# ACCOUNTING FOR IMPUTATION WHEN ESTIMATING VARIANCES IN THE ECONOMIC SURVEYS AT THE CENSUS BUREAU

**Patrick J. Cantwell**
**Bureau of the Census, SRD, Washington, DC 20233-9100**[1]
patrick.j.cantwell@census.gov

**Key words: inflation factor; within-unit imputation; ratio imputation, regression imputation.**

## 1. Introduction

In most surveys, we encounter missing data of one or more types. Some sample units leave some data items blank--item nonresponse. The usual approach is to impute all missing, inconsistent, or otherwise invalid items. This paper considers the effect on the estimates of variance from treating imputed values as if they were reported, and compares strategies to address it.

Data processing for many of the economic surveys conducted by the Census Bureau has been moved onto a generalized system called the Standard Economic Processing System (StEPS). Methods for estimating variances available in many systems, including StEPS, treat all processed data as if they were reported, ignoring the fact that the imputed values were not observed. The result is that this "naïve" estimator of variance typically is biased; often it underestimates the true variance.

Under StEPS, a survey is not restricted to one method of imputation. Some surveys apply a primary method, and, if the variable(s) required is not available, they revert to a second or even a third method. We call this multi-phased procedure *mixed imputation*. We have not seen it addressed often in the literature beyond Shao and Steel (1999) and Full (2000).

Our goal is to develop the capability in StEPS to obtain an approximate estimate of variance that takes into account the component due to the imputation of missing or invalid values. However, in producing this variance estimate, several considerations are to be balanced: (1) the accuracy of the resulting variance estimates, (2) the ability to generalize the procedure to various types of imputation, (3) the procedure's robustness to the use of mixed imputation, and (4) the ease of implementing the procedure within StEPS.

Weighing these constraints, we considered several procedures discussed in the literature, and concentrate on two of them. The first is a simple procedure that inflates the naïve variance estimate by a factor that depends on the amount and type of imputation. Under the second procedure (Kim 2001), we create "pseudo-data," a second set of responses perturbed enough that commonly applied variance estimation formulae or software will pick up the variability caused by the imputation.

Section 2 contains a brief review of the literature and the methods. Some options for imputation under StEPS are described in Section 3. In Sections 4 and 5, respectively, we explore the inflation-factor approach and Kim's method. Finally, in Section 6 we provide results of a simulation study.

Many derivations, citations, and other details are left out of this paper, but can be obtained from the author in a separate technical report.

## 2. Brief Review of Relevant Literature

In recent years continual advances have been made in estimating the variance of an estimator that accounts for the imputation of missing values. Rubin (1978) introduced the method of multiple imputation as a way of representing the variability created by the imputation.

For the case of hot-deck imputation, Rao and Shao (1992) presented a method that allows one to apply the usual jackknife variance procedure to approximate the result. Their "adjusted jackknife variance estimator" uses only one imputation, rather than several; it requires making a slight adjustment in each of the imputed values whenever a responding unit is the one that is deleted in the jackknife computations.

Directly related to the developments in this paper are Korn and Graubard (1999) and Kim (2001). In the former, the authors apply a factor to inflate the estimate of the variance derived from the reported and imputed values. They consider mainly the cases of simple mean and hot-deck imputation based on one or more imputation cells. Kim proposes a procedure whereby missing values are imputed, then the observed values are perturbed sufficiently so that the resulting set of processed values correctly approximates the variability of the underlying

---

[1] This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau publications. This report is released to inform interested parties of research and to encourage discussion.

population. With these two approaches, the existing software for variance estimation could be used. The inflation factor approach has several issues to resolve.

## 3. Imputation Under StEPS

The Census Bureau's Standard Economic Processing System (StEPS) allows for a variety of imputation methods to be applied across surveys or even within the same survey (Luery 2001). Some of the more important types are the following. Each is accompanied here by an example from an annual survey; the target value, $y_{t,i}$, is sales in the year 2000.

1. Ratio (of identicals): $$\hat{y}_{t,i} = x_i \cdot \frac{\sum_{resp} y_{t,j}}{\sum_{resp} x_j} \qquad (3.1)$$

The "identicals" are the units for which information is available at two times or from two sources, the second typically being the current survey of interest. They provide an estimate of the ratio of change from the first source to the second. For example, one might compute the ratio of change in sales from 1999 to 2000 based on units that respond to the survey each time. This ratio is applied to 1999 sales from the nonresponding unit.

2. Regression: $$\hat{y}_{t,i} = \beta_0 + \beta_1 x_{1,i} + \ldots + \beta_m x_{m,i} \qquad (3.2)$$

Based on respondents to the survey, one or more known variables, such as the value of inventories in 1999 or "census equivalent sales" are used to fit a regression model for sales in 2000. The known variables can be from current or past periods. The unknown y value is then predicted by the model.

3. "Within-unit": $$\hat{y}_{t,i} = x_i^{/}, \text{ e.g.,} = y_{t-h,i} \cdot \frac{x_{t,i}}{x_{t-h,i}} \qquad (3.3)$$

Some variable, $x_i^{/}$, known for the nonresponding unit i, is substituted for the unknown value $y_{t,i}$. For example, $x_i^{/}$ may be the product of the unit's sales in 1999 and the change (ratio) in payroll from 1999 to 2000. The payroll information might be available from other sources, such as administrative (tax) records. The label "within unit" is given to indicate that all the information for the imputation arises from the unit itself; no summary statistics, such as means or ratios from the responding units, are used.

StEPS can accommodate other types of imputation: donor methods, such as hot-deck and cold-deck imputation, and taking the mean from a cell defined, perhaps, by strata or other known information. Surveys can also specify their own method of imputation.

## 4. Applying an Inflation Factor

To estimate the true variance of an estimator when values are imputed, we try to multiply the usual variance estimator that ignores imputation by a factor that depends on the amount and type of imputation. The amount can be measured in several ways. In what follows, we'll use the unweighted nonresponse rate. One may wish to incorporate the sampling weights. To describe our strategy, we define two expectations:

V1: The *expected value* of the "naïve" variance estimator, that is, the estimator that treats all imputed values as if they were reported and valid, conditional on the amount of imputation.

V2: The true variance of the estimator in the presence of imputation, again, conditional on the amount of imputation.

When the applied variance estimator is unbiased, it is often the case that V1 $\leq$ V2, with equality only when there is 100% response. The plan then is to determine or approximate the value of the inflation factor V2/V1 for a specific type of imputation as a function of the amount of imputation. This factor is then applied to the naïve variance estimate, as estimated by a variance formula or variance estimation software.

Several problems can arise with this approach. In practice, there may be many cells or groups for imputation, making the procedure complex. As we see below, the inflation factor may be a function of population parameters, such as a correlation or coefficients of variation, that must be estimated. Further, with a certainty stratum, that is, one whose units are all selected with probability 1, V1 is equal to 0 if the variance estimator applied incorporates the finite population correction (fpc). Strategies to deal with this are not covered in this paper.

To apply this approach, we consider a population, U, of size N. To estimate the total of some variable Y for the population, we take a sample, S, of size n. In this paper, we assume the sample is a simple random sample taken without replacement (srswor). In the sample, for the item y there are r respondents and n-r nonrespondents. These sets are represented by R and NR, respectively, with $R \cup NR = S$.

Suppose further that all units in the sample are equally likely to respond. Although this response mechanism is unrealistic, it is assumed in much of the research on the effects of imputation (Rao and Shao 1992, Korn and Graubard 1999, Kim 2001)--at least within a specifed stratum or cell--and presents a starting point for this method.

### 4.1 Regression Imputation

Often we have auxiliary information available to use in the imputation in the form of a vector of variables **X** known for all units in the sample, perhaps for all members in the population. To demonstrate the inflation method here, we assume that there is only one X variable; the method can easily be extended to cover multivariate **X**. We define $\bar{x}_R = (1/r)\,\Sigma_R\, x_i$, and $\bar{x}_S = (1/n)\,\Sigma_S\, x_i$, as the means of x from only the respondents and from the entire sample, respectively; and

$$b \;=\; \frac{s_{xy,R}}{s^2_{x,R}} \;=\; \frac{\sum_R (y_i - \bar{y}_R)(x_i - \bar{x}_R)}{\sum_R (x_i - \bar{x}_R)^2}\,, \qquad (4.1)$$

the usual sample estimate of the least squares regression coefficient. We define the processed values as

$$y_i^I \;=\; \begin{cases} y_i\,, & \text{if } i \in R \\ \hat{y}_i \;=\; \bar{y}_R \;+\; b\,(x_i - \bar{x}_R), & \text{if } i \in NR \end{cases} \qquad (4.2)$$

The simple weighted estimator in the presence of regression imputation can be written as

$$\hat{Y}_{REGR} \;=\; \Sigma_S\,(N/n) y_i^I \;=\; N\,(\bar{y}_R + b\,(\bar{x}_S - \bar{x}_R)) \quad (4.3)$$

To determine V1, we evaluate the variance, $s_I^2$, of the processed values, $y_i^I$, for all i in the sample. It can be shown that, for a fixed set of population X values, and conditional on the number of nonrespondents,

$$E\,(\,(n-1)\,s_I^2\,) \;=\; E\,(\,\Sigma_S\,(y_i^I - \bar{y}^{\,I})^2$$

$$=\; (\,(r-1)(1-\rho^2) + (n-1)\rho^2)\,S_{y,U}^2\,, \qquad (4.4)$$

where $\bar{y}^{\,I}$ is the mean of the $y_i^I$'s, and ρ is the population correlation coefficient between X and Y. Ignoring fpc,

$$V1 \;\approx\; N^2\,\{\frac{r-1}{n-1}(1-\rho^2) + \rho^2\}\,\frac{S^2_{y,U}}{n}\,. \qquad (4.5)$$

Cochran (1977, p. 340) derives the true variance as

$$V2 \;\approx\; N^2\,\{\frac{n}{r}(1-\rho^2) + \rho^2\}\,\frac{S^2_{y,U}}{n}\,. \qquad (4.6)$$

To inflate the naïve variance estimate to better measure the true variance, we apply the ratio

$$\frac{V2}{V1} \;\approx\; \frac{n(n-1)(1-\rho^2) + r(n-1)\rho^2}{r(r-1)(1-\rho^2) + r(n-1)\rho^2}\,. \qquad (4.7)$$

In practice, ρ must be estimated. Current or past data are usually sufficient for this purpose.

### 4.2 Ratio Imputation

Let w be the ratio of the sample totals or means of Y and X from the respondents, $w = \bar{y}_R \,/\, \bar{x}_R$, and let W be the population ratio. (W and w are used in place of R and r to avoid confusion with the set of respondents and its size.) The processed variable is now

$$y_i^I \;=\; \begin{cases} y_i\,, & \text{if } i \in R \\ \hat{y}_i \;=\; w\,x_i \;=\; \dfrac{\bar{y}_R}{\bar{x}_R}\,x_i\,, & \text{if } i \in NR \end{cases} \qquad (4.8)$$

With ratio imputation, the simple weighted estimator, $\hat{Y}_{RATIO}$, is $N\,w\,\bar{x}_S \;=\; N(\bar{y}_R/\bar{x}_R)\,\bar{x}_S$. For a fixed set of population X values, conditional on the number of nonrespondents, assuming r is large enough and that $w^2$ and $s_{x,R}^2$ are approximately independent, and ignoring the fpc and terms that are of order 1/r times those that remain, we can show that

$$V1 \;\approx\; N^2\,\frac{1}{n}\,\{\,\frac{r-1}{n-1}\,S_{y,U}^2 \;+\; \frac{n-r}{n-1}\,W^2\,S_{x,U}^2\,\} \qquad (4.9)$$

Again, the true variance is found in Cochran (1977, p. 344):

$$V2 \;\approx\; \frac{N^2}{n}\,\{\frac{n}{r}S_{y,U}^2 \;+\; \frac{n-r}{r}[W^2 S_{x,U}^2 \;-\; 2WS_{xy,U}\,]\} \qquad (4.10)$$

where $S_{xy,U}$ is the population covariance term between y and x. Writing W as $\bar{Y}_U/\bar{X}_U$, and expressing the coefficient of variation of the variable x, $S_{x,U}/\bar{X}$, as λ times that of y, the inflation ratio becomes

$$\frac{V2}{V1} \;\approx\; \frac{n(n-1) + (n-1)(n-r)(\lambda^2 - 2\rho\lambda)}{r(r-1) + r(n-r)\lambda^2}\,. \qquad (4.11)$$

As before, ρ and λ must be estimated from the data.

### 4.3 Within-Unit Imputation

When applying within-unit imputation, a missing $y_i$ value is merely replaced in the estimator by some variable $x_i'$ obtained only from unit i. If the expected value of the imputed $x_i'$'s is different from that of the missing $y_i$'s, the estimator for Y, $\hat{Y}_{WIU}$, is typically biased. But, more to our interest in this paper, in that case the naïve variance estimator may *overestimate* the variance. This can occur because the imputation does not use summary statistics from the responding units, statistics that tend to draw the imputed values toward the center of the observed y values or toward some regression line or curve.

For simplicity of notation, we will drop the prime on the $x_i'$ values. The processed values under within-unit imputation are $y_i^I = y_i$, if $i \in R$, or $x_i$, if $i \in NR$.

Then $\hat{Y}_{WIU} = \Sigma_S (N/n)\, y_i^I = (N/n)\, \{\Sigma_R\, y_i + \Sigma_{NR}\, x_i\}$

$$= (N/n)\,\{\, r\,\bar{y}_R + (n-r)\,\bar{x}_{NR}\} \tag{4.12}$$

Similar to the prior development, it can be shown that

$$E(s_I^2) = \left(\frac{r}{n}\right)\left(1 - \frac{n-r}{(n-1)N}\right) S^2_{y,U} + \left(\frac{n-r}{n}\right)\left(1 - \frac{r}{(n-1)N}\right) S^2_{x,U} + \frac{2}{(n-1)N} S_{xy,U} + \frac{r(n-r)}{n(n-1)} (\bar{Y} - \bar{X})^2 \tag{4.13}$$

If $\bar{Y}$ and $\bar{X}$ are not about the same, the final term here could be nontrivial. In that case, one may want to reconsider the use of X as an imputation variable for Y. But if $\bar{Y} \approx \bar{X}$ and we can ignore terms of the same order as the sampling fraction, we can write V1 as

$$V1 \approx \frac{N^2}{n^2}\,[\, r\,S^2_{y,U} + (n-r)\,S^2_{x,U}\,] \tag{4.14}$$

The true variance of $\hat{Y}_{WIU}$ does not depend on the difference between the population means, $\bar{Y}$ and $\bar{X}$:

$$V2 = \frac{N^2}{n^2}\,\{\, r\left(1 - \frac{r}{N}\right) S^2_{y,U} + (n-r)\left(1 - \frac{n-r}{N}\right) S^2_{x,U} - 2\,\frac{r(n-r)}{N}\, S_{xy,U}\} \tag{4.15}$$

If we drop terms of the same order as the sampling fraction, it is easy to see that V1 and V2 are about the same. In this situation, the naïve variance estimate is approximately unbiasd; there is no need to inflate it.

## 5. Applying Kim's Method

The approach taken by Kim (2001) is to create a second set of data that reflects the added variability due to the use of imputation. The advantage over other procedures mentioned in Section 2 is that this set of "pseudo-data" can then be put through the existing variance estimation software to produce a valid estimate of the variance; no extra steps are required within the variance estimation. It should be mentioned that Shao and Steel (1999) also proposed such a procedure in the case of ratio estimation.

As an example, let us return to regression imputation with a set of auxiliary variables, **X**. The regression model for $y_i$ is $\hat{y}_i = \bar{y}_R + b\,(x_i - \bar{x}_R)$. The estimator for Y was given in equation (4.3), and its variance in (4.6). An estimator for this variance based on regression theory can be given as

$$\hat{V}_{REGR} = \frac{1}{n(n-1)}\,\Sigma_S\,(\hat{y}_i - \bar{y}^{\,I})^2 + \frac{1}{r(r-2)}\,\Sigma_R\,(y_i - \hat{y}_i)^2 \tag{5.1}$$

The first sum represents the variability among the modeled y's, $\hat{y}_i$, while the second sum measures the error in the regression model. Kim shows that this variance can be rewritten simply as

$$\frac{1}{n}\,s^{2*} = \frac{1}{n(n-1)}\,\Sigma_S\,(y_i^* - \bar{y}^{\,I})^2\,, \tag{5.2}$$

where the $y_i^*$'s are the "pseudo-data" defined as

$$y_i^* = \begin{cases} c_{REGR}\,y_i + (1 - c_{REGR})\,\hat{y}_i\,, & \text{if } i \in R \\ \hat{y}_i\,, & \text{if } i \in NR \end{cases}, \tag{5.3}$$

with $c_{REGR} = (n(n\text{-}1)/(r(r\text{-}2))^{1/2}$.

Kim shows that this value of $c_{REGR}$ is the appropriate factor to introduce the proper amount of perturbation of the reported values under regression imputation. In his paper, he considers other situations and sampling schemes, and derives other perturbation factors, c. It is not clear that this strategy produces such a close approximation to the true variance under all types of imputation. However, among other results, he considers complex sampling designs, and shows that the standard jackknife variance estimator applied to the pseudo-data is approximately unbiased for the variance.

## 6. A Simulation Study

We conducted a simulation study to measure the performance of several variance estimators under three types of imputation and mixtures of the three. The goal was to estimate the total retail sales in the year 2000 from new auto dealers from a frame of units constructed as follows.

*The Frame and Sample.* We obtained a subset of the file of sample units from the 2000 Annual Retail Trade Survey, conducted by the Bureau of the Census, in the North American Industry Classification System (NAICS) code 441110, new auto dealers. From this file, we extracted all units with a sampling weight greater than 10, that is, the smaller firms, as determined by their measure of size in the 1997 Census of Retail Trade. After deleting units with specific fields missing, there were N = 779 dealers, which were treated as the frame. From this frame, we drew simple random samples of size n = 40 without replacement (srswor). The target variable, Y, is sales in 2000.

*The Imputation Procedures.* Each unit selected into the sample was given a 70% chance of responding, independently from unit to unit. If the unit was a nonrespondent, we applied one of three imputation procedures, as listed in Table 1. The three procedures are the same as those described in Section 3, using the same variables. To apply the inflation methods more realistically, when computing the parameters ρ and λ, we used data from the prior (available) period. For example, for the ratio of identicals, ρ, the correlation between sales in 2000 and sales in 1999, was estimated using data from 1999 and 1998 (where ρ = .88).

**Table 1. Imputation Variables and Parameters Used in the Simulation**

| Type of Imputation | Imputation Variable | True Correl. with Y | Used for Inflation Method |
|---|---|---|---|
| 1. Ratio of identicals | Sales in 1999 | ρ = .89<br>λ = 1.08 | ρ = .88<br>λ = 1.02 |
| 2. Regression | Inventories in 1999 | ρ = .78 | ρ = .77 |
| 3. Within-Unit | Sales in 1999 × Payroll in 2000/1999 | ρ = .95 | --- |

*The Mix of Imputation.* To evaluate the procedures under mixed imputation, a nonrespondent's imputed value was determined randomly, according to one of the three imputation procedures. For each simulation (row) in Tables 2 and 3, the imputation mix is denoted by (a, b, c), where a, b, and c represent the probabilities of using procedures 1, 2, and 3 (as denoted in Table 1), respectively. In the first three simulations (rows of the tables), all the imputation was completed using a single method. For surveys using mixed imputation, a realistic mix might be something like (.75, .20, .05), where the variable required for the primary imputation methods is missing 25% of the time, forcing one to use a second or third procedure. To help show what can happen under mixed imputation, we looked at more extreme mixtures, such as (.50, .25, .25).

*The Variance Estimation Procedures.* Several methods are compared: the naïve variance estimator; Kim's estimator that perturbs the data for ratio (Kim 1) or regression (Kim 2) imputation; and the inflation estimator developed for ratio (Inflation 1) or regression (Inflation 2) imputation.

*Results of the Simulation.* Results are provided in Tables 2 and 3. In the tables, each row represents 10,000 simulations of srswor of size n = 40 from a frame of size N = 779. The bias of the variance estimation methods is addressed in Table 2. Each entry in the table is the ratio of the expected value of the variance estimator divided by the "true" variance under the imputation procedure. Here, the expectation is approximated by taking the average of the variance estimate over the simulations, while the "true" variance is approximated by taking the sample variance of the 10,000 estimates of $\hat{Y}$. A ratio closer to 1 implies a smaller bias.

In Table 3, each entry is the ratio of the root mean squared error (RMSE) of the given method divided by that of the naïve variance estimator. Thus, better performance is indicated by a smaller entry.

**Table 2. The Bias of the Variance Procedures**

$$\text{Table entry} = \frac{E\,(\text{variance estimator})}{\text{"true" variance}}$$

| Actual Imputation | naïve var. estm. | Kim method 1 | Kim method 2 | Inflation method 1 | Inflation method 2 |
|---|---|---|---|---|---|
| (1, 0, 0) | 0.86 | 1.00 | 1.08 | 0.94 | 1.17 |
| (0, 1, 0) | 0.81 | 0.93 | 1.00 | 0.88 | 1.09 |
| (0, 0, 1) | 0.99 | 1.13 | 1.22 | 1.08 | 1.34 |
| Mixed: | | | | | |
| (.50, .25, .25) | 0.89 | 1.03 | 1.11 | 0.98 | 1.21 |
| (.25, .50, .25) | 0.87 | 1.00 | 1.08 | 0.96 | 1.18 |
| (.25, .25, .50) | 0.92 | 1.06 | 1.14 | 1.01 | 1.25 |
| (1/3, 1/3, 1/3) | 0.90 | 1.03 | 1.11 | 0.98 | 1.22 |

*Observations.* From Table 2, it is clear that, under the circumstances of the simulation, the naïve variance estimator has a downward bias under imputation when using only a ratio of identicals (13.7%) or only regression imputation (19.2%). Kim's procedures for ratio (Kim 1) and regression imputation (Kim 2) eliminate the bias effectively. The inflation procedures are less successful. With ratio imputation, Inflation 1 doesn't go far enough, leaving a downward bias of 5.5%; with regression imputation, Inflation 2 goes too far, producing an upward bias of 9.3%.

When within-unit imputation is used exclusively, the naïve variance estimator is essentially unbiased. This is not surprising, as no summary statistics from the sample respondents are used to drive the imputed values to some central point or line; only information from the nonresponding unit is used. As expected, Kim's procedures and the inflation method do not work well here, leaving a large upward bias. They were developed for different types of imputation.

**Table 3. The Root Mean Squared Error**

Table entry = $\dfrac{\sqrt{\text{MSE of variance estimator}}}{\sqrt{\text{MSE of naïve variance}}}$

| Actual Imputation | naïve var. estm. | Kim method 1 | Kim method 2 | Inflation method 1 | Inflation method 2 |
|---|---|---|---|---|---|
| (1, 0, 0) | 1 | 0.99 | 1.10 | 0.91 | 1.37 |
| (0, 1, 0) | 1 | 0.90 | 0.90 | 0.92 | 1.12 |
| (0, 0, 1) | 1 | 1.37 | 1.65 | 1.19 | 2.20 |
| Mixed: | | | | | |
| (.50, .25, .25) | 1 | 1.08 | 1.22 | 0.98 | 1.59 |
| (.25, .50, .25) | 1 | 1.01 | 1.11 | 0.96 | 1.45 |
| (.25, .25, .50) | 1 | 1.16 | 1.35 | 1.05 | 1.78 |
| (1/3, 1/3, 1/3) | 1 | 1.09 | 1.23 | 1.00 | 1.61 |

When variance is added to the evaluation, the RMSE (Table 3) provides a somewhat different picture. The naïve estimator, although biased downwards under ratio and regression imputation, still has the smallest variance. This is to be expected; the other procedures eliminate the bias by perturbing the data or by multiplying by a factor greater than 1. Combining bias and variance, the RMSE under ratio imputation is about the same for the naïve estimator and the Kim 1 (ratio) method. Inflation 1, however, has the smallest RMSE with ratio imputation, about a 9% improvement over the others. With regression imputation, Kim 1 and 2, and Inflation 1 have a RMSE about 8% to 10% below that of the naïve estimator. Inflation 2, between overcompensating for the bias and having a larger variance, performs poorly.

Under mixed imputation, when the frequency of the second (or third) imputation procedure is no longer negligible, one hopes that the variance estimator works well with each type of imputation. Among the methods studied under the circumstances here, the Kim 1 procedure appears most robust for eliminating the bias, while Inflation 1 delivers the smallest RMSE provided the proportion of within-unit imputation is small. If the latter is large, the naïve estimator may have the smallest RMSE.

*Drawbacks of the Study.* It is appropriate to mention some drawbacks of this simulation study. (i) The sample design is srswor. Results under stratified or probability proportional to estimated size sampling may be different. (ii) The response mechanism is over-simplified and somewhat unrealistic. (iii) The sampling fraction used here is very small. Although that is common, it doesn't permit an evaluation of the methods with smaller strata and larger sampling fractions.

*Current and Future Work.* We are preparing simulations under more complex sample designs, under more complex types of response mechanisms, and using strata of different sizes and sampling fractions (including certainty sampling). We have also extended the inflation procedure to accommodate within-unit imputation, and to produce more robust factors under mixtures.

## Acknowledgment

The author thanks Jae-Kwang Kim for helpful discussions on his procedure, Bill Davie for providing the data set used in the simulation, and Paul Siegel and Jeff Stratton for providing very helpful comments.

## References

Cochran, W.G. (1977). *Sampling Techniques,* 3rd Ed., J. Wiley, New York, N.Y.

Full, S. "A Comparison of Jackknife and Bootstrap Methods for Variance Estimation in the Presence of Imputation in ONS Business Surveys." Proceedings of the Second International Conference on Establishment Surveys, 1423-1428.

Kim, J.-K. (2001). "Variance Estimation After Imputation." *Survey Methodology,* **27**, No. 1, 75-83.

Korn, E.L., and Graubard, B.I. (1999). *Analysis of Health Surveys.* J. Wiley, New York, N.Y.

Luery, D. (2001). "General Imputation." Unpublished documentation for the Standard Economic Processing System, Economic Statistical Methods and Programming Division. Bureau of the Census, Washington, DC.

Rao, J.N.K. and Shao, J. (1992). "Jackknife Variance Estimation With Survey Data Under Hot Deck Imputation." *Biometrika,* **79**, 4, 811-822.

Rubin, D. (1978). Multiple Imputation in Sample Survey--a Phenomenological Bayesian Approach to Nonresponse." *Proceedings of the Section on Survey Research Methods, Am. Stat. Assoc.,* 20-34.

Shao, J. and Steel, P. (1999). "Variance Estimation for Survey Data with Composite Imputation and Nonnegligible Sampling Fractions." *Journal of the Am. Stat. Assoc.,* **94**, 254-265.