MULTIPLE IMPUTATION IN A COMPLEX SAMPLE SURVEY

Eugene M. Burns*, Energy Information Administration
Energy Information Administration EI-651, Washington, DC 20585

KEY WORDS: Item nonresponse, jackknife, replication, hot-deck

1. INTRODUCTION

Multiple imputation for missing survey data is relatively new concept. As defined by one of its leading proponents, "multiple imputation is the technique that replaces each missing or deficient value with two or more acceptable values representing a distribution of possibilities" (Rubin 1987, p.2). Multiply-imputed data reflects the uncertainty contained in the imputation process in a way not possible with singly-imputed data.

Incorporating multiple imputation into a survey, both as an imputation method and in subsequent estimation, does present some practical problems. So far, most discussions of multiple imputation have been limited to surveys with some relatively tractable sample design (or have disregarded the sample design (Oh and Scheuren 1980)). However, complex sample surveys are another important set of surveys. In addition to the need for large-scale imputation for many missing items, complex sample surveys are characterized by features such as multistage stratified cluster sampling (possibly list-supplemented), reweighting for unit nonresponse, and possible post-stratification adjustments. Incorporating multiple imputation into such a survey would seem to be a formidable task.

This paper describes proposed procedures, and presents results, for incorporating multiple imputation into one complex sample survey, the Energy Information Administration's Commercial Buildings Energy Consumption Survey (CBECS) (Energy Information Administration 1988, 1989). The CBECS, conducted triennially, has a multistage area probability sample (list-supplemented) of approximately 7,000 buildings. In the 1986 CBECS, one Primary Sampling Unit (PSU) was selected from each of 129 sampling strata. Variances were estimated for the 1986 survey using a jackknife estimator. For variance estimation, the 129 PSUs were collapsed into 44 strata, each with two units, plus a separate stratum consisting of buildings selected with certainty, for a total of 89 units. The first replicate for variance estimation consisted of all units except one unit, designated at random, from the first collapsed stratum. Replicates 2 through 44 were defined in an analogous fashion, randomly omitting one of the two units from strata 2 through 44, respectively. (The jackknife estimator did not make use of the 44 complementary replicates, those which could have been created by including the units originally omitted and excluding the units originally selected.)

All CBECS surveys (1979, 1983, and 1986) have dealt with the problem of unit nonresponse by adjusting sampling weights. For the 1986 CBECS, both overall and replicate nonresponse adjustment factors were calculated. Replicate adjustment factors were computed by using only those cases included in the replicate, according to the jackknife design. Survey point estimates were calculated, as in prior survey cycles, by using the overall, full sample, adjustment factors. However, the survey variances were calculated using the replicate, "pseudosample," adjustment factors. By using nonresponse adjustments calculated separately for each replicate, the estimates of variance could account for the effects of the unit nonresponse adjustment procedure.

The standard CBECS practice in dealing with item nonresponse has been to impute for missing values from the building survey with a hot-deck procedure. The distinction between imputed and reported values has been ignored in variance estimation.

The CBECS building questionnaire contains about 300 items. In 1986, about two-thirds of these items had at least one case with a missing value. Most of the item nonresponse rates were low, but two important items, building square footage and number of employees, were among

*The author thanks Dwight French, Paul Gargiullo, and Miriam Goldberg (Energy Information Administration) for their comments on this paper. The opinions expressed herein are solely those of the author and should not be construed as representing the opinions or policy of any agency of the United States Government.

those with higher rates. These two items were selected to assess the impact of multiple item imputation on the CBECS estimates. For square footage, a very important analysis variable, the nonresponse rate was 28.7 percent, and for number of employees the nonresponse rate was 11.6 percent. Anticipating a nonresponse problem, the CBECS questionnaire follows both the square footage question and the employment question with another question which requires only a categorical response. The nonresponse rate was much lower, 1.6 percent, for each of the categorical versions of the questions. In imputing for missing numeric responses, the corresponding categorical items (and other related items) have been used to define the hot-deck cells.

This paper presents the results of an empirical investigation of multiple imputation for square footage and number of employees, using data from the 1986 CBECS. The next section of this paper describes the methods used for the incorporation of multiple imputation, and the results obtained. The following section presents the results of an item nonresponse simulation, and the paper concludes with a discussion of the results.

2. MULTIPLE IMPUTATION

The basic approach chosen to incorporate multiple imputation into the CBECS was to impute for missing cases both in the full sample and separately in each of the 44 replicates (pseudosamples) employed in the jackknife variance calculation. The full sample imputed values are to be used to produce survey point estimates, while the pseudosample imputed values are to be used to calculate variances. This proposed use of full sample and pseudosample values is analogous to the use made of full sample and pseudosample unit nonresponse weight adjustments. (A similar suggestion can be found in Ford (1983).) Both the unit and the item nonresponse are handled within cells defined over all units included in the replicate, rather than within each unit, due to the limited sample sizes available with each unit.

The step to replicate weight adjustment was a relatively easy one, given the fact that the CBECS was already employing replication methods for variance calculation. The use of replicate unit nonresponse adjustment, in turn, facilitated the choice of replicate item nonresponse adjustment. Thus, it was conceptually easier to find a way to incorporate multiple item imputation into a complex sample survey, which had already adopted replication methods to deal with intractable variance estimation problems, than it might have been in a simpler survey.

The hot-deck algorithm used in this study was the same one used to impute for the 1986 CBECS data. For item X, the algorithm proceeded as follows:

1. a uniform random number, U_i, was assigned to each case, and the cases were partitioned into a set of donors (with X reported) and receivers (with X missing);
2. the donors and the receivers were sorted separately into cells by square footage category (9 levels), employment category (12 levels), and principal building activity (10 types), and by U_i within cells;
3. donors and receivers were matched one-to-one within cells until either (i) all receivers found donors or (ii) the supply of donors was exhausted;
4. if all receivers were matched with donors in Step 3, then the hot-deck procedure was finished; otherwise, Steps 1 through 3 were repeated omitting the principal building activity as a sorting criterion.

The hot-deck algorithm was first applied to the full sample, and then separately to each of the 44 pseudosamples.

After the imputations were completed, four quantities were estimated, first with the use of singly- imputed values and then with the use of multiply-imputed values. Two quantities were totals: the total square footage and the total number of employees. One quantity, square footage per building, was a ratio mean, and the fourth quantity was a ratio, square footage per employee. These quantities were selected to represent the types of statistics presented in the CBECS reports (Energy Information Administration 1988, 1989).

The quantities were first estimated with singly-imputed values and then with multiply-imputed values. The quantities estimated with singly-imputed values used the full sample imputed values for both point estimates and variances, as has been the standard CBECS practice. The quantities estimated with multiply-imputed values used the full sample imputed value for the point estimate, and the replicate imputed values for the variances. All calculations were performed using a SAS program written by the author. The program results were validated against Westat's

Table 1. Comparison of Jackknife Variances and Relative Standard Errors (RSEs) Calculated Using Single and Multiple Item Imputation

Item Estimated	Point Estimate	RSEs: Single Imputation	RSEs: Multiple Imputation	RSE(multiple) - RSE(single): RSE Diff.	RSE(multiple) - RSE(single): Pct.Diff.
Square footage (million sq.ft.)	58,215	3.031	3.513	0.481	15.9
Number of employees (thousand)	73,613	3.940	4.759	0.818	20.8
Square feet per building	14,014	3.554	3.991	0.437	12.3
Square feet per employee	791	3.164	4.601	1.438	45.4

WESVAR program (Flyer and Mohadjer 1988), which calculates jackknife variances with replicate weights, for the singly-imputed case.

Table 1 shows the overall effects of multiple imputation on estimated variances and relative standard errors (RSEs). (The relative standard error is the standard error expressed as a percent of the quantity estimated, and is the form in which variances are presented in CBECS publications.) As Rubin (1987) has demonstrated, hot-decking from an empirical distribution of sample respondents understates the amount of variation to be found in the population. However, the RSEs in Table 1 are correct for the procedure used. The problem lies with the imputation procedure, which understates variation, rather than with the variance estimator.

Results had been expected to bear some relationship to the item nonresponse rates. However, although the item nonresponse rate for square footage was over twice as large as the rate for employment, the effects of multiple imputation were similar for the two items. The RSE for square footage was 15.9 percent higher with multiple imputation than with single imputation, while the RSE for employment was 20.8 percent higher with multiple imputation. Not surprisingly, the RSE for square feet per employee increased the most (45.4 percent), reflecting the increases in both numerator and denominator variation, while the RSE for square feet per building increased the least (12.3 percent), reflecting the increase in the numerator variation only.

Different patterns of item nonresponse could be responsible for the difference observed between square footage and employment. Square footage tends to have a higher nonresponse rate for the smaller (and narrower) square footage categories, while employment nonresponse is higher for the larger (and broader) employment categories. Therefore, the square footage imputed values, hot-decked within categories, are more tightly bounded than the number of employees imputed values.

3. THE ITEM NONRESPONSE SIMULATION

An item response simulation was designed to address some additional questions. The preceding comparison (of RSEs estimated using singly-imputed values versus RSEs estimated using multiply-imputed values) could not answer the question of whether the hot-deck procedure produces biased point estimates, nor of how the RSEs based on either singly- or multiply-imputed values would compare with RSEs based on the true values, if they had been available. Furthermore, the results may have been dependent on the particular set of item respondents and nonrespondents found in the data.

The two items, square footage and employment, were assumed to be missing at random within cells defined by the cross-classification of square footage category, employment category, and principal building activity. This assumption appeared to be valid. Data from the complete CBECS sample were used to estimate the proportion of cases within each cell with (i) both items reported, (ii) employment reported but not square footage, (iii) square footage reported but not employment, and (iv) neither item reported.

To form the simulated populations with item nonresponse, any case with nonresponse to either item was discarded. Item nonresponse was then

simulated by assigning a uniform random number to each remaining, fully-reported, case. Depending on the value of the random number, and the square footage-employment-building activity cell to which the case belonged, the case was simulated to have neither, either, or both items missing. A total of thirty simulated populations were formed.

Imputation and estimation were accomplished for each simulated population using the same procedures as had been used for the complete sample, with one exception. For a few cases in some of the replications, donors were not found for some cases with missing data, even after collapsing cells over building activities. This problem was anticipated, since the sample size had been decreased by nearly one-third through the elimination of cases missing either item from the complete sample. These final few cases were hot-decked within cells defined only by the categorical version of the missing item.

Table 2 is the simulation version of Table 1, and shows roughly the same pattern of effects of multiple imputation on the estimated RSEs. The percent increases in the estimated RSEs for number of employees was about 20 percent for both cases. The mean percent increase for square footage was lower in the simulated nonresponse case than in the actual nonresponse case (10 percent rather than 15 percent) and, accordingly, the increases for square feet per employee were also lower (32 percent rather than 45 percent).

Table 3 compares the means of the point estimates from the nonresponse simulations to the point estimates based on actually reported data. On the average, the estimates from the simulation were slightly higher.

Table 4 is an interesting summary table. The RSEs obtained using singly-imputed values are shown to be very similar to, but slightly lower than, the RSEs based on the fully-reported data. In both the fully-reported data and the singly-imputed data, one value represented each item for

Table 2. Comparison of Jackknife Variances and Relative Standard Errors (RSEs) Calculated Using Single and Multiple Item Imputation, Based on Thirty Simulation Runs

Item Estimated	Point Estimate	Mean RSEs: Single Imputation	Mean RSEs: Multiple Imputation	RSE(multiple) - RSE(single): RSE Diff.	RSE(multiple) - RSE(single): Pct.Diff.
Square footage (million sq.ft.)	33,978	4.496	4.975	0.480	10.7
Number of employees (thousand)	47,229	4.903	5.857	0.954	19.5
Square feet per building	13,525	4.496	4.996	0.500	11.1
Square feet per employee	719	4.281	5.624	1.344	31.5

Table 3. Accuracy of Point Estimates for Hot-deck Item Imputation Procedure, Based on Thirty Simulation Runs

Item Estimated	Point Estimate (Actual Reported Data)	Simulation Point Estimate: Mean	Simulation Point Estimate: Std.Dev.	Point Estimate Error: Mean	Point Estimate Percent Error: Mean	Point Estimate Percent Error: Minimum	Point Estimate Percent Error: Maximum
Square footage (million sq.ft.)	33,881	33,978	137	97	0.3	-0.6	1.0
Number of employees (thousand)	47,135	47,229	223	93	0.2	-1.1	1.0
Square feet per building	13,486	13,525	55	39	0.3	-0.6	1.0
Square feet per employee	719	719	4	1	0.1	-0.9	1.2

Table 4. Comparison of Jackknife RSEs Calculated Using Single and Multiple Item Imputation With Those Obtained Using Actual Reported Data, Based on 30 Simulation Runs

Item Estimated	RSE Estimate (Actual Reported Data)	Single Imputation					Multiple Imputation				
		RSE	RSE Diff	Pct. RSE Difference			RSE	RSE Diff	Pct. RSE Difference		
		Mean	Mean	Mean	Min	Max	Mean	Mean	Mean	Min	Max
Square footage (million sq.ft.)	4.490	4.496	0.005	0.1	-2.2	2.0	4.975	0.485	10.8	-5.4	35.8
Number of employees (thousand)	5.020	4.903	-0.117	-2.3	-6.8	4.3	5.857	0.837	16.7	-2.7	60.3
Square feet per building	4.544	4.496	-0.048	-1.0	-3.3	0.7	4.996	0.452	9.9	-3.3	30.7
Square feet per employee	4.393	4.281	-0.112	-2.6	-12.3	3.2	5.624	1.231	28.0	0.2	80.8

each case, and so the calculated RSEs were unable to distinguish levels of uncertainty between reported and imputed data. However, the RSEs based on the multiply-imputed values are all considerably larger than those based on the fully-reported data, indicating that uncertainty involved in the imputation procedure is being captured by the jackknife RSE estimates.

4. DISCUSSION

This study has demonstrated the feasibility, and the desirability, of incorporating multiple imputation into a complex sample survey. However, some additional issues need to be discussed.

One set of issues involves computer resources: both the space needed to store 45 copies of a fairly large data set, and the time required to produce imputed values and to calculate variances. In the computing environment available for CBECS data processing, these are not serious barriers to the implementation of multiple imputation. A more serious computer-related issue for the implementation of multiple imputation is the capabilities of the software, TPL-VARIANCE (Gargiullo and Goldberg 1989), used to produce CBECS tables of variances and estimates. TPL-VARIANCE already incorporates replicate unit nonresponse adjustments into its estimates. However, multiple item imputation would require a different programming approach from that taken to incorporate replicate unit nonresponse. One large benefit of reprogramming to handle replicate data sets is that this approach would allow many different sources of survey processing variation, not just multiple item imputation, to be reflected in estimates of survey variances. In principle, TPL can handle replicate data sets (using the TPLMULGN procedure), but work on this project is in its initial stages.

A second issue is whether the approach described in this paper, replicate item nonresponse imputation, should be called "multiple imputation" or not. It might be argued that the above approach should not be called "multiple imputation," because "true" multiple imputation would seem to require the generation of multiple imputed values within each replicate (or unit within replicate). However, this is really a side issue, since the important question is whether the proposed approach is an appropriate way of dealing with the problem of reflecting imputation variance within the CBECS framework. The above approach definitely captures the spirit of multiple imputation, and represents a feasible solution to the problem of adapting this valuable concept to the complex sample survey situation.

Finally, multiple imputation was tested in this study because it allows estimates of variance to reflect variability in the imputation procedure. However, the fact that imputation variability can be reflected so directly in the estimates focuses attention back onto the imputation procedure. The survey practitioner is less likely to be satisfied with suboptimal imputation procedures once their effects can be directly observed in survey estimates. Multiple imputation thus provides a

new spur for the development and improvement of imputation procedures.

REFERENCES

Energy Information Administration 1988. *Nonresidential Buildings Energy Consumption Survey: Characteristics of Commercial Buildings, 1986.* DOE/EIA-0246(86).

_____ 1989. *Nonresidential Buildings Energy Consumption Survey: Commercial Buildings Consumption and Expenditures, 1986.* DOE/EIA-0318(86).

Flyer, P. and Mohadjer, L. 1988. *The WESVAR Procedure.* Rockville, MD: Westat, Inc.

Ford, B.L. 1983. "An Overview of Hot-Deck Procedures," in *Incomplete Data in Sample Surveys, Volume 2, Theory and Bibliographies.* New York: Academic Press.

Gargiullo, P.M. and Goldberg, M.L. 1989. "A Modified Table Producing Language (TPL) System for Producing Tables of Survey Statistics with Variances," *Proceedings of the Bureau of the Census Fifth Annual Research Conference.*

Oh, H.L. and Scheuren, F.J. 1980. "Estimating the Variance Impact of Missing CPS Income Data," *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

Rubin, D.B. 1987. *Multiple Imputation for Nonresponse in Surveys.* New York: Wiley.