer-valued participant counts are used in the calculation of the estimates.

Table 3 illustrates the impact of rounding and post-imputation edits for a given cell. All plans that fall within the given cell are represented in Table 3. For each plan, the pre-edit participation ratio and post-edit participation ratio are given. The effect of rounding is illustrated in the pre-edit participation ratios. For the cell mean method, were it not for rounding, these values would all be identical for plans with imputed participation ratios. The effect of the post-imputation edits is demonstrated in the post-edit participation ratios. For the plans in this cell, the non-zero imputed participation ratios are generally scaled back due to the edit constraints. Prior to the edits, the variance in participation ratios within the cell was smallest under the cell mean method. After the edits, the within-cell variance for the cell mean method was higher than the within-cell variance for the random within-cell hot-deck method.

In addition to presenting estimates and standard errors based on usable and imputed data, Table 2 gives the estimates and standard errors when only the original usable data were used. No adjustment was made for observations having missing participants.

Comparing the completed data estimates to the incomplete data estimates in Table 2 yields someinteresting results. First, the estimates based on original usable data differ from the estimates based on all usable and imputed data. The Medical Care Overall, Medical Care Employee Coverage Partly Employer Financed, Dental Care Employee

Table 2. Estimates and Standard Errors for Completed Data Set Under Each Imputation Method And for Incomplete Data Set (Usable Data Only)

	RANDOM HOT DECK IMPUTATION	NEAREST NEIGHBOR HOT DECK IMPUTATION	CELL MEAN IMPUTATION	USABLE DATA ONLY
HEALTH CARE (BENEFIT AREA 01)				
Medical Care Overall	0.8325642 (0.0086870)	0.8344175 (0.0089335)	0.8267384 (0.0096782)	0.7264706 (0.0143597)
Employee Coverage				
Wholly Employer Financed	0.4090410 (0.0138016)	0.4116196 (0.0140500)	0.4062935 (0.0139738)	0.3793521 (0.0140962)
Partly Employer Financed	0.4235232 (0.0129650)	0.4227978 (0.0130656)	0.4204449 (0.0135476)	0.3471185 (0.0131725)
Family Coverage				
Wholly Employer Financed	0.2600487 (0.0117892)	0.2595629 (0.0118144)	0.2585617 (0.0118841)	0.2425420 (0.0118516)
Dental Care				
Employee Coverage	0.6084625 (0.0143194)	0.6054571 (0.0145191)	0.6050754 (0.0148655)	0.5158463 (0.0155983)
Family Coverage	0.2650852 (0.0119643)	0.2612195 (0.0117805)	0.2594663 (0.0118775)	0.2399664 (0.0118266)
LONG-TERM DISABILITY INSURANCE (BENEFIT AREA 04)				
Long-term Disability Ins.	0.4041517 (0.0140258)	0.4031598 (0.0140335)	0.4024101 (0.0140112)	0.3802355 (0.0138561)
DEFINED BENEFIT AND DEFINED CONTRIBUTION PLANS (BENEFIT AREA 05)				
Defined Benefit Pension				
Wholly Employer Financed	0.5569876 (0.0150619)	0.5628944 (0.0151861)	0.5545407 (0.0152015)	0.4883613 (0.0159638)
Partly Employer Financed	0.0240989 (0.0040001)	0.0236726 (0.0039658)	0.0240302 (0.0040139)	0.0208462 (0.0035801)
Defined Contribution				
Wholly Employer Financed	0.1404187 (0.0098199)	0.1355685 (0.0096765)	0.1409745 (0.0100744)	0.1310603 (0.0096029)
Partly Employer Financed	0.2499399 (0.0112146)	0.2534026 (0.0110894)	0.2467732 (0.0108638)	0.1983440 (0.0108575)
Capital Accumulation				
Wholly Employer Financed	0.0220757 (0.0035056)	0.0219510 (0.0035026)	0.0221452 (0.0035089)	0.0217074 (0.0034981)
Partly Employer Financed	0.0740874 (0.0055900)	0.0732949 (0.0055616)	0.0738513 (0.0055536)	0.0608907 (0.0052621)

NOTE: The figures in parentheses are standard errors of the corresponding estimates.

Table 3. Comparison of Participation Ratios Across Methods for a Given Cell
(Pre-Edit Participation Ratio / Post-Edit Participation Ratio)

Plan	Status	CELL MEAN IMPUTATION	NEAREST NBR. HOT DECK IMPUTATION	RANDOM HOT DECK IMPUTATION
1	Usable	1.00000 / 1.00000	1.00000 / 1.00000	1.00000 / 1.00000
2	Usable	1.00000 / 1.00000	1.00000 / 1.00000	1.00000 / 1.00000
3	Usable	1.00000 / 1.00000	1.00000 / 1.00000	1.00000 / 1.00000
4	Usable	1.00000 / 1.00000	1.00000 / 1.00000	1.00000 / 1.00000
5	Imputed	0.75000 / 0.00000	1.00000 / 0.25000	0.00000 / 0.00000
6	Imputed	0.75000 / 0.00000	0.75000 / 0.25000	1.00000 / 0.25000
7	Imputed	0.80000 / 0.00000	0.00000 / 0.00000	1.00000 / 0.40000
8	Imputed	0.80000 / 0.00000	0.00000 / 0.00000	0.00000 / 0.00000
9	Imputed	0.82609 / 0.08696	0.00000 / 0.00000	0.00000 / 0.00000
10	Imputed	0.82609 / 0.08696	0.00000 / 0.00000	1.00000 / 0.30435
11	Imputed	0.84211 / 0.10526	0.00000 / 0.00000	0.00000 / 0.00000
12	Imputed	0.84211 / 0.10526	1.00000 / 1.00000	1.00000 / 0.47368
13	Imputed	0.81818 / 0.09091	1.00000 / 0.18182	0.63636 / 0.09091
14	Imputed	0.81818 / 0.09091	1.00000 / 0.18182	0.63636 / 0.09091
Mean*		0.81817 / 0.39678	0.63193 / 0.46970	0.66131 / 0.45544
Std. Deviation*		0.16067 / 0.54956	0.57960 / 0.55148	0.51673 / 0.49276

*Weighted mean and weighted standard deviation.

Coverage, Defined Benefit Pension Wholly Employer Financed, and Defined Contribution Partly Employer Financed estimates based on usable data only are significantly different (at α=0.05 level) from each of the corresponding estimates when both usable and imputed data were used..

The estimates based on usable data only are lower than the estimates based on usable and imputed data. There are two reasons for this. Firstly, when only usable data are used, for occupations having at least one usable and one unusable plan, disregarding the unusable plan is equivalent to acting as if it had zero participants. After participant imputation, the unusable plan often has a nonzero imputed participant figure. Secondly, there difference is likely to be a reflection of item nonresponse bias in the estimates based on original usable data. Since each of the cell-defining auxiliary variables has a significant effect on participation, imputing from within the imputation cells should reduce the item nonresponse bias.

V. CONCLUSIONS

Survey methodologists are often faced with having to decide what approach to use to impute for missing responses. We found that for each of the imputation methods studied, the choice of imputation method did not significantly affect the estimates. Also, the variance estimates obtained did not appear to vary much across imputation methods. We provided a comparison of the estimates obtained when the "original usable data only" dataset was used and when the "usable and

imputed data" dataset was used; differences among some of these estimates are likely to be a reflection of bias due to missing data, as well as underestimation of participation due to ignoring possible participation in unusable plans.

There are other issues to address. There are several practical issues that involve the ease of implementation, such as ease of programming, amount of collapsing, and cost of executing. For our particular implementation, all three methods appeared to be relatively equivalent in their difficulty to program. The nearest neighbor hot-deck and random hot-deck methods required more collapsing of cells than the cell mean method, since those methods attempted to use each usable plan at most once as a donor. The cell mean method was the least costly to implement in our case, mostly due to the fact that this method involved less collapsing than the other two methods.

VI. ISSUES FOR FURTHER RESEARCH

An important issue for future consideration is the comparison of subdomain estimates when different imputation methods are used. Often, analysts are interested in looking at subdomain estimates, e.g. estimates for particular occupations and/or industries. Since occupation and industry were used as cell-defining variables, these subdomain estimates may differ across imputation methods.

We discussed the fact that the variance estimates under the cell mean method were not always smaller than the variance estimates under the hot-deck methods. We gave reasons why this phenomenon occurs with our survey data. Although edits are survey-specific and generalizations may be difficult if not impossible, we would like to learn more about the impact edit constraints have on imputation procedures and estimates.

In this study, we have compared three imputation methods commonly used in establishment surveys. However, there are other methods that are currently being used, as described in Section III.D. We would like to test these other methods on the same dataset and compare the results with those presented in this paper

REFERENCES

BLS Handbook of Methods (Bulletin 2414,September 1992), Washington, D.C.: Bureau of Labor Statistics, 67-77.

Fay, R.E. (1992), "When Are Inferences from Multiple Imputation Valid?" *Proceedings of theSurvey Research Methods Section,* Washington, D.C.: American Statistical Association.

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

_____ (1986), "The Treatment of Missing Survey Data," *Survey Methodology,* 12, 1-16.

Kalton, G., and Kish, L. (1981), "Two Efficient Random Imputation Procedures," *Proceedings of the Survey Research Methods Section,* Washington, D.C.: American Statistical Association, 146-151.

Little, R.J.A. (1986), "Missing Data in Census Bureau Surveys," in *Proceedings of the Second Annual Census Bureau Research Conference,* Washington, D.C.: Bureau of the Census, 442-454.

Oh, H.L., and Scheuren, F. (1980), "Estimating the Variance Impact of Missing CPS Income Data," *Proceedings of the Survey Research Methods Section,* Washington, D.C.: American Statistical Association, 408-415.

Platek, R., and Gray, G.B. (1978), "Nonresponse and Imputation," *Survey Methodology,* 4, 144-177.

Rao, J.N.K., and Shao, J. (1992), "Jackknife Variance Estimation with Survey Data under Hot Deck Imputation," *Biometrika,* 79, 811-822.

Rubin, D.B. (1978), "Multiple Imputations in Sample Surveys--A Phenomenological Bayesian Approach to Nonresponse," *Proceedings of the Survey Research Methods Section,* Washington, D.C.: American Statistical Association, 20-34.

_____ (1987), *Multiple Imputation for Nonresponse in Surveys,* New York: John Wiley & Sons, Inc.

Sedransk, J. (1985), "The Objective and Practice of Imputation," in *Proceedings of the First Annual Research Conference,* Washington, D.C.: Bureau of the Census, 445-452.

Shah, B.V., Barnwell, B.G., Hunt, P.N., and LaVange, L.M. (1992), *SUDAAN User's Manual, Release 6.0,* Research Triangle Park, N.C.: Research Triangle Institute.

West, S.A., Butani, S., Witt, M., and Adkins, C. (1989), "Alternate Imputation Methods for Employment Data," in *Proceedings of the Survey Research Methods Section,* Washington, D.C.: American Statistical Association, 227-232.