# Estimating the Variance Due to Hot Deck Imputation for Product Value Estimates in the 2017 Economic Census

Katherine J. Thompson[1], Matthew Thompson[2]
[1,2]Economic Statistical Methods Division, U.S. Census Bureau, Washington DC 20233

**Abstract**
The Economic Census collects data on the revenue obtained from products from all sampled units. Product data collection is quite challenging; establishments can report values from a long list of potential products in a given industry. Moreover, product descriptions are quite detailed, many products are mutually exclusive, and reported products are subjected to strict additivity constraints. Consequently, legitimate missing values occur frequently and nonresponse is quite high. Auxiliary data are not available, and other predictors such as total receipts are often weakly related. In the 2017 Economic Census, missing product data will be imputed using hot deck imputation, and variance estimates for product data will be published for the first time. The variance estimator must account for sampling variance, calibration weighting, and imputation variance. Thompson, Thompson, and Kurec (2016) present results of a simulation study that examines the first two factors. We focus on the estimation of the imputation variance component. Using a simulation study, we compare the statistical properties of this component estimated under both model-assisted and Bayesian frameworks.

**Key Words**: hot deck, nearest neighbor, variance estimator

## 1. Introduction

The Economic Census is the U.S. Government's official five-year measure of American business and the economy. The term “Economic Census” is a bit of a misnomer; the majority of sectors samples the small single-unit establishments and surveys all of the multi-unit establishments[2]. The Economic Census collects a core set of data items from each establishment called general statistics items: examples include total receipts or value of shipments (“receipts”), annual payroll and number of employees in the first quarter. In addition, the Economic Census collects data on the revenue obtained from product sales (hereafter referred to as “products”). With the exception of the construction sector, all sectors construct a *complete universe of general statistics values* by using administrative data in place of respondent data for unsampled units. However, product data are collected from only the sampled establishments. In most sectors, weighted sample estimates are further calibrated to the industry totals for receipts.

In the 2017 Economic Census, missing product data will be imputed using hot deck imputation (Thompson and Liu 2015; Knutson and Martin 2015), and variance estimates for product totals will be published for the first time. Depending on the industry, random

[1]Any views expressed are those of the author(s) and not necessarily those of the U.S. Census Bureau.

[2] A single-unit (SU) establishment owns or operates a business at a single location, whereas multi-unit (MU) establishments comprise two or more establishments that are owned or operated by the same company.

hot deck or nearest neighbor hot deck imputation will be implemented (Tolliver and Bechtel 2015; Bechtel, Morris, and Thompson 2015). The variance estimator must account for sampling variance, calibration weighting, and imputation variance. Thompson, Thompson, and Kurec (2016) present results of a simulation study that examines the first two factors. In this paper, we focus solely on the estimation of the imputation variance component for product totals.

Product data collection is challenging. The Economic Census collects information on over 8,000 different products defined by the North American Product Classification System (NAPCS); see https://www.census.gov/eos/www/napcs/more.html. However, many products are rarely reported. Establishments can report values from a long list of potential products in a given industry (some lists span more than 50 potential products), and consequently, many establishments choose not to report any product data (complete product nonresponse). These lists vary by industry and can in fact differ within broader sector[3]. Furthermore, product descriptions are quite detailed and some products are mutually exclusive. Finally, all reported product values within a given establishment are expected to sum to the total receipts value reported earlier in the questionnaire. Missing product data can occur when an establishment does not respond to the census, when a responding establishment provides no product information, or when a responding establishment provides product information that does not sum within an allowable percentage tolerance to its total receipts.

Product value imputation is even more challenging. There are no auxiliary data sources. Moreover, other predictors such as total receipts are often weakly related. In most industries, the frequently reported products are highly correlated with total receipts and generally make up the majority of the total value of receipts, whereas the remaining products are not. Thus, the best predictors of an establishment's products are the industry assigned to the establishment from the sampling frame (which may change after collection) and the total receipts value (Ellis and Thompson, 2015). Finally, legitimate zero values are expected for the majority of eligible products in an industry, at both the individual establishment and total industry level.

From a variance estimation perspective, most of the challenge lies with the poor predictors and high expected zero rates for many products, although discounting the high nonresponse rate would be very optimistic as the possibility of a low donor-to-recipient ratio for hot deck is quite high. It is not unreasonable to expect to find a variance estimation method that produces estimates with good statistical properties in terms of bias and stability for the well-reported products. It may be unreasonable to hold similar hopes for the remaining products.

[3] Prior to the 2017 Economic Census, a list of products specific to each industry was provided directly on the industry-specific questionnaire along with designated space for product "write-ins." However, the U.S. Census Bureau will implement the North American Product Classification System (NAPCS) in the 2017 Economic Census. NAPCS allows the collection of the same product from different industries, thus facilitating cross-sector product tabulations. In the upcoming census, data collection will be electronic, and the respondents will have greater flexibility in designating their products due to the instrument design and the new product classification system.

Using a simulation study, Thompson, Thompson, and Kurec (2016) explore two different approaches to estimating product data variances under complete response with calibration: a design-based replication approach and a Bayesian (model-based) approach. Likewise, we study the statistical properties of the nonresponse variance (imputation) component estimated under a design-based (model-assisted) framework and under a model-based framework. This is the first step in our variance estimation project. The final recommendation will combine the results from both studies in a simulation "cook-off" that compares the fully assembled variance estimators, simultaneously accounting for sampling variance, calibration weighting, and imputation variance. For now, the objective is to find viable estimators under either framework, where viable refers to statistical properties over repeated samples and feasible processing resources/time.

Section 2 describes the simulation study data and design. Section 3 introduces the studied imputation and variance estimation methodologies. In Section 4, we present results on the properties of the imputed product estimates with both hot deck random and hot deck nearest neighbor imputation for both single and multiple-imputed estimates. Section 5 presents the results of the variance estimation evaluation. We conclude in Section 6 with a few general comments and a brief outline of the next steps in our study.

## 2. Study Data and Simulation Approach

To compare alternative imputation and variance estimation methods on the same outcome variables over repeated samples, an accepted practice is to (1) Create a realistic population (complete response); (2) Induce nonresponse and apply the considered imputation methods to the selected outcome variables in each replicate; and (3) Compute the pre-determined evaluation criteria and compare the results.

Nordholt (1998) and Charlton (2004) provide examples of excellent large-scale applications. We found this first step to be the most challenging. Subject matter and classification experts provided microdata from the 2012 Economic Census from a set of North American Industry Classification System (NAICS) industries whose eligible products were expected to remain consistent under the introduction of NAPCS; hereafter, we refer to these data as our test industries. The studied industries cover 11 sectors, excluding construction. These industries are not meant to be representative of the entire Economic Census. This is the same set of test industries used to evaluate imputation methods on products (Thompson and Liu 2015; Ellis and Thompson 2015; Tolliver and Bechtel 2015; Bechtel, Morris, and Thompson 2015; Knutson and Martin 2015).

The test industries' microdata are subject to complete product nonresponse and contain only sampled units. Using the donor record criteria provided by the subject matter experts, we filled in missing product values using nearest neighbor hot deck imputation[4] for the sampled cases. Then, we used the SIMDAT algorithm (Thompson 2000) to create completed records for the unsampled single-unit establishments in each industry. This nonparametric "nearest neighbors" simulation technique creates simulated data with the

[4] The simulation study presented by Knutson and Martin (2015) created four different "completed" populations per industry: one with ratio imputation, one with random hot deck imputation, one with nearest neighbor imputation, and one with sequential regression multivariate imputation. However, the earlier study found no correlation between the "best" method used to complete the missing data induced in the population and the method that was originally used to create the population. The imputation bias results presented in Section 4 are similar.

same correlation structure as the sample survey (training) data and similar quantile values. To implement the algorithm, we had to limit the number of simulated products in each industry to the four best-reported products in each industry in terms of number of establishments that reported the product plus an "all other product values" item containing the balance of the difference between the establishment total receipts and summed top four products. This final "catch all" product does not resemble the collected data and is excluded from our analyses.

After creating 21 industry populations, we independently induced complete product nonresponse (i.e. no product data reported from establishment) using the logistic regression response propensity models described in Ellis and Thompson (2015) in 5,000 replicates. The propensity models tend to give higher probabilities of product response to multi-unit establishments and to "larger" establishments in terms of total receipts.

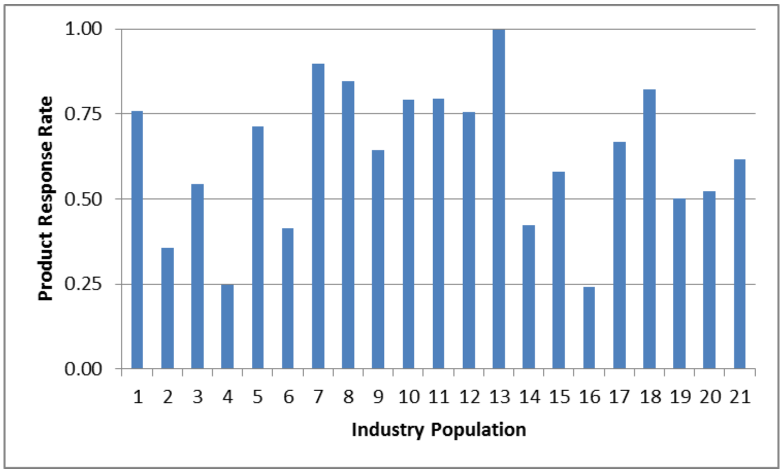


**Figure 1: Averaged Product Response Rates (Reported at Least One Valid Value of a Product) in the Test Industry Populations at the U.S. Level (Source: 2012 Economic Census)**

Figure 1 presents the averaged product response rates (PRR), computed as the ratio of the number of establishments providing at least one valid product value to the total number of eligible establishments in the test industries. Although the median PRR is 0.64, the majority of establishments provide data for one or two reported products. Five of the 21 industries have a PRR that is less than 0.50 i.e., less than half of the establishments in the industry are reporting *any* product values.

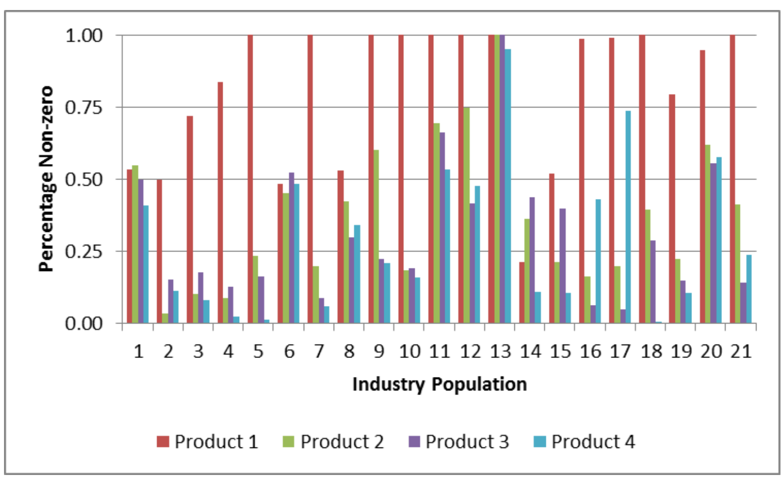


**Figure 2: Percentage of Non-zero Reported Values of Products in the Test Industry Population for U.S. (Source: 2012 Economic Census)**

Figure 2 presents the percentage of positive (non-zero) values reported by product in each industry population. Although the majority of the studied industries have a high incidence of non-zero reported Product 1 values (median =0.99), the other three products do not have similarly high ratios (median values of 0.36, 0.22, and 0.21, respectively).

Product estimates in the Economic Census are computed at the industry by state level. However, the population in these categories can often be quite small and the sample size is greatly reduced after inducing product nonresponse. To obtain stable imputed estimates over repeated samples at this level would require a very large number of replicates for the poorly reported products (3 and 4). As a processing compromise, we produced imputed estimates and variance estimates at the industry by region level, aggregating the region estimates to the industry level to produce industry totals. Hereafter, all presented statistics are provided at the U.S. level by industry; corresponding industry by region measures are available upon demand.

In each replicate, we used random and nearest neighbor imputation to create completely imputed populations under both single and multiple imputation; this is discussed further in Section 3. We obtained empirical bias and MSE values (truth) from the 5,000 replicates as

Relative Bias $$RB\left(\hat{Y}_{ij}^{hm}\right) = \left[\sum_{r=1}^{5,000}\left(\hat{Y}_{ijr}^{hm}/5,000\right)\Big/Y_{ij}\right] - 1 = \left[\frac{\hat{\bar{Y}}_{ij}^{hm}}{Y_{ij}}\right] - 1$$

MSE $$MSE\left(\hat{Y}_{ij}^{hm}\right) = \left[\sum_{r=1}^{5,000}\left(\hat{Y}_{ijr}^{hm} - \hat{\bar{Y}}_{ij}^{hm}\right)\Big/5,000\right] + \left[\hat{\bar{Y}}_{ij}^{hm} - Y_{ij}\right]^2$$

where *i* indexes the industry, *j* indexes the product, *h* indexes the hot deck imputation method, *m* indexes the number of implicates (single or multiple), and *r* indexes the replicate.

In 1,000 of the 5,000 replicates, we obtained variance estimates for both the random and nearest neighbor imputed product values estimates for each of the considered variance estimators discussed in Section 3. We used these 1,000 replicates to assess the relative bias of each considered variance estimator over repeated samples as

$$RBV\left(\, v^{u}\left(\,\hat{Y}_{ij}^{hm}\,\right)\right) = \left[\sum_{r=1}^{1,000} v^{ur}\left(\hat{Y}_{ijr}^{hm}\right)/1,000\Big/MSE\left(\hat{Y}_{ij}^{hm}\right)\right] - 1$$

where $v^{ur}$ indexes the variance estimate obtained using variance estimator $u$ in replicate $r$.

The simulation study is a complete block design with repeated measures, where the industries could be considered random effects. Many of the evaluations presented in Sections 3 and 4 are fixed or mixed ANOVA models, where the outcome is the absolute relative bias of the studied treatment (imputation method, variance estimator). We used the SAS PROC MIXED® procedure (SAS/STAT(R) 9.2 User's Guide, Second Edition) to fit and evaluate the following models:

**Model 1:** $Y_i^l = \tau_l + \epsilon_i, \epsilon_i \sim N(0, \sigma^2)$

**Model 2:** $Y_{ij}^l = \tau_l + \beta_k + \gamma_{lk} + \epsilon_{ij}, \epsilon_i \sim N(0, \sigma^2), \beta_k \sim N\left(\beta, \sigma_\beta^2\right)$

Model 1 is a fixed effect model where the industry estimates are subjects. Model 2 is a mixed model where the industry by region estimates are subjects using the compound symmetry covariance structure (other options failed to yield positive-definite matrices). All F-tests are conducted at $\alpha$=0.05. A significant treatment effect for both Model 1 and 2 is strong evidence of treatment effect. A significant effect for only Model 1 is weak evidence (an indication) of a difference, demonstrating a vulnerability to the select set of populations. On a few occasions, there were significant effects for Model 2, but not Model 1. Although the industry by region-level estimates are much more unstable than their industry counterparts due to small sample sizes, we consider this as strong evidence of a fixed effect difference as well.

## 3. Hot Deck Imputation

Hot deck imputation selects a 'similar' unit from a donor pool and uses its data to impute a group of missing values, thus preserving existing relationships between items. For product data imputation, we select a single donor for each recipient and impute values for all products as $\tilde{y}_{ijk}^{hm} = x_{ik}(y_{ijl}^{d}/x_{il}^{d})$, where $y_{jkj}$ is the value of product $j$ in industry $i$ for *recipient* establishment $k$ and $x_{ik}$ is the unit's value of total receipts (always available). The imputation ratios are obtained from the donor record's corresponding product and total receipts value, thus preserving the establishment level multivariate distributions and ensuring additivity.

In random hot deck imputation, the donor record is randomly selected, usually with replacement (Brick and Kalton 1996). This particular application of hot deck is optimal when both the response propensities and the expected value(s) of the variable(s) of interest are homogenous within an imputation cell (Andridge and Little 2010). Due to the random selection, this method "preserves the distributional properties of the imputed dataset; that is, the distribution function for imputed data within a cell differs from the distribution function for the respondents in the cell only because of the randomness of imputation" (Kim and Fuller 2004).

Nearest neighbor hot deck imputation uses a distance measure to select the donor. The distance measure can be a function of one or more auxiliary variables and can have several functional forms. Distance measures based on a variable (or variables) available for all units are calculated for all donors compared to all recipients. The donor that is closest to a particular recipient within an imputation cell is selected, with a donor randomly selected in the event of a tie. This hot deck method is optimal when the variable(s) used for the distance measure is highly correlated with the variable(s) of interest and the response propensities are homogenous within an imputation cell. Nearest neighbor imputation is deterministic imputation and does not have the same asymptotic properties as random imputation. That said, Chen and Shao (2001) describe many advantages of the nearest neighbor imputes including: (1) reasonable values with little or no chance of "nonsensical" imputes; (2) asymptotically unbiased and consistent estimators of population means and quantiles; and (3) employment of a robust nonparametric model that relates outcome (to be predicted) to matching variables. Nearest neighbor hot deck imputation is especially attractive for business surveys with skewed populations, as it guards against selecting a donor record with a very different variable distribution (assuming that reporting patterns are correlated with size of business). We use the absolute difference between donor receipts value and recipient receipts value. In the event of a donor tie, we randomly choose a donor.

Cell collapsing is necessary when there are insufficient donors in an imputation cell (our cell minimum was five). We use similar imputation cells as the 2017 Economic Census. The finest level was industry /unit type (Single or Multi)/region, where the industry code could incorporate a further sub-classification by legal form of operation (LFO) or type of operation, depending on the sector. In 2017, the Economic Census will use state instead of region to be consistent with the sample design and publication requirements. If fewer than five donors are available in a cell, we dropped the region classification, then the unit type. Collapsing occurs very rarely in our applications.

### 3.1 Model Assisted Variance Estimator

Beaumont and Bocci (2009) propose a model-assisted variance estimator that can be used with either random or nearest neighbor imputation, "as long as the donor identification is available on the data file for each recipient." Their estimator comprises two components: an imputation variance component and a sampling variance component. Since our simulation does not incorporate sampling, we do not estimate either the sampling variance component or the covariance component.

For a given industry *i*, let the population total for product $y_j$ in domain *d* be given by $T_{dy_j} = \sum_{k \in i} d_k y_{ijk}$ where *k* indexes the establishment. There are $n_i$ eligible establishments in industry *i*, of which $n_{ri}$ respond with eligible product data. The donor set is indicated by $s_r$ and the recipient set is indicated by $s_m$. As this is a census, all units are included with certainty. The imputed estimator for product $y_j$ in domain *d* is

$$\hat{T}^I_{dy_j} = \sum_{k \in s_r} d_k y_{ijk} + \sum_{k \in s_m} d_k x_{ik} (y_{ijl(k)} / x_{il(k)})$$

where $l(k) \in s_r$ is the donor used to impute the recipient *k* and $x_{ij}$ is the value of total receipts. This can be rewritten in linear form as

$$\hat{T}^I_{dy_j} = \sum_{k \in s_r} \frac{W_{idk}}{x_{ik}} y_{ijk}$$

where $W_{idk} = d_k x_{ik} + \sum_{p \in s_{m,k}} d_k x_{ip}$ and $s_{m,k} = \{p\text{: } p \in s_r \text{ and } l(p) = k\}$. Since $x_{ik}$, the value of total receipts, is known for all establishments in industry *i* ,the expression is in linear form. Again, we are only estimating the nonresponse (imputation) variance component for this study. So, we decompose the total error as $\hat{T}^I_{dy_j} - T_{dy_j} = \sum_{k \in s_r} \big((W_{idk} - d_k x_{ik}) y_{ijk}\big) / x_{ik} - \sum_{k \in s_m} d_k y_{ijk}$, i.e., the difference between the missing values that were imputed using the (adjusted) hot deck method and the true values in the population. This can be rewritten as $\hat{T}^I_{dy_j} - T_{dy_j} = \sum_{k \in s_r} \big((\sum_{p \in s_{m,k}} d_k x_{ip}) / x_{ip}\big) y_{ijp} - \sum_{k \in s_m} d_k y_{ijk}$. Beaumont and Bocci (2009) propose the following mean squared error (MSE) estimator

$$V_q(\hat{T}^I_{dy_j} - T_{dy_j}) = \sum_{k \in s_r} ((W_{idk} - d_k x_{ik}) / x_{ik})^2 \, \hat{\sigma}^2_{ijk} - \sum_{k \in s_m} d_k \, \hat{\sigma}^2_{ijk}$$

where $\hat{\sigma}^2_{ijk}$ in an estimator of the true variance under imputation model *q:*

$$E_q(y_{ijk}|X) = \mu_{ijk},\ V_q(y_{ijk}|X) = \sigma^2_{ijk}, cov_q(y_{ijk}, y_{ijl}) = 0.$$

Note that the estimator in Beaumont and Bocci (2009) includes a covariance component in the imputation variance estimator that is not necessarily expected to be zero when the computations are applied to a probability sample. However, in the case of a census, this component is exactly equal to zero. In our simulation study, **X** contains indicators for imputation cell variables and total receipts value (available for all establishments).

Following Beaumont and Bocci (2009), we consider two regression models to estimate $\hat{\sigma}_{ijk}^2$:

| | Description | Model | Estimators |
|---|---|---|---|
| PAR1 | Ratio | $y_{ijk} = \beta_{ij}x_{ijk} + \varepsilon_{ijk}, \varepsilon_{ijk} \sim (0, x_{ijk}\sigma_{ij}^2)$ | $\hat{\beta}_{ij} = \sum_{k \epsilon s_r} y_{ijk} / \sum_{k \epsilon s_r} x_{ik}$ <br> $\hat{\sigma}_{ij}^2 = \sum_{k \epsilon s_r} (y_{ijk} - \hat{\beta}_{ij}x_{ik})^2 / \sum_{k \epsilon s_r} x_{ik}$ <br> $\hat{\sigma}_{ijk}^2 = \hat{\sigma}_{ij}^2 x_{ik}$ |
| PAR2 | OLS | $y_{ijk} = \beta_{ij}x_{ijk} + \varepsilon_{ijk}, \ \ \varepsilon_{ijk} \sim (0, \sigma_{ij}^2)$ | $\hat{\beta}_{ij} = \sum_{k \epsilon s_r} y_{ijk}x_{ik} / \sum_{k \epsilon s_r} x_{ik}^2$ <br> $\hat{\sigma}_{ijk}^2 = \sum_{k \epsilon s_r} \left(y_{ijk} - \hat{\beta}_{ij}x_{ik}\right)^2 / (n_r - 1)$ |

The PAR1 model is the ratio estimator model commonly used in sample surveys and frequently used with business surveys. However, total receipts ($x_{ij}$) is generally a very poor predictor of product values for all but the most frequently reported products in a given industry (products 1 and 2 in our application). Unfortunately, as mentioned earlier), this is the only auxiliary variable available for product data imputation (along with industry code, used to define imputation cells). The PAR2 model underlies stratified sample designs. In the Economic Census – and our simulation –the majority of units are included with certainty and the constant mean or variance assumption is unlikely. We did try to fit several other models, such as the penalized spline model used in Beaumont and Bocci (2009) and a one-way ANOVA model with industry as the only covariate (i.e. the cell mean model), but model fits were either comparable or worse.

Table 1 presents the adjusted-$R^2$ values for both no-intercept regression models by population (no product nonresponse) at the U.S. level. The industry *x* region level statistics have a very similar overall pattern , with the best predictive models for product 1 and weak predictions for the less well-reported products (indeed, in some cases no predictive power). Thompson and Ellis (2012) present similar findings as measured by correlation. Often, one model has higher adjusted $R^2$ values for the same item within industry, but (1) there is no clear pattern of when this occurs and we could not find good predictors to explain this phenomenon and (2) the same model did not always work best for different products within the same industry.

Given these generally poor fits (in general) of the PAR1 and PAR2 regression models for three of our four products, it was unlikely that the model assisted variance estimation approach would produce unbiased estimates. On the other hand, the method is certainly easy to program and has been tailored to the hot deck imputation procedures that we are using – and the program runs very quickly. The latter is an advantage in a production setting with such a high volume of data. Consequently, it seemed worth at least examining this approach, although we were only optimistic about its feasibility with products 1 and 2 (and were perhaps naïve in our hopes).

Table 1: Adjusted $R^2$ Values for Variance Estimation Models

| | Product 1 | | Product 2 | | Product 3 | | Product 4 | |
|---|---|---|---|---|---|---|---|---|
| **Industry** | **PAR1** | **PAR2** | **PAR1** | **PAR2** | **PAR1** | **PAR2** | **PAR1** | **PAR2** |
| 1 | 0.42 | 0.47 | 0.69 | 0.81 | 0.60 | 0.72 | 0.22 | 0.39 |
| 2 | 0.41 | 0.53 | 0.30 | 0.45 | 0.07 | 0.02 | 0.09 | 0.30 |
| 3 | 0.80 | 0.87 | 0.16 | 0.13 | 0.06 | 0.02 | 0.07 | 0.06 |
| 4 | 0.88 | 0.96 | 0.08 | 0.02 | 0.03 | 0.01 | 0.09 | 0.09 |
| 5 | 0.97 | 0.98 | 0.52 | 0.61 | 0.20 | 0.33 | 0.02 | 0.02 |
| 6 | 0.52 | 0.40 | 0.46 | 0.33 | 0.17 | 0.30 | 0.17 | 0.22 |
| 7 | 0.98 | 0.98 | 0.05 | 0.06 | 0.03 | 0.04 | 0.02 | 0.02 |
| 8 | 0.48 | 0.79 | 0.25 | 0.11 | 0.18 | 0.05 | 0.16 | 0.05 |
| 9 | 0.96 | 1.00 | 0.17 | 0.01 | 0.05 | 0.00* | 0.04 | 0.00* |
| 10 | 0.86 | 0.86 | 0.33 | 0.39 | 0.37 | 0.56 | 0.21 | 0.20 |
| 11 | 0.99 | 0.99 | 0.20 | 0.21 | 0.31 | 0.34 | 0.10 | 0.11 |
| 12 | 0.97 | 0.98 | 0.58 | 0.66 | 0.16 | 0.10 | 0.30 | 0.48 |
| 13 | 0.94 | 0.95 | 0.96 | 0.96 | 0.95 | 0.95 | 0.91 | 0.91 |
| 14 | 0.25 | 0.55 | 0.27 | 0.12 | 0.33 | 0.25 | 0.11 | 0.13 |
| 15 | 0.55 | 0.44 | 0.17 | 0.37 | 0.21 | 0.14 | 0.09 | 0.09 |
| 16 | 0.61 | 0.23 | 0.39 | 0.82 | 0.09 | 0.06 | 0.19 | 0.11 |
| 17 | 0.98 | 0.98 | 0.19 | 0.22 | 0.17 | 0.60 | 0.18 | 0.06 |
| 18 | 0.94 | 0.94 | 0.40 | 0.32 | 0.28 | 0.29 | 0.00 | 0.00 |
| 19 | 0.69 | 0.58 | 0.16 | 0.07 | 0.08 | 0.04 | 0.05 | 0.03 |
| 20 | 0.61 | 0.74 | 0.31 | 0.63 | 0.22 | 0.35 | 0.24 | 0.29 |
| 21 | 0.97 | 0.98 | 0.31 | 0.24 | 0.05 | 0.06 | 0.18 | 0.19 |
| **Median** | 0.86 | 0.87 | 0.30 | 0.32 | 0.17 | 0.14 | 0.11 | 0.11 |

* = not significant at $\alpha$ =0.05

### 3.2 Approximate Bayesian Bootstrap

The Approximate Bayesian Bootstrap (ABB) proposed is frequently used to obtain multiply imputed variance estimates with hot deck imputation. ABB is a non-Bayesian method that approximates a Bayesian procedure (Rubin and Schenker 1986; Rubin 1987) ABB involves drawing a random sample of respondents (donors) with replacement and imputing values for missing data using the sample of respondents drawn in the first step as the imputation base.

Each round of the ABB procedure creates one complete dataset (implicate). This procedure is then repeated $M$ times – 20 times in this research – to obtain multiply imputed datasets (implicates). The multiply imputed variance with $M$ implicates is estimated as $\hat{V}_M = W_M + (1 + M^{-1})B_M$, where $W_M = \sum_{v=1}^{M} \hat{V}_v(\hat{T}^I_{dy_j(v)})$ and $B_M = \frac{1}{M-1}\sum_{v=1}^{M}(\hat{T}^I_{dy_j(v)} - \hat{T}^I_{dy_{j(v)}})^2$.

ABB is straightforward to implement. If the number of implicates is small (say 20 or fewer), the approach is appealing for a large program with many variables (8,000+

products) in numerous domains. Furthermore, the ABB approach can be seamlessly integrated with the Finite Population Bayesian Bootstrap proposed by Zhou, Raghunathan, and Elliot (2012), which is being considered for the Economic Census.

## 4. Imputation Evaluation

Table 2 presents the ratio of the empirical MSEs (computed from 5,000 replicates) of the multiply imputed totals with 10 and 20 implicates to the singly imputed hot deck estimates for random and nearest neighbor imputation. In general, the MSEs are all very close, although there are a few exceptions where multiple imputation outperforms single imputation and a few large exceptions where single imputation outperforms multiple imputation for all items.

Table 2: Ratio of Empirical MSEs of Multiply Imputed Totals to Singly Imputed Totals

| Industry | Random | | | | | | | | Nearest Neighbor | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Product 1 | | Product 2 | | Product 3 | | Product 4 | | Product 1 | | Product 2 | | Product 3 | | Product 4 | |
| | 10 | 20 | 10 | 20 | 10 | 20 | 10 | 20 | 10 | 20 | 10 | 20 | 10 | 20 | 10 | 20 |
| 1 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.99 | 0.99 |
| 2 | 1.00 | 1.00 | 0.99 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.99 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 |
| 3 | 1.01 | 1.00 | 1.00 | 1.00 | 1.04 | 1.04 | 0.99 | 0.99 | 1.00 | 1.00 | 1.06 | 1.06 | 1.00 | 1.00 | 0.97 | 0.97 |
| 4 | 1.00 | 1.00 | 1.02 | 1.03 | 0.95 | 0.95 | 0.98 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 | 1.02 | 1.02 | 1.42 | 1.40 |
| 5 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.01 | 1.00 | 1.00 | 1.00 | 1.00 | 1.01 | 1.01 | 0.99 | 0.99 | 1.00 | 1.00 |
| 6 | 0.99 | 0.99 | 1.00 | 1.00 | 0.94 | 0.94 | 0.96 | 0.95 | 0.98 | 0.98 | 0.98 | 0.98 | 1.10 | 1.09 | 1.01 | 1.00 |
| 7 | 1.00 | 1.00 | 1.00 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.98 | 0.98 | 0.85 | 0.85 |
| 8 | 1.00 | 0.99 | 1.00 | 1.00 | 0.99 | 0.99 | 1.00 | 1.00 | 1.02 | 1.01 | 1.00 | 1.00 | 1.00 | 1.00 | 0.96 | 0.96 |
| 9 | 0.99 | 0.99 | 0.96 | 0.96 | 0.98 | 0.97 | 0.96 | 0.95 | 1.00 | 1.00 | 1.00 | 1.01 | 1.02 | 1.02 | 1.15 | 1.15 |
| 10 | 1.00 | 1.00 | 1.01 | 1.01 | 1.01 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 | 0.99 | 0.99 | 0.96 | 0.95 | 0.99 | 0.99 |
| 11 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.99 | 0.99 |
| 12 | 1.00 | 1.00 | 1.00 | 1.00 | 1.03 | 1.02 | 1.00 | 1.00 | 1.00 | 1.00 | 1.01 | 1.01 | 0.99 | 0.99 | 1.02 | 1.02 |
| 13 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| 14 | 2.20 | 1.29 | 1.27 | 1.59 | 1.46 | 1.93 | 1.11 | 1.18 | 2.31 | 1.09 | 3.79 | 2.81 | 3.13 | 2.87 | 2.26 | 2.63 |
| 15 | 1.16 | 1.08 | 2.04 | 1.11 | 0.78 | 0.64 | 1.03 | 0.99 | 1.01 | 0.99 | 1.25 | 0.94 | 1.30 | 1.21 | 2.41 | 1.89 |
| 16 | 0.99 | 1.00 | 0.99 | 1.00 | 0.99 | 0.99 | 0.99 | 0.99 | 1.27 | 1.27 | 1.23 | 1.25 | 1.39 | 1.39 | 0.86 | 0.85 |
| 17 | 1.00 | 1.00 | 0.97 | 0.96 | 1.01 | 1.01 | 1.02 | 1.02 | 1.00 | 1.00 | 1.43 | 1.43 | 1.00 | 1.00 | 1.01 | 1.01 |
| 18 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.93 | 0.93 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.22 | 1.22 |
| 19 | 1.00 | 1.00 | 0.99 | 0.99 | 0.99 | 0.98 | 1.02 | 1.02 | 0.99 | 0.99 | 1.01 | 1.01 | 0.93 | 0.91 | 1.01 | 1.01 |
| 20 | 1.00 | 1.00 | 0.98 | 0.98 | 1.00 | 1.01 | 1.00 | 1.00 | 1.03 | 1.02 | 1.00 | 1.00 | 1.00 | 1.00 | 1.03 | 1.03 |
| 21 | 1.00 | 1.00 | 0.99 | 0.99 | 1.01 | 1.01 | 1.00 | 1.00 | 1.00 | 1.00 | 1.01 | 1.01 | 1.06 | 1.06 | 1.00 | 1.00 |
| Median | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |

Tables 3 and 4 present the relative biases for random and nearest neighbor hot deck for single (S) and multiple imputation (10 and 20). Absolute relative biases greater than 0.10 are shaded for emphasis.

Table 3: Relative Biases of Products 1 through 4 with Random Hot Deck Imputation

| | Product 1 | | | Product 2 | | | Product 3 | | | Product 4 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Industry** | **S** | **10** | **20** | **S** | **10** | **20** | **S** | **10** | **20** | **S** | **10** | **20** |
| 1 | -0.01 | -0.01 | -0.01 | 0.01 | 0.01 | 0.01 | 0.00 | 0.00 | 0.00 | -0.01 | -0.01 | -0.01 |
| 2 | 0.01 | 0.00 | 0.00 | -0.12 | -0.11 | -0.11 | 0.11 | 0.11 | 0.11 | 0.05 | 0.05 | 0.05 |
| 3 | -0.04 | -0.05 | -0.05 | 0.00 | 0.01 | 0.01 | 0.42 | 0.43 | 0.44 | -0.02 | -0.01 | -0.01 |
| 4 | 0.00 | 0.00 | 0.00 | -0.21 | -0.21 | -0.21 | 1.03 | 1.00 | 1.00 | -0.55 | -0.54 | -0.54 |
| 5 | 0.01 | 0.01 | 0.01 | 0.00 | 0.00 | 0.00 | -0.12 | -0.12 | -0.12 | -0.01 | 0.01 | 0.01 |
| 6 | 0.01 | 0.01 | 0.01 | 0.02 | 0.02 | 0.02 | -0.17 | -0.17 | -0.17 | -0.04 | -0.04 | -0.04 |
| 7 | 0.00 | 0.00 | 0.00 | 0.03 | 0.03 | 0.03 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 |
| 8 | -0.04 | -0.04 | -0.04 | -0.02 | -0.02 | -0.02 | 0.03 | 0.02 | 0.02 | 0.00 | 0.00 | 0.00 |
| 9 | -0.05 | -0.05 | -0.05 | 0.44 | 0.43 | 0.43 | 0.22 | 0.21 | 0.21 | 0.51 | 0.48 | 0.48 |
| 10 | 0.04 | 0.04 | 0.04 | -0.10 | -0.10 | -0.10 | -0.11 | -0.11 | -0.11 | -0.09 | -0.09 | -0.09 |
| 11 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | -0.02 | -0.02 | -0.02 |
| 12 | 0.00 | 0.00 | 0.00 | -0.03 | -0.03 | -0.03 | 0.08 | 0.09 | 0.09 | -0.02 | -0.02 | -0.02 |
| 13 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 14 | -0.02 | -0.15 | -0.08 | 0.05 | 0.09 | 0.12 | -0.07 | -0.10 | -0.13 | 0.13 | 0.15 | 0.16 |
| 15 | -0.04 | -0.10 | -0.07 | -0.05 | 0.25 | 0.09 | 0.19 | 0.12 | 0.06 | -0.04 | -0.06 | -0.02 |
| 16 | 0.44 | 0.43 | 0.43 | -0.57 | -0.57 | -0.57 | -0.65 | -0.64 | -0.64 | -0.04 | -0.04 | -0.04 |
| 17 | 0.02 | 0.02 | 0.02 | -0.29 | -0.29 | -0.28 | -0.52 | -0.53 | -0.53 | 0.56 | 0.57 | 0.57 |
| 18 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | -0.01 | -0.01 | -0.01 | -0.47 | -0.45 | -0.45 |
| 19 | -0.01 | -0.01 | -0.01 | -0.01 | -0.01 | -0.01 | 0.14 | 0.14 | 0.14 | 0.10 | 0.10 | 0.10 |
| 20 | -0.01 | -0.01 | -0.01 | 0.06 | 0.06 | 0.06 | -0.06 | -0.06 | -0.06 | 0.01 | 0.01 | 0.01 |
| 21 | -0.01 | -0.01 | -0.01 | 0.04 | 0.03 | 0.03 | 0.04 | 0.04 | 0.04 | 0.02 | 0.02 | 0.02 |

Notice that the direction of the corresponding relative biases (within industry, product, and method) is often the same. Recall from Figure 2 that products 2 through 4 have high incidence of reported zeros. With Product 1 – the most frequent non-zero reported product value – the relative biases are generally close to zero. As the reported zero rate increases, the incidence of biased imputed estimates increases. In most cases, it does not appear that adding implicates reduces the size of the relative biases. Indeed, the F-tests for differences in absolute relative bias due to the number of implicates (S, 10, 20) failed to find any evidence of a treatment effect. Table 4 presents the relative biases for the same estimates obtained using nearest neighbor imputation.The nearest neighbor imputed estimates tend to be less biased than the random imputed counterparts. Overall the number of implicates does not seem to affect the level of relative bias.

To summarize, within hot deck variation, the single and multiply imputed estimates generally have comparable relative biases. With random hot deck, the level of bias is fairly minimal for the most frequently reported product and is not necessarily negligible otherwise, thus leading to confounding in the variance estimation analysis of the remaining products. With nearest neighbor, the level of bias is fairly low for the majority of products.

Table 4: Relative Biases of Products 1 through 4 with Nearest Neighbor Imputation

| | Product 1 | | | Product 2 | | | Product 3 | | | Product 4 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Industry** | **S** | **10** | **20** | **S** | **10** | **20** | **S** | **10** | **20** | **S** | **10** | **20** |
| 1 | -0.01 | -0.01 | -0.01 | 0.01 | 0.01 | 0.01 | 0.02 | 0.02 | 0.02 | -0.03 | -0.02 | -0.02 |
| 2 | 0.03 | 0.03 | 0.03 | -0.03 | -0.02 | -0.02 | -0.07 | -0.07 | -0.07 | 0.01 | 0.02 | 0.02 |
| 3 | 0.00 | -0.01 | -0.01 | 0.04 | 0.06 | 0.06 | -0.08 | -0.08 | -0.08 | -0.04 | 0.02 | 0.02 |
| 4 | 0.00 | 0.00 | 0.00 | -0.04 | -0.04 | -0.04 | -0.10 | -0.11 | -0.11 | 0.18 | 0.26 | 0.25 |
| 5 | 0.00 | 0.00 | 0.00 | 0.02 | 0.03 | 0.03 | -0.12 | -0.12 | -0.12 | -0.01 | 0.00 | 0.00 |
| 6 | 0.03 | 0.03 | 0.03 | -0.02 | -0.01 | -0.01 | -0.08 | -0.09 | -0.09 | 0.00 | -0.01 | -0.01 |
| 7 | 0.00 | 0.00 | 0.00 | 0.02 | 0.02 | 0.02 | 0.04 | 0.03 | 0.03 | 0.10 | 0.08 | 0.08 |
| 8 | -0.01 | -0.01 | -0.01 | 0.02 | 0.03 | 0.03 | 0.01 | 0.01 | 0.01 | -0.02 | -0.01 | -0.01 |
| 9 | 0.00 | 0.00 | 0.00 | 0.02 | 0.03 | 0.03 | 0.00 | -0.02 | -0.02 | -0.05 | -0.08 | -0.08 |
| 10 | 0.00 | 0.00 | 0.00 | -0.01 | -0.01 | -0.01 | -0.03 | -0.02 | -0.02 | -0.02 | -0.01 | -0.01 |
| 11 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | -0.02 | -0.01 | -0.01 |
| 12 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 0.01 | -0.02 | -0.02 | -0.02 | 0.01 | 0.02 | 0.02 |
| 13 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 14 | -0.03 | -0.17 | -0.05 | 0.06 | 0.27 | 0.22 | -0.07 | -0.18 | -0.17 | 0.03 | -0.20 | -0.22 |
| 15 | 0.03 | -0.04 | -0.02 | -0.06 | 0.14 | -0.02 | 0.03 | -0.11 | -0.09 | -0.03 | 0.33 | 0.26 |
| 16 | 0.16 | 0.20 | 0.20 | -0.27 | -0.31 | -0.31 | -0.10 | -0.21 | -0.21 | 0.39 | 0.33 | 0.33 |
| 17 | -0.01 | -0.02 | -0.02 | 0.06 | 0.22 | 0.22 | -0.01 | -0.01 | -0.01 | 0.01 | 0.02 | 0.02 |
| 18 | 0.00 | 0.00 | 0.00 | 0.01 | 0.01 | 0.01 | 0.00 | 0.00 | 0.00 | -0.14 | -0.21 | -0.21 |
| 19 | -0.01 | -0.01 | -0.01 | 0.06 | 0.06 | 0.06 | -0.07 | -0.06 | -0.06 | 0.02 | 0.02 | 0.02 |
| 20 | -0.02 | -0.02 | -0.02 | 0.01 | 0.01 | 0.01 | 0.00 | 0.01 | 0.01 | 0.03 | 0.04 | 0.03 |
| 21 | 0.00 | 0.00 | 0.00 | 0.02 | 0.02 | 0.02 | -0.04 | -0.05 | -0.05 | 0.02 | 0.02 | 0.02 |

# 4. Variance Evaluation

## 4.1 Model Assisted Variance Estimator

With the model-assisted variance estimator, we use a regression model to estimate the unit variance. Table 5 presents the relative biases of each variance estimate obtained using random hot deck imputation with the PAR1 and PAR2 models for each product.

Clearly, the relative biases of the PAR2 variance estimates are at unacceptable levels. Interestingly, the ANOVA tests for only three of the four products provide strong evidence of differences due to estimating model (Product 1: Model 1 p-value=0.02, Model 2 p-value=0.04; Product 2 Model 1 p-value=0.17, Model 2 p-value=0.03; Product 3: Model 1 p-value=0.01, Model 2 p-value=0.04). That said, the PAR1 variance estimates are far from unbiased, as the magnitude of the *least* biased variance estimate is 0.17 (17%).

Table 5: Relative Biases of the Variance Estimates for Products 1 through 4 Obtained Via the Beaumont-Bocci Model-Assisted Method with Random Hot Deck Imputation

| Industry | Product 1 | | Product 2 | | Product 3 | | Product 4 | |
|---|---|---|---|---|---|---|---|---|
| | PAR1 | PAR2 | PAR1 | PAR2 | PAR1 | PAR2 | PAR1 | PAR2 |
| 1 | -0.77 | 31.27 | -0.85 | 22.76 | -0.73 | 61.04 | 1.05 | 253.10 |
| 2 | 5.90 | 543.84 | 8.80 | 803.02 | 2.37 | 169.28 | 1.95 | 204.28 |
| 3 | -0.30 | 531.05 | 37.07 | 21327.31 | 2.88 | 4391.24 | 32.63 | 2337.91 |
| 4 | 2.89 | 4627.28 | 99.76 | 135834.25 | 3.00 | 3942.21 | 72.11 | 121808.34 |
| 5 | -0.97 | -0.21 | 0.23 | 36.21 | 0.37 | 39.10 | 6.29 | 182.30 |
| 6 | 4.29 | 2258.80 | 1.25 | 654.65 | 8.78 | 3178.92 | 13.21 | 2546.91 |
| 7 | -1.00 | -0.93 | -0.17 | 23.62 | 0.74 | 35.99 | 2.56 | 183.67 |
| 8 | 25.77 | 7174.59 | 25.43 | 9853.29 | 20.31 | 5029.79 | 62.61 | 12486.40 |
| 9 | 4.81 | 259.04 | 66.34 | 3253.15 | 113.16 | 14188.26 | 33.16 | 2163.49 |
| 10 | 1.79 | 838.23 | 8.65 | 3000.38 | 19.17 | 4979.84 | 20.80 | 6623.63 |
| 11 | -1.00 | -0.88 | -0.91 | 38.02 | -0.96 | 16.24 | -0.87 | 57.97 |
| 12 | -1.00 | -0.96 | -0.93 | 0.08 | -0.79 | 2.47 | -0.85 | 1.43 |
| 13 | -1.00 | -1.00 | -1.00 | -1.00 | -1.00 | -1.00 | -1.00 | -1.00 |
| 14 | 63.87 | 769.77 | 19.04 | 358.97 | 13.03 | 283.38 | 13.57 | 311.04 |
| 15 | 1.65 | 64.64 | 25.53 | 429.74 | 1.83 | 115.21 | 17.29 | 374.79 |
| 16 | 0.58 | 290.19 | 2.11 | 975.57 | 14.52 | 22444.84 | 42.81 | 17936.26 |
| 17 | -0.97 | 7.51 | 4.15 | 1257.72 | 1.46 | 23.83 | -0.53 | 80.51 |
| 18 | -0.99 | -0.79 | -0.80 | 3.16 | -0.76 | 3.50 | 0.21 | 29.14 |
| 19 | 26.50 | 2454.78 | 86.42 | 9982.56 | 51.70 | 5706.08 | 78.06 | 7380.88 |
| 20 | 98.48 | 6754.40 | 58.45 | 3898.87 | 370.76 | 21789.49 | 291.83 | 17156.16 |
| 21 | -0.93 | 3.80 | 1.34 | 159.20 | 8.92 | 709.46 | 4.96 | 407.30 |

Table 6 presents the relative biases of each variance estimate obtained using random nearest neighbor imputation with the PAR1 and PAR2 models for each product.

Although the level of relative bias is greatly decreased from the respective levels obtained from random hot deck imputation, these variance estimates are far from unbiased, and the direction of the biases has no clear pattern. With Products 2 and 4, the PAR2 model variances have larger absolute average relative biases than their PAR1 counterparts (Product 2: Model 1 p-value=0.10, Model 2 p-value=0.05; Product 4: Model 1 p-value=0.04, Model 2 p-value=0.02).

Table 6: Relative Biases of the Variance Estimates for Products 1 through 4 Obtained Via the Beaumont-Bocci Model-Assisted Method with Nearest Neighbor Hot Deck Imputation

| Industry | Product 1 | | Product 2 | | Product 3 | | Product 4 | |
|---|---|---|---|---|---|---|---|---|
| | PAR1 | PAR2 | PAR1 | PAR2 | PAR1 | PAR2 | PAR1 | PAR2 |
| 1 | -0.97 | -0.86 | -0.98 | -0.90 | -0.97 | -0.67 | -0.74 | -0.11 |
| 2 | -0.64 | 2.50 | -0.42 | 4.97 | -0.72 | 0.80 | -0.83 | 0.47 |
| 3 | -0.97 | -0.86 | 1.30 | 10.15 | -0.21 | 2.49 | 2.08 | 15.26 |
| 4 | -0.94 | -0.92 | 2.08 | 3.31 | 1.79 | 2.59 | 3.96 | 7.52 |
| 5 | -1.00 | -0.97 | -0.77 | 0.36 | -0.67 | 0.99 | -0.08 | 5.84 |
| 6 | -0.34 | 3.09 | -0.44 | 1.71 | 1.50 | 12.26 | 1.88 | 12.06 |
| 7 | -1.00 | -1.00 | -0.87 | -0.87 | -0.73 | -0.74 | -0.73 | -0.72 |
| 8 | -0.76 | -0.89 | -0.86 | -0.89 | -0.83 | -0.90 | -0.43 | -0.65 |
| 9 | -0.98 | -1.00 | 1.65 | -0.27 | 1.84 | 0.05 | 1.26 | -0.25 |
| 10 | -0.96 | -0.96 | -0.85 | -0.84 | -0.46 | -0.56 | -0.73 | -0.69 |
| 11 | -1.00 | -1.00 | -0.99 | -0.96 | -0.99 | -0.98 | -0.98 | -0.94 |
| 12 | -1.00 | -1.00 | -0.97 | -0.97 | -0.84 | -0.85 | -0.95 | -0.95 |
| 13 | -1.00 | -1.00 | -1.00 | -1.00 | -1.00 | -1.00 | -1.00 | -1.00 |
| 14 | 0.74 | 0.61 | -0.07 | 0.14 | 0.30 | 0.57 | 0.15 | 0.64 |
| 15 | -0.88 | -0.83 | 0.35 | 1.48 | -0.46 | 0.08 | -0.12 | 1.22 |
| 16 | -0.35 | -0.10 | 0.18 | 0.77 | 1.57 | 6.26 | -0.34 | 0.96 |
| 17 | -1.00 | -1.00 | -0.53 | -0.83 | -0.29 | -0.81 | -0.71 | -0.94 |
| 18 | -1.00 | -1.00 | -0.95 | -0.94 | -0.94 | -0.93 | -0.27 | -0.10 |
| 19 | -0.25 | -0.11 | 0.50 | 1.05 | 2.30 | 3.49 | 4.77 | 6.64 |
| 20 | 0.06 | 0.56 | 0.08 | 0.73 | 3.20 | 4.26 | 1.28 | 2.78 |
| 21 | -0.99 | -0.98 | -0.67 | -0.50 | 0.66 | 1.58 | -0.33 | 0.02 |

Collectively, these results are quite disappointing, although not unexpected given the poor predictive power of the regression models used for Products 2 through 4 (see Table1). Regardless of hot deck variation or assumed error model, the variance estimates are extremely biased, and the direction of the biases is not consistent. One could make a very weak argument for improved performance using the PAR1 error model, but it would not be convincing, given the magnitude of the relative biases and the inconsistency of direction.

## 4.2 Multiple Imputation (ABB) Results

Recall that the ABB is a form of replication variance estimation. Processing time can be quite long and the procedure can be computer resource-intensive. So, we are interested in using the fewest number of implicates feasible to obtain approximately unbiased variance estimates – and are willing to sacrifice possible gains in stability as a consequence. This sacrifice in stability should be primarily manifested with the rarely reported products (i.e. those with a high reported zero rate). Table 7 presents the relative biases of each variance estimate for each product obtained using random hot deck imputation with 10 and 20 implicates.

Table 7: Relative Biases of the Variance Estimates for Products 1 through 4 Obtained Via the ABB with 10 and 20 Implicates with Random Hot Deck Imputation

| Industry | Product 1 | | Product 2 | | Product 3 | | Product 4 | |
|---|---|---|---|---|---|---|---|---|
| | MI10 | MI20 | MI10 | MI20 | MI10 | MI20 | MI10 | MI20 |
| 1 | 0.01 | 0.27 | 0.61 | 0.16 | 2.13 | 1.25 | 6.25 | 4.18 |
| 2 | 0.10 | 0.22 | 0.41 | 0.00 | 0.67 | 0.76 | 0.54 | 0.67 |
| 3 | 0.14 | 0.38 | 11.33 | 7.86 | 0.90 | 0.36 | 9.50 | 6.53 |
| 4 | 0.46 | 0.62 | 0.35 | 0.04 | 0.47 | 0.62 | 0.67 | 0.77 |
| 5 | 0.89 | 0.35 | 3.82 | 2.45 | 2.85 | 1.75 | 11.16 | 7.71 |
| 6 | 0.89 | 0.35 | 0.04 | 0.26 | 2.66 | 1.63 | 9.22 | 6.40 |
| 7 | 0.88 | 0.92 | 0.65 | 0.18 | 2.14 | 1.25 | 4.56 | 2.97 |
| 8 | 3.64 | 2.35 | 0.19 | 0.15 | 0.40 | 0.00 | 2.72 | 1.66 |
| 9 | 3.25 | 2.02 | 0.17 | 0.17 | 5.33 | 3.56 | 3.20 | 2.04 |
| 10 | 0.11 | 0.37 | 0.21 | 0.43 | 0.71 | 0.24 | 0.53 | 0.09 |
| 11 | 0.93 | 0.95 | 0.81 | 0.87 | 0.88 | 0.91 | 0.72 | 0.80 |
| 12 | 0.91 | 0.93 | 0.85 | 0.89 | 0.73 | 0.81 | 0.73 | 0.80 |
| 13 | 0.97 | 0.98 | 0.98 | 0.99 | 0.98 | 0.98 | 0.97 | 0.98 |
| 14 | 2.89 | 3.69 | 0.01 | 0.43 | 0.29 | 0.61 | 2.39 | 1.27 |
| 15 | 0.69 | 0.76 | 1.15 | 1.68 | 0.37 | 0.46 | 2.09 | 1.32 |
| 16 | 0.96 | 0.97 | 0.72 | 0.80 | 0.70 | 0.79 | 1.93 | 1.09 |
| 17 | 0.94 | 0.96 | 0.04 | 0.31 | 0.70 | 0.79 | 0.73 | 0.80 |
| 18 | 0.76 | 0.83 | 0.45 | 0.61 | 0.28 | 0.49 | 1.32 | 0.66 |
| 19 | 0.12 | 0.37 | 1.48 | 0.77 | 1.01 | 0.44 | 2.33 | 1.37 |
| 20 | 0.60 | 0.14 | 0.45 | 0.04 | 1.74 | 0.94 | 2.13 | 1.24 |
| 21 | 0.69 | 0.78 | 0.60 | 0.71 | 1.78 | 0.98 | 0.16 | 0.17 |

Here, the variance estimates are always positively biased. However, the magnitude of the bias is greatly reduced from the model-assisted results presented in Section 5.1. In all cases, the computed average absolute relative bias obtained with 20 implicates is smaller than the average of the 10-implicate estimates, although this difference is only weakly significant for Products 3 and 4 (Product 3: Model 1 p-value=0.01, Model 2 p-value=0.90; Product 4: Model 1 p-value = 0.01, Model 2 p-value = 0.81). Originally, these results appeared to be counterintuitive. Kim (2002) proved that the multiply imputed variance estimator with random hot deck imputation is always negatively biased under the cell mean model, and that the magnitude of the bias increases as the number of implicates increases. However, the cell mean model does not adequately represent any of our test data as discussed above. Moreover, it is possible that an atypical set of donors could be resampled. Table 8 presents the relative biases of each variance estimate for each product obtained using nearest neighbor hot deck imputation with 10 and 20 implicates.

Table 8: Relative Biases of the Variance Estimates for Products 1 through 4 Obtained Via the ABB with 10 and 20 Implicates with Nearest Neighbor Hot Deck Imputation

| Industry | Product 1 | | Product 2 | | Product 3 | | Product 4 | |
|---|---|---|---|---|---|---|---|---|
| | MI10 | MI20 | MI10 | MI20 | MI10 | MI20 | MI10 | MI20 |
| 1 | 0.02 | 0.27 | 0.62 | 0.16 | 2.02 | 1.17 | 7.00 | 4.71 |
| 2 | 0.03 | 0.26 | 0.52 | 0.09 | 0.49 | 0.64 | 0.55 | 0.68 |
| 3 | 0.21 | 0.43 | 9.62 | 6.62 | 11.11 | 7.69 | 1.84 | 1.04 |
| 4 | 0.48 | 0.63 | 1.74 | 0.96 | 11.96 | 8.27 | 0.04 | 0.31 |
| 5 | 0.90 | 0.36 | 3.54 | 2.25 | 3.02 | 1.88 | 2.24 | 1.32 |
| 6 | 0.77 | 0.27 | 0.36 | 0.03 | 3.86 | 2.50 | 3.65 | 2.33 |
| 7 | 0.88 | 0.92 | 0.83 | 0.31 | 2.09 | 1.21 | 1.21 | 0.58 |
| 8 | 4.26 | 2.78 | 0.01 | 0.28 | 0.66 | 0.19 | 1.15 | 0.53 |
| 9 | 3.19 | 1.99 | 13.76 | 9.51 | 52.98 | 37.70 | 40.70 | 28.83 |
| 10 | 0.24 | 0.11 | 0.21 | 0.14 | 3.96 | 2.59 | 5.32 | 3.51 |
| 11 | 0.93 | 0.95 | 0.81 | 0.87 | 0.88 | 0.92 | 0.92 | 0.95 |
| 12 | 0.91 | 0.93 | 0.84 | 0.89 | 0.48 | 0.63 | 0.86 | 0.90 |
| 13 | 0.97 | 0.98 | 0.98 | 0.99 | 0.98 | 0.98 | 0.99 | 0.99 |
| 14 | 2.09 | 3.62 | 0.72 | 0.73 | 0.63 | 0.71 | 0.83 | 0.17 |
| 15 | 0.69 | 0.77 | 2.49 | 2.16 | 0.03 | 0.26 | 0.41 | 0.52 |
| 16 | 0.88 | 0.92 | 0.17 | 0.41 | 0.02 | 0.27 | 0.68 | 0.77 |
| 17 | 0.94 | 0.96 | 0.31 | 0.51 | 0.26 | 0.47 | 0.03 | 0.26 |
| 18 | 0.76 | 0.83 | 0.45 | 0.61 | 0.31 | 0.51 | 1.00 | 1.00 |
| 19 | 0.08 | 0.34 | 0.73 | 0.24 | 4.41 | 2.92 | 3.01 | 1.85 |
| 20 | 0.34 | 0.04 | 1.04 | 0.46 | 1.47 | 0.77 | 0.25 | 0.10 |
| 21 | 0.69 | 0.78 | 0.62 | 0.73 | 2.10 | 1.21 | 0.39 | 0.56 |

Again, the ABB variance estimates are consistent overestimates. Although the average relative biases appear to be smaller for the 20-implicate estimates, there is only weak evidence of a treatment effect for Product 2 (Model 1 p-value=0.04, Model 2 p-value =0.85).

Recall that the ABB procedure resamples from the donor records. We hypothesized that accuracy of the variance estimates could be affected by the level of product response or the percentage of non-zero reported values [Note: Ellis and Thompson (2015) found that unit type (single or multi-unit ) was predictive of the number of reported products, but not of the probability of reporting any product]. Unfortunately, we could not find any evidence of either relationship in these data sets.

Although the ABB variance estimates are not unbiased, these results are not unpromising. The direction of the bias is consistent regardless of imputation method. Given the rarity of many products, it is conservative to overestimate the variance. Moreover, it appears that the bias of the ABB variance estimates is decreased as the number of implicates increases. Of course, this will not be true for the rarely reported products (as there are few donor cases to resample), but does hold promise for the more frequently reported products.

## 5. Conclusion

The objective of this study was to find an approximately unbiased estimate of the imputation variance component for the majority of studied products in both the design-based (model-assisted) and Bayesian (model-based) framework. Ideally, our recommended method would work equally for both hot deck variations, as the type of hot deck method will vary by industry in the 2017 Economic Census. More important, the adopted methods should provide unbiased variance estimates for the majority of products. Lastly, the computation time and resources should be fairly undemanding, given the number of reported products and the size of the Economic Census in terms of number of establishments.

In one sense, the second objective is not particularly obtainable for many products when random imputation is used. Random hot deck imputation works well on our product data sets under very restrictive conditions such as small number of donor records, limited range of unit size, and few collected products; Bechtel, Morris, and Thompson (2015) provides more details. Otherwise, the random hot deck imputation can yield severely biased product estimates, as seen in Section 4.

Finding unbiased variance estimates of biased estimates is challenging, although not impossible. The considered model-assisted estimation approach is extremely appealing, but did not prove viable with our datasets. We suspected this from the beginning given the poor predictive power of the available ratio model for the majority of studied products. We likewise doubted that the stratification model would prove superior, as the Economic Census stratification is not particularly fine and has limited strata. Indeed, the lack of a consistent ratio (or regression) model is what led to the original (model free) hot deck recommendation for products. Our results confirmed our intuition, sadly.

Ironically, the considered Bayesian approach uses direct replication and does not assume a model. These resampling results were certainly more consistent and yielded less biased estimates. It is likely that the level of bias could be further reduced with additional implicates for the well-reported products, although the difference in bias appears to be fairly trivial with the less well-reported products and might not justify the additional computation resources. Indeed, we note that variance estimates for the nearest neighbor imputed products are not totally unacceptably biased, regardless.

Ultimately, we need to find a variance estimator that accounts for sampling errors, nonsampling errors, and post-stratification. The results presented here provide evidence for pursuing resampling methods. Thompson, Thompson, and Kurec (2016) recommend using the Finite Population Bayesian Bootstrap (FPBB) described in Zhou, Raghunathan, and Elliot (2012) to estimate the sampling error of the post-stratified estimates, which can be easily integrated with the ABB. Our challenge will be minimizing the computer-resources needed (in terms of number of implicates) and perhaps convincing our stakeholders that it would be preferable to limit the publication of product estimates to the largest contributors in an industry.

## Acknowledgements

We thank William Davie, Jr., Carma Hogue and Broderick Oliver for their comments on earlier versions of the paper.

## References

Andridge, R. R., & Little, R. J. (2010). A Review of Hot Deck Imputation for Survey Non-response. *International Statistical Review, 78*(1), 40-64. doi:10.1111/j.1751-5823.2010.00103.x

Beaumont, J., & Bocci, C. (2009). Variance estimation when donor imputation is used to fill in missing values. *Canadian Journal of Statistics Can. J. Statistics, 37*(3), 400-416. doi:10.1002/cjs.10019

Bechtel, L., Morris, D.S., and Thompson, K.J. ( 2015). Using Classification Trees to Recommend Hot Deck Imputation Methods: A Case Study. *Proceedings of the FCSM Research Conference.*

Brick, J., & Kalton, G. (1996). Handling missing data in survey research. *Statistical Methods in Medical Research, 5*(3), 215-238. doi:10.1177/096228029600500302

Charlton, J. (2004). Editorial: Evaluating automatic edit and imputation methods, and the EUREDIT project. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 167*(2), 199-207. doi:10.1111/j.1467-985x.2004.02051.x

Chen, J., & Shao, J. (2001). Jackknife Variance Estimation for Nearest-Neighbor Imputation. *Journal of the American Statistical Association, 96*(453), 260-269. doi:10.1198/016214501750332839

Ellis, Y. and Thompson, K.J. (2015). Exploratory Data Analysis of Economic Census Products: Methods and Results. *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Kim, J. K. (2002). A note on approximate Bayesian bootstrap imputation. *Biometrika, 89*(2), 470-477. doi:10.1093/biomet/89.2.470

Kim, J. K., & Fuller, W. (2004). Fractional hot deck imputation. *Biometrika, 91*(3), 559-578. doi:10.1093/biomet/91.3.559

Knutson, J. and Martin, J. (2015). Evaluation of Alternative Imputation Methods for Economic Census Products: The Cook-Off. *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Nordholt, E. S. (1998). Imputation: Methods, Simulation Experiments and Practical Examples. *International Statistical Review, 66*(2), 157-180. doi:10.1111/j.1751-5823.1998.tb00412.x

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

Rubin, D. B., & Schenker, N. (1986). Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse. *Journal of the American Statistical Association, 81*(394), 366-374. doi:10.1080/01621459.1986.10478280

SAS/STAT(R) 9.2 User's Guide, Second Edition. Retrieved April 14, 2016, from https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#mixed_toc.htm

Shao, J., & Sitter, R. R. (1996). Bootstrap for Imputed Survey Data. *Journal of the American Statistical Association, 91*(435), 1278-1288. doi:10.2307/2291746

Thompson, J. R. (2000). *Simulation: A modeler's approach*. New York: Wiley.

Thompson, K.J. and Liu, X. (2015). On Recommending a Single Imputation Method for Economic Census Products. *Proceedings of the Section on Government Statistics*, American Statistical Association.

Thompson, M. Thompson, K.J., and Kurec. R. (forthcoming in 2016). Variance Estimation for Product Value Estimates in the 2017 Economic Census Under the Assumption of Complete Response. *Proceedings of the Governments Statistics Section*, American Statistical Association.

Tolliver, K. and Bechtel, L. (2015). Implementation of Hot Deck Imputation on Economic Census Products. *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Zhou, H., Raghunathan, T., and Elliot, M. (2012). A Semi-Parametric Approach to Account for Complex Designs in Multiple Imputation. *Proceedings of the Proceedings of the FCSM Research Conference.*