Hot Deck Imputation of Multinomial Distributions When There Are Fewer Donors than Recipients

Katherine Jenny Thompson, U.S. Census Bureau[1]
Rebecca Andridge, the Ohio State University
Laura Bechtel, U.S. Census Bureau
Natasha McCarthy, U.S. Census Bureau

Detailed breakdowns of totals items are collected in surveys. Detail proportions can vary greatly by sample unit, and the multinomial distributions can likewise vary by imputation cell. Consequently, although it might be feasible to develop viable parametric imputation models for the total, it is challenging for the collective set of detail items. Instead, a common practice is to use some form of hot deck imputation to match donor and recipient records, then impute the donor's complete set of proportions. Nearest neighbor imputation is useful when the set of proportions is correlated with unit size. This approach preserves the correlation between the detailed items within imputation cell, as long as the number of donors is greater than or equal to the number of recipients. Unfortunately, this condition often does not hold in practice. Collapsing imputation cells is not an attractive alternative. We explore unrestricted usage of the donor records in the original cell versus the use of a random draw from the donor record's multinomial distribution via a limited simulation study.

Key words: nearest neighbor hot deck imputation, multinomial distribution

1. Introduction

Detailed breakdowns of totals items are collected in surveys. The detail proportions can vary greatly by sample unit, and their multinomial distributions may be related to different predictors than the associated total. This creates two separate but related missing data challenges: (1) to develop viable imputation models for the total and (2) to develop viable imputation models for the *set* of associated detail items. A preliminary step for many statistical imputation methods is to partition the sample into disjoint cells that either contain units with the same response propensity or have the same cell means for the key characteristics (Kalton and Kaspryzk, 1986). With imputation, it may be preferable to develop cells whose covariates are related to the conditional expectation of the variable(s) of interest, given the imputation model (Haziza and Beaumont, 2007). Covariates are categorical – continuous variables can be "binned" into size categories – and imputation cells are formed by cross-classifying the selected covariates or nesting the size category cells within the more definitive classifier (e.g., size category within industry). Having subdivided the sample, the next step is to determine the appropriate imputation models for each outcome variable. Models are developed from the respondents' data in the imputation cells (donors) and applied to create "substitute values" for the nonrespondents' (recipients) missing data. In practice, the complete set of defined imputation cells may not be used. Instead, "ad hoc methods are often applied to collapse small imputation cells[2]" (Fang, Hong, and Shao, 2009). Cell collapsing procedures generally attempt to pair cells with common response propensities or common conditional expectation (or cell means) to the maximum extent possible.

[1] Any views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

[2] Imputation cells that contain fewer donors than recipients or whose donor count is less than a predetermined threshold.

Further complicating the situation, the appropriate imputation cells for the total and for the set of details may differ. Consider the fictional pair of industries depicted below. Both industries have a similar range of businesses in terms of sales, although the jewelry industry distribution is more skewed. Each business is requested to report the amount of their total sales obtained by three mutually exclusive sources. The multinomial distributions of these sets of details differ by the industry and business size category as measured by sales.

Industry	**Store Size**	**Sales Range**	**True Underlying Multinomial Distribution**		
			Inside Store Credit Card	Inside Store Cash	Online Purchase
Jewelry	Large	$300,000 +	75%	25%	0%
	Medium	$25,001 - $300,000	40%	40%	20%
	Small	$1 - $25,000	50%	50%	0%
Books	Large	$100,000 +	50%	5%	45%
	Medium	$20,001- $100,000	50%	20%	30%
	Small	$1 -$20,000	5%	20%	75%

If sales is a good predictor of response propensity (e.g., the larger stores are more likely to report a value then the smaller stores) or if the mean value of sales greatly differs between within-industry size category, then it might be worthwhile to further subdivide the industry-level imputation cells. Even so, collapsing to the industry level for imputation when insufficient cell-level donor (respondent) records are available would not be entirely unpalatable. Indeed, it might even be possible to collapse industry cells that have similar sales distributions without overly biasing the estimates. Given that sales is a continuous variable, it may be possible to approximate its distribution with a parametric model. In this case, missing values could be imputed via mean imputation or with multiple imputation with random draws from the specified parametric model. If strongly related covariates are available, other imputation options could be considered as well, such as pseudo empirical likelihood estimators or regression or ratio imputation with single or multiple imputation.

In contrast to sales, the multinomial distributions of source of sales are quite different in the six distinct categories. In the jewelry industry, neither the largest nor the smallest stores offer an online purchasing option; however, the breakdowns between credit card and cash purchases are very dissimilar. The medium-size stores offer an online option, and the percentages of credit and cash sales are not very different from that of the small store category. However, collapsing the medium and small store categories for imputation would be ill-advised, as it would induce an artificial probability of online sales in the small store category, as would collapsing the large and medium size store category. In the book industry, the multinomial distributions in the three size categories are very different, and there is no way to combine the size classes that would correctly preserve any distribution. Moreover, collapsing by similar size category across industry would be unwise.

With these multinomial distributions, a response is defined as providing "reasonable" information on the ***complete set*** of details; for example, requiring the details to add to the total within a pre-specified tolerance (e.g.± 10%). Response propensity might be related to different factors for the set of detail items than for the total, or the relationship might be less direct. Small businesses might not provide the detailed breakdowns because the information is not part of their recordkeeping; in this case, business size is a good predictor

of response. However, response might be a consequence of the placement of the inquiry on the questionnaire and could be missing at random (MAR) or missing completely at random (MCAR). And of course, the probability of response might be directly related to the queried information and would therefore be missing-not-at-random (MNAR).

Moreover, there are few viable imputation model options for the collective set of detail items. Instead, a common practice is to use some form of hot deck imputation to match donor and recipient records, then impute the donor's complete set of proportions (Andridge and Little 2010, Beaumont and Bocci 2009). Nearest neighbor imputation is useful when the set of proportions is correlated with unit size; random hot deck imputation could be otherwise useful. Either approach preserves the correlation between the detailed items within imputation cell and can yield (nearly) unbiased estimates, as long as the number of donors is greater than or equal to the number of recipients. Unfortunately, this condition often does not hold in practice, and for the reasons outlined above, collapsing imputation cells is not an attractive alternative. However, using a small number of donors in hot deck imputation can create very inefficient estimates, as the same donors may be used multiple times. The effect of donor overuse in imputation cells with proportions of donor records can be especially pronounced with unrestricted nearest neighbor hot deck imputation since a single donor can be used several times rather than use a "more distant" neighbor.

In this paper, we explore alternative variations of nearest neighbor hot deck imputation, considering unrestricted usage of the donor records in the original cell versus the use of a random draw from the donor record's multinomial distribution. This research is motivated by the Economic Census conducted by the U.S. Census Bureau discussed in Section 2. We examine the statistical properties of the alternative imputed estimates over repeated samples from a census, independently randomly inducing response in the same population, yielding R replicates. Using a census greatly simplifies the simulation and eliminates confounding with the effects of different sample designs on the imputed estimates. As discussed in Section 3, we utilize a multiple imputation (MI) estimator, using Approximate Bayesian Bootstrap (ABB), a non-Bayesian method that approximates a Bayesian procedure (Rubin and Schenker 1986; Rubin 1987).

Section 2 describes our motivating problem. Section 3 presents the variations of nearest neighbor hot deck imputation explored in this paper. Section 4 presents a limited simulation study using data generated from a theoretical distribution and modeled using historic data from selected industries from the 2012 Economic Census. We conclude in Section 5 with general observations and ideas for future research.

2. The Motivating Problem

The Economic Census collects information on the revenue obtained from product sales. Often, product descriptions are quite detailed, and many products are mutually exclusive. The reported product dollar values are expected to sum to the total receipts reported earlier in the questionnaire (within a tolerance). Total receipts is available for each unit, and the set of reported product values are the associated details. Fink, Beck, and Willimack (2015) report that the same handful of products (≈ 3) are reported by all establishments in an industry, with random variation in reporting for the remaining products provided on the industry questionnaire. Auxiliary data on products are not readily available, and total receipts are often weakly related to the distribution of detail items. Consequently, legitimate missing product values occur frequently and product distribution nonresponse is quite high.

Although the same product can potentially be produced in different industries under the North American Product Classification System (NAPCS), product reporting is intertwined with industry classification. Certainly the product distributions will differ between industries, even when the same products are reported. As a result, imputation cells cannot be collapsed beyond the 6- to 8-digit industry category depending on the sector and it is undesirable to drop the unit type classification. The 2017 Economic Census will be the first incidence of product-reporting under NAPCS, so no historical data collected under this classification system are available yet.

In the 2017 Economic Census, missing product data will be imputed using hot deck imputation. Hot deck imputation provides a flexible approach to dealing with missing data that retains multivariate relationships without making explicit parametric model assumptions. Instead, hot deck methods impute missing values to recipient units using reported values (donors) from a similar unit.

Ellis and Thompson (2015) present the empirical response propensity analysis used to determine the hot deck imputation cells. In general, they reported very few covariates that are predictive of product distributions besides industry, although product distributions within the same industry do often differ by unit type (single or multi-unit establishment where a single-unit establishment owns or operates a business at a single location and multi-unit establishments comprise two or more establishments that are owned or operated by the same company). In our own exploratory data analysis, we have found no evidence against a missing-at-random (MAR) response mechanism within the designated imputation cells. With business surveys, unit size is often highly correlated with response; larger units are more likely to respond than smaller units (Thompson and Oliver 2012; Thompson, Oliver, and Beck 2015). However, we were unable to find a similar relationship between unit size and product reporting.

Thompson and Liu (2015) gives an overview of the large scale research project conducted to determine an imputation method for Economic Census products: more details are provided in Ellis and Thompson (2015), Tolliver and Bechtel (2015), Bechtel, Morris, and Thompson (2015), and Knutson and Martin (2015). Because the majority of establishments in an industry report often tend to report the same products, these studies focused on the statistical properties of the alternatively imputed estimates of the two most frequently reported products per industry. In other words, none of these studies examined the imputation cell-level multinomial distributions of products. Given the symbiotic relationship between industry classification and product reporting, it is not unreasonable to assume that there is a single multinomial distribution of products within each imputation cell, in contrast to individual establishment-level multinomial distributions perhaps related to unit size.

For these reasons, we sought a hot deck imputation method that minimizes or completely sidesteps cell collapsing while producing efficient estimates in terms of precision or coverage and accurate estimates in terms of bias, at least for the well-reported products. We considered a variety of options for selecting donors. Initially, we considered predictive mean matching across imputation cells as an alternative to cell collapsing. We attempted to use receipt totals and payroll totals to predict the proportions in each of the two most commonly-reported products (using simulated data that mimic the distributions seen in historical data), with the goal of using these models to create predictive means on which to match. However, the predictive ability of these models was extremely low ($R^2 < 0.1$ for

more than 90% of industry/unit type combinations studied). Quality matches could not be made, and there was no advantage to using the predictive mean instead of a nearest neighbor approach based on receipt totals. Thus we dismissed predictive mean matching as a viable option and settled on a nearest neighbor approach (based on total receipts), despite having reservations about a statistical relationship between unit size and detail distribution in our studied datasets.

The Economic Census implements both random and nearest neighbor hot deck imputation, depending on the industry. Either method would yield the same expected product distributions under the assumption of a single multinomial distribution in an imputation cell. One advantage of random hot deck imputation is that there is more control over the donor base. For example, restrictions can be placed on how many times a donor is used and donors can be randomly drawn from the donor pool with or without replacement. On the other hand, random hot deck imputation can add noise to imputed estimates, obscuring differences with small sample or respondent cell sizes. Alternatively, nearest neighbor hot deck imputation produces less noisy estimates and would be less biased than random hot deck imputation if indeed unit size were related to the product multinomial distribution. However, one disadvantage of nearest neighbor hot deck imputation is that adding restrictions on the usage of donors yields donors that can be very far away from the recipient. Accordingly, the Economic Census hot deck imputation methodology does not impose restrictions on how many times a donor was used. This limits imputation cell collapsing but may fail to preserve the microdata distribution if the donor count is small (say, less than 5). Historically, the product response rate tends to be quite low, so imputation cells with small number of donors are a frequent occurrence. Consequently, we focus on nearest neighbor hot deck imputation, leaving other variations of hot deck imputation for future research.

3. Imputation Methods

Hot deck nearest neighbor imputation uses auxiliary variable(s) available for both donors and recipients. Ideally, these variables should be highly correlated with the variables that are being predicted. A distance function determines the distance of each donor from each recipient; there are several different distance functions used in practice, and the selection generally depends on factors such as the number of available variables, the number of items to be imputed, and the form of the predictive relationship. The donor that is the smallest distance from the recipient is selected for imputation. In our implementation, we had very little auxiliary data to use for a distance function, and we used the Euclidian distance (absolute value of the difference) between each recipient and each donor's total receipts values.

We considered five variations of nearest neighbor hot deck imputation. Each variation selects a "donor" multinomial distribution and applies these percentages to the recipients' value of total receipts as described in Section 1. The procedures are outlined below.

Collapse: We collapsed the imputation cells to the highest category when there were fewer than *five* donors in at least one of the subcategories. In our application, industry is the highest category, and each industry is further subdivided by unit type. The nearest neighbor donor is selected using the Euclidean distance within the collapsed cell.

Unrestrict: Essentially the same procedure, without cell collapsing. This allows unrestricted use of donors.

Draw: Uses a random draw from the donor record's multinomial distribution instead of the donor distribution, where the donor is selected from the same imputation cell (no collapsing)

Cluster-Draw: When an imputation cell contains fewer than five donors, average the multinomial distributions of all donor records in the cell and randomly draw the multinomial distribution from the cluster. Each average is obtained as product values (averaged over donors) divided by the total receipts (averaged over donor). Use the **Draw** procedure when there are at least five donors in an imputation cell.

Cluster-All: Obtain averaged the multinomial distribution of the cluster of the five nearest neighbor donors when the imputation cell contains at least five donors; otherwise, use the average of the available donor records in the imputation cell.

Using a draw instead of directly applying the donor ratio should variability to the imputed data. We expect that this should improve coverage, as the **Unrestrict** procedure can overly smooth the data when the same donor is used multiple times.

The two uncollapsed procedures (**Unrestrict** and **Draw**) can artificially increase the probability of a zero value of a rarely reported product if the actual probability is small (close to zero) and the donor record contains a zero for the item. By selecting a cluster of nearest neighbor donors and using the average value within "details" to obtain the donor probabilities, we hope to alleviate this problem.

With each of these methods, we utilize a multiple imputation (MI) estimator, using the Approximate Bayesian Bootstrap (ABB), 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 as the nearest neighbor hot deck imputation base and applying each studied imputation method using the resampled donor pool. Our applications create 20 multiply imputed data sets per replicate; the implicate estimates were combined to estimate total values of each studied item along with their associated variance estimates. Because we are restricting this research to a census, we select simple random samples with replacement. If these imputation methods were applied to sample survey data, some consideration on resampling would need to be given for unequal probability sample signs; see Andridge and Little (2009) for example. In a similar vein, the appropriate MI variance estimator for a sample survey would need to account for both the sample survey design and the imputation/nonresponse variance and the combining rules would need to be modified from those provided in Rubin (1987); for example, see Zhou, Raghunathan, and Elliot (2012).

4. Simulation Study

4.1. Simulation Study Design

To evaluate the proposed nearest neighbor hot deck imputation variations, we created three different simulated populations, using the models and parameters provided in Table 1.

The first population is designed to work well with nearest neighbor imputation when the Euclidean distance measure is defined by value of total receipts. Population 1 comprises two disjoint imputation cells (no overlap in total receipts values) with entirely different multinomial distributions of five products in each imputation cell. In each cell, total receipts was generated from a gamma distribution, with shape and scale parameters selected to ensure no overlap in values between cells. Product proportions were randomly drawn from the multinomial distributions shown in Table 1, so that the expected value of each product is the same within imputation cell (i.e. no relationship between unit size and expected proportion of a product). Product values (in $1000) were obtained for each simulated unit by multiplying each product proportion by the unit's simulated value of total receipts and rounding the product to a single digit. The second two populations were modeled from 2012 Economic Census data. In these populations, products 1 through 4 are specific items, and product 5 contains the balance of the reported product values. Here, the imputation cells are defined by unit type, not total receipts, so that the nearest neighbor donor for a recipient record in a collapsed cell may be drawn from a different cell. Total receipts was generated within **imputation cell** from a lognormal distribution, using sample moments from the original census data files. In both of these industries, there is some overlap in the size of total receipts between the imputation cells. Figure 1 presents side-by-size boxplots of the distribution of total receipts by imputation cell within simulated industry.

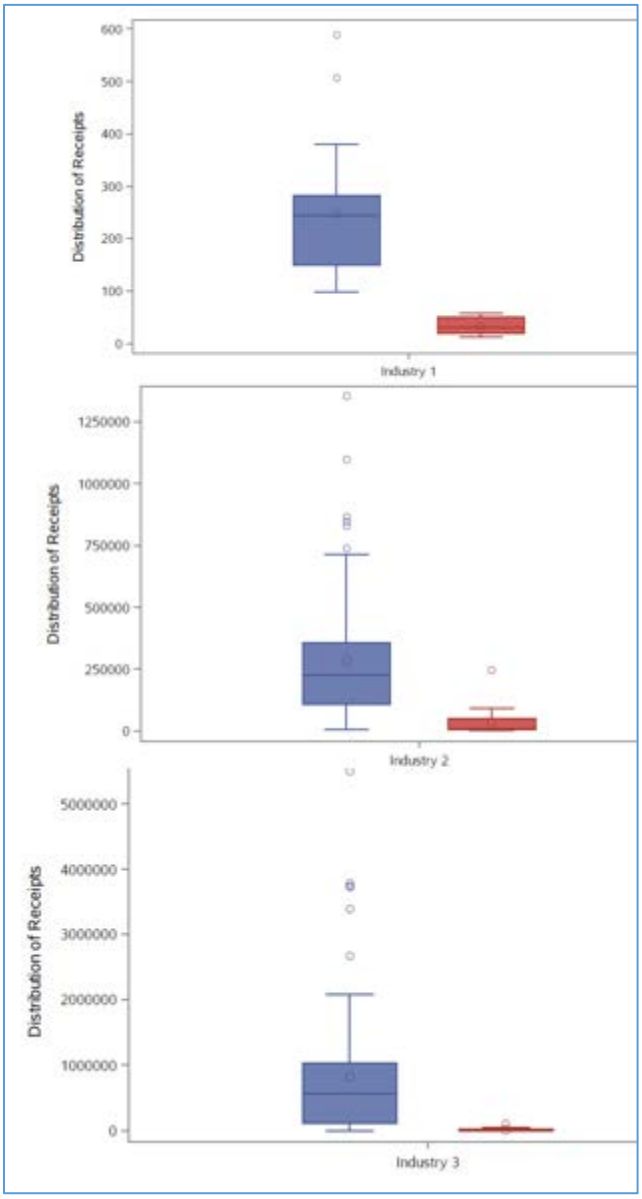


Figure 1: Side-by-Side Boxplots of Distributions of Total Receipts by Imputation Cell within Industry with Imputation Cell 1 in Blue (Left) and Imputation Cell 2 in Red (Right)

In contrast to the Population 1 procedure, the unit-level multinomial distributions were not drawn from a single multinomial distribution within imputation cell. Instead, the unit-level establishment proportions were obtained from the original Economic Census data. These proportions were randomly assigned to each simulated unit and the same procedure described above was applied to obtain final product value estimates. This added noise to the simulated data, although the product proportions for the first two products tend to be fairly stable within imputation cell. Table 1 presents the mean and median cell proportions by imputation cell.

Table 1: Simulation Models and Parameters by Imputation Cell

Industry	Imputation Cell	Total Receipts		Average Cell Percentages (Median in parenthesis)				
		Model	Parameters	Product 1	Product 2	Product 3	Product 4	Product 5
1 (Imagined)	1	Gamma	α=4, β=0.6	34.62 (35.01)	35.66 (37.08)	13.6 (12.95)	10.26 (9.89)	5.86 (6.14)
	2	Gamma	α=4, β=0.1	67.38 (66.67)	6.38 (6.45)	4.8 (5.26)	16.04 (15.79)	5.4 (5.88)
2 (31211100)	1	Lognormal	μ = 11.30, σ=1.12	25.47 (25.53)	21.82 (22.20)	13.18 (0.73)	16.83 (1.03)	22.71 (21.86)
	2	Lognormal	μ =8.66, σ=1.28	20.09 (0.00)	10.89 (0.00)	20.83 (0.00)	18.66 (0.00)	29.53 (11.69)
3 (51121000)	1	Lognormal	μ = 8.48, σ=2.14	42.55 (40.00)	22.96 (4.13)	13.02 (0.00)	6.67 (0.00)	14.8 (8.00)
	2	Lognormal	μ =6.54, σ=1.77	61.71 (64.90)	5.47 (0.00)	0.58 (0.00)	3.06 (0.00)	29.19 (7.30)

The product distribution in the first industry is fairly homogeneous by design as the product distributions were randomly generated from multinomial distributions. Industries 2 and 3 are quite different. With Industry 2, units in the first imputation cell tend to report ***either*** Product 1 or Product 2 ***or*** Product 3, and the product averages are influenced by a few large cases. Units in the second imputation cell tend ***not*** to report Products 1 through 4. In this case, using the cell averages to estimate the underlying product distribution yields misleading results. In Industry 3, Product 1 is reported by a high proportion of units, but Products 2 through 4 are very rarely reported in the second imputation cell. Thus, for Industries 2 and 3, collapsing imputation cells for nearest neighbor imputation should result in overly large (and overly frequent) imputed product values.

Table 2 provides the sample sizes under full response and proportion of units that reported a non-zero value for the product. All of the industries have more sample in their 1st imputation cells. Furthermore, the proportion of units in the 1st imputation cell is substantively larger than that in the other cell in Industries 2 and 3. Recall that overlap in nearest neighbor units is likely in collapsed cells in these two populations. Consequently, we expect increases in relative bias in the ***second*** imputation cell with cell collapsing procedures over the other considered nearest neighbor hot deck variations.

Table 2: Population Sample Sizes and Percent Units Reporting a Non-Zero Value for the Product

Industry	Imputation Cell	Number of Units	Establishment Counts				
			Product 1	Product 2	Product 3	Product 4	Product 5
1 (Imagined)	1	25	100.00	100.00	100.00	100.00	100.00
	2	15	100.00	100.00	93.33	100.00	86.67
2 (312111000)	1	102	63.73	61.76	64.71	62.75	71.57
	2	22	27.27	18.18	31.82	36.36	63.64
3 (51121000)	1	82	82.93	54.88	42.68	42.68	79.27
	2	19	89.47	15.79	10.53	10.53	63.16

To avoid confounding results with a specific sample survey design, we treat each population as a census. Given population POP_i (i = 1, 2, 3), we independently repeat the following procedure in replicate s (s =1, 2, …,1240):

- Induce "product nonresponse" using a MCAR response mechanism with response propensity (r) = 0.25, 0.50, and 0.75, resulting in three different sets of respondents per replicate and population labeled $POP_{i,r,s.}$ We deleted all samples that contained one or more imputation cells with zero donors to prevent the need for a back-up method and to simplify the interpretation of the results;
- In each $POP_{i,r,s}$, resample the donors using the Approximate Bayesian Bootstrap with 20 implicates;
- Apply nearest neighbor hot deck imputation variation v (v = 1, 2, …,5) to the ABB implicate within $POP_{i,r,s}$ to obtain the complete data set of products. Note that each unit has a nonmissing value of total receipts;
- For each product, obtain the multiply-imputed (MI) estimate and variance estimate of the total;
- For each pair of products, obtain the MI correlations and associated variance estimates, using the Fisher's z transformation with no bias adjustment available in the SAS PROC CORR.

Of course, a MCAR response mechanism is not realistic. However, using the same response propensity in each imputation cell greatly simplifies the interpretation of the results. Our focus is on the results with response propensities of 50% and 25%; the results obtained using a 75% response rate serve as a baseline. Cell collapsing is very infrequent with the 50% product response rate in all industries, as is likewise rare with a 25% response rate in Industry 2. When the product response rate is 25%, collapsing occurs frequently in Industry 3, and a high proportion of samples collapse imputation cells in Industry 1 with a 25% product response rate. In that industry, however, the values of total receipts do not overlap between imputation cells, so that the selected nearest neighbor in the "collapsed cell" will be from the recipient's imputation cell, essentially making the collapsing procedure equivalent to the **Unrestrict** procedure.

We assume that the primary function of the survey is to produce estimates of totals. Consequently, minimizing the bias is very important. That said, an imputation method that is known to be approximately unbiased over repeated samples can certainly yield biased estimates for rare characteristics. Moreover, the bias of any hot deck imputation procedure will be a function of the donor-to-recipient ratio (preferably greater than one) and by the number of sampled units in the imputation cell. Lastly, the performance of any imputation method is related to the response rate. For example, if the nonresponse rate is quite high,

then the variance estimates may be quite large. We use 95% confidence interval coverage rates to jointly assess the combined precision and accuracy of the imputed totals.

An advantage of hot deck imputation over mean or ratio/regression imputation is that it can preserve the underlying distribution of the microdata. If the imputed microdata are being used in other analyses such as regression modeling, then it is important to preserve the correlation structure. Again, with very small samples and high nonresponse, it may be impossible to exactly recreate through imputation the population level of correlation for each pair of items. However, it is important to preserve the sign of the correlation coefficients (negative or positive) along with the "significance" (zero or nonzero).

We compute the relative bias and 95% confidence interval coverage of each nearest neighbor hot deck variation *v* estimate of Product *p* (p = 1, 2, 3 in Population 1; p = 1,2,…,5 in Populations 2 and 3) in Population *i* under response propensity *r* ($\hat{Y}_{irv}^{p}$) as

Relative bias $RB\left(\hat{Y}_{irv}^{p}\right) = \frac{\sum_{s=1}^{1240} \hat{Y}_{irvs}^{p}}{Y_i^p} - 1$

Coverage Percentage of samples whose 95% confidence intervals (normally distributed) constructed with the MI estimate and MI variance contain the true population value for product *p* (Y_i^p).

Each industry estimate is computed overall and by imputation cell. With the exception of the estimates obtained with the collapsed cell procedures, we expected to attain approximately the same percentage of bias for each estimate within population and response propensity. Moreover, we did not expect to see increases in relative bias under cell collapsing for the larger-sample imputation cell estimates. Intuitively, we expected to see increased coverage rates when random draws are employed instead of direct application of the donor ratio.

To evaluate the performance of the hot deck method on correlation, we categorized the non-zero Fishers-z transformed correlations ($\alpha = 0.05$) by sign (positive or negative) and grouped the remaining pairs into the "not correlated" category, then repeated the same classification procedure on the multiply imputed Fishers-z transformed correlations for each imputation method. Table 3 provides summary counts of population correlations. Often, positive or negative correlations are very weak, albeit statistically significant.

The Appendix provides scatterplot matrices for each pair of items within industry and imputation cell. Within industry, there are distinct differences in the paired item relationships between the imputation cells besides the obvious discrepancy in number of units. In imputation cell 1, total receipts is often a viable predictor of a product, and there are occasional linear relationship between a pair of products. In contrast, there is very little evidence of any linear relationship for the majority of paired items in imputation cell 2, regardless of industry – with the possible exception of total receipts and Product 1. None of the plots provide any evidence of a negative association between any two items. Consequently, we consider any change from positive or uncorrelated to negative to be extremely misleading. Given the small number of donor units in many of imputation cell 2 (all industries), we expect a fair amount of "switching" of correlation status in these cells (from uncorrelated to positive); we are more concerned from a practical perspective about similar switching in imputation cell 1 (all industries).

Table 3 provides summary counts for the correlations in the population (total of 15 correlations) by imputation cell.

Table 3: Summary Counts for the Population Correlations

Population	Imputation Cell	Positive Correlation	Negative Correlation	Uncorrelated
Industry 1	1	14	0	1
	2	10	0	5
Industry 2	1	8	2	5
	2	2	0	13
Industry 3	1	10	0	5
	2	2	0	13

To compare the imputed distributions to the population in terms of preserving linear relationship between items, we constructed the contingency table shown in Figure 2 by imputation method at the industry and imputation cell levels.

		Population		
		Positive Correlation	Negative Correlation	Uncorrelated
Sample	Positive Correlation			
	Negative Correlation			
	Uncorrelated			
	Not Applicable			

Figure 2: Sample Contingency Table for Assessing Correlation Effects

Occasionally multiply imputed correlations were not available because a non-zero value for one or more products was not present in any of the randomly resampled donor records in an ABB implicate. Replicates that contain at least one missing MI correlation are classified as "not applicable" in the contingency table. If the imputation pattern matched exactly for a particular replicate, the off diagonal cells and the entire Not Applicable row would be equal to zero.

4.2 Results

4.2.1. Totals

Figures 3 through 5 present the relative biases obtained for each procedure for product response rates of 75%, 50%, and 25% respectively by imputation cell within industry. The effects of cell collapsing on relative bias depends on three factors: (1) the number of units in the larger (uncollapsed) imputation cell relative to the number of units in the smaller imputation cell; (2) the proximity of the nearest neighbor in the collapsed cell; and (3) the product response rate. In the scenarios where there are at least as many donors as recipients (product response rates of 75% and 50%), cell collapsing occurs very infrequently -- if at all. Recall that imputation cell collapsing is rare in Industry 2, even with a product response rate of 25%, whereas it occurs frequently in the other industries.

Our simulation is designed to mimic frequently encountered scenarios in economic programs, with one imputation cell containing more sample units than the other. In this simulation, all units have the same weight, so that the larger imputation cell measures are

almost identical to the overall population (industry) measures, which are consequently omitted. If the imputation cells contain approximately the same number of units or units are unequally weighted, then the aggregate results could differ from those of either imputation cell.

In all Figures, black font indicates a deterministic imputation method and red indicates a draw from a multinomial distribution. Nearest neighbor hot deck imputation variations are indicated as follows: Square = Collapse; Circle = Unrestrict; Triangle = Draw Plus = Cluster-draw; Cross = Cluster-all. Product Response Rate is abbreviated to "PRR."

In Industry 1, there is no overlap in the value of total receipts between imputation cells. Consequently, the donor nearest neighbor in the collapsed cells will be selected from the recipient units' original imputation cell, theoretically correctly preserving the imputation cell product distributions. In this scenario, the **Collapse** and **Unrestrict** procedures are essentially the same. Of course, it is not generally true that these two procedures would be this similar. Usually, when donor pools are collapsed, the donor could be drawn from the other imputation cell (e.g., random hot deck, nearest neighbor with a different distance function). In our Industry 1 scenario, the **Draw** and **Cluster-draw** procedures are also essentially equivalent as there is a very low probability of any zero-valued reported product. Accordingly, the comparable performance of the five considered methods when there are at least as many donors as recipient is reassuring. The **Collapse-all** procedure tends to reduce the bias over the other methods when there are fewer donors than recipients. This is an expected artifact of the simulation design, as the draws are made from unbiased estimators and averaging increases their precision. In Industries 2 and 3, the values of total receipts overlap between imputation cells. Collapsing the imputation cells is expected to increase magnitude of the relative bias especially for the rarely reported products because the donor nearest neighbor might not be selected from the recipient's imputation cell. Of course, the probability of this increases as the product response rate decreases.

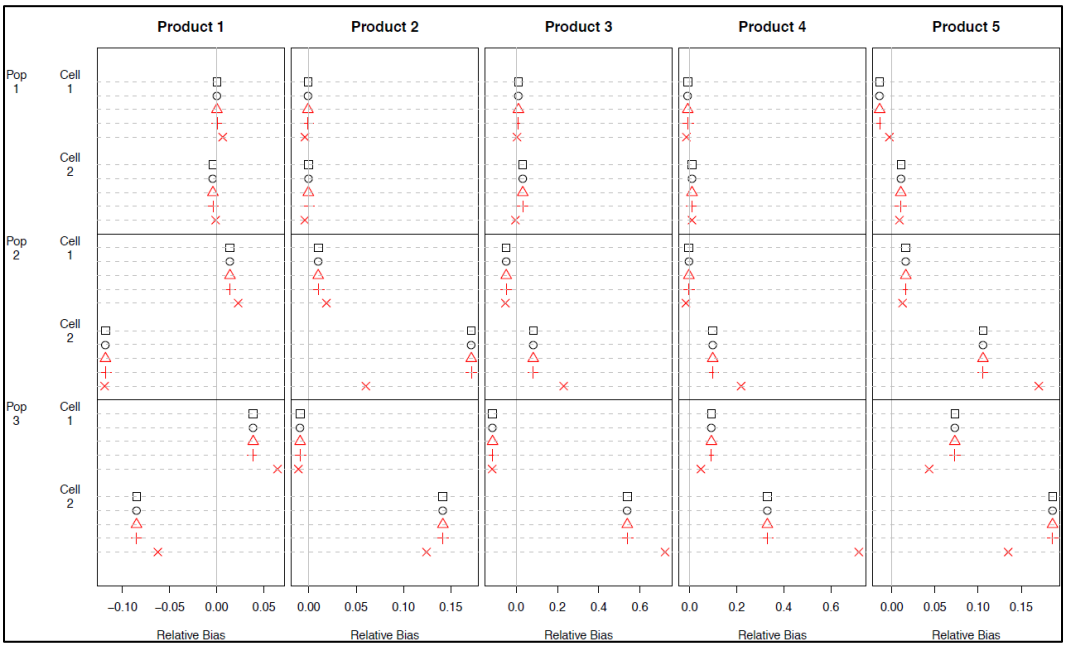


Figure 3: Relative Bias Results for All Products by Industry and Imputation Cell with PRR = 75%

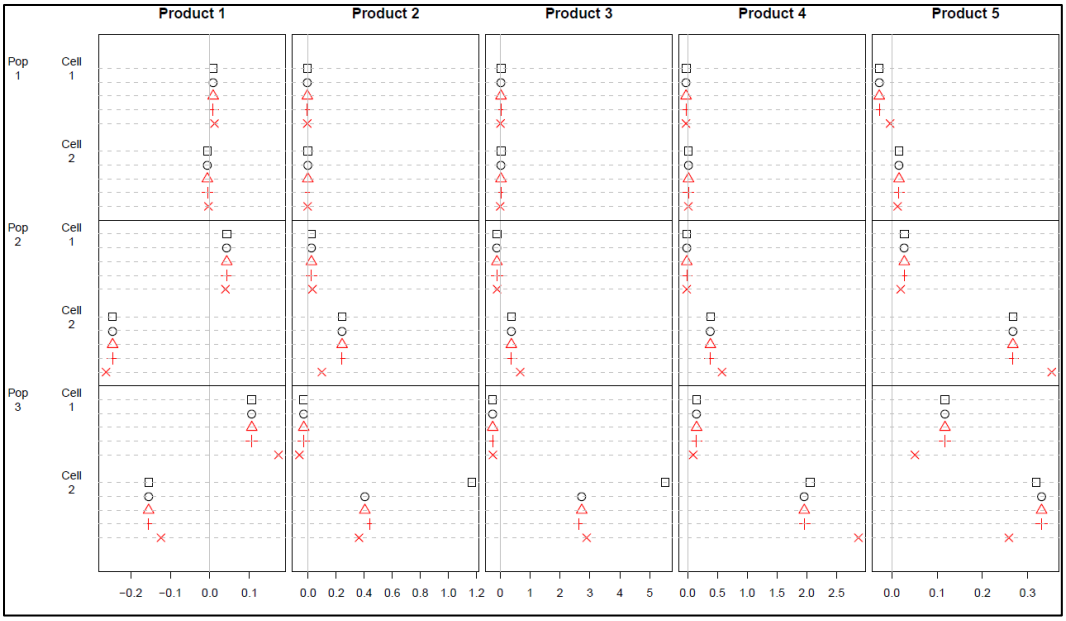


Figure 4: Relative Bias Results for All Products by Industry and Imputation Cell with PRR = 50%

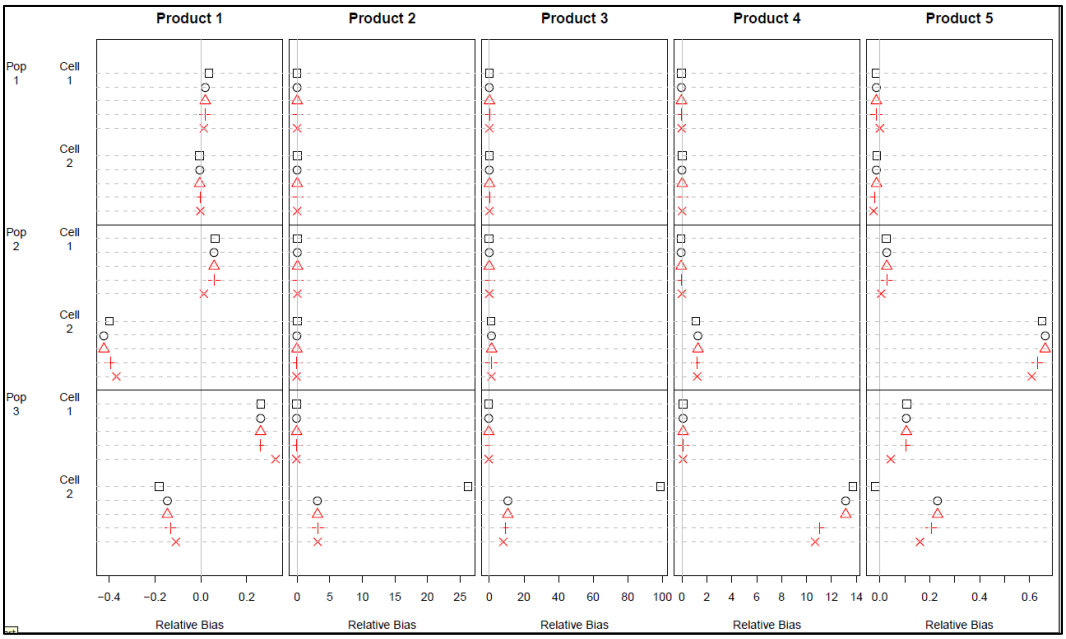


Figure 5: Relative Bias Results for All Products by Industry and Imputation Cell with PRR = 25%

Given the cell proportions in Table 2, we expected unbiased estimates of Products 1, 2, and 5 in imputation cell 1 in Industry 2 along with unbiased estimates of Product 1 in imputation cell 2. This was not the case in imputation cell 2, even with a 75% PRR, probably due to the very small sample size and the skewness of the units' product distributions in this imputation cell. Recall that cell-collapsing occurs very rarely in this industry, even when there are fewer donors than recipients (expected number of respondents = (22) × (0.25) = 5.5 > cell minimum of 5 respondents). When there are as many donors as recipients, the **Collapse** procedure is usually but not always the same as the **Unrestrict** and **Draw** procedures. However, when there are fewer donors than recipients *and* the realized number of donors in imputation cell 2 is less than five, the **Collapse** procedure increases the relative

bias of the majority of products in imputation cell 1. In this case, nearest neighbor donors are obtained from the other imputation cell, contaminating the product distributions. Since the range of total receipts is much smaller in imputation cell 2, the nearest neighbor is usually obtained from the same imputation cell, so that cell collapsing has little effect on the bias. We note that the effects on bias of cell collapsing could be quite different with another distance function.

In this industry, the majority of the units in imputation cell 2 report zero values for Products 1 through 4. However, a small proportion of the units report very large values of selected products. As a result, the cell average is a poor estimate of the underlying multinomial distribution, and the **Cluster-draw** procedure is not improving the bias over the **Collapse** procedure. The **Cluster-all** procedure, which uses an averaged distribution for imputation, produces substantially different results from the other four procedures. Since it always uses an averaged distribution, bias is reduced for some products in imputation cell 2 (e.g, product 2) and inflated for other products (e.g., products 3, 4, 5). In contrast, this pattern is not seen in imputation cell 1, and in fact bias is somewhat reduced for most products. Here the average is a better representation of the underlying multinomial distribution.

Industry 3 illustrates the combined effect of all three factors on relative bias. Even with a high product response rate, the products estimates in imputation cell 2 are highly biased, likely due to the very small sample size combined with the very low frequency of non-zero reported values. Collapsing imputation cells greatly increases the bias, even for the well-reported Product 1. There are improvements with the **Collapse-draw** or **Collapse-all** procedures over the other procedures for the rarely reported products when the donor to recipient ratio decreases, especially in imputation cell 2. The improvements are not "across the board," especially for the well-reported Products 1 and 2 in imputation cell 2. Again the **Cluster-all** procedure yields decreased bias for some products and increased bias for other products within an imputation cell, since it uses an averaged distribution which is not a good representation of the true underlying distribution.

Figures 6 through 8 present the 95% confidence interval coverage rates with product response rates of 75%, 50%, and 25% respectively.

When the donor to recipient ratio is high (75%), none of the proposed methods have a substantial advantage over the others, though there does appear to be higher coverage for some imputation cells using the **Cluster-All** procedure. All these procedures yield severe overcoverage for all products in Industry 1, regardless of imputation cell, as well as in Industries 2 and 3 overall and in imputation cell 1, with exceptions for the rarely reported products in imputation cell 2 (Industries 2 and 3). The **Draw** procedure was designed to improve the coverage over the **Unrestrict** by adding variability to the imputation procedure; when there are very few donors in an imputation cell and no cell collapsing, the **Unrestrict** procedure is very similar to mean imputation. However, the coverage improvement obtained using the draw is fairly trivial compared to the unrestricted use of a donor record. By and large, the **Collapse-all** procedure often improves the coverage over the other methods when there are fewer donors than recipients, even though nominal coverage is almost never achieved.

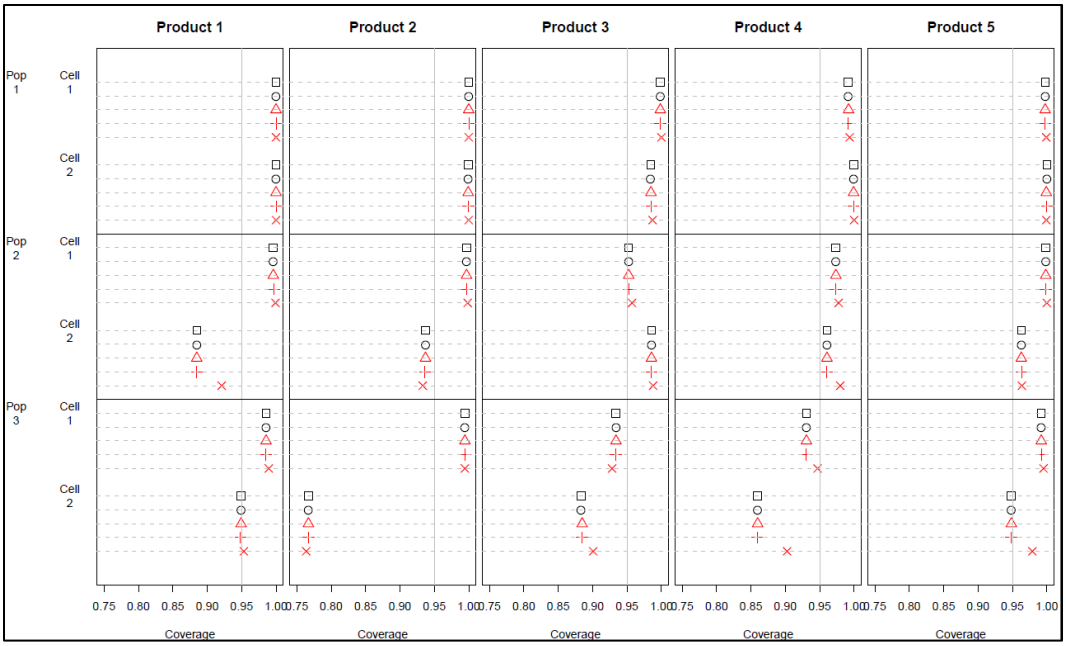


Figure 6: Coverage Results for All Products by Industry and Imputation Cell with PRR = 75%

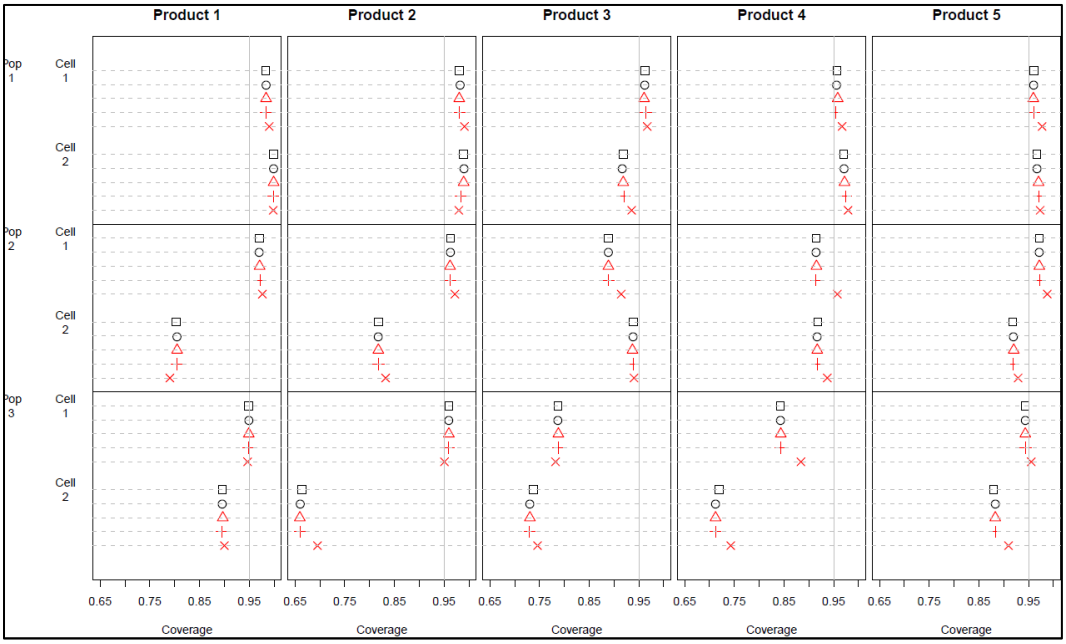


Figure 7: Coverage Results for All Products by Industry and Imputation Cell with PRR = 50%

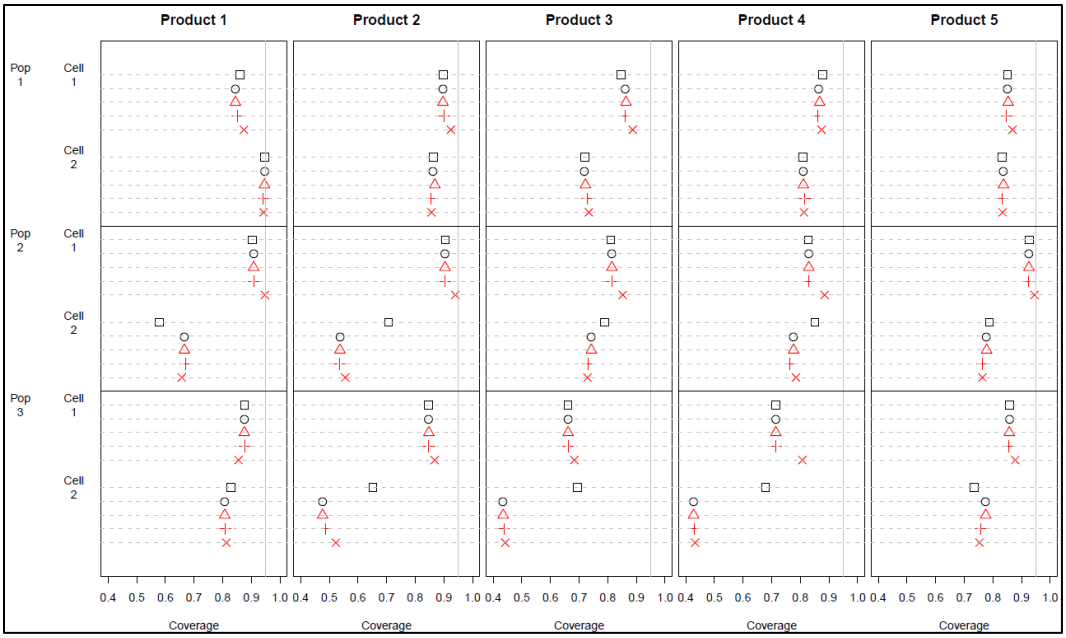


Figure 8: Coverage Results for all Products by Industry and Imputation Cell with PRR = 25%

To summarize,

- If the nearest neighbor is drawn from a different imputation cell, then the bias in the imputed multinomial distribution induced by cell collapsing can be unacceptably severe.
- If there is a high incidence of zero-reported values, then averaging a cluster of nearest neighbor donor records for the random draw can reduce the bias. This procedure implicitly assumes that the imputation cells perfectly delineate the disjoint multinomial distributions. If there is a missing latent variable in the imputation cell definition – as may be the case in Industries 2 and 3 – then the averaging can be detrimental, especially for rarely reported products.
- When there are fewer donor records than recipients and sample sizes are small, using a random draw can improve coverage.

4.2.2 Correlations

Figures 9 through 11 compare the realized (transformed) sample correlations to the population statistics for each industry when there are more donors than recipients (product response rate = 75%) and when there are fewer donors than recipients (product response rate = 0.25%). Similar comparisons when number of donors equals the number of recipients are available upon demand, but are omitted to conserve space.

In Industry 1, the first two products are frequently reported in imputation cell 1, and Product 1 is frequently reported in imputation cell 2. The studied population correlations are generally positive (14 of 15 in imputation cell 1; 10 of 15 in imputation cell two), with the remaining pairs of items uncorrelated. When there more donors than recipients, the **Cluster-all** procedure best preserves the direction of the correlation in both imputation cells. As the ratio of donors to recipients decreases, the **Cluster-all** procedure outperforms the other methods in terms of preserving positive correlation, at a slight cost of creating correlated item pairs that should be uncorrelated. Although cell collapsing occurs frequently in this industry, there is no overlap between the distance measure, so that the

Collapse and **Unrestrict** procedures are essentially equivalent. In this scenario, both procedures do a very poor job of preserving the actual linear relationships in the imputed microdata, especially when there are fewer donors than recipients. However, these procedures are less likely to induce correlation in the imputed microdata than the others.

Figure 9: Summary of Industry 1 Correlation Comparisons by Imputation Cell

In Industry 2, a high percentage of the establishments report values for several products in imputation cell 1. However, the scatter plots in the Appendix indicate that units tend to either report most of the value for a single product or allocate the majority of their total receipts to products 1 and 2. In imputation cell 2, there are no strong linear relationships with the exception of total receipts and product 1, and there is likely a missing latent variable in the imputation cell definition, as the majority of units either report product 1 or from the completely different set of less-frequently reported products contained in product 5. In Industry 2, most of the significant correlations in imputation cell 1 are positive (8), with 2 negative correlations and the remainder uncorrelated, whereas most of the item pairs in imputation cell 2 are uncorrelated with very weak positive correlation for two pairs of items. When there are more donors than recipients, the **Cluster-all** procedure tends to preserve the correlation, with a slight exception for the negative correlations that represent a small fraction of the studied statistics. There is no clear advantage of any procedure in imputation cell 2 in this scenario, although the **Cluster-all** procedure is slightly worse in terms of creating positive correlations in the imputed data. As the donor to recipient ratio decreases, the **Cluster-all** procedure continues to best preserve the correlations in imputation cell 1. However, the performance is less consistent in imputation cell 2.

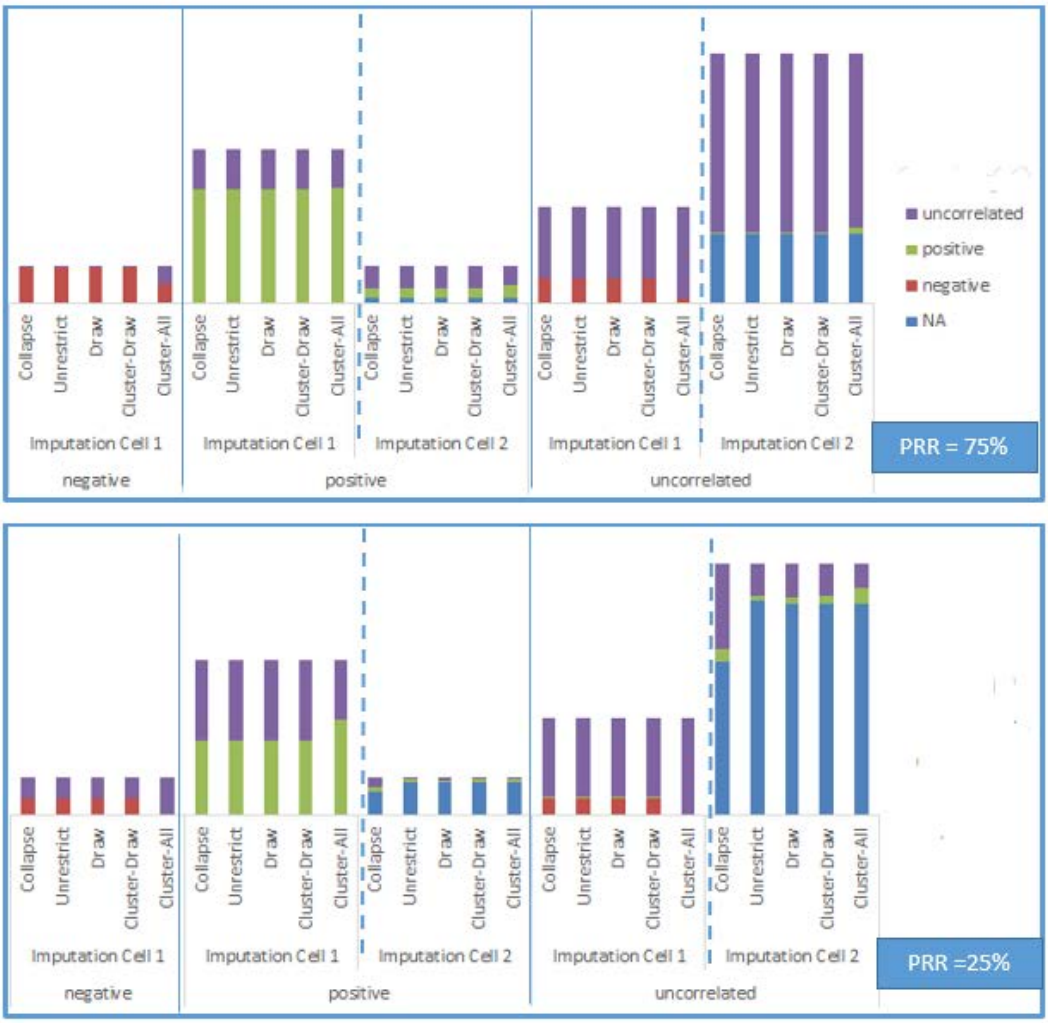


Figure 10: Summary of Industry 2 Correlation Comparisons by Imputation Cell

In Industry 3, there are no strong linear relationships in the population, regardless of imputation cell. Again, the **Cluster-all** procedure does the best job of preserving the correlation structure in imputation cell 1, although it does tend to induce a positive correlation when none exists more frequently than with the other procedures. There is some evidence that the **Cluster-all** procedure outperforms the other methods in imputation cell 2, with the caveats that (1) there are very few strongly correlated pairs of items in the imputation cell and (2) the correlation often cannot be estimated. Certainly, there is no overwhelming compelling evidence of improvements from the **Cluster-All** procedure over the others, especially as the donor-to-recipient ratio decreases.

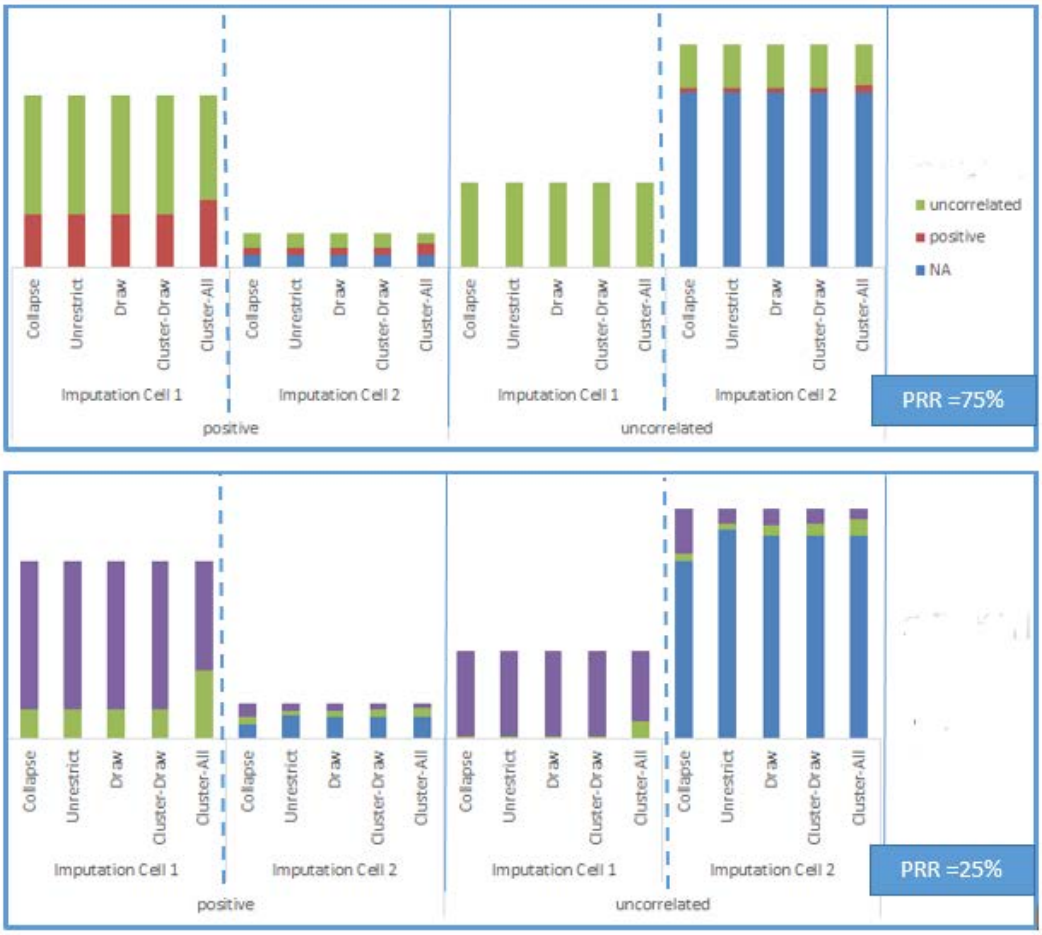


Figure 11: Summary of Industry 3 Correlation Comparisons by Imputation Cell

Of the five considered procedures, the **Cluster-all** procedure is the most promising in terms of preserving linear relationships between items, even when the donor to recipient ratio is low. There is a microdata preserving cost in that this method can induce correlation between variables in the imputed data when there is none. However, this shortcoming is confounded with imputation cell size, the number of donors, and the high incidence of missing MI-imputed correlations.

5. Conclusion

Many different factors contribute to the performance of the five imputation methods considered, and thus it is difficult to draw an overall conclusion as to the "best" method. However, a few results are clear. Firstly, when there is overlap between imputation cells such that collapsing could lead to a donor being selected from a different cell, the **Collapse** procedure can lead to very biased estimates. The other methods, which avoid collapsing, in general produce less biased estimates across all products. The exception is the **Cluster-all** procedure, which, due to the use of an averaged distribution, can cause increased bias for some products and decreased bias for others, depending on whether there really is a single underlying multinomial distribution that well-represents the entire imputation cell. A second observation is that using a draw from a distribution can improve coverage compared to imputing directly from a donor record, though improvements are small in most cases. And a third observation is that the **Cluster-all** procedure performs better than the

other four methods in preserving linear correlations among items, even when there are fewer donors than recipients. This method will occasionally induce correlations where there are none, since an averaged distribution is used for imputation, and for a set of recipients in a cell, the averaged distributions used for imputation will tend to be more similar than if a single donor were selected for each recipient.

Overall, the **Collapse-all** method is a promising option for the case where there are very low response rates, and survey analysts are both resistant to using the same donor repeatedly and resistant to collapsing cells. When an averaged distribution is used, this avoids overuse of a donor, thus improving microdata quality and alleviating subject matter concerns. The version of **Collapse-all** that we implemented was relatively basic, as we took a simple unweighted average of the five closest donors. One could envision modifying the procedure to improve its performance, for example using a weighted average of donors so that "closer" donors are given higher weight.

There are many future directions suggested by our results. Firstly, our simulation considered only the relatively simple case of a census, and future work should consider sampling effects. In addition, we used a small set of fixed cell sizes with a limited number of distributions for product data. While these were chosen to reflect "typical" product data from the Economic Census, we only considered a very small subset of industries, and it would be worthwhile to consider other cell sizes and distributions. We also fixed the minimum number of donors at five; future work could evaluate whether differences between the methods arise if this minimum is set lower or higher. Similarly, in the methods that averaged over sets of donors, we only considered the case of averaging five donors; the optimal number of donors may be higher or lower, and probably depends on characteristics of the imputation cells. Future work could also consider alternative distance functions, or ways to pick donors. Unfortunately, in the motivating application there is limited auxiliary data on which to "match" donors to recipients (e.g., total receipts is the only variable) and it is a weak predictor of the multinomial distribution in many populations. Thus, the ability to find "good" matches will be limited, regardless of the distance function.

There are several other directions for future work that may lead to differing conclusions about the relative performance of the methods. Firstly, we only considered MCAR mechanisms. However, MAR mechanisms may be likely in practice, and should be considered. In general, we do not have a strong rationale for believing data are MNAR in the motivating data, though this could be considered in future simulations. A second point to consider is that we generated data in each imputation cell from a single multinomial distribution. This may not be the case in the real data, as there may be dependence between the auxiliary variable (receipts) and the multinomial distribution, even within a cell. For example, units with smaller receipts might have more imbalanced multinomial distributions. If we were to incorporate this dependence in the data generation, then the **Cluster-all** procedure might not perform as well in certain scenarios, as a naïve average distribution would not necessarily be the "best" distribution from which to impute.

Regardless of where the future work takes us in terms of ultimate recommendations, this study demonstrates that it is possible to obtain hot deck imputed data sets that share many of the underlying properties of the generating population, even when the number of donors is much smaller than the number of recipients. This is reassuring, as our study provides strong examples of the disadvantages of allowing cell collapsing when imputing a complete

set of proportions – and the best performing methods here rely on some form of the sample data from the recipients' imputation cell.

Acknowledgements

The authors thank Carma Hogue and Justin Nguyen for their careful review of earlier versions of this manuscript.

References

Andridge, R.R. and Little, R.J. (2009). The Use of Sample Weights in Hot Deck Imputation. *JOS*, **25**, pp. 21-36.

Andridge, R. R. and 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.F. and 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.*

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.

Fang F, Hong, Q. and Shao, J. (2009). A Pseudo Empirical Likelihood Approach for Stratified Samples With Nonresponse. *Annals of Statistics*, 37 (1), pp. 371–393 DOI: 10.1214/07-AOS578.

Fink, E.B., Beck, J.L. and Willimack, D.K. (2015). Data-Driven Decision Making and the Design of Economic Census Data Collection Instruments. *Proceedings of the FCSM Research Conference.*

Haziza, D. and Beaumont, J.F. (2007). On the Construction of Imputation Classes in Surveys. *International Statistical Review*, **75**, pp. 25-43.

Kalton, G. and Kasprzyk, D. (1986). The Treatment of Missing Survey Data. *Survey Methodology*, **12**, pp. 1-16.

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.

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.

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, K.J. and Oliver, B.E. (2012). Response Rates in Business Surveys: Going Beyond the Usual Performance Measure. *Journal of Official Statistics* 28: 221-237. Available at: http://www.jos.nu/Articles/abstract.asp?article=282221.

Thompson, K.J., Oliver, B.E., and Beck. J. (2015). An Analysis of the Mixed Collection Modes for Two Business Surveys Conducted by the US Census Bureau. *Public Opinion Quarterly* 79 (3): 769-789. DOI: 10.1093/poq/nfv013.

Thompson, M. Thompson, K.J., and Kurec. R. (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.

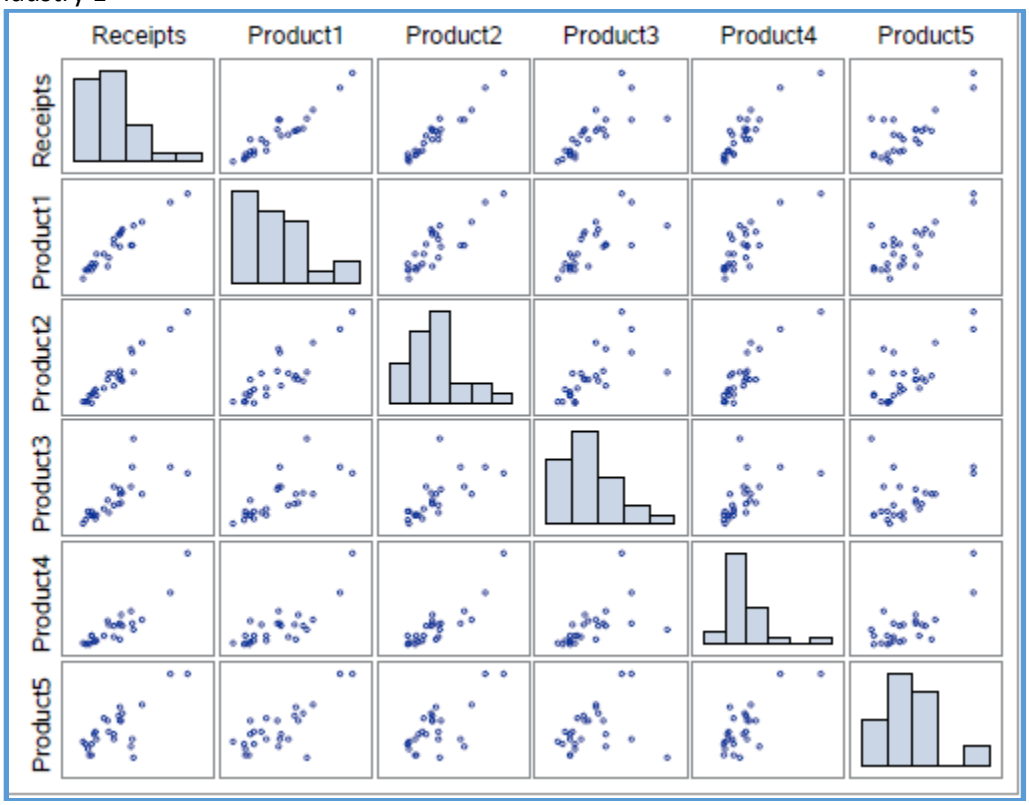
Tolliver, K. and Bechtel, L. (2015). Implementation of Hot Deck Imputation on Economic Census Products. *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

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.*

Appendix

Histogram Scatterplots of By Industry and Imputation Cell

Industry 1



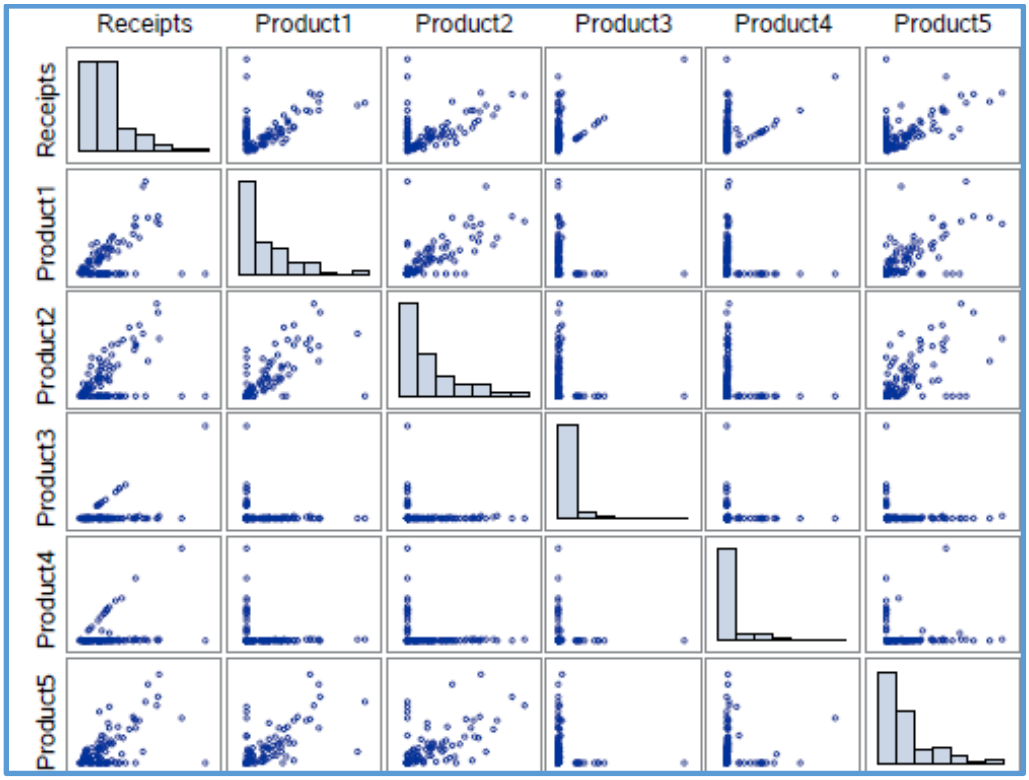
Imputation Cell 1

Imputation Cell 2

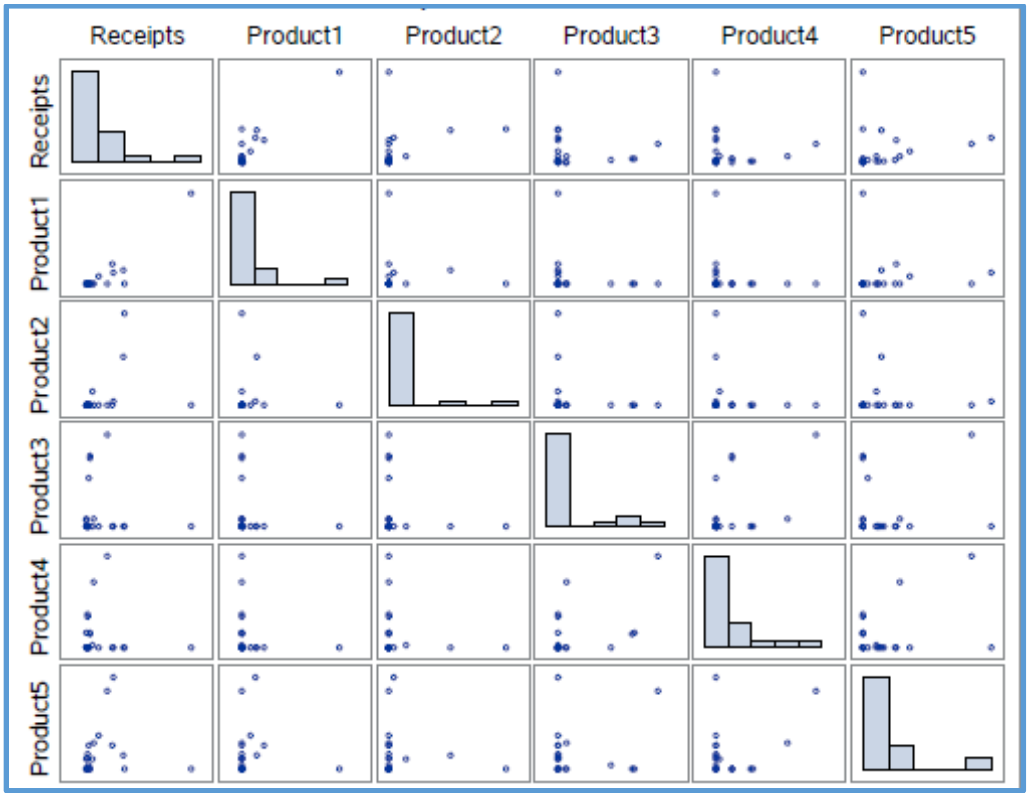
Appendix

Histogram Scatterplots of By Industry and Imputation Cell

Industry 2



Imputation Cell 1

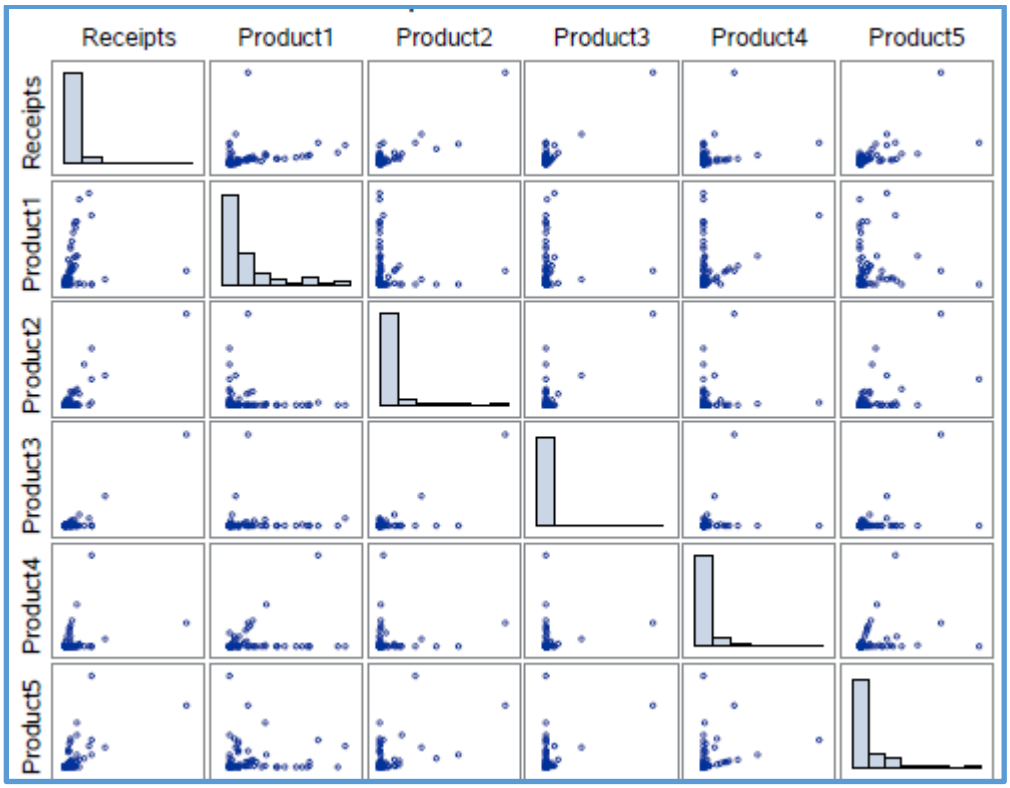


Imputation Cell 2

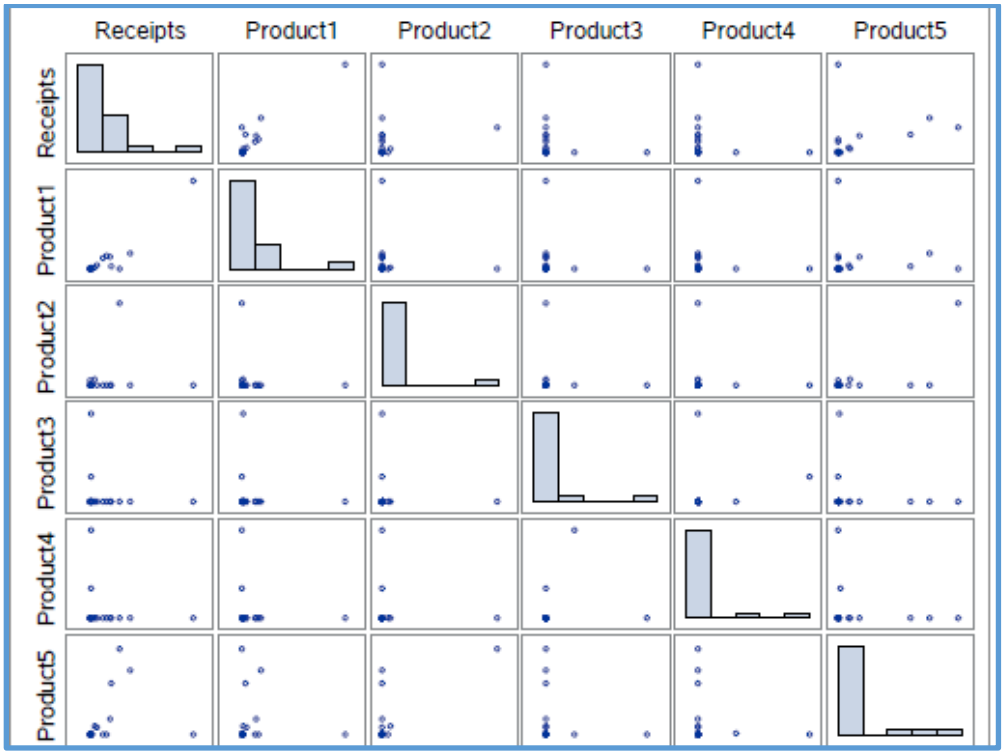
Appendix

Histogram Scatterplots of By Industry and Imputation Cell

Industry 3



Imputation Cell 1



Imputation Cell 2