Assessing Several Hot Deck Imputation Methods Using Simulated Data from Several Economic Programs

Laura Bechtel[1], Yarissa Gonzalez, Matthew Nelson, Roberta Gibson
U.S. Census Bureau, Washington, D.C. 20233

Abstract

Economic programs conducted by the U.S. Census Bureau often use ratio imputation models to impute missing or erroneous values. These methods are designed to yield consistent estimated totals at the cost of failing to preserve the underlying distribution of the micro data. Hot deck imputation procedures classify units into disjoint groups based on variables assumed to be correlated with the missing values. Donor values are then matched from respondents to nonrespondents within the classification group, thus preserving the within-unit between-item multivariate relationships. In this paper, we present the results of a simulation study that evaluates the performance of hot deck imputation on data modeled from three very different economic programs, considering three different hot deck methods (both performed with and without micro-level adjustments.)

Key Words: hot deck imputation, simulation, Kolmogorov-Smirnov tests

1. Introduction

Economic programs conducted by the U.S. Census Bureau often use ratio imputation models to impute missing or erroneous values. These methods are designed to yield consistent estimated totals at the cost of failing to preserve the underlying distribution of the micro data. In this paper, we evaluate hot deck imputation as an alternative imputation method that may improve the ratio imputation procedures, by preserving the underlying distribution of the micro data in addition to yielding consistent estimated totals.

Hot deck imputation procedures use reported values from the current sample to impute for missing values. Sample units are classified into disjoint groups (imputation cells) based on variables ***available for all units in the sample*** that are correlated with the missing values. By classifying the sample units in this way, it is reasonable to assume that within each classification group, nonrespondents follow the same distribution as respondents (Ford, 1983). Donor values from respondents (donors) are then matched to nonrespondents (recipients) within an imputation cell. In theory, this approach preserves the expected cell totals and preserves the within-unit between-item multivariate

[1] Any views expressed on statistical or methodological issues are those of the authors and not necessarily those of the U.S. Census Bureau.

relationships. Hot deck procedures can be used to account for both unit and item nonresponse.

In this paper, we consider three hot deck imputation methods: Random Hot Deck (RHD), Sequential Hot Deck (SHD), and Backward-Forward Hot Deck (BFHD) as defined in Kalton and Kasprzyk (1986). Prior to applying any imputation method, the units' data are evaluated (edited), and units are marked as donors (contain valid data for all inspected items) or recipients (require valid replacement items).

The RHD method randomly selects a donor within an imputation cell and assigns the donor value(s) to the unit that has missing or unusable data. The number of times a variable can be used is predetermined by the statistician. If there are fewer donors than recipients, a donor may need to be used more than once or an alternative imputation method may need to be used e.g., a mean or median calculated using historic or current data.

The SHD method sorts the units in an imputation cell by an auxiliary variable(s) – available for all sampled units – that is correlated with the variable(s) being imputed. Prior to applying the hot deck procedure to the imputation cell data, an initial donor value is stored in the "donor deck." The SHD method then goes through the sorted cell from top to bottom. If the first observation is flagged for imputation, the "donor deck" value will be used. If the first observation is a donor, the donor deck value is replaced with the first observation's value. The algorithm proceeds through the remaining imputation cell data in similar manner; either replacing the donor deck value if the observation is a donor or imputing using the donor deck value if it is a recipient. This imputation method clearly lends itself to using a donor more than once, especially when there is a cluster of recipients next to each other in the sorted imputation cell.

Like the SHD method, the BFHD method sorts the units within the imputation cell by an auxiliary variable(s) that is correlated with the item(s) being imputed. However, instead of starting with the first observation in the imputation cell after sorting, BFHD imputation starts with the first ***recipient*** observation in the imputation cell and attempts to find a donor from the sorted data above, going backward. It stops at the first available donor found and uses that value. If it does not find a donor searching backward, the BFHD algorithm goes forward (toward the bottom of the cell) and uses the first available donor value(s) to impute. Like with the RHD method, the statistician must decide how many times a donor may be used and what alternative value will be used if no donor is found.

Hot deck imputation is appealing in that it can be used to find a replacement value for a single item or for a group of items. An alternative approach is to develop an item-specific (univariate) imputation model. This is the general practice for the majority of current economic surveys, which attempt to use direct substitution imputation or ratio imputation models (Ozcoskun and Hayes, 2009). Direct substitution methods (also known as logical edits) replace the missing or edit-failing data item value with information (for the same unit) obtained from another source. The ratio imputation methods used for imputation in this research utilize the following prediction model:

Auxiliary $$y_{tj} = \beta x_{tj} + \varepsilon_{tj},\ \varepsilon_{tj} \sim (0, x_{tj}\sigma^2) \qquad (1.1)$$

Under the ***auxiliary*** model, the Best Linear Unbiased Estimator (B.L.U.E.) of β within imputation cell *p* is given by $\hat{\beta} = \sum_j w_{tj} y_{tj} I_{tj} / \sum_j w_{tj} x_{tj} I_{tj}$, where I_{tj} is a response indicator variable and w_{tj} is the sampling weight (the inverse of probability of selection) for unit *j* in statistical period *t* (Magee, 1997). In practice, imputation parameters are developed separately for each imputation cell. The auxiliary variable (*x*) differs by item (*y)* and is available for all sampled units.

There are several advantages to using these univariate imputation models in production. With direct substitution, model validation is not necessary. With the ratio models, model validation is fairly straightforward using standard residual analysis techniques or cross-validation. Since each data item is considered separately, it is possible to offer a hierarchy of imputation options to be attempted in order of expected reliability. Lastly, ratio regression models are explicitly designed to predict consistent values, so imputed values do not perturb the expected totals. However, there are several disadvantages. First and most important, item-by-item imputation fails to preserve multivariate relationships. This failure can affect the precision of ratio estimators by reducing the between-item correlation coefficient. The lack of between-item correlation in modeling likewise affects balance complexes (sets of items comprised of a total and associated details). Second, the distribution of the imputed microdata resulting from univariate ratio imputation can be very different from the population distribution, which affects the dataset's utility for subsequent analyses.

The evaluation of hot deck imputation presented here uses simulated data develop from three different economic programs: the Annual Capital Expenditures Survey (ACES), the Plant Capacity Utilization Survey (PCU), and the Services Annual Survey – Information Sector (SAS - I). Each program accounts for nonresponse differently. The PCU and the SAS-I perform item imputation using a combination of imputation methods including a ratio imputation model, whereas the ACES uses a unit-nonresponse weighting adjustment. For these simulations, we restrict our evaluations to performance on unit nonrespondents and make the broad --and unrealistic -- assumption that the remaining data are valid. This assumption avoids confounding, thus facilitating analysis of the procedures under consideration while rendering it impossible to draw direct conclusions about the studied programs' procedures.

In Section 2, we provide a brief overview of each program's key items collected, sampling methodology, and ratio imputation procedures. In Section 3, we present the simulation studies for each program and discuss our results. We finish in Section 4 with some concluding remarks and recommendations for future research.

2. Program Background

2.1 Annual Capital Expenditure Survey (ACES)

The Annual Capital Expenditures Survey (ACES) collects data about the nature and level of capital expenditures in non-farm businesses operating within the United States. Respondents report capital expenditures, broken down by type (expenditures on Structures and expenditures on Equipment) for the calendar year in all subsidiaries and divisions for all operations within the United States.

Each year, the ACES selects an independent stratified simple random sample without replacement from two subpopulations: employer companies (ACE-1) and non-employer (ACE-2) companies. Separate forms are mailed to businesses with employees (employer) and without employees (non-employer). Our research is restricted to the employer statistics that are collected on the ACE-1 forms.

In the ACE-1 design, units are stratified into size-class strata within each industry on the sampling frame. There are five separate ACE-1 strata in each industry, consisting of one certainty stratum and four noncertainty strata defined by company size within industry, ranked from largest to smallest within industry. While the ACES samples are independent from year-to-year, the certainty portion of the sample does have a large overlap of units from sample to sample.

The ACES publishes totals and year-to-year change estimates. Detailed capital expenditures data are collected from each sampled company in more than one item on the ACE-1 questionnaire. Total capital expenditures are first reported in survey item 1. Item 2 requests that the company-level value reported in item 1 be further broken down by type of capital expenditures (structures or equipment) cross-classified by new and used as shown by in Figure 1 below. The respondent company reports the same information for each industry in which the company operated and had capital expenditures by completing a separate row for each industry in Item 6 of the questionnaire. For ease in modeling, we restrict our analysis to the company-level capital expenditures variables represented by the marginal column totals in Figure 1 (Total capital expenditures, capital expenditures on structures, and capital expenditures on equipment).

Figure 1: Company Level Capital Expenditures (Item 2)

	Structures	Equipment	Total
New	X_{NS}	X_{NE}	$X_{N\bullet}$
Used	X_{US}	X_{UE}	$X_{U\bullet}$
Total	$X_{\bullet S}$	$X_{\bullet E}$	$X_{\bullet\bullet}$

The collected data are required to add to consistent values (e.g., total capital expenditures collected in Item 1 should equal $X_{\bullet\bullet}$ from Item 2 and the overall total in Item 6). Items are subjected to exact equality edits, and analysts resolve edit failures manually. This is a time-consuming process.

Although the ACE-1 survey design is fairly typical for a business survey, the collected data are not. Smaller companies often report legitimate values of zero for capital expenditures, and consequently the majority of the estimates are often obtained from the certainty and large non-certainty companies. As the capital expenditures are further cross-classified, the incidence of reported zeros (especially among smaller companies) increases. Moreover, because the ACE-1 samples are independently selected each year, historic data cannot be used when accounting for nonresponse. Instead, ACES uses an adjustment-to-sample weighting adjustment procedure that assumes the auxiliary model (1.1) with the frame payroll value as the auxiliary variable[2]. Weighting cells are the

[2] The ACE-1 weight adjustment procedure is mathematically equivalent to applying the same auxiliary imputation model to each separate item.

design strata, provided that there is at least one respondent in the cell. The adjustment cell weighting procedure has been shown to demonstrate good statistical properties overall for the survey (Haziza et al, 2010), but has some problematic model assumptions for the smallest size strata within industry (Smith and Thompson, 2009). In this paper, we evaluate if hot deck imputation can be used to produce similar or improved tabulated estimates while improving the imputation procedure for smaller strata. We also keep an eye toward other potential applications of these imputation methods for ACES data to correct inconsistent reported data.

More information on the ACES sample design, estimation procedures, and variance estimation can be found at http://www.census.gov/econ/aces/index.html.

2.2 Plant Capacity Utilization (PCU)

The PCU analysis uses data from a survey that is no longer being conducted, but was selected for inclusion in this research because of its sample design and key estimates. In 2007, the Quarterly Survey of Plant Capacity Utilization (QPC) replaced the annual Plant Capacity Utilization (PCU) survey. At that time, several improvements to the prior survey's sampling and imputation methodologies were introduced, as well as other methodological changes. Consequently, our analysis may or may not be applicable to the QPC.

The final PCU sample was selected towards the end of 2004, and the final PCU data collection was conducted in 2006. The initial frame consisted of manufacturing and publication establishments from the 2002 Economic Census and was stratified by 6 digit NAICS (North American Industry Classification System). To reduce coverage bias, additional strata were added to represent establishments that came into business (were born) after 2002. The primary portion of the PCU sample was selected using a pps-sample design with census total receipts used as the measure of size for each establishment in the frame. Establishments actually selected for the sample were assigned sampling weights equal to the inverse of their respective probabilities of selection. For a detailed description of the PCU data and sample design see the publication appendices: http://www.census.gov/prod/2007pubs/mqc1-06.pdf.

The PCU published industry level estimates of plant capacity utilization rate, defined as the ratio of actual production to full production capability. The PCU uses direct substitution and auxiliary ratio imputation for each of these totals. We examine whether hot deck imputation can be used to preserve the within-unit correlation structure for these items, which in turn could lead to more precise estimation.

2.3 Services Annual Survey – Information (SAS-I)

The Service Annual Survey – Information (SAS-I) collects and publishes estimates of revenue, expenses, and inventories for information service industries. SAS-I uses a stratified simple random sample design. Stratification is performed by industry group, which is then further stratified by a measure of size related to estimated sales (or revenues). For each stratum there is one certainty sub-stratum and up to 12 non-certainty sub-strata. The sampling units are either companies for certainties or Employee Identification Numbers (EINs) for noncertainties. Each sampling unit represents one or more establishments owned or controlled by the same firm. Imputation cells are defined by industry code and tax status.

The key items collected by SAS-I are total revenue and total expenses. For both revenue and expenses, there are many detail revenues and expenses reported that sum up to their respective totals. In this paper, we focus on the expense detail items.

For balance complexes, SAS-I first imputes the totals. For this, there is a hierarchy of imputation methods designed to minimize modeled replacement values or to preserve a previous within-item relationship before resorting to cell-level ratio imputation. After obtaining a valid value for total expenses, the associated the expense detail items are imputed using the detail-to-total auxiliary ratio model presented in (1.1), where x_i is the total expense value for item i and y_i is the detail expense item being imputed. See **http://www.census.gov/services/index.html** for more details on the SAS-I methodology.

In this paper, we examine whether imputing a set of detail values simultaneously from one donor provides improvements over the currently employed ratio imputation methods.

3. Simulation Studies

3.1 Populations and Samples

Each program modeled a complete population from existing sample data. The ACES and SAS-I populations described below were created expressly for this research project. The PCU populations were developed for the research reported in Steel at al (2009).

Initial ACE-1 and SAS-I populations were generated using the nonparametric nearest neighbor SIMDAT algorithm (Thompson, 2000) using the program described in McNerney and Adeshiyan (2006). The PCU populations were modeled as multivariate lognormal by industry using the lognormal program described in McNerney and Adeshiyan (2006).

Both simulation approaches require an input (training) dataset that clearly identifies the modeling cell and provides a unit-level weight that sums to the population size. For ACES and SAS-I, we used the auxiliary ratio model (1.1) to obtain a nonresponse weight adjustment in each stratum, with payroll as the auxiliary variable for ACES and receipts as the auxiliary variable for SAS-I. Nonresponse weight adjustment was not necessary for PCU because the population size was modeled from fully-imputed dataset. If there were fewer than five sample observations within a stratum, we collapsed strata for modeling.

After generating the complete populations, we selected 1,000 repeated samples using each program's sampling methodology. To be consistent with the underlying imputation cell development principle of a missing at random response mechanism (Särndal and Lundström (2005)), unit level nonresponse was randomly induced in each sample using the imputation cell level response probabilities obtained from the program's respective weighted response rates .Next, we ran each program's ratio imputation method, random hot deck, sequential hot deck, and backward-forward hot deck on the ***non-certainty*** component of each sample.

In Section 3.3, we present the results of each hot deck imputation variant. We considered two donor values per hot deck imputation variant: an unadjusted donor value that directly substitutes the donor value into the record; and an adjusted donor value obtained via a

unit-level adjustment procedure for measure-of-size to the recipient value ($y_i' = x_i^d \times (M_i / M_d)$, where x_i^d is the donor unit value of the item substituted for unit *i*, and M_i and M_d are the measures-of-size associated with the recipient and donor units, respectively[3]).

We used the programs' imputation cells as defined for ratio imputation (or nonresponse weighting adjustment cells) for all applications, without performing any assessment of whether they satisfied the properties of a response homogeneity group as defined in Särndal and Lundström (2005). Specifically, units assigned to the same imputation cell are assumed to be highly homogeneous, and each imputation cell mean should differ. This property is necessary for using hot deck imputation, but is less necessary for ratio imputation under (1.1). Additionally, our goal was to use a donor no more than once, because using a donor several times makes no improvement to the micro data over the ratio imputation methods. Once each hot deck method was applied to the data, we calculated the estimates for evaluation.

3.2 Evaluation Methodology

To evaluate the simulation results, we analyzed the effect of each imputation method on both the micro data and macro data. We chose to focus on existing estimation procedures and did not estimate the imputation component of the variance. This is consistent with current economic programs variance estimation procedures. However, we note that if we develop imputation methods that produce less biased estimates, then this omitted variance component should not contribute excessively to the mean squared error.

To assess the macro data effects of each imputation method, we considered total estimates, quartile estimates, and a mean per unit estimates. Let

x_i^e be the population value of statistic *e* (total, quartiles, and mean per unit) for item *i*.

$\hat{x}_{sim}^e$ be sample estimate for sample *s* using imputation method *m*.

To assess the estimation properties over repeated samples of each considered imputation method *d*, we computed relative bias, mean square error, and mean absolute error.

The relative bias of each estimate for each imputation procedure is given by

$$B(x)_{mi}^e = \frac{\sum_{s=1}^{1000} \frac{\hat{x}_{sim}^e}{1000} - x_i^e}{x_i^e}$$

The mean squared error (MSE) of each estimate for each imputation procedure is given by

$$MSE\ (x)_{mi}^e = \frac{\sum_{s=1}^{1000} (\hat{x}_{sim}^e - x_i^e)^2}{1000}$$

[3] For all of our programs the adjustment did not improve the considered hot deck imputation methods, so we do not present the results. However, the results are available upon request.

The mean absolute error (MAE) of each estimate for each imputation procedure is given by

$$MAE\ (x)_{mi}^{e} = \frac{\sum_{s=1}^{1000} | \hat{x}_{sim}^{e} - x_{i}^{e} |}{1000}$$

Hot deck imputation is frequently used in an effort to create micro-data with distributional properties that are similar to the population's distribution (Ford, 1983). To assess the goodness-of-fit of our imputed complete data sets, we performed Kolmogorov-Smirnov tests using a 10% significance level within each sample, comparing the empirical CDF of the imputed distribution to the empirical EDF of the sample with complete response. Ideally, we do not want to reject this hypothesis. Finally, we looked at averaged inter-item correlations without performing any tests of significance. We wanted to get a general idea of how each imputation method maintained the overall correlation structure. Note that the MAE results are omitted in the subsequent sections; they are generally parallel to the MSE results. Any differences did not alter our conclusions. MAE results are available from the authors upon request.

3.3 Results

3.3.1. ACES Simulation Results

For this study, we used the ACES weighting adjustment cells as the imputation cells. In this study, we focused on the noncertainty portion of ACE-1 and did not induce nonresponse for certainties. Since some cells had insufficient donors, we allowed the programs to use a donor twice instead of once. Payroll was used as the sort variable for SHD and BFHD.

Table 1 presents summary level results for Total Capital Expenditures averaged over the 1000 samples[4]. Table 2 presents the corresponding relative biases. In both tables, the first two columns provide the summary statistics from the population and from the original samples (no missingness) to provide (1) an assessment of the representativeness of the complete sample and (2) an (unreachable) target value for each imputation method. The remaining columns provide estimates using each considered imputation method. We do not include the corresponding statistics for the adjusted hot deck imputation methods, since results were nearly equivalent.

Table 1: Key Statistics for Total Capital Expenditures (In Thousands of Dollars)

Statistic	Complete Response		Missing at Random (Sample)			
	Population	Sample	Ratio Method	RHD	SHD	BFHD
Total	1.08E+09	1.08E+09	1.08E+09	1.08E+09	1.08E+09	1.08E+09
Q1	0.00	0.00	0.00	0.00	0.00	0.00
Median	0.00	0.05	0.06	0.07	0.05	0.05
Q3	9.86	9.77	9.69	9.73	9.60	9.65
Mean per unit	193.15	193.08	192.28	193.14	192.49	192.66

[4] Similar results were obtained for the other studied items and are available upon request.

Table 2: Relative Bias for Total Capital Expenditures

Statistic	Complete Response	Missing at Random			
		Ratio Method	RHD	SHD	BFHD
Total	-0.04%	-0.02%	-0.03%	-0.34%	-0.28%
Q1	0.00%	0.00%	0.00%	0.00%	0.00%
Median	0.05*	0.06*	0.07*	0.05*	0.05*
Q3	-1.04%	-1.82%	-1.40%	-2.75%	-2.23%
Mean per unit	-0.04%	-0.45%	0.00%	-0.34%	-0.25%

* Value is the bias, not relative bias

The results in Table 1 are very promising for all variants of hot deck imputation, with the imputed distributions appearing to be very closely aligned with the population and full sample distributions. The relative biases presented in Table 2 serve several purposes. First, they show that our study uses sufficient samples, since the complete response sample estimates are essentially unbiased. Second, they demonstrate very promising properties for the RHD method, although SHD and BFHD tend to be more biased. This pattern – improved biases and similar summary statistics with RHD over the ratio method – is repeated for the other studied items. We suspect that the SHD and BFHD methods are "overkill" for ACES, since the imputation cells are strictly delineated by industry and size, and the units within each cell are already fairly homogeneous.

Table 3 presents the MSE results for Total Capital Expenditures. The patterns are similar for both the MSE and MAE: all imputation methods overestimate the MSE and the MAE, and the errors obtained with the ratio method are closest to those from the complete response sample. All of the hot deck methods have very similar values and are not too far from the ratio method. The same general results were found when we looked at the averaged inter-item correlations; the correlations did not appear to be any better or any worse for any of the evaluated (ratio method or hot deck) imputation methods.

Table 3: Mean Squared Error for Total Capital Expenditures (In Thousands of Dollars)

Statistic	Complete Response	Missing at Random			
		Ratio Method	RHD	SHD	BFHD
Total	9.7737E14	1.0131E15	1.0368E15	1.0361E15	1.0359E15
Q1	0.0000	0.0000	0.0000	0.0000	0.0000
Median	0.0075	0 .0111	0.0150	0.0094	0.0100
Q3	0.1432	0.2781	0.3041	0.3839	0.3358
Mean per unit	31.2243	33.7745	33.1356	33.1000	33.0681

The micro-level comparisons for the ACES hot-deck imputed data sets are equally promising. In all cases, the hot deck imputed datasets appeared to conform well to the complete response sample data sets, failing to reject the null hypothesis of the Kolmogorov-Smirnov (goodness of fit) tests. In contrast, the distributions obtained from the ratio method adjusted datasets had very poor fits, rejecting the null hypothesis of the Kolmogorov-Smirnov tests approximately 75% of the time.

3.3.2. PCU Results

For this study, we used the 6-digit industry as imputation cells, as outlined in Section 2.2. Units within these cells are quite heterogeneous in size, and thus not all requirements for

response homogeneity are met (Note: the existing imputation cells are well designed for weighted imputation regression). Payroll was used as the sort variable for SHD and BFHD.

This simulation uses only the auxiliary ratio model for imputation instead of the hierarchical imputation procedure that was implemented for the program. Unfortunately, by simulating complete unit nonresponse, we lose both items' and are forced to substitute the deleted auxiliary variable value, which yields artificially perfect ratio imputation values by replacing the missing value with its original – exact – value. Tables 4, 6, and 8 present summary level results for the plant capacity utilization ratio, actual production capacity, and full production capacity averaged over the 1000 samples[5]. Tables 5, 7, and 9 present the corresponding relative biases.

Table 4: Key Statistics for Plant Capacity Utilization Ratio

Statistic	Complete Response		Missing at Random (Sample)			
	Population	Sample	Ratio Method	RHD	SHD	BFHD
Q1	0.51306	0.53878	0.57886	0.53910	0.53960	0.53881
Q2	0.63777	0.67932	0.66489	0.67931	0.67898	0.67889
Q3	0.79501	0.83999	0.78038	0.84100	0.83818	0.83910
Mean/unit	0.65923	0.65987	0.66275	0.66990	0.66067	0.66253

Table 5: Relative Bias for Plant Capacity Utilization Ratio

Statistic	Complete Response	Missing at Random (Sample)			
		Ratio Method	RHD	SHD	BFHD
Q1	5.014%	12.825%	5.075%	5.174%	5.019%
Q2	6.515%	4.253%	6.514%	6.462%	6.448%
Q3	5.658%	-1.840%	5.785%	5.430%	5.546%
Mean/unit	0.097%	0.533%	1.619%	0.218%	0.501%

With the ratio estimator, there is a pronounced bias, and the bias is inconsistent in direction and magnitude for each estimate when compared with the complete response sample. The hot deck methods achieve similar biases as the complete response sample measures.

[5] Similar results were obtained for the other studied items and are available upon request.

Table 6: Key Statistics for Actual Production Capacity

Statistic	Complete Response		Missing at Random (Sample)			
	Population	Sample	Ratio Method	RHD	SHD	BFHD
Total	11706239	11866224	11867695	11860981	12679508	35480252
Q1	404.15	419.37	412.92	419.24	418.17	422.15
Q2	981.55	1008.34	998.07	1020.71	1014.49	1149.01
Q3	2371.76	2425.31	2408.10	2492.28	2484.86	3257.83
Mean/unit	2151.58	2180.95	2181.24	2179.97	2332.89	6515.20

Table 7: Relative Bias for Actual Production Capacity

Statistic	Complete Response Sample	Missing at Random (Sample)			
		Ratio Method	RHD	SHD	BFHD
Total	0.029%	1.396%	32.132%	1.352%	8.346%
Q1	1.197%	5.008%	20.836%	4.975%	4.706%
Q2	0.822%	3.574%	26.344%	4.845%	4.206%
Q3	0.220%	2.483%	29.942%	5.313%	4.999%
Mean/unit	0.144%	1.511%	32.353%	1.465%	8.583%

Table 8: Key Statistics for Full Production Capacity

Statistic	Complete Response		Missing at Random (Sample)			
	Population	Sample	Ratio Method	RHD	SHD	BFHD
Total	17788153	17778257	18024923	23094056	18019144	19068631
Q1	657.93	662.09	660.74	780.84	683.83	680.75
Q2	1559.19	1554.12	1558.95	1892.30	1606.10	1601.70
Q3	3633.71	3623.06	3648.76	4590.11	3791.20	3790.52
Mean/ unit	3265.67	3267.49	3312.84	4246.77	3311.72	3508.37

Table 9: Relative Bias for Full Production Capacity

Statistic	Complete Response Sample	Missing at Random (Sample)			
		Ratio Method	RHD	SHD	BFHD
Total	-0.056%	1.319%	29.828%	1.299%	7.199%
Q1	0.632%	2.021%	18.681%	3.936%	3.469%
Q2	-0.325%	1.081%	21.364%	3.008%	2.726%
Q3	-0.293%	1.050%	26.320%	4.334%	4.316%
Mean/unit	0.056%	1.431%	30.042%	1.410%	7.432%

With the two total estimates, the ratio method has superior performance over repeated samples, thus validating the usage of ratio imputation methods for obtaining estimates of totals. The SHD method results in less bias than the ratio methods when looking at the total and mean per unit. However, for the quartiles the SHD biases are generally somewhat higher in magnitude.

Tables 10 through 12 present the MSE of the three studied statistics. The MSE and MAE results parallel the relative bias results, with the hot deck methods yielding more precise ratio estimates, but the ratio method yielding considerably more precise totals estimates.

Table 10: Mean Squared Error for Plant Capacity Utilization Ratio

Statistic	Complete Response Sample	Missing at Random (Sample)			
		Ratio Method	RHD	SHD	BFHD
Q1	0.00072	0.00451	0.00078	0.00082	0.00077
Q2	0.00179	0.00092	0.00184	0.00183	0.00181
Q3	0.00215	0.00038	0.00237	0.00212	0.00218
Mean/unit	0.00010	0.00012	0.00024	0.00016	0.00016

Table 11: Mean Squared Error for Actual Production Capacity

Statistic	Complete Response Sample	Missing at Random (Sample)			
		Ratio Method	RHD	SHD	BFHD
Total	1.19E+11	1.57E+11	1.59E+13	3.07E+11	2.35E+12
Q1	2450.66	3039.39	11208.30	4203.43	3894.36
Q2	8022.63	9186.60	81166.45	14334.71	12578.57
Q3	26426.30	31230.99	580003.75	66901.74	57250.03
Mean/unit	10131.39	11560.79	546771.13	16526.95	90185.27

Table 12: Mean Squared Error for Full Production Capacity

Statistic	Complete Response Sample	Missing at Random (Sample)			
		Ratio Method	RHD	SHD	BFHD
Total	2.81E+11	3.56E+11	3.20E+13	7.04E+11	4.45E+12
Q1	6440.01	6833.66	24803.24	10069.55	8971.78
Q2	16116.55	16881.32	143062.97	25878.75	24131.06
Q3	54517.21	57452.65	1078175.43	138660.97	122459.80
Mean/unit	22875.57	25895.67	1103278.57	37482.60	173562.70

Turning our focus micro-data distribution comparisons, we first examined the univariate distributions of the individual totals, again using the Kolmogorov-Smirnov tests. For totals, the ratio imputation method performed well; failing to reject the null hypothesis 100% of the time. As expected, however, the univariate ratio imputation approach does not work well for preserving the multivariate characteristics of the data. First, none of the empirical distributions developed from ratio-imputed data match the complete response distribution (100% rejection rate), whereas all of the hot deck imputed data sets do (100% acceptance rate). This conjecture was further validated by examining the sample

correlation structures. The averaged correlations from the hot deck imputed data were closer to the complete response sample correlations, where the correlation estimates from the imputed data using either of the ratio methods were consistently further away from their complete response sample counterparts and overestimated the correlation.

3.3.3. SAS-I Simulation Results

For this study, we used the existing SAS-I imputation cells (industry code by tax status). For sequential and backward-forward hot deck simulation methods, the total expenses variable was used for sorting and matching.

The imputation procedures in the SAS-I simulation study differ somewhat from the other two presented studies. Of interest is the balance complex of total expenses (see King and Bogle, 2003), which varies by industry and has 13-15 associated detail items. In the proposed SAS-I application, the total expenses will be imputed first. A unit-level distribution of the detail item values are obtained via hot deck imputation and are raked to the original total value. The measure of size adjustment did not make sense for imputing the distribution of details, so these statistics are unavailable for SAS-I.

Below, we present results for lease and rental payments, an expense detail that is included in the balance complex for all SAS-I industries. Table 13 presents the population total, quartiles, and mean per unit for lease and rental payments averaged over the 1000 samples[6]. Tables 14 and 15 present the corresponding relative biases and MSE values.

Table 13: Key Statistics for Lease and Rental Payments

Statistic	Complete Response		Missing at Random			
	Population	Sample	Ratio Method	RHD	SHD	BFHD
Total	5.78E+09	5.77E+09	5.37E+09	6.22E+09	5.92E+09	5.79E+09
Q1	3107.57	3109.01	2800.35	2481.73	2632.22	2802.89
Median	14402.84	14553.63	12057.61	12838.57	14387.37	13732.20
Q3	42403.46	43042.09	39682.38	42322.19	44810.14	43285.55
Mean/unit	72939.94	73047.36	67929.85	78741.42	75071.25	73310.27

Table 14: Relative Bias for Lease and Rental Payments

Statistic	Complete Response Sample	Missing at Random			
		Ratio Method	RHD	SHD	BFHD
Total	-0.12%	-7.13%	7.67%	2.44%	0.24%
Q1	0.05%	-9.89%	-20.14%	-15.30%	-9.80%
Median	1.05%	-16.28%	-10.86%	-0.11%	-4.66%
Q3	1.51%	-6.42%	-0.19%	5.68%	2.08%
Mean/unit	0.15%	-6.87%	7.95%	2.92%	0.51%

[6] Similar results were obtained for the other studied items and are available upon request.

Table 15: Mean Squared Error for Lease and Rental Payments

Statistic	Complete Response Sample	Missing at Random			
		Ratio Method	RHD	SHD	BFHD
Total	1.5668E+17	3.0458E+17	4.5464E+17	2.3782E+17	1.7689E+17
Q1	577172.37	281894.85	842729.42	916925.74	747236.74
Median	3397380.45	7375137.91	5535302.59	4837728.13	4173218.76
Q3	27256441.67	26433414.66	29716856.76	42113602.34	32384949.65
Mean/ unit	28977208.31	51485161.06	79842916.29	43153626.79	32958520.95

With the SAS-I samples, the BFHD method is outperforming the other methods. The bias effects of over-using the same donor are obvious with the SHD results, as is the over-correction to the mean with the ratio method. The random hot deck method seems to be taking donors from observations that are very different from the recipient observations, which is evidenced by high relative biases and overestimation of the MSE and MAE.

Looking at the Kolmogorov-Smirnov tests[7] for lease and rental payments, we find promising results for the various hot deck imputation methods. For all hot deck methods, we fail to reject the null hypothesis 100% of the time, whereas for the ratio method we reject the null hypothesis approximately 85% of the time. Finally, when we looked at the averaged correlation matrices without performing any statistical tests, the correlation structure generated from data imputed using the ratio method did not appear to represent the full sample correlation structure as well as the data imputed using the various hot deck methods.

4. Conclusion

In this study, we applied a variety of hot deck imputation methods to several different simulated economic program data sets, obtaining somewhat mixed results.For hot deck imputation, the "ideal" survey setting has the imputation cells correspond to strata and the stratification ensures homogeneous within-cell data and heterogeneous between-cell means. In our ACES simulations, the results were extremely promising in terms of both estimate level and distributional properties. For the SAS-I simulation, the results are also quite promising, with one variant of hot deck imputation maintaining the within-unit distribution of the detail items while also providing slightly better estimates of totals than the currently used ratio imputation method. Again, the data within the SAS-I imputation cells tend to be fairly homogeneous in terms of measure of size. Finally, for the PCU simulations, where the units are very heterogeneous with respect to measure of size within imputation cell, the hot deck imputation results for totals were inferior to those obtained using a ratio imputation model.

From this study, one could "take away" some validation of ratio imputation methods. That would not be a bad conclusion, but it is not the only possible conclusion. In examining correlation structures of hot deck imputed simulated data, we demonstrated that hot deck methods, even poorly applied, can improve over univariate imputation methods in preserving key multivariate data characteristics. For a survey whose primary statistic of interest is a rate, this could be viewed as a major improvement. We

[7] Kolmogorov-Smirnov Tests were only performed on 500 of 1000 repeated samples.

recommend further exploration of hot deck imputation after performing additional data analysis to determine alternative imputation cells that incorporate the differences in unit size within industry.

Acknowledgments

The authors thank Katherine Jenny Thompson, Victoria McNerney, Ben Neely, Steve Riesz, Xijian Liu, and Suzanne Dorinski for their valuable contributions to this research project.

References

Ford, Barry L. (1983). An Overview of Hot-Deck Procedures. *Complete Data Sample Surveys* (Vol. 2), pp185 – 207. City: Academic Press, Inc.

Haziza, D., Thompson, K.J., and Yung, W. (2010). The Effect of Nonresponse Adjustments On Variance Estimation. *Survey Methodology*, **36**, pp. 35-43.

Kalton, G. and Kasprzyk, D. (1986). The Treatment of Missing of Survey Data. *Survey Methodology*, **12**(No. 1), pp. 1-16.

King, Carol S. and Bogle, Rebecca D. (2003). Using Hot Deck Donor Imputation Methodology in the Service Annual Survey. *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Magee, L. (1998). Improving survey-weighted least squares regression. *Journal of the Royal Statistical Society*, **60**(No. 1), pp. 115-126.

McNerney, V.G. and Adeshiyan, S.A. (2006). User Guide for Generalized Population Simulation Programs. Internal memorandum, Office of Statistical Methods and Research for Economic Programs of the U.S. Census Bureau.

Ozcoskun, L. and Hayes, M. (2009). *The Economic Directorates Editing and Imputation Inventory*. OSMREP, U.S. Bureau of the Census, Washington, DC.

Särndal, C.-E., and Lundström, S. (2005). Estimation in Surveys with Nonresponse. New York: John Wiley & Sons, Inc.

Smith, J.Z., & Thompson, K.J. (2009). Nonresponse Bias Study for the Annual Capital Expenditures Survey. *Proceedings of the Section on Government Statistics*, American Statistical Association.

Steel, P., McNerney, V., and Slanta (2009). An Investigation of Stratified Jackknife Estimators Using Simulated Establishment Data Under an Unequal Probability Sample Design, *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Thompson, James R. (2000). Simulation: A Modeler's Approach. New York: John Wiley & Sons, 87-110.