Implementation of Ratio Imputation and Sequential Regression Multivariate Imputation on Economic Census Products[1]

Maria M Garcia, Darcy Steeg Morris, and L. Kaili Diamond
US Census Bureau, 4600 Silver Hill Rd., Washington D.C., 20233

Abstract

The Economic Census[2] collects general items from business establishments such as total receipts, as well as detailed items such as product data. Although product data are an important component of the Economic Census, they vary by establishment and across trade areas, can be difficult to collect, and are characterized by low item response rates. Beginning in 2017, the Census Bureau will begin using the North American Product Classification System (NAPCS) for economy-wide product tabulations. Under NAPCS, products are no longer linked to industry, so we seek a single imputation method for all products. We present two regression models for these data: ratio imputation and sequential regression multivariate imputation (SRMI). The ratio estimator uses a simple linear regression model with total receipts as the single predictor and product receipts as the estimated value. The SRMI method proposed by Raghunathan et al. (2001) imputes missing values consecutively by fitting a sequence of regression models to estimate product receipts conditioning on observed and imputed variables. We present the methodologies, implementation, and application to missing product data imputation.

Key Words: Economic Census, Missing Data, Imputation, Multiple imputation

1. Introduction

The Census Bureau conducts an Economic Census every five years to collect information from business establishments that produce goods and services. Data collected include general items such as annual payroll and total receipts, as well as detailed items such as product data. Product data are an essential component of the Economic Census. They vary by establishment and across trade areas, can be difficult to collect, and are characterized by low item response rates. Prior to 2017, Economic Census questionnaires contained a list of products specific to each industry. For the 2017 Census, the Economic Census Reengineering Project is implementing the North American Product Classification System (NAPCS), which allows the collection of the same product across different industries. This major change allows production of cross-sector product data statistics for the first time. We take this opportunity to investigate methods for missing product data imputation with the goal of identifying a **single imputation method** for all products that can be used in all trade areas for producing economy-wide product tabulations.

[1] This report is released to inform parties of ongoing research and to encourage discussion. The views expressed are those of the authors and not necessarily of the US Census Bureau.

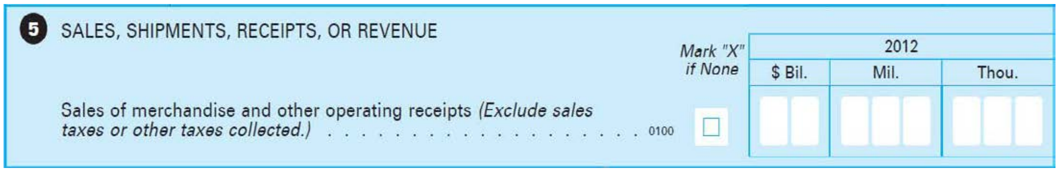
[2] Starting with the 2017 collection, the Economic Census will be published as the Census of U.S. Businesses.

Missing product data can occur when an establishment does not file a census questionnaire (unit nonresponse), a respondent provides no detailed product data (item nonresponse), or the reported product data do not sum to the reported total receipts (partial product data). Current methods for treating missing product data vary across the different trade areas. In the manufacturing and mining sectors, each establishment's aggregated products are compared to the final total receipts, the difference of the sum of the reported product values and the reported total receipts is published as "Not specified by Kind (NSK)" and no attempt is made to impute for missing values. The construction area uses the hot deck nearest neighbor method while the service sectors retain product data that is within a pre-specified raking tolerance and "expand" the retained aggregated data using a ratio estimator (see Section 3). We researched and evaluated the application of four separate imputation methods. Tolliver and Bechtel (2015) report on the application of the hot deck (nearest neighbor and random) imputation method. In this paper, we present two alternative methodologies for missing product data imputation: Ratio imputation (expansion) and the sequential regression multivariate imputation (SRMI). In Section 2, we provide background on Economic Census product data, Sections 3 and 4 present the ratio estimator and the SRMI methods respectively, results appear in Section 5, and we close with a brief summary in Section 6.

2. Background on Economic Census Product Data

2.1 Product Data Collection

The 2012 Economic Census questionnaire collects information on the value of sales, shipments, receipts, or revenue in separate items of the questionnaire. Respondents first report their overall total value in survey Item 5. Detailed values broken down by the type of products likely to be reported within the industry are collected in Item 22 and must balance to the total value provided in Item 5. Figure 1 displays an example for the collection of survey Item 5 for establishments operating in the retail trade area.

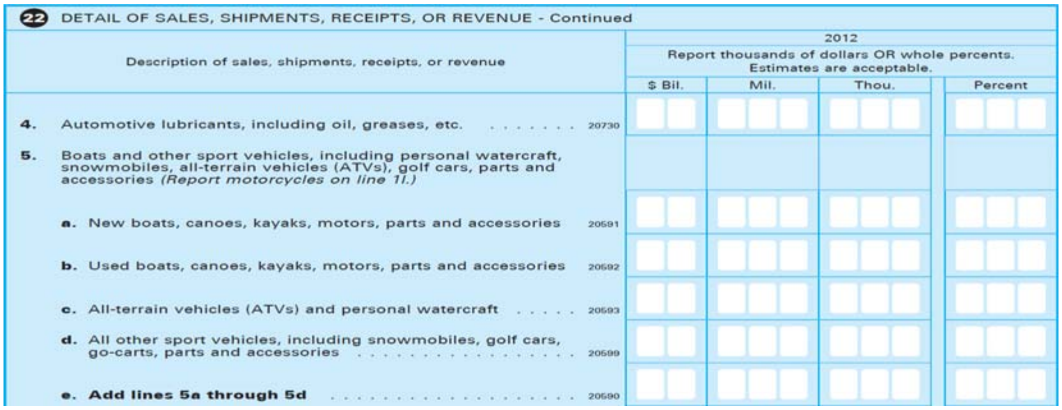


5 SALES, SHIPMENTS, RECEIPTS, OR REVENUE

	Mark "X" if None	2012 $ Bil.	 Mil.	 Thou.
Sales of merchandise and other operating receipts *(Exclude sales taxes or other taxes collected.)* 0100	□			

Figure 1: Example of Economic Census Collection Instrument for Item 5

Figure 2 provides an excerpt for the detailed collection of product data (Item 22) for establishments in the "Automobile Dealers Industry" within the retail trade sector. Establishments report totals either in thousands of dollars or in percentages for products grouped into ten separate "broad" products occupying about five pages in the printed instrument, with the constraint that the sum of all broad products must balance to the total value of receipts (or percentages) reported in Item 5. This extract displays broad products (4) "Automotive lubricants, including oil, greases, etc." and (5) "Boats and other sport vehicles, including personal watercraft, snowmobiles, all-terrain vehicles (ATVs), golf cars, parts and accessories." Note that Line (5.e) is an additional balance constraint requiring the value of total receipts reported in broad product (5) sums up to the reported product details (Lines (5.a) - Line (5.d)). Item response rates for detail items are lower than those for broad items; consequently, our analysis is limited to broad product data.



22 DETAIL OF SALES, SHIPMENTS, RECEIPTS, OR REVENUE - Continued

Description of sales, shipments, receipts, or revenue		2012			
		Report thousands of dollars OR whole percents. Estimates are acceptable.			
		$ Bil.	Mil.	Thou.	Percent
4. Automotive lubricants, including oil, greases, etc.	20730	□□	□□□	□□□	□□□
5. Boats and other sport vehicles, including personal watercraft, snowmobiles, all-terrain vehicles (ATVs), golf cars, parts and accessories *(Report motorcycles on line 11.)*					
a. New boats, canoes, kayaks, motors, parts and accessories	20591	□□	□□□	□□□	□□□
b. Used boats, canoes, kayaks, motors, parts and accessories	20592	□□	□□□	□□□	□□□
c. All-terrain vehicles (ATVs) and personal watercraft	20593	□□	□□□	□□□	□□□
d. All other sport vehicles, including snowmobiles, golf cars, go-carts, parts and accessories	20599	□□	□□□	□□□	□□□
e. Add lines 5a through 5d	20590	□□	□□□	□□□	□□□

Figure 2: Extract from 2012 Economic Census Collection Instrument for Retail Trade Establishments in the Automobile Dealer Industry (Item 22)

2.2. Statistical Challenges

Economic Census product data do not easily lend themselves to statistical models. In prior censuses, the questionnaire included a list of likely products within the industry, which could change under NAPCS. It is a challenge to develop imputation models given a set of products, and difficult to develop good models when the covariates themselves (the products) can change. Moreover, product data have very low item response rates for all but the most frequently reported products. Thus there is little available data for modeling, and it is not unlikely that the response mechanism is non-ignorable i.e., that product data are not reported because it is difficult or tedious for respondents to provide it. If the response mechanism is non-ignorable, then it is difficult to find an imputation method that would work well under the additivity constraint.

For evaluation purposes, we cannot use existing tabulated data as a gold standard. First, adjustment methods vary by trade area; we are concerned that using available tabulated data for evaluating alternative methods would compromise the other method's results by treating the data obtained using one of those methods as "true." Furthermore, the percentage of eligible sampled units that provide at least one valid product varied across trade areas, but was often quite low in our test data sets. It is possible that product respondents could differ systematically from product nonrespondents on an unobserved criterion (e.g. a latent class variable). The data sets were also quite noisy due to sampling error (in many sectors, the single unit establishments are sampled). Historical product data were generally not available to "fill in the gaps," and we agreed that even if such data were available, it would not greatly reduce the nonresponse.

2.3 Test Data

We used 2012 Economic Census product data from seven trade areas: Finance, Insurance, and Real Estate (FIRE), Manufacturing (MAN), Mining (MIN), Services Industries (SER), Retail Trade (RET), Wholesale Trade (WHO), and Transportation, Communication, and Utilities (UTL), and 2007 Economic Census product data in the construction (CON) trade. Classification experts selected ten to 30 industries (except in construction) with common products under NAPCS. These industries were included in the exploratory data analyses and response propensity analyses completed prior to developing imputation models. Because of project schedule and processing time constraints, we selected five industries per trade area for the simulation study. The construction sector collects some items for which there is no direct translation to products under NAPCS, so the construction test data present a “worst case” scenario and are included only for completeness. All test data have undergone post-collection editing and imputation and specialty edits (See Sigman and Wagner (1997) and Wagner (2000)). We restrict the missing data adjustment procedures to sampled units that are full year reporters, have positive final total receipts values, and were used for product estimation.

2.4. Preliminary Data Analyses

We conducted analyses to gain insight into the characteristics of reported and missing product data. The purpose of these analyses was twofold: To understand reported product data to inform the selection of imputation methods and to understand the nature of missing product data to assess existing imputation cells and inform refinements. Results of these analyses provided insight on several issues including choosing a set of predictors, identifying covariates that are related to establishment reporting of usable product data, identifying sorting variables within imputation cells, and choosing raking tolerances. Details and results of these analyses appear in Ellis and Thompson (2015).

3. Ratio Expansion Estimator (EXP)

The ratio imputation procedure (referred to in-house as the “expansion method”) assumes a no-intercept simple linear regression model to estimate values for a variable with missing values. The weighted model described below takes into account both unequal sampling and unit size in the parameter estimation. The weighted least squares estimate of the regression parameter (β) is the best linear unbiased estimator under this model (Magee 1998). Cochran (1977, Chapter 6) and Lohr (2010, Chapter 4.6 and 11.4) demonstrate how the ratio estimator that employs the weighted regression model is also optimal. The Services Sector produces estimates of tabulated products using this model.

3.2 Model and Methods

For a given imputation cell, let y_j^p denote the value of the p^{th} product for the j^{th} establishment and x_j the value of its total receipts. The underlying ratio imputation model is a simple no-intercept regression model with total receipts x_j as the single predictor and product value y_j^pas the estimated value, $y_j^p = \beta^p x_j + \varepsilon_j$, where $\varepsilon_j \sim N(0, w_j x_j \sigma^2)$. Let R denote the set of all establishments reporting at least one usable product within the imputation cell. The estimated weighted least squares regression parameter is $\hat{\beta}^p = \sum_{j\in R} w_j y_j^p \,/\, \sum_{j\in R} w_j\, x_j$. Magee (1998) describes this estimator as a quasi-Aitken WLS (QWLS) estimator, proving its consistency for a super-population regression coefficient when probability sampling is performed independently by strata. At the aggregate imputation cell level, we used the estimated regression parameters, summed y_j^p over all establishments that reported the product, and obtained the resulting ratio estimator for product p as $\hat{Y}^p = \left(\frac{\sum_{j\in R} w_j y_j^p}{\sum_{j\in R} w_j x_j}\right) \sum_j w_j x_j = \left(\frac{\sum_j w_j x_j}{\sum_{j\in R} w_j x_j}\right) \sum_{j\in R} w_j y_j^p$. We then multiply each establishment’s total receipts by each product ratio at the aggregate level to arrive at an imputed value for each case.

The ratio estimate is a best linear unbiased estimate if the following conditions are met:

1. The relation between product values and total receipts is a line through the origin.
2. The product and total receipts have strong positive correlation.
3. The variance of products about the regression line is proportional to total receipts.

Figure 3 displays an example of data that satisfy all three assumptions. Notice that the variability increases as the size of the independent variable (total receipts) increases, and there are a few outlying values that may slightly affect overall fit.

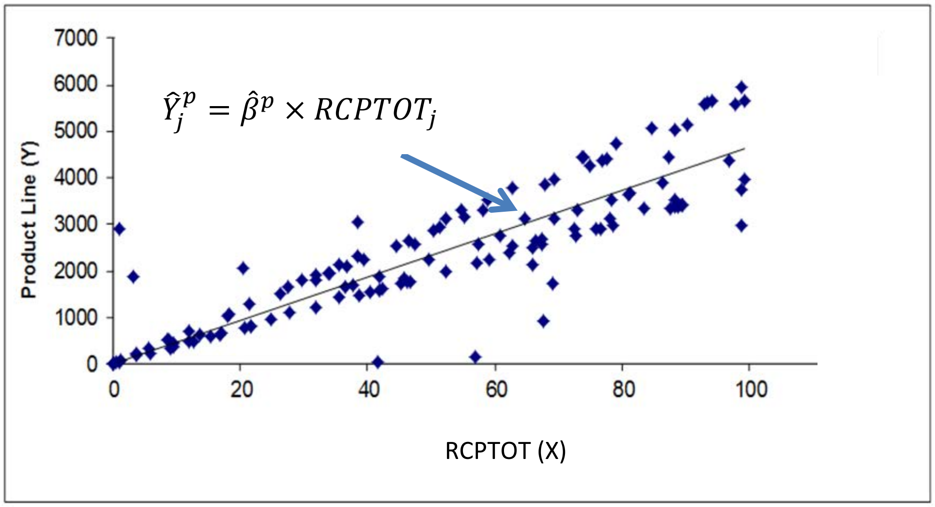


Figure 3: Example Illustration of the Ratio Imputation Model

There are several advantages when imputing using the ratio imputation procedure. It is an intuitive, verifiable, imputation model, based on industry averages that is easy to implement, and can retain additivity if we incorporate the sample coverage adjustment. It has some disadvantages as well. First, product data often do not satisfy the model assumptions. For example, for less-reported products, the assumption of the straight line through the origin can be tenuous. Second, the imputation method does not preserve the multivariate distribution of products within an establishment. Third, it imputes values for all reported products in the imputation cell for product non-respondents. Some of these products might not be relevant in reality. Finally, there is very little variability in the imputed values based on this procedure; imputed values are close to the regression line.

4. Sequential Regression Multivariate Imputation (SRMI)

In this section, we describe the model-based procedure sequential regression multivariate imputation (SRMI) proposed by Raghunathan et al. (2001). SRMI is a general method for multiple imputation (Rubin, 1987) of missing data. The procedure does not require the explicit joint distribution of the variables; instead, it specifies a conditional distribution for each variable separately. The method imputes missing values consecutively by fitting a sequence of regression models conditioning on all the observed and imputed variables. The imputations are random draws from the posterior predictive distribution; regression coefficients are drawn from their current posterior and imputes are drawn from the regression equation conditional on other variables and the new regression coefficients.

4.2 Model and Method

Let the set of responses to a sample survey be denoted by a matrix Y. Let n be the number of variables and let Y_p denote the column corresponding to variable p. Let l represent the number of variables with missing values and $n - l$ the number of fully observed variables so that $Y = (Y_1, Y_2, \dots, Y_l, Y_{l+1}, \dots, Y_n) = (Y_{miss}, Y_{obs})$. In the context of missing product data imputation, the columns in Y_{miss} correspond to valid products within the trade area that may include missing data and are candidates for imputation. The columns in Y_{obs} correspond to other variables in the data file that are available for all sampled units.

Following Raghunathan et al. (2001), assume product data are ordered according to the amount of missing values, from least to most (i.e. Y_1 is the variable with the fewest missing values, followed by Y_2 , and so forth.) For each $p = 1, 2, \dots, l,$ we partition the set of responses into observed and missing values, such that each $Y_p = (Y_{p,obs}, Y_{p,miss})$. SRMI is an iterative procedure; on each iteration, imputations for $Y_{p,miss}$ are generated as random draws from the posterior predictive distribution specified by the regression model $P(Y_p | Y_q, q \neq p; \theta_p)$, conditioning on the other variables and unknown set of parameters θ_p as follows:

1. Draw θ_p from $P(\theta_p | Y_{p,obs}; Y_q, q \neq p)$
2. Draw $Y_{p,miss}$ from $P(Y_{p,miss} | Y_q, q \neq p; \theta_p)$

This procedure works in a Bayesian framework and a diffuse prior is placed on the parameters θ_p. The iterative process imputes values Y_p^t for $t = 1, 2, \dots, tol$ for some pre-specified tolerance. At each iteration, Y_q represents the fully observed regressors that have been updated with imputed values and thus are now considered "observed".

By nature of these data, not all products are reported in all industries, thus our model performs Steps 1 and 2 within imputation cells. First, we regress variables with missing product values using only observations where the product variable to be imputed is non-missing; this regression produces a set of parameters θ. These parameters are estimates of the population parameters and thus have distributions. In Step 1, we take draws from the distribution of the regression parameters. In Step 2, we calculate a predicted value using the draws for the regression coefficients and the observed values for all predictors.

There are several computer software packages for implementing SRMI. IVEware is a suite of macros callable from SAS (Raghunathan et al., 2002) that performs single or multiple imputations of missing values using the SRMI method. The R packages MICE (Buuren et al., 2011) and *mi* (Su et al., 2011) implement a chained equations approach to iterative regression imputation. Our approach to missing product data imputation uses the SRMI model as it is implemented in the IVEware software of Raghunathan et al. (2002).

4.3 SRMI Implementation for Economic Census Product Data

IVEware is a SAS callable software that implements SRMI as described in Raghunathan et al. (2001). IVEware has many built-in tools such as automatic model selection and regression diagnostics and it allows the user control of parameters related to model specification, model selection and convergence, including capability to restrict imputation to certain subpopulations, restrict imputations within specified bounds, etc.

We run IVEware by imputation cell as not all products are reported in all industries. This not only ensures that only eligible products are imputed, but controls the imputation by building the model with data within the same industry. Establishments report product values for multiple products within each trade area, hence there are multiple (distinct) product values that might be missing ($prodval_1, prodval_2, \dots, prodval_l$). We classified these variables as mixed variables within IVEware; mixed variables are both categorical (value of zero) and continuous (nonzero). We use a two-stage model to impute missing product values. First, we use a logistic regression model to impute zero vs. non-zero status (i.e., $I(prodval = 0)$ or $I(prodval = 1)$). Conditional on imputing a nonzero status for the value of $prodval$, we model the product value as a function of the covariates using a linear regression model. In addition, because product values must be nonnegative, we specified a BOUNDS statement in IVEware restricting the range of imputed values to be greater than or equal to zero. Below we describe other issues regarding the implementation of IVEware for missing product data imputation.

4.3.1 Data Issues

The 2012 Economic Census data files have one record for each unique combination of establishment and product. We created one record per establishment, including all possible products within the industry as required for the sequential framework of the regression. In addition, we reformat the data to distinguish missing product values from products with zero value to prevent IVEware from imputing a valid zero product value.
Given that Economic Census data tend to be right skewed, we applied the log transform $\ln(Z + 1)$ to total receipts and product values. Note that applying the log transform to $(Z + 1)$ ensures zero-valued products are mapped to zero-valued product consistent with the definition of product values as mixed variables having a zero/non-zero status.

Product data have low item response rates; some products are reported frequently, but many products are rarely reported, and there are products where only one establishment reported a nonzero value. This can lead to poor estimates of the regression coefficients for that particular product value. In those situations, IVEware imputes that single reported value for all eligible recipients. There is no clear method of dealing with the scarcity of these products. Our solution is to **not impute** for the scarce products; instead, we create a remainder product that contains the sum of all scarce products, thus reducing the number of poorly estimated coefficients while maintaining a “good” estimate of the total.

4.3.2 Model Specification and Selection

In addition to incorporating the correlations between product values via conditional models, SRMI allows the user to incorporate other covariates in the imputation model. We choose the covariates according to the exploratory data analyses and response propensity models results: total receipts, geographic region, and sample weight as covariates in all trade areas; state FIPS (Federal Information Processing Series) code in the RTL, WHO, SER, FIRE and UTL sectors and operating expenses in the WHO and SER sectors. We selected the stepwise regression option in IVEware and specified an inclusion rule based on the R^2, where only variables with a marginal R^2 of 0.01 or higher are included as predictors.

4.3.3 Ratio Adjustment

IVEware does not have the built-in capability to ensure the sum of imputed and observed detailed products equals the receipt total. After applying the inverse transformation to the imputed data, we do a simple ratio adjustment of the receipt total to the sum of products to ensure the sum of products balances to the total value of receipts.

4.4 Ratio Expansion and SRMI

SRMI imputes missing values consecutively by fitting a sequence of regression models while ratio expansion imputes using a single regression model with total receipts as the sole predictor. It takes into consideration that each product value is related to other items, including the product values for the other products reported for the establishment. Moreover, available auxiliary information, such as administrative records, can be incorporated into the model. In contrast, ratio expansion uses a single predictor (total receipts) on a univariate outcome (product value). The model fills-in plausible values in most cases, but it fails to include other covariates in the model, imputing (one variable at a time) without taking into account correlations between the variables (see Little, 2013.)

Ratio expansion and SRMI are both regression methodologies. The advantages of SRMI can be translated to ratio expansion by formulating an extension of ratio expansion in the context of SRMI. Little (2013) proposed models to extend ratio imputation to an iterative sequential regression procedure for the case in which the size measure (i.e., total receipts) is fully observed which could be implemented using IVEware. Table 1 displays a tabular comparison of the two regression models we implemented (SRMI and EXP) and ratio expansion in the context of SRMI (EXP-SRMI). The column labeled EXP-SRMI describes the model proposed by Little (2013) to extend the ratio imputation method to an iterative procedure that includes other covariates, as is done in the SRMI.

Table 1: Comparison of SRMI and Ratio Expansion Models

	EXP	SRMI	EXP - SRMI
Dependent Variable(s)	$y_j^p \big/ x_j^{1/2}$	$I(y_j^p = 0)$ y_j^p	$y_j^p \big/ x_j^{1/2}$
Intercept	No	Yes	No
Independent Variable(s)	$x_j^{1/2}$	x_j $I(y_j^{q\neq p} = 0)$ $y_j^{q\neq p}$ Other variables	$x_j^{1/2}$ $y_j^{q\neq p} \times x_j^{1/2}$ Other variables
Iterative	No	Yes	Yes

Note: y_j^p= value of product p for establishment j, $x_j =$ total value of receipts for establishment j

Table 2 displays reported, missing (denoted by **.**), and imputed product data for four establishments in the retail trade area. For illustration purposes, we selected establishments with data for six of the top ten products in addition to data for the remainder product (sum of all scarce products). Note that consistent with our data reformatting we impute only missing product values; products with reported zero values are considered valid and are not candidates for imputation. The ratio expansion imputes data for **all** missing product values. SRMI uses a model-based approach to imputation; **not all** valid products within the imputation cell are filled-in with nonzero values, as the specified model does not require nonzero imputes for all valid products with missing data. Whether a product has a zero or nonzero value depends on the imputed value for the indicator (i.e., $I(prodval_p = 0)$) according to the logistic regression model.

Table 2: Example of Missing and Imputed Values Using EXP and SRMI for Four Establishments in the Retail Trade Area

	Total			Products				Remainder
Establishment	**Receipts**	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**Product**
E1	2303	0	0	0	0	0	0	2303
E2	14378	0	0	0	1610	252	0	12516
E3	16030	.	.	.	.	.	.	.
E4	25130	.	.	.	.	.	.	.
				EXP				
E1	2303	0	0	0	0	0	0	2303
E2	14378	0	0	0	1610	252	0	12516
E3	16030	77	448	56	644	119	630	14056
E4	25130	126	707	84	1015	189	987	22022
				SRMI				
E1	2303	0	0	0	0	0	0	2303
E2	14378	0	0	0	1610	252	0	12516
E3	16030	0	0	0	1442	0	3521	11067
E4	25130	0	0	602	532	0	0	23996

5. Results

In this section, we summarize the evaluation methodology and results; the overall project results and recommendations appear in Knutson and Martin (2015.) Evaluating imputation methods usually involves generating artificial populations, repeatedly (typically upward of 1,000 repetitions) drawing samples and randomly generating missing values on the samples (see Schafer and Graham, 2002). We could then impute using our candidate methods, and compare the resultant fully imputed data sets on some predetermined statistical criteria. Designing such a study is a challenge with these data. Dr. Trivellore Raghunathan (University of Michigan visiting Summer at Census Scholar, 2014) recommended applying the four separate candidate imputation methods to the original dataset to create four complete populations.

Because we are multiply-imputing using the SRMI, our evaluation analysis requires setting-up the ratio expansion (and the two hot deck methods) in a multiple imputation framework. Rubin and Schenker (1986) proposed the Approximate Bayesian Bootstrap (ABB), as a method for creating multiple imputations when missing data are ignorable. The ABB is a two-stage resampling procedure. The first step is to draw a random sample of respondents with replacement followed by imputing values for missing data by taking draws from this sample. The ABB draws imputations from a resample of the observed data instead of drawing directly from the observed data. This extra step introduces additional variation making the ABB method into a “proper multiple imputation” according to Rubin's theory (1987). Multiple repetitions of this process produces multiple imputed datasets. We used the ABB to multiply-impute using the ratio expansion method.

We selected five industries from each of our data sets, each with at least two well-represented products and generated four complete populations by applying the four separate imputation methods: ratio expansion (EXP), nearest neighbor hot deck (HDN), randomized hot deck (HDR), and SRMI. We then randomly induced unit nonresponse in each population using the fitted unit level response probabilities reported in Ellis and Thompson (2015), repeating the process 50 times to produce 50 replicates. For each of the 50 replicates, we multiply-impute (100 times) using each method.

5.1 Evaluation Statistics

Rubin (1987) provided procedures for multiple imputation inference. With multiple completed datasets, we compute multiple point and variance estimates for a parameter of interest. In our study, the statistic of interest is the multiply-imputed estimated total for each product within each imputation cell.

Let *i* index imputation cells, *p* index products, *r* index replicates, *v* index the implicates, and *m* index the four separate imputation methods. The multiply-imputed estimated total for product *p* in imputation cell *i* from replicate *r* using imputation method *m* is, $\bar{Y}_r^{ipm} = \frac{1}{N_v}\sum_{v=1}^{N_v} \hat{Y}_{rv}^{ipm}$, 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 in the implicate and $N_v = 100$ is the number of implicates.

We use Rubin's formulae (Rubin, 1987) to compute the between, within, and total imputation variances. The within imputation variance for each replicate across the implicates is $\bar{\mathrm{U}}_r^{ipm} = \frac{1}{N_v}\sum_{v=1}^{N_v} \hat{V}\,(\hat{Y}_{rv}^{ipm})$, and $B_r^{ipm} = \frac{1}{N_v - 1}\sum_{v=1}^{N_v}(\hat{Y}_{rv}^{ipm} - \bar{Y}_r^{ipm})^2$ is the variance between the implicates. We use the between and within imputation variances to calculate the estimated total variance, $T_r^{ipm} = \bar{U}_r^{ipm} + \left(1 + \frac{1}{N_v}\right) B_r^{ipm}$.

The **fraction of missing information** (FMI) (Rubin, 1987) for product *p* in imputation cell *i* from replicate *r* obtained with imputation method *m* on the $N_v = 100$ implicates is calculated as $FMI_{\bar{Y}_r^{ipm}} = \left(1 + \frac{1}{N_v}\right)\frac{B_r^{ipm}}{T_r^{ipm}}$.

We define the **imputation error (**IE) of product *p* within imputation cell *i* obtained using imputation method *m* in replicate *r* as $IE_r^{ipm} = \bar{Y}_r^{ipm} - Y^{ip}$, where Y^{ip} is the trade area population total of product *p* within imputation cell *i*. The **absolute imputation error** (AIE) measures the magnitude of the IE and is computed as $AIE_r^{ipm} = \left|IE^{ipm}\right|$.

5.2 Comparative Analyses of AIE and FMI

We present comparative analyses of the AIE and FMI for the EXP and SRMI. Our goal is to assess the ratio expansion imputation and SRMI procedures in terms of AIE and FMI; not to compare the two methodologies in order to select the best imputation method (comparison of all methods is reported in Knutson and Martin, 2015). Recall that we assume product data is missing at random (MAR); for what follows, we make the following assumptions:

1. Imputation cells are independent
2. Replicates are independent (by simulation design)
3. Products are independent (true between industries, strong within industries for the "top two" products)

5.2.1 Comparison of the Imputation Error

We compare the magnitude of the imputation errors when imputing product data using EXP and the SRMI. Because the ratio imputation model imputes values that are "close" to the industry means, we expected the imputation errors obtained using EXP to be **smaller** than those obtained using the SRMI, and use sign and binomial tests (see Conover, 1999) to investigate this assumption. First, we conduct paired-sign tests by product and imputation cell, testing H_0: $AIE_{EXP}^{p} \leq AIE_{SRMI}^{p}$ against the one-sided alternative, where AIE^{p} is the absolute imputation error for product *p* in a given trade area population and industry cell obtained using either the EXP or SMRI method. The sign test examines the differences in paired-AIE values within replicates to determine if there is a larger than expected number of pairs with negative values. After conducting the sign tests for each product/imputation cell combination in the trade area population, we perform a binomial test on the counts of "successes" (null hypothesis is not rejected). We do this test to determine if the total number of trials where the absolute imputation error obtained using EXP is smaller than the absolute imputation error obtained using SRMI is larger than what would be expected under the null hypothesis. As before, we let *p* index products, *i* index imputation cells, and *r* index replicates.

Method for AIE comparisons

In each trade area population:

1) For each imputation cell and product, use sign tests[3] to test
   H_0: $AIE_{EXP} \leq AIE_{SRMI}$
   H_A: $AIE_{EXP} > AIE_{SRMI}$
   Procedure (Paired-data Sign Test):
   a. $DIFF\_IE_{ir}^{p}{=}\left|BIAS_{EXP,ir}^{p}\right| - \left|BIAS_{SRMI,ir}^{p}\right|$
   b. Let 1 = + = $DIFF\_IE_{ir}^{p} < 0$ $(AIE_{EXP} \leq AIE_{SRMI})$. A <u>small</u> number of +'s is evidence that the EXP imputed data tend to have higher absolute imputation error than the SRMI-imputed data for product *p* in imputation cell *i*.
   c. Conduct the sign test for each product within imputation cell (50 independent observations/replicates for each test.)
   Let $I_i^p = \begin{cases} 1 & \text{if the null hypothesis is not rejected} \\ 0 & \text{otherwise} \end{cases}$
2) Pool <u>all</u> product results in a trade area population and perform a binomial test to see if more than the expected 50% are exhibiting the tested behavior (valid given Assumptions 1 and 3 above.)

Table 3 summarizes the paired-data sign tests results comparing the absolute imputation error for data imputed using the ratio imputation method to the absolute imputation error for data imputed using SRMI for each trade area population. Recall that we suspect that

[3] The Wilcoxon signed-rank test is a more powerful test, assuming that the distribution of paired differences within replicates is symmetric. However, to keep the testing parallel with the testing discussed for FMI, we are opting to use the less powerful test.

the AIE for the EXP-imputed data tends to be smaller than the AIE for the SRMI-imputed data. To confirm this, we expect to see a large number of cells where AIE$_{\text{EXP}}$ tested as smaller than AIE$_{\text{SRMI}}$ and p-values that are larger than 0.10.

Table 3: Comparison of AIE for EXP and SRMI

		Number of Cells Where AIE$_{\text{EXP}}$ Tested as Smaller than AIE$_{\text{SRMI}}$			
		Trade Area Population			
Trade Area	**Trials**	**EXP**	**HDR**	**HDN**	**SRMI**
CON	**50**	29	24	21	20
MAN	**10**	9*	7*	7*	6
RET	**45**	43*	41*	39*	35*
SER	**19**	14*	11	12*	18*
WHO	**42**	42*	36*	38*	35*
FIRE	**13**	12*	11*	12*	9*
UTL	**19**	16*	15*	16*	14*

* Significant at α = 0.10

The results confirm the imputation errors obtained using EXP are **smaller** than their SRMI counterparts on the same data sets, regardless of trade area population. The exception is the Construction trade area, where a large number of cells exhibit this behavior, but not a sufficient amount to conclude that the pattern is not random.

5.2.2 Comparison of the Fraction of Missing Information

We compare the magnitude of the FMI when imputing product data using the EXP and the SRMI methods. Since the SRMI-imputed values tend to be more dispersed than their EXP counterparts are, we believe that they yield more realistic imputed data and that therefore the FMIs obtained using EXP should be **larger** than those obtained using the SRMI. Each FMI has a variance, thus we account for the variance of the FMIs in our analysis. We use paired-z tests within replicate to determine the validity of our hypothesis. In doing this, we want to account for the correlation due to the repeated measures design within replicates, which is not easily obtained. Instead, we conduct sensitivity tests with $\rho = 0$, 0.5, and 1. We then perform a binomial test to determine if the total number of trials where the FMI obtained using SRMI is smaller than the FMI obtained using EXP is larger than what would be expected under the null hypothesis.

Method for FMI comparisons

In each trade area population:

1) Within each replicate, perform a z-test for each product within imputation cell to test
$H_0: FMI_{EXP} \leq FMI_{SRMI}$
$H_A: FMI_{EXP} > FMI_{SRMI}$

Let $$z_{ir}^{*p} = \frac{FMI_{EXP,ir}^{p} - FMI_{SRMI,ir}^{p}}{\sqrt{Var(FMI_{EXP,ir}^{p}) + Var\left(FMI_{SRMI,ir}^{p}\right) - 2\rho SE(FMI_{EXP,ir}^{p})SE\left(FMI_{SRMI,ir}^{p}\right)}}.$$

Under $H_0, z_{ir}^{*p} \sim N(0,1)$. Since we are conducting a one-sided test, large positive values indicate that H_A is more appropriate. Reject H_0 at α = 0.10 if $z_{ir}^{*p} > 1.29$.

2) For each imputation cell and product, use sign tests to test
$H_0: FMI_{EXP} \leq FMI_{SRMI}$
$H_A: FMI_{EXP} > FMI_{SRMI}$

Procedure (Paired-data Sign Test):

a. Let 1 = + = indicator of failure to reject H_0 in replicate r for product p in imputation cell i (FMI$_{\text{EXP}}$ ≤FMI$_{\text{SRMI}}$). A small number of +'s is evidence that the EXP imputed data tend to have higher FMI than the SRMI-imputed data for product p in imputation cell i.

b. Conduct the sign test for each product within imputation cell (50 independent observations/replicates for each test).

$$\text{Let } I_i^p = \begin{cases} 1 & \text{if the null hypothesis is not rejected} \\ 0 & \text{otherwise} \end{cases}$$

3) Pool all product results in a trade area population. In this case, we are interested in determining whether a larger-than-expected count of product estimates has smaller FMIs when imputing with SRMI than EXP. Therefore, we subtract the counts obtained from 2.b above from the total number of cells and perform a binomial test to see if more than the expected 50% are exhibiting the tested behavior.

Table 4 summarizes the results comparing the FMIs within each trade area population. Recall that we suspect that the FMIs for the data imputed using EXP tend to be larger than the FMIs for data imputed using the SRMI. To confirm this, we expect to see a large number of cells where FMI$_{\text{EXP}}$ tested as larger than FMI$_{\text{SRMI}}$ and p-values that are smaller than 0.10. We are also concerned about the sensitivity of our z-tests at the replicate level to assumed levels of correlation and would like the results to be the same (or nearly the same) regardless of assumed value of ρ. As ρ approaches 1, we should see increasing counts of rejections of the null hypothesis that FMI$_{\text{EXP}}$ ≤FMI$_{\text{SRMI}}$.

Table 4: Comparison of FMI for EXP and SRMI

Number of Cells Where FMI$_{\text{EXP}}$ Tested as Larger than FMI$_{\text{SRMI}}$

		Trade Area Population											
		ρ=0				ρ=0.5				ρ=1			
Trade Area	Trials	EXP	HDR	HDN	SRMI	EXP	HDR	HDN	SRMI	EXP	HDR	HDN	SRMI
CON	50	37*	32*	37*	32*	37*	33*	37*	32*	38*	33*	38*	34*
MAN	10	8*	9*	9*	10*	8	9*	9*	10*	10*	9*	9*	10*
RET	45	0	0	0	0	0	0	0	0	0	0	0	0
SER	19	14*	13*	13*	15*	14*	14*	13*	15*	14*	14*	14*	15*
WHO	42	24	29*	29*	30*	25*	30*	30*	31*	25*	31*	31*	32*
FIRE	13	9*	10*	10*	10*	9*	10*	10*	11*	10*	11*	10*	12*
UTL	19	6	15*	15*	13*	6	15*	15*	13*	7	15*	16*	13*

* Significant at α = 0.10

As mentioned above, the counts displayed in Table 4 represent the counts of cells where the **alternative hypothesis** was accepted, although the form of the binomial test was equivalent to those used for the AIE tests. However, it is easier to interpret the table by displaying counts of accepted alternatives. Note that **none** of these results is sensitive to the assumed correlation coefficient. The tests results showed that the data imputed using the SRMI tend to have **lower** FMIs than the data imputed using EXP on the repeated measures pairs, with one exception, the FMI for the retail trade area EXP population. This is consistent with the results from the overall evaluation study; Knutson and Martin

(2015) noted that the SRMI often outperformed the ratio expansion and the two versions of the hot deck with respect to the FMI.

6. Concluding Remarks

In this paper, we presented the application of the model-based imputation method SRMI and a multiple imputation analogue of the ratio expansion to missing data imputation for Economic Census products. The ratio expansion procedure is an intuitive model, based on industry averages, that has the added advantage of automatically satisfying the additivity constraints. The SRMI imputes missing values consecutively by fitting a sequence of regression models conditioning on observed and imputed values which we implemented using the software application IVEware (Raghunathan et al., 2002).

We reported results of a simulation study of multiple product data imputation for 50 replicates of four separate populations. We presented analyses for two separate measures, the absolute imputation error and the fraction of missing information. Our results showed that the ratio imputation yields lower imputation errors than SRMI on repeated measured pairs, for most trade areas/population combinations. In contrast, SRMI imputation yields **lower** FMIs on the repeated measures pairs.

These results demonstrate that the model-based procedure SRMI is a feasible method for imputing missing product data. The SRMI creates multiple imputed datasets, which gives us a framework to compute the variance due to uncertainty in the data. Although using existing, well-tested software, with several built-in tools is ideal, it is important to acknowledge several challenges and limitations. Implementing SRMI requires preliminary analyses for model specification and selection in addition to identification of appropriate covariates for each trade area. Furthermore, we are concerned that it would be difficult to develop a fixed set of legal products by industry under NAPCS, which would greatly complicate the modeling procedures. Additionally it is not possible to guarantee the imputed data satisfy the additivity constraints; we do a ratio adjustment of total receipts to sum of products to ensure the balancing constraints hold. Moreover, the SRMI implementation includes a modeled remainder term (“All Other Products”) which would need to be allocated over other products in production. Finally, the SRMI is computationally intensive; running times are slow when compared to the ratio expansion and the two versions of the hot deck. This was a concern during the simulation study when multiply-imputing 50 replicates for each of four separate populations. However, run times should not be a problem in a production setting. On balance, we feel that the SRMI is a feasible methodology for product data imputation that could further be explored for other Economic Census variables and possibly other economic surveys.

Acknowledgements

The authors thank all project team members; Dr. Raghunathan for his advice on model building, simulation, and implementation; Katherine Jenny Thompson for her collaboration on the comparative analyses; and Cheryl Merkle, Yves Thibaudeau, and Martin Klein for their reviews and comments.

References

Buuren, S. V., and Groothius-Oudshoorn, K. 2011. "Mice: Multivariate Imputations By Chained Equations In R." *Journal of Statistical Software*: 45(2): pp. 1–31. http://www.jstatsoft.org/v45/i03/

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

Conover, W.J., 1999. Practical Nonparametric Statistics. 3rd ed. New York: John Wiley and Sons, Inc.

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

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

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

Little, R. J. A. 2013. "A Multivariate Extended Ratio Model for Multiple Imputation of Item Nonresponse in Establishment Surveys". US Census Bureau Manuscript.

Magee, L. 1998. "Improving Survey-Weighted Least Squares Regression." *Journal of the Royal Statistical Society (B)*: 60(1): pp. 115-126.

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.

Raghunathan, T. E., Solenberger, P., and Van Hoewyk, J. 2002. *IVEware: Imputation and Variance Estimation Software User Guide*. University of Michigan, Survey Methodology Program Web site: http://www.isr.umich.edu/src/smp/ive/

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.

Schafer, J. L. and Graham, J. W. 2002. Missing Data: Our View of the State of the Art. Psychological methods, 7(2):147-177.

Sigman, R. S. and Wagner, D. 1997. Algorithms For Adjusting Survey Data That Fail Balance Edits. *Proceedings of the Section on Survey Research Methods*: American Statistical Association.

Su, Y., Gelman, A., Hill, J., and Yajima, M. 2011. "Multiple Imputation With Diagnostics (Mi) In R: Opening Windows Into The Black Box." *Journal of Statistical Software*: 45(3): pp. 1–67. http://www.jstatsoft.org/v45/i02/

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

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