Evaluation of Alternative Imputation Methods for 2017 Economic Census Products[1]

Jeremy Knutson and Jared Martin

Abstract

In preparation for the 2017 change to the North American Product Classification System (NAPCS), Economic Census staff was tasked with determining a single imputation method to treat missing product data collected from all trade areas. To objectively compare four proposed imputation methods, we conducted a simulation study to obtain two evaluation measures: imputation error, to measure the accuracy of the overall estimate, and the fraction of missing information (FMI), to measure the precision of the imputed estimate. For the “cook-off,” we generated complete pseudo populations by applying each imputation method to missing sample data, inducing product nonresponse in each population, and applying each imputation method to the missing data. Nonresponse was induced independently in each pseudo population, yielding 50 replicates. Each imputation procedure was multiply-imputed within replicate. Imputation methods (“treatments”) are evaluated within trade area using the average imputation error and FMI. This evaluation approach is generalizable to other programs with similar missing data problems.

1. Introduction

Choosing the best method to correct for nonresponse is not a simple task for any data collection activity, let alone the Economic Census, which is the U.S. Government's official five-year measure of American business and the economy. Prior to this 2014 study, the strategy of correcting for nonresponse varied by subject matter area within the Economic Census. This paper details the difficult process of making an objective recommendation of a single method for use in eight diverse trade areas of the Economic Census. The evaluation focuses on the performance of four chosen imputation methods on product estimates in selected industries with common products under North American Product Classification System (NAPCS) at the national and industry level.

One of the goals of the Economic Census Reengineering project is to fully implement the NAPCS in the 2017 Economic Census, a process which began in the 2002 Econ Census. Unlike previous census collections, product information obtained using NAPCS allows for cross-trade area tabulation of products. The goal of this research is to recommend a single methodology for imputing missing product data collected using the NAPCS for the 2017 Economic Census. Research was conducted in two phases: (1) an exploratory data analysis phase to study data characteristics; and (2) a simulation study to assess statistical properties and performance of selected imputation methods.

The Economic Census is processed in eight different trade areas: Construction (CON), Finance, Insurance, and Real Estate (FIR), Manufacturing (MAN), Mining (MIN), Services Industries (SER), Retail Trade (RET), Wholesale Trade (WHO), and Transportation, Communication, and Utilities (UTL). Each trade area is composed of similar industries; and within each trade area a core set of data items is collected from each establishment called general statistics items. In addition, the Economic Census collects information on the revenue obtained from product sales. Prior to the 2017

[1] Disclaimer: This report is released to inform interested parties of research and to encourage discussion. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

Economic Census, a list of products specific to each industry was provided directly on the industry questionnaire. Beginning with the 2017 Economic Census, data collection will be electronic, and the respondents will have greater flexibility in reporting products. Moreover, NAPCS allows the collection of the same product in different industries.The methods of treating missing product data in the 2012 Economic Census (and prior censuses) varied greatly by trade area.

2. Product Data Collection

The Economic Census attempts to collect a total value for sales, shipments, receipts, or revenue from all sampled establishments. Product data (labeled as "Details of sales, shipments, receipts, or revenue") are collected towards the end of the questionnaire. The types of products that an establishment is expected to produce or to sell are strongly related to the primary industry in which the establishment operates. As mentioned in the introduction, the sum of the product values reported should add to the value of total receipts provided. In-house, each of the individual collected products on the form are referred to as "product lines", hereafter referred to as "products". The products are expected to sum up to the total receipts value.

This research used selected 2012 Economic Census product data from seven trade areas: FIR, MAN, MIN, SER, RET, WHO, UTL and selected 2007 Economic Census product data in the construction (CON) trade, since 2012 CON data was still in processing. All data have undergone post-collection editing and imputation (Plain Vanilla (PV) and specialty edits. See Sigman and Wagner (1997) and Wagner (2000)).

In all trade areas except CON, classification experts selected ten to thirty industries per trade area with common products under NAPCS. These industries were included in the phases of exploratory data analyses and response propensity analyses. We selected five industries per trade in the final simulation evaluation phase as described in Section 3. Unfortunately, there is no direct translation of Kind-of-Business (KOB) and Type of Construction (TOC) to NAPCS construction products, so the CON analyses present a "worst case" scenario at best and are included only for completeness.

3. Imputation Methods

Three types of imputation method were considered for this project: ratio (expansion) imputation, hot deck imputation (random and nearest neighbor), and sequential regression multivariate imputation (SRMI). See Garcia, Morris, and Diamond (2015) for a discussion of the EXP imputation and SRMI implementations, see Tolliver and Bechtel (2015) for a discussion of HDN and HDR implementations.

4. Evaluation Statistics

For each product within an imputation cell and trade area population, each imputation method was evaluated using two statistics: imputation error and the fraction of missing information (FMI). For each imputation method, we obtain these summary measures by product and within imputation cell.

Since the Economic Census produces product tabulations and does not release corresponding micro-data, the evaluation criteria have been calculated at an industry tabulation level. An imputation method that produces realistic micro-data is not required (although desirable), whereas estimate accuracy is necessary. Thus comparisons between establishment-level imputed product values (within replicate and implicate) to their corresponding population values to measure error as done in Charlton (2004) will not be performed.

4.1. Imputation Error

We define the ***imputation error*** (IE) of product p in imputation cell i obtained using imputation method m in replicate r in a given trade area population as $IE_r^{ipm} = \left(\bar{Y}_r^{ipm} - Y^{ip}\right)$, where Y^{ip} is the trade area population total of product p in imputation cell i.

The ***absolute imputation error*** (AIE) measures the magnitude of the imputation error (ignoring direction) and is computed as $AIE_r^{ipm} = |IE_r^{ipm}|$.

4.2. The Fraction of Missing Information

To avoid the possibility that the imputation methods would tie with respect to IE, a second evaluation criteria, the ***fraction of missing information*** (FMI), was also evaluated. FMI is a measure of “the level of uncertainty about the values one would impute for current nonrespondents” (Wagner, 2010). The FMI for product p in imputation cell i from replicate r obtained with imputation method m on v implicates (i.e. each final imputed data set will be constructed by averaging across v multiple imputed data sets) is given by

$$FMI_{\bar{Y}_r^{ipm}} = (1 + \frac{1}{v})\frac{B_r^{ipm}}{T_r^{ipm}}$$

where B_r^{ipm} and T_r^{ipm} are the multiple-imputation between and total variances defined in Section 5 using v = 100 implicates in our applications. If the imputation method tends to yield consistent distributions, then the between-implicate component will be very small, and the FMI will be close to zero. If the imputation method performs inconsistently, then the FMI value will approach one.

Since the FMI is a random variable with a measurable variance. Wagner (2010) and Harel (2007) note that a large number of implicates are required to estimate the FMI with reasonable precision; Wagner (2010) uses 100 implicates, and Harel (2007) recommends using between 50 – 200 implicates, depending on the level of precision desired and the “true” (but unknown) value of the FMI. Harel (2003) provides an approximate expression for the variance of the FMI, which we use in Section 6:

$$\hat{V}\left(FMI_{\bar{Y}_r^{ipm}}\right) \approx \frac{FMI_{\bar{Y}_r^{ipm}}\left(1 - FMI_{\bar{Y}_r^{ipm}}\right)}{\sqrt{\frac{V}{2}}}.$$

5. Simulation Study Procedure

After implementing the four different imputation methods and testing each on Economic Census data, we now employ a data-driven procedure, which produces the necessary results to objectively compare each imputation method to all of the others. This comparison procedure, known in-house as the “cook-off”, is summarized as follows:

1. Impute to create complete pseudo populations
2. Randomly induce nonresponse using a pre-specified propensity model
3. Impute missing values using specified model(s)
4. Compare the resultant fully imputed dataset(s) on predetermined (statistical) criteria (evaluation is covered in Section 6)

Figure 1 depicts steps 1-3 of our simulation procedure. We independently repeat the process 50 times to create 50 *replicates*. Nordholt (1998) reports invariant results using a similar procedure with 50 replicates. Within replicate, we applied each imputation method to the missing data to obtain complete datasets, using *multiple imputation* to obtain the statistics needed for evaluation (*v* = 100 implicates per replicate).

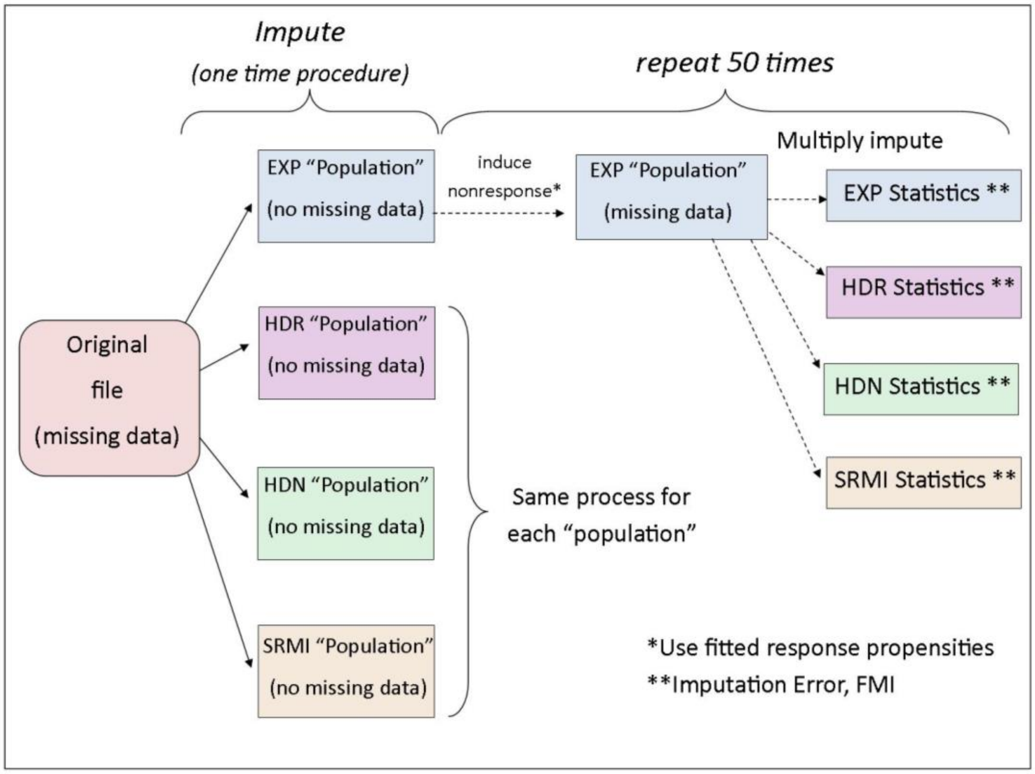
Impute
(one time procedure)
repeat 50 times
Multiply impute
Original file (missing data)
EXP "Population" (no missing data)
HDR "Population" (no missing data)
HDN "Population" (no missing data)
SRMI "Population" (no missing data)
induce nonresponse*
EXP "Population" (missing data)
EXP Statistics **
HDR Statistics **
HDN Statistics **
SRMI Statistics **
Same process for each "population"
*Use fitted response propensities
**Imputation Error, FMI

Figure 1: Simulation Cook-Off Procedure

This procedure permits the robustness of the imputation method to be evaluated over repeated samples and under alternative response mechanisms (for two excellent large-scale applications, see Northolt (1998) and Charlton (2004)). Frequently, similar evaluations obtain the population data by simulating realistic complete population data, restricting the study data to unit respondent data, or "imputing" missing values with historic data from the same units. Similar data simulation approaches were infeasible for our data sets. Each industry collects different products – with little overlap in products. It is difficult to develop reasonable multivariate models to generate simulated data, since many products are reported by only a few establishments. The available data are insufficient for developing parametric models or for resampling methods for the rarely reported products. Moreover, the low item response rates and the possibility that the response mechanism could be non-ignorable (related to the products collected) make it unwise to treat the product respondent data as a good representation of the available universe. Finally, there was consensus from the subject matter experts that any matched historical data would likely be to be unrealistic.

Rather than attempt to develop a single "realistic" population for each trade area, we selected five industries, each with at least two well-represented products. Then we generated four complete "populations" by applying each candidate imputation method to replace the missing data as suggested by Dr. Trivellore Raghunathan (University of Michigan), which is the "Impute (one time procedure)" in Figure 1 above. This was done to mitigate any possible interaction between the imputation method used to produce the "population" and the imputation method being tested.

After developing four complete "populations" in each trade area, we randomly induced unit nonresponse in each population using fitted unit level response propensity probabilities. We fit logistic regression models to find covariates that significantly contribute to the probability that a unit respondent will provide usable product data.

The conditional probability that an establishment reports usable product data is estimated by the logistic function of a linear combination of the explanatory covariates:

$$\Pr\left(Y_{kj} = 1 \middle| \boldsymbol{X}_{kj}^{w}\right) = \pi\left(\boldsymbol{X}_{\boldsymbol{kj}}^{\boldsymbol{w}}\right) = \frac{\exp\left(\beta^{w}\boldsymbol{X}_{kj}^{w}\right)}{1+\exp\left(\beta^{w}\boldsymbol{X}_{kj}^{w}\right)} = \frac{\exp(\beta_0+\beta_1 x_{kj1}+\cdots+\beta_w x_{kjw})}{1+\exp(\beta_0+\beta_1 x_{kj1}+\cdots+\beta_w x_{kjw})},$$

where

$$Y_{kj} = \begin{cases} 1 & \text{if the establishment j in industry k provided any usable product line data} \\ 0 & \text{otherwise} \end{cases}$$

and

$\boldsymbol{X}_{kj}^{w}$= ($x_{\mathrm{kj}1}$, $x_{\mathrm{kj}2}$, … , x_{kjw}) denotes the vector of *w* potential explanatory covariates of unit response from establishment *j* in industry *k*.

We performed response propensity modeling by trade area using a forward selection procedure derived by Wang and Shin (2011). Each additional covariate must be statistically significant given those already in the model in the forward selection. We use the likelihood-ratio test to measure overall goodness-of-fit for each candidate model, whose test statistic is

$$D = -2\ln\left[\frac{\text{likelihood of the fitted model}}{\text{likelihood of the saturated model}}\right].$$

Under the null hypothesis $\beta\boldsymbol{X} = 0$, and *D* has an approximate chi-squared distribution. Each variable in the forward selected model must be statistically significant using the Wald statistic.

Ideally, we want to minimize the number of covariates. Furthermore, any categorical variable must have a sufficient number of respondents per imputation cell, if we are to consider it as a possible covariate. In addition to considering the goodness-of-fit test results described above, we examined the Rescaled R^2 from Tjur (2009). We calculated the mean predicted probability of an event for each of the two categories of the dependent variable and calculated the difference between those two means. Like the "traditional" R^2 used in linear regression, the upper bound is 1.0 and the interpretation is analogous.

As discussed in the Section 4, one of our evaluation criteria is the FMI. The FMI is a ratio of two variance estimates (between and total) that are usually obtained using multiple imputation. The Sequential Regression Multivariate Imputation (SRMI) applications easily adapt to multiple imputation, as advertised. However, the hot deck and expansion methods require multiple imputation analogues.

Furthermore, these multiple imputation analogues for hot deck and expansion must “incorporate appropriate variability among the repetitions of the model” (Rubin 1988); this is an imputation property referred to as “proper” (as defined in Rubin 1987). A proper multiple imputation method will ensure that the resulting fully imputed datasets represent the sampling uncertainty in the imputed values as well as estimation uncertainty associated with the underlying model parameters. Without both types of variability, the imputation procedure is not proper in that it will underestimate the overall variability of the imputation procedure. Rubin (1987) explicitly addresses the underestimation of variability in a simple multiple imputation hot deck: one that simply repeats random draws from respondents. As an example of a proper multiple imputation procedure, consider a standard linear regression model. We would want to (1) draw the parameters of the model from their associated posterior distribution and (2) draw missing values from their posterior distribution conditional on the parameters drawn in step (1). Such a two-stage strategy for multiply imputing datasets with the appropriate amount of variability is not straightforward for all methods.

Within each replicate (out of total *R* replicates), we obtain multiply-imputed estimates of total and variances for the ten selected products in each trade area. The *multiply-imputed estimated total* for product *p* in imputation cell *i* from replicate *r* obtained with imputation method *m* is:

$$\bar{Y}_r^{ipm} = \frac{1}{100}\sum_{v=1}^{100} \hat{Y}_{rv}^{ipm},$$

where $\hat{Y}_{rv}^{ipm} = \sum_{j\in i} w_j\, \ddot{y}_{rvj}^{ipm}$and $\ddot{y}_{rvj}^{ipm}$ is the *j*th establishment’s value of the product (reported or imputed) in the implicate.

For each replicate we find the corresponding multiple imputation variance. The *within imputation variance* is the average of the *v* =100 complete data variances:

$$\bar{\mathrm{U}}_r^{ipm} = \frac{1}{100}\sum_{v=1}^{100} \hat{V}\left(\hat{Y}_{rv}^{ipm}\right).$$

The *between imputation variance* is the variance between the *v*=100 complete data estimates:

$$B_r^{ipm} = \frac{1}{99}\sum\nolimits_{v=1}^{100} (\hat{Y}_{rv}^{ipm} - \bar{Y}\,_r^{ipm})^2.$$

Finally, the *total variance* is a weighted sum of the two aforementioned variances:

$$T_r^{ipm} = \bar{U}_r^{ipm} + \left(1+\frac{1}{100}\right) B_r^{ipm},$$

for more details, see Rubin (1987) and Zhang (2003).

Next, we describe how the replicate estimates statistics are used to obtain the evaluation statistics. Rubin and Schenker (1986) and Rubin (1987) propose the Approximate Bayesian Bootstrap (ABB) as a tool for introducing appropriate variability into a multiple imputation procedure. ABB is a non-Bayesian method that approximates a Bayesian procedure and adjusts for the uncertainty in the distribution parameters resulting in a proper imputation procedure. ABB involves:

1. Drawing a random sample of respondents with replacement, and
2. 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 results in one complete dataset. This procedure is then repeated 100 times to obtain multiple imputed datasets.

ABB is a natural and straightforward way to implement multiple imputation for the hot deck methodology. Typically, for the expansion method, variability is introduced via draws from the distribution of the parameters. However, we decided that using a two-stage model that involves drawing from the error distribution of the model for the expansion method would inherently change the methodology. Thus, we chose to use ABB for expansion to keep the methodology and model intact while incorporating the additional variability by altering the sample for analysis.

Because of the skewed population data, we implemented a slight modification of ABB for both the hot deck and expansion methods. In the first step of the ABB procedure – randomly sampling respondents with replacement – we used probability proportion to size (PPS) sampling with replacement in order to take into account sampling probabilities. This is a simpler case of the adaptation of ABB for complex survey design presented in Dong et. al. (2014).

Ultimately, we repeat each imputation method 5,000 times per *population* for the expansion (EXP), hot deck nearest neighbor (HDN), and hot deck random (HRD), and 50,000 times per *population* for sequential regression multiple imputation (SRMI). This design is a complete block design applied to each population where each product is a block[2] and each imputation method is a treatment, with repeated measures on each of the 50 sets of nonrespondent establishments (one set per replicate).

6. The Evaluation Procedure (using Manufacturing trade area as example)

Given the evaluation statistics described in Section 4, we define the most accurate imputation method *within a trade area* as having

- The lowest IE (closest to zero) for the majority of products ("unbiased")
- The lowest FMI (closest to zero) for the majority of products ("precise")

The evaluation statistics described in this section are rank-based, and the statistical tests are nonparametric. Using rank-based procedures will allow use to choose a "best" imputation method without assuming that the data have any particular distribution. That said, performance information is lost, especially when all imputation methods perform equally well or badly for one evaluation measure but display great disparities in performance between the four methods for the other evaluation measure. These procedures were independently applied to the simulation study results in each trade area.

[2] Our evaluation is restricted to ten products per industry. However, the imputation procedures that we apply to the replicates with missing data consider all potential products (not just the top ten), with the exceptions for the SRMI implementation.

After completing the trade area analyses, the final recommendation was developed jointly.

6.1. Product Level Analysis (within trade area population)

The first step of the analysis is to evaluate the imputation methods' relative performances for each product within industry. This was done separately by trade area population (i.e., [TRADE]-EXP population, [TRADE]-HDN population, [TRADE]-HDR population, [TRADE]-SRMI population, where [TRADE] is one of the eight Econ Census trade areas). Recall that for our study, the trade area populations cover five selected industries within from that trade area. Our analyses used their "top ten" products, defined as the most frequently reported products (by number of establishments) in the selected industries. Unfortunately, our analysis was limited to this subset of products, because for iterative analyses the models do not converge with the less-reported products. Each of these products have been reported by establishments within one or more of the selected industries. Analysis of IE and FMI are conducted separately by **product**.

Within a trade area population, we obtain a single score (rank) that describes the IE performance of each imputation method for product *p* in industry *k* using this procedure:

1. Obtain the ***median*** *absolute imputation error* (RANK_AIE) of product *p* in the imputation cell over the fifty replicates. Rank the four values (one per imputation method) by ascending value, using the mean rank for tied ranks (e.g. for a two way tie for rank "2", assign each the rank (2+3)/2 = 2.5, and the remaining methods are assigned ranks 1 and 4).
2. Obtain the ***range*** *of the imputation error* (RANK_RANGE) of product *p* in the imputation cell over the *R* replicates. Note that we use the actual range of the IE (largest – smallest) for this criterion, not the absolute IE. Rank the four values of the ***range*** of IE by ascending value, using the mean rank for tied ranks.
3. Obtain the weighted average over the two ranked values for each treatment: COMBINED_RANK=0.70*RANK_AIE + 0.30*RANK_RANGE. These weights were developed heuristically, so that the magnitude of the IE has more influence on the rank than the range of the magnitudes (over replicates), and yet the method that yields large outliers is still penalized.
4. Aggregate COMBINED_RANK by product within trade area population and divide by number of imputation cells containing product to obtain an averaged COMBINED_RANK (the product may be reported in more than one imputation cell within industry or may be reported in more than one industry).
5. Rank to obtain FINAL_RANK, using the mean rank for tied ranks.

Table 1 provides an example of this ranking procedure performed on a single product (PRODUCT1) in the MFG imputation cell 3273200 from the SRMI trade area population. If PRODUCT1 had been reported in more than one imputation cell – in this case it was reported by only one of the selected industries – another four rows per reporting imputation cell would be added to the following table, and final rank would be an average of ranks across multiple industries.

Table 1: Illustration of Ranking Procedure for Imputation Error for a PRODUCT1

METHOD	MEDIAN (AIE)	IE RANK	RANGE (IE)	RANGE-RANK	COMBINED RANK	FINAL RANK
EXP	31401	1	64653	1	1	1
HDN	33426	2	75078	4	2.6	2
HDR	33566	3	66815	2	2.7	3
SRMI	83990	4	73602	3	3.7	4

FMI, in contrast to the IE measures, has a variance that is maximized when the FMI = 0.50 and is minimized when the FMI equals zero or 1. In other words, for a given number of implicates, the variance of the FMI is minimized when either the imputation method is performing extremely well or extremely poorly. Although it is important to incorporate the FMI's variance into the analysis, it would be unwise to use the corresponding variance as a comparative method in this case.

Thus, to incorporate the variance of the FMI into our comparison, we test a general linear hypothesis on the minimum and maximum of the *average value* of the FMI. The general linear hypothesis is performed for each product *p* over the *R* replicates at $\alpha = 0.10$. In a given trade area population and imputation cell, let

μ = $\boldsymbol{R} \times \mathbf{1}$ **vector** of FMI values for product *p* and imputation method *m*

Σ = $\boldsymbol{R} \times \boldsymbol{R}$ **matrix** of FMI variances for the product and imputation method with off-diagonal values

(covariance between replicates) = 0

$\boldsymbol{K} = \boldsymbol{1} \times \boldsymbol{R}$ **vector** of known constants. Since we are testing the average FMI, *K* =(1/R 1/R …. 1/R)

$\boldsymbol{K_0}$ = a value in [0,1], representing a hypothetical FMI value.

Note: the matrix product $\boldsymbol{K}\boldsymbol{\mu}$ = the average FMI for product *p* and imputation method *m* over *R* replicates.

The hypothesis test of interest is:

H_0: $\boldsymbol{K\mu} = \boldsymbol{K_0}$ (Note that the product)

H_A: $\boldsymbol{K\mu} \neq \boldsymbol{K_0}$

The test statistic is given by $(K\mu - K_0)^T (K \Sigma \; K^T)^{-1}(K\mu - K_0) \sim \chi^2_1$ under $H_{0.}$ Iterating over values of K_0 for each test provides a range of values that satisfy the null hypothesis. Thus, the values of K_0 immediately below and above these values provide lower and upper bounds (not a confidence interval) on the average FMI for each product within imputation cell and population over all replicates.

Within a trade area population, we obtain a single score (rank) that examines the FMI performance of each imputation method on product *p* in industry *k* using the following procedure.

1. Find MIN_K_0 and MAX_K_0, which are the minimum and maximum possible values of average FMI, according to the general linear hypothesis test.
2. Summarize MIN_K_0 and MAX_K_0 by the single value: MIDPOINT_FMI = (MIN_K_0 + MAX_K_0)/2.
3. Within imputation cell, rank the four values of MIDPOINT_FMI for product *p* to obtain RANK_MIDPOINT.

4. If the given product appears in multiple imputation cells within the trade area, aggregate RANK by product within trade area and divide by number of imputation cells containing product.
5. Rank to obtain FINAL_RANK, using the mean rank for tied ranks.

Table 2 provides an example of this ranking procedure performed on single product (PRODUCT1) in the same MFG imputation cell, 3273200, from the SRMI trade area population.

Table 2: Illustration of Ranking Procedure for FMI for a Single PRODUCT1

METHOD	MIN_K_0	MAX_K_0	MIDPOINT FMI	MIDPOINT RANK	FINAL RANK
EXP	0.5216	0.6079	0.56475	4	4
HDN	0.4797	0.5666	0.52315	2	2
HDR	0.4961	0.5828	0.53945	3	3
SRMI	0.3275	0.4100	0.36875	1	1

6.2. Imputation Method Selection, Within MAN Trade Area Population

As mentioned in Section 5, the simulation study conducts a complete block design experiment independently in each trade area population. In our design, the ten studied products within trade area represent the blocks, and the treatments are the imputation methods (repeated measures on each establishment). Each treatment is ranked within block (Section 6.1.), with ties represented by means and the lowest rank representing the method with the best performance. Typically, a complete block repeated measures design is analyzed using a two-way analysis of variance (ANOVA). At a minimum, ANOVA assumes that that the residuals have the same variances (homoscedasticity), but inferences that use the F-test require that that variances are i.i.d. normal.

The Friedman Test (Friedman, 1940) is a two-way analysis of variance that uses rank as the measure of interest (i.e. is the nonparametric analog to the two-way ANOVA). There are two assumptions for this test: (1) the results between block are approximately independent (i.e. the results for one product do not influence the results for the other products), and (2) within block, the observations can be ranked in order of interest. Technically, we may not have complete independence among products collected within the same industry. However, we believe that the number of products is large enough within industry to offset the dependence. Demsar (2006) recommends a minimum of five treatments to attain comparable power to the ANOVA test; Conover (1999, Chapter 5.8) does not provide a similar limit on number of treatments or number of blocks, but does note that the power of the tests is directly affected by both.

The omnibus test determines whether all four treatments exhibit the same performance.
H_0: All treatments have equal average rank ($R^1 = R^2 = R^3 = R^4$)
H_A: At least one treatment has a different performance from the others

Let $A = \sum_p \sum_m (R^{pm})^2$, the sum of the squares of the (average) ranks
$C = \frac{PM(M+1)^2}{4} = \frac{10\ \times 4(4+1)^2}{4}$, the "correction factor" for ties in rank

$$T_1 = (M-1)\sum_m \left(R^m - \frac{P(M+1)^2}{2}\right)^2 \Big/ (A-C) = 3\sum_m \left(R^m - \frac{10(4)^2}{2}\right)^2 \Big/ (A-C)$$

$$T_2 = \frac{(P-1)T_1}{P(M-1)-T_1} = \frac{9T_1}{10\times 3 - T_1}$$

Friedman (1940) proposed the T_1 measure; the T_2 is the two-way analysis of variance statistics on ranks recommended by Iman and Davenport (1980).

Under H_0, $T_2 \sim F(M-1,(P-1)(M-1)) = F(3,27)$. Reject H_0 if $T_2 > F(3,27,\alpha=0.10)$.

If the omnibus test is rejected, then it is appropriate to perform pairwise comparisons of rank, adjusted for multiple comparisons. We use the method outlined in Conover (1999, Ch. 5.8), Note that several other options are provided in Demsar (2006). The recommended test is adjusted for ties (as in the omnibus test statistic). At α = 0.10, a pair of summary ranks$(R^p, R^{p'})$ is significantly different when

$$|R^p - R^{p'}| > t_{1-\frac{\alpha}{2}}\sqrt{\frac{2P(A-C)}{(P-1)(M-1)}\left[1-\frac{T_1}{P(M-1)}\right]} = t_{1-\frac{\alpha}{2}}\sqrt{\frac{20(A-C)}{(9)(3)}\left[1-\frac{T_1}{10(3)}\right]}$$

The examples below illustrate these procedures. Table 3 continues our earlier example, presenting the complete set of ranked IE results in MAN_{SRMI} trade area population for the ten products.

Table 3: Ranked Imputation Error Results within Product for SRMI Population, Manufacturing Industry

Blocks	Treatment			
	EXP	HDN	HDR	SRMI
PRODUCT1	1	2	3	4
PRODUCT2	3	1	4	2
PRODUCT3	4	3	2	1
PRODUCT4	1	1	3	4
PRODUCT5	2	3	1	4
PRODUCT6	3	4	2	1
PRODUCT7	2	1	3	4
PRODUCT8	2	4	1	3
PRODUCT9	2	3	4	1
PRODUCT10	2	3	4	1
SUM	22	25	27	25

The omnibus hypothesis tests whether at least one treatment has different results from the others using the sum of the ranks across the products within treatment ($R^{EXP} = 22$, $R^{HDN} = 25$, $R^{HDR} = 27$, $R^{SRMI} = 25$). Here, the test statistic (T_2) = 0.2593. The critical value of this test is F(3,27,α=0.10) = 2.2906. Since $T_2 < F(3,27,\alpha=0.10)$, we fail to reject the null hypothesis. There is not a significant difference between the performances of the different methods. No further testing is appropriate for IE (in this population and trade area) and all cell entries for this row (trade area population/statistic) are represented by (1+2+3+4)/4 = 2.5 (a four way tie) in the trade areas summary table.

Table 4 presents the complete set of FMI results in the MAN_{SRMI} trade area population for the ten products.

Table 4: Ranked FMI Results within Product for SRMI Population, Manufacturing Industry

Blocks	Treatment			
	EXP	HDN	HDR	SRMI
PRODUCT1	4	2	3	1
PRODUCT2	3	2	4	1
PRODUCT3	4	2	3	1
PRODUCT4	3	2	4	1
PRODUCT5	3	2	4	1
PRODUCT6	3	4	2	1
PRODUCT7	4	2	3	1
PRODUCT8	4	2	3	1
PRODUCT9	3	2	4	1
PRODUCT10	4	2	3	1
SUM	35	22	33	10

The omnibus test statistic for this set of summed ranks is $T_2 = 35.1176$, which is greater than $F(3,27,\alpha=0.10) = 2.2906$. Since the null hypothesis is rejected with this set of summed ranks, we conclude that at least one of the treatments has a significantly different result than the others.

In order to find the treatment(s) with the lowest rank, we must examine the pairwise comparisons. For these tests, $t_{1-\frac{\alpha}{2}}\sqrt{\frac{20(A-C)}{(9)(3)}\left[1-\frac{T_1}{10(3)}\right]} = 4.6811$, according to the pairwise test, described above. Table 5 presents the pairwise comparison test results.

Table 5: Pairwise Comparisons for FMI in SRMI Population, Manufacturing Industry Population Differences in Summed Ranks

Difference	$\left\lvert \text{EXP} - \text{HDN} \right\rvert$	$\left\lvert \text{EXP} - \text{HDR} \right\rvert$	$\left\lvert \text{EXP} - \text{SRMI} \right\rvert$	$\left\lvert \text{HDN} - \text{HDR} \right\rvert$	$\left\lvert \text{HDN} - \text{SRMI} \right\rvert$	$\left\lvert \text{HDR} - \text{SRMI} \right\rvert$
Value	13	2	25	11	12	23
Significant	Yes	No	Yes	Yes	Yes	Yes

These results demonstrate no statistical difference between the results for EXP and HDR. However, EXP and HDR have significantly worse results with respect to FMI than those obtained from HDN or SRMI. In our summary table for this trade area population, the method with the lowest rank sum, SRMI, is assigned rank 1, HDN is assigned rank 2, and the tied methods, EXP and HDR, are assigned the average of ranks 3 and 4, 3.5.

6.3. Trade Area Recommendations

The Friedman testing and treatment scoring procedures described in Section 6.2. are performed independently in each trade area population. After the simulation study is completed in all four populations of a given trade area, we created a summary table, like the example depicted in Table 6, to examine the relative performance of the imputation methods on both statistics within trade area in the studied industries and products.

The H_0 P-value column presents the results of the omnibus test for differences by treatment within trade area population for the studied statistic (IE or FMI). The other columns present the imputation method's score within trade area population for the studied statistic. Table 6 presents the trade area recommendation process, using Manufacturing trade area scores. The last row of Table 6, SRMI Population, are the ranks that were found in 7.2, the other ranks in the table are found in a similar fashion.

In our recommendation, in addition to performance, it was necessary that we consider the challenges of implementing each imputation method. In the following tables, it is shown that SRMI is frequently the best performer with respect to FMI. Despite this, we were hesitant to recommend SRMI because of the large implementation challenges associated with imputing the many seldom-reported products.

Table 6: Summary scores for Manufacturing Industries

Population	Imputation Error				FMI			
	EXP	HDN	HDR	SRMI	EXP	HDN	HDR	SRMI
EXP Population	2	2	2	4	3.5	1.5	3.5	1.5
HDN Population	2.5	2.5	2.5	2.5	3.5	2	3.5	1
HDR Population	2.5	2.5	2.5	2.5	3.5	2	3.5	1
SRMI Population	2.5	2.5	2.5	2.5	3.5	2	3.5	1

In MAN, the EXP, HDN, and HDR methods have no statistical difference in performance with respect to IE, with SRMI performing worse than the others in one population. The SRMI method has the lowest FMI rank in three of the four populations, tying with HDN in the other. HDN performs better than both EXP and HDR with respect to FMI. Since HDN avoids the aforementioned difficulties of extending SRMI to all products in the trade area, we recommend HDN as the best compromise, since we are trying to simultaneously balance the objectives of low IE and low FMI.

6.4. Summary and Discussion

Similarly, this simulation study was performed for the other seven trade areas. In all of the studied industries, a form of hot deck was chosen as the best compromise of the considered methods. However, the recommended hot deck variation was split between trade areas. HDN was recommended for MAN, MIN, SER, and CON, and HDR was recommended for RET, WHS, FIR and UTL. However, the studied industries were not a probability sample and may not be representative of the larger trade areas.

7. Conclusion

When assigned the difficult task of recommending a single "best" imputation method to correct for nonresponse in all trade areas of the Economic Census, Census staff devised an imputation "cook-off" process to aid in making an objective recommendation. It was necessary for us to base this recommendation on statistical criteria.

Three separate missing data treatments (ratio (expansion) imputation, hot deck imputation (random and nearest neighbor), and sequential regression multivariate imputation) were chosen as possible candidates to become the single method to be used across all Econ Census trade areas. We developed statistical criteria for evaluation that balanced total IE (i.e., accurate tabulations) and nonresponse bias correction. To remain impartial, we developed an evaluation procedure that objectively considered both factors' importance, but perhaps downplayed major advantages within a measure (for example, one method might have a much lower IE than another).

According to this evaluation, hot deck imputation appears to be the best compromise of the methods considered. We found that different variations of hot deck performed better in different situations. For example, an imputation cell that contains a large number of products and a fairly homogeneous population in terms of total receipts would probably have better results with nearest neighbor imputation, whereas an imputation cell with very few donor records would be better off using random hot deck. Keeping in mind that we examined a limited number of products in a limited number of industries, we strongly recommend retaining this flexibility of hot deck choice in production.

In 2017, the Economic Census will be using all-electronic data collection and will be collecting and publishing products under NAPCS. Our research uses historical data, and although we tried to mitigate the effects of NAPCS changes on the studied products by our industry selection, we cannot fully predict the extent of the differences on the new collected data, especially in situations where products can be reported in multiple industries. More importantly, it is impossible to predict what effects the electronic data collection will have. By implementing hot deck imputation, we hope to be able to quickly resolve production problems related to these changes, perhaps by revising matching criteria or using coarser imputation cells. Certainly, we can avoid relying on model assumptions that we cannot validate.

This recommendation is only the first step. Implementation will require not only further development of SAS code, but cell collapsing criteria, distance functions, and a cold deck or an alternative back-up method for the rare case where no donor record exists. In addition, research on producing establishment counts is needed, as is research on calibration of product data to industry total receipts.

References

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), pp. 199-207.

Cochran, W. 1977. Sampling Techniques. 3rd ed. New York: John Wiley and Sons, Inc.

Conover, W. 1999. Practical Nonparametric Statistics. New York: John Wiley.

Demsar, J. 2006. "Statistical Comparisons of Classifiers over Multiple Data Sets." *Journal of Machine Learning Research*: 7, pp. 1–30.

Dong, Q., Elliott, M. R., and Raghunathan, R. E. 2014. "A Nonparametric Method To Generate Synthetic Populations To Adjust For Complex Sampling Design Features." *Survey Methodology*: 40(1): pp. 29-46.

Friedman, M. 1940. A Comparison of Alternative Tests Of Significance For The Problem Of M Rankings. *Annals of Mathematical Statistics:* 11: pp. 86–92.

Garcia, M., Morris, D., and Diamond, L.K. (forthcoming in 2015). Implementation of Ratio Imputation and Sequential Regression Multiple Imputation on Economic

Census Products. Proceedings of the Section on Survey Research Methods: American Statistical Association.

Harel, O. 2003. “Strategies for Data Analysis with Two Types of Missing Values.” *PhD thesis from the Pennsylvania State University Graduate School Department of Statistics.*

Harel, O. 2007. “Inferences On Missing Information Under Multiple Imputation And Two-Stage Multiple Imputation.” *Statistical Methodology*: 4, pp.75-89.

Iman, R.L. and Davenport. J.M. 1980. “Approximations Of The Critical Region Of The Friedman Statistic.” *Communications in Statistics*: pp. 571–595.

Lohr, S. L. 2010. Sampling: Design and Analysis. 2nd ed. Boston: Brooks/Cole.

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., and Solenberger, P. 2001. “A Multivariate Technique For Multiply Imputing Missing Values Using A Sequence Of Regression Models.” *Survey Methodology*: 27(1): pp. 85–95.

Roberts, G., Rao, J.N.K., and Kumar, S. 1987. “Logistic Regression Analysis of Sample Survey Data.” Biometrika 74: pp. 1–12.

Rubin, D.B. 1988. An Overview of Multiple Imputation. *Proceedings of the Section on Survey Research Methods*: American Statistical Association.

Rubin, D.B. 1987. Multiple Imputation for Nonresponse in Surveys. Hoboken, NJ: John Wiley & Sons.

Rubin, D.B., and Schenker, N.1986. “Multiple Imputation For Interval Estimation From Simple Random Samples With Ignorable Nonresponse.” *Journal of the American Statistical Association*: 81(394): pp. 366-374.

Tjur, T. 2009. “Coefficients of Determination In Logistic Regression Models—A New Proposal: The Coefficient Of Discrimination.” *The American Statistician* 63: 366-372.

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

Wagner, D. 2000. Economic Census General Editing – Plain Vanilla. *Proceedings of the 2nd International Conference on Establishment Surveys*.

Wagner, J. 2010. “The Fraction of Missing Information as a Tool for Monitoring the Quality of Survey Data.” *Public Opinion Quarterly*: 74(2), pp. 223-243.

Wang, F. and Shin, H.. 2011. “Model Selection Macros for Complex Survey Data Using PROC SURVEYLOGISTIC/SURVEYREG.” *MWSUG Proceedings*.

Zhang, P. 2003. “Multiple Imputation: Theory and Method.” *International Statistical Review*: 71(3), pp. 581-592.