On Recommending a Single Imputation Method for Economic Census Products

Katherine Jenny Thompson[1] and Xijian Liu[1]
[1]Economic Statistical Methods Division, U.S. Census Bureau,
4600 Silver Hill Road, Washington, DC 20233

Abstract
The U.S. Census Bureau conducts an Economic Census every five years, producing key measures of American business and the economy. Besides collecting a set of common items from *all* eligible establishments, the Economic Census collects detailed information on each establishment's products. Beginning in 2017, the Economic Census will use the North American Product Classification System (NAPCS) to produce economy-wide product tabulations from cross-sector collections. This marks a major departure from the current collection – which explicitly links products to industry – and makes the trade-area specific missing data adjustment practices impossible. Instead, we sought a single imputation method for all industry products that is statistically defensible and operationally practical. The research is complicated by the nature of product data, which are characterized by poor item response rates, few available predictors, additivity-within-establishment requirements, and different units of collection. In this paper, we describe our decision-making process, briefly presenting selected empirical analyses before focusing on the evaluation methods used for a comprehensive simulation study.

Key Words: response propensity analysis, rank-based statistics, imputation error, nonparametric tests

1. Introduction

The Economic Census is the U.S. Government's official five-year measure of American business and the economy. The Economic Census collects a core set of data items from each establishment called general statistics items: examples include annual payroll, total receipts or shipments, and number of employees in the first quarter. In addition, the Economic Census collects information on the revenue obtained from product sales (hereafter referred to as "products"). Prior to the 2017 Economic Census, a list of products specific to each industry was provided directly on the industry questionnaire (although establishments did have the opportunity to write-in additional products). However, the U.S. Census Bureau plans to implement the North American Product Classification System (NAPCS) in the 2017 Economic Census[2]. In the upcoming census, data collection will be electronic, and the respondents will have greater flexibility in designating their products. Moreover, NAPCS allows the collection of the **same** product from different industries, thus facilitating cross-sector product tabulations.

[1] Contact Katherine.J.Thompson@census.gov or Xijian.Liu@census.gov. This report is released to inform interested parties of research and to encourage discussion. Any views expressed on statistical issues are those of the authors and not necessarily those of the U.S. Census Bureau.
[2] Starting with the 2017 collection, the Economic Census will be published as the Census of U.S. Businesses.

Although the Economic Census is a single program, the Economic Census sectors (comprising industries) are processed in eight trade areas: Construction (CON), 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). The methods of treating missing product data in the 2012 Economic Census and prior censuses varied greatly by trade area, with no adjustments to reported values performed in MAN and MIN, nearest neighbor hot deck imputation used in CON, and ratio imputation performed in the other trade areas.

With NAPCS, products are no longer assigned to unique industries. Consequently the trade area-specific adjustment procedures are no longer a viable option. In anticipation, the Portfolio Management Governing Board (PMGB) of the Economic Directorate authorized the creation of a dedicated team to recommend a **single** missing data treatment method for producing product receipts totals under NAPCS in future Economic Censuses, providing statistically defensible justification for the recommendation based on data-driven results. The commissioned team members included methodologists, subject matter experts, and classification experts. The subject matter experts and classification experts developed the test data used for all analyses and provided expertise on the 2012 Economic Census procedures. The methodologists' familiarity with the subject matter and expertise on the current procedures ranged from completely novice (the majority) to extremely knowledgeable about a selected subset of trade area procedures. Both team leads were methodologists who were familiar with Economic Census processing procedures and methods in general but had little or no experience with the specific procedures used in product processing.

The team was initially given six months for this research project to allow time for implementation into the production system. During this time, the team had to learn about product data and distributions, develop the research methodology, develop applications, conduct data processing, analyze the results, and prepare a report containing a final recommendation. The team was expected to brief the project stakeholders on progress at regular intervals and to ensure "buy-in" throughout the process. This latter requirement was particularly challenging, as any recommended single method would result in at least one trade area changing its long-accepted procedures, although the PMGB acknowledged that the "best" imputation method could possibly differ by trade area or not exist at all. If the former proved true, then the team was instructed to recommend the best compromise. If the latter proved true, then the team had to defend to the project stakeholders why a single imputation method could not be recommended, providing research-based justification. Finally, the research project was conducted during a peak production processing time for the 2012 Economic Census, which meant that the team's computing resources were shared with other production processes, adding another layer of difficulty.

To summarize, the team was given a project with a large scope, long term implications, shared computing resources, and no universal "gold standard." The team was charged with making an objective recommendation, based on impartial research. The atmosphere was charged, as the trade-area experts believed that their procedures were already optimal. The deadline was tight, as 2017 Economic Census planning was already underway and anticipated changes needed to be incorporated into the larger schedule. At the same time, census leadership had a goal of using common processes in the Economic Census whenever supported by research. In this paper, we describe the decision-making process used throughout the project, focusing primarily on the selected methods. Section 2 provides background information on the product data characteristics and motivates the choice of studied imputation methods. Section 3 describes our evaluation process in detail. We conclude in Section 4 with some general comments on the decision making processes used throughout the project.

2. Product Data Background

The Economic Census collects information on over 8,000 different products in total. However, many products are rarely reported. Product data are characterized by poor item response rates for all but the most frequently reported products, additivity-within-establishment requirements, few available predictors, and different units of collection. Respondents are provided with a list of potential products; these lists vary by industry and can in fact differ within broader trade area. Often, product descriptions are quite detailed and many products are mutually exclusive. In some industries, respondents are asked to provide data in broad product categories along with more detailed (sub) product information. Depending on the trade area, respondents are asked to report the dollar amount (value) of each product, the percentage of the total receipts from the product sales, or both. [Note: this inconsistency in collection is going away with the 2017 Economic Census.] The reported product dollar values are expected to sum to the total receipts reported earlier in the questionnaire (percentages are expected to sum to 100%). Missing product data can occur when an establishment does not respond to the census (unit nonresponse), when a responding establishment provides no product information, or when a responding establishment provides product information which does not sum to its total receipts (partial product information).

The list of potential products is driven by the industry. In most industries, the frequently reported products are highly correlated with total receipts (and generally make up the majority of the total value of receipts), whereas the remaining products are not. Thus, the best predictors of an establishment's products are the industry assigned to the establishment from the sampling frame (which may change after collection), the total receipts value, and the value or percentage distribution of the other reported products in the same questionnaire. Besides the additivity constraints, there is a reasonable expectation that the majority of establishments in an industry report common products; these products should be imputed more frequently than other, more rarely reported, products. We considered four different imputation approaches that easily accommodate those requirements. The single ratio imputation method currently used in several trade areas is a no-intercept weighted least square regression model that uses total receipts as the single predictor for each product, taking into account both unequal sampling and unit size in the parameter estimation; hereafter we denote this as EXP.

Ratio imputation is easy to implement and preserves the industry reported-product distributions. The weighted least squares estimate of the regression parameter (β) is the best linear unbiased estimator (B.L.U.E.) under this model (Magee 1998). However, it is not a particularly strong prediction model for products that are poorly correlated with total receipts, as is often the case (see Ellis and Thompson 2015). To address the simple ratio model deficiencies, we also considered the Sequential Regression Multivariate Imputation (SRMI) described in Raghunathan et al (2001) and hot deck imputation (random and nearest neighbor). Both of these methods preserve multivariate distribution of products within establishment. However, the SRMI method relies on parametric models and allows the inclusion of additional independent predictors in the imputation model. The random hot deck (HDR) and nearest neighbor hot deck methods (HDN) models are nonparametric: Andridge and Little (2010) provide an excellent overview. See Garcia, Morris, and Diamond (2015) for a discussion of the EXP and SMRI implementation procedure in this study; see Tolliver and Bechtel (2015) for a discussion of the HDN and HDR implementations.

Finally, it should be noted that in many trade areas, the Economic Census is a bit of a misnomer. The majority of trade areas select a subsample of small single-unit establishments, while surveying all multi-unit establishments, with the exceptions being construction trade (which selects a probability proportional to size (PPS) sample of all establishments) and wholesale trade (which conducts a complete census). With the exception of the construction sector, all trade areas construct a complete universe of general statistics values using administrative data. However, product information is collected from only the sampled establishments.

3. Project Framework

The team divided the project into three separate components, each lasting approximately two months. The classification experts provided industries from each trade area whose products did not change by census year under NAPCS. These study industries varied in size and had a wide variety of products. However, they are not necessarily representative of all industries. Subject matter experts extracted the test data and provided classification rules for donor records (can be used to imputation) and recipient records (need an imputed value). The 2012 Economic Census micro-data processing in the construction trade area was not completed, so the construction test data were extracted from the 2007 Economic Census. The other trade areas' test data were extracted from the 2012 Economic Census. Once the empirical test data were available, the methodologists conducted the series of exploratory data analyses described in Ellis and Thompson (2015). Besides providing inputs for the simulation study discussed below – imputation cells and response propensity models – these analyses were valuable tools for knowledge gaining. Results were discussed with the subject matter team experts (and in some cases, analyses modified) before presentation to the project stakeholders.

Component	Purpose	Leaders
Test Data Preparation and Knowledge Sharing	• Find test data with comparable products under 2012 EC and NAPCS • Define donors/recipients • Bring staff "up to speed" on data collections	Subject Matter and Classification Experts
Exploratory Data Analysis (Empirical Data)	• Understand the "nature" of reported data to assess potential imputation methods • Understand the "nature" of missing data to assess potential imputation cells and to develop response propensity models.	Methodologists
Evaluation Study	• Evaluate the performance of considered imputation methods over repeated samples	Methodologists

After completing the exploratory data analyses, only two months remained to make a "data-based recommendation." Fortunately, the team developed the more comprehensive evaluation plan presented in Section 3.2 in parallel with the earlier empirical analyses. The plan was refined over a one month period, without looking at the simulation study results. This ensured objectivity and facilitated a quick analysis. Some facets of the evaluation plan were slightly modified after the simulation study was completed, but mostly, the original evaluation plan remained intact.

3.1. Evaluation Design

3.1.1. Evaluation Measures

Making an objective "data-based decision" hinged on obtaining relevant and computable criteria. To avoid the possibility of a tie, we sought measures that provided information on different aspects of the imputed estimates, in particular focusing on *unbiasedness* and *precision*. For both, we condition on the realized sample (census), so that the studied measures examine the properties of the error component due to the choice of imputation method. Of course, in practice, these errors may be dwarfed by the products' sampling errors and nonresponse errors.

Lohr (2009, Chapter 2.3) defines an estimator as *unbiased* when its expected value equals the population value. Of course, the census data are sample-based, and the true population value of any product is therefore unknown. Instead, we focus on minimizing the "**imputation error**" (IE) of a given product, i.e. the error induced by the employed imputation method holding sampling errors fixed. Following Sarndal and Lundstrom (2005, Ch 12), we define the imputation error as the difference between an imputed product estimate (obtained from a given sample) and the population (frame) estimate.

The imputation error served as a proxy for measuring the degree of nonresponse bias in the product value tabulations. Because we selected imputation methods that were expected to perform well on the product data, we expected trivial differences between corresponding imputed totals of well-reported products. So, while imputation error was certainly important, it could not serve as the sole evaluation criteria.

Moreover, because the levels of each product total vary greatly, it is not advisable to compare the values of imputation error or absolute imputation error between different products in the same subdomain (e.g., industry, trade area). If between-product comparisons are preferred, it is wiser to compute a relative imputation error or relative absolute imputation error. This mitigates the comparability problem without fully alleviating it. For example, although a relative imputation error rate near zero is desirable, a relative imputation error rate greater than 50% might be acceptable for some products with very small aggregate totals and might be unacceptable for others.

Lohr (2009, Chapter 2.3) defines an estimator as *precise* when the variability of the estimates over repeated samples is small. To measure precision, we use the **fraction of missing information** (FMI) which "measures the level of uncertainty about the values one would impute for current nonrespondents" (Wagner, 2010). The FMI is a natural byproduct of a multiple imputation approach and is estimated as the ratio of the between-imputation variance to the total variance of a specific estimator, with an adjustment factor based on v for the finite number of imputations (Little and Rubin, 2002). The FMI is always bounded between zero and one. If the imputation method is precise, then the between-implicate component will be very small, and the FMI will be close to zero. If the imputation method performs inconsistently, then the FMI will approach one.

For our evaluation, we defined the most accurate imputation method as having

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

3.1.2. Simulation Study Design

The choice of evaluation measures naturally informed the evaluation study requirements, specifically (1) assessment over different respondent sets (replication) and (2) usage of multiple imputation. To compare alternative imputation methods on the same outcome variables over repeated samples, an accepted practice is to:

- Create a realistic population (complete response)
- Apply the considered imputation methods to the selected outcome variables in each replicate
- Compute the pre-determined evaluation criteria and compare the results

Northolt (1998) and Charlton (2004) provide examples of excellent large-scale applications. Our simulation approach mimicked the spirit of this approach, but was modified due to the limitations of our study data. In the cited studies, the population data for simulation studies are obtained by simulating realistic complete population data, restricting the study data to unit respondent data, or "imputing" missing values with historic data from the same units. Similar data simulation approaches were infeasible for our data sets. First, the percentage of eligible sampled units that provided at least one valid product varied across trade areas but was often quite low. Moreover, it is possible that product respondents could differ systematically from product nonrespondents on an unobserved criterion (e.g. a latent class variable). We could not dismiss the possibility of non-ignorable response mechanisms for product data, as the collections are quite detailed and often burdensome. The respondent data sets were subject to sampling error in many sectors. Lastly, historic product data from the same establishments were generally not available to "fill in the gaps." In any case, there was little convincing evidence that the historic reported data would not be a representative sample of the full census.

Instead, we created four complete sets of trade area "populations" from each of the original test datasets by applying each candidate imputation method to replace the missing data as suggested by Dr. Trivellore Raghunathan (University of Michigan), denoted $POP^{EXP,<TRADE>}$, $POP^{HDN,<TRADE>}$ $POP^{HDR,<TRADE>}$ $POP^{SRMI,<TRADE>}$ where <TRADE> refers to the trade area. This suggestion was a pivotal point in our evaluation: previously, we had searched analysis methods that relied entirely on the historic sample data and not found any satisfactory methods (all assumed much higher rates of response and ignorable response mechanisms). This approach satisfied concerns about robustness to model assumptions, was thorough, and made sense to both the methodologists and the program stakeholders.

After creating the four "populations" for each trade area, we randomly induced unit nonresponse in each population using the empirical unit level response propensity models outlined in Ellis and Thompson (2015), independently repeating the process in 50 *replicates* as suggested in Nordholt (1998). Within replicate, we applied each imputation method to the missing data to obtain complete datasets, using *multiple imputation* to obtain the IE and FMI. Wagner (2010) and Harel (2007) note that a large number of implicates are required to estimate the FMI with reasonable precision when multiple imputation is used to obtain the FMI; Wagner (2010) uses 100 implicates, and Harel (2007) recommends using between 50 – 200 implicates, depending on the level of precision desired and the “true” (but unknown) value of the FMI. We used 100 implicates, balancing computation run-time and accuracy requirements. The SRMI applications were implemented using IVEWare, which performs multiple imputation (Raghunathan et al. 2002). For EXP, HDR, and HDN, we implemented a slightly modified variation of the Approximate Bayesian Bootstrap (ABB) proposed by Rubin and Schenker (1986) and Rubin (1987). As is typical with business data, our test industries were highly right-skewed, with a few larger establishments accounting for the majority of a tabulated industry total. To accommodate this phenomenon, we used PPS sampling with replacement instead of simple random sampling to account for the right-skewed population data, a modification of the ABB adaptation for complex survey design presented in Dong et al (2014).

With the exception of Mining trade area, we selected five industries (each containing at least two well-represented products) from each trade area to limit the processing demands; the classification experts had provided only four industries for the Mining trade area. We included industries of various sizes while maintaining the total number of records in each trade within a manageable level with the object of achieving a wide variety of industries and products in each trade containing sufficient data for reliable results while allowing the programs to run without errors in a realistic processing time-frame. Because the reporting rates for products can be quite inconsistent, we restricted our **evaluation** to the two best-reported products in each selected industry (in terms of number of establishments that reported the product).

Even with a limited number of industries and a limited number of selected products for analysis, the simulation procedure was quite resource intensive: a single replicate required 100 implicates apiece for EXP, HDR, and HDN imputation (one imputation per implicate) and 1000 runs for SRMI (10 iterations of the model-fitting per implicate, yielding 100 implicates). Moreover, although the **analysis** was limited to the best represented products in the studied industries, the imputation procedures produced full product distributions for each establishment.

3.2.The Evaluation Procedure

The evaluation procedure consisted of the following steps:

1. *Product-level analyses* within trade area population
2. *Imputation method selection* within trade area population
3. *Imputation method selection* between trade area population

All of these steps use rank-based procedures, thus avoiding some subjectivity and completely sidestepping any distributional assumptions. That said, performance information is lost, especially when all imputation methods perform equally well or badly for one evaluation measure but display great disparities in performance between the four methods for the other evaluation measure. For example, four imputation methods for a product might have equivalent performance on IE, but one method might have had consistently smaller FMI values.

3.2.1. Product-level analysis within trade area population

Within each trade area population, the analyses of the imputation (IE) and FMI were conducted separately by product (ten products per trade area), comparing the relative performance of each imputation method on each product. Both ranking procedures use the TIES=MEAN option, which is required to implement the Friedman Tests discussed in Section 3.2.2.

The single score that describes the imputation error performance of each imputation method on product p in industry k accounts for two properties – magnitude of the IE (ignoring direction) and the spread of the IEs – with more importance placed on the first property. To measure magnitude, we computed the ***median absolute IE*** for each product p in each imputation cell over all R=50 replicates. Using the resistant median with a 25-value breakdown point in 50 replicates over the mean (with its single breakdown point) provided some "insurance" against one or two outliers in the collected set of replicates. For each product, we ranked the four values (one per imputation method) by ascending value, assigning a single RANK_AIE to each product. To measure spread, we obtained the ***range*** *of the imputation error* of product p in the imputation cell over the R replicates within each imputation cell, using the actual range of the imputation error (largest – smallest) for this criterion. Then, we independently similarly ranked the four values of the ***range*** of imputation error within product by ascending value, assigning a RANK_RANGE value to each product.

Next, we obtained the **weighted average of the two ranked values** for each imputation method so that we had four separate measures for each product within imputation cell: COMBINED_RANK=0.70*RANK_AIE + 0.30*RANK_RANGE. These weights were developed heuristically, so that the magnitude of the imputation error has more influence on the rank than the range of the magnitudes over all replicates.

RANK_AIE is based on proximity to zero, discounting direction. RANK_RANGE is based on spread. The COMBINED_RANK places emphasis on the first property, but penalizes a method that yields large outliers (over the 50 repeated samples) by taking the range of values into account. Because a given product may be reported in more than one imputation cell within industry or may be reported in more than one industry, we averaged the COMBINE_RANK values for each product within a trade area population and repeated the ranking procedure. These final ranks were rescaled within product such that their sum equals 10.

In contrast to the imputation error measures, each FMI has an associated variance. Unfortunately, the variance is maximized when the FMI = 0.50 and is minimized when the FMI equals zero or 1 i.e., the variance of the FMI is minimized when either the imputation method is performing extremely well or extremely poorly. To incorporate the FMI's variance into our analyses, we performed general linear hypothesis tests on the ***average value*** of the FMI for each product p over the R replicates at $\alpha = 0.10$ (the U.S. Census Bureau standard). In a given trade area population and imputation cell, let

μ = **1 × *R* vector** of FMI values for product p and imputation method m
$\sum$ = ***R* × *R* matrix** of FMI variances for the product and imputation method with off-diagonal values =0
K = ***1* × *R* vector** of known constants. Since we are testing the average FMI, K =(1/R 1/R …. 1/R)

K_0_= a known constant representing a realized FMI value (e.g., 0.01, 0.65, 0.66,…,0.70,…,0.99)

The hypothesis test of interest is H_0: $K'\mu = K_0$ (average FMI for product *p* and imputation method *m* over R replicates= K_0). The test statistic is given by $\mathbf{(K\mu - K_0)^T (K\Sigma\ K^T)^{-1}(K\mu - K_0)} \sim \chi^2_1$ under $H_{0.}$ Iterating over values of K_0 for each test provides a range of values that satisfy the null hypothesis. Thus, the values of K_0 immediately below and above these values provide lower and upper bounds on the average FMI for each product within imputation cell and trade area population over all replicates. After obtaining these bounds, we obtained a single score (rank) by ranking the four values of the midpoint of the average FMI bounds for product *p* in the imputation cell (RANK_MIDPOINT), averaging the ranks for the same product across imputation cells, and obtaining a FINAL_RANK over the averaged values, again using Ties = Mean. Again, the final ranks are assigned within product such that their sum equals 10.

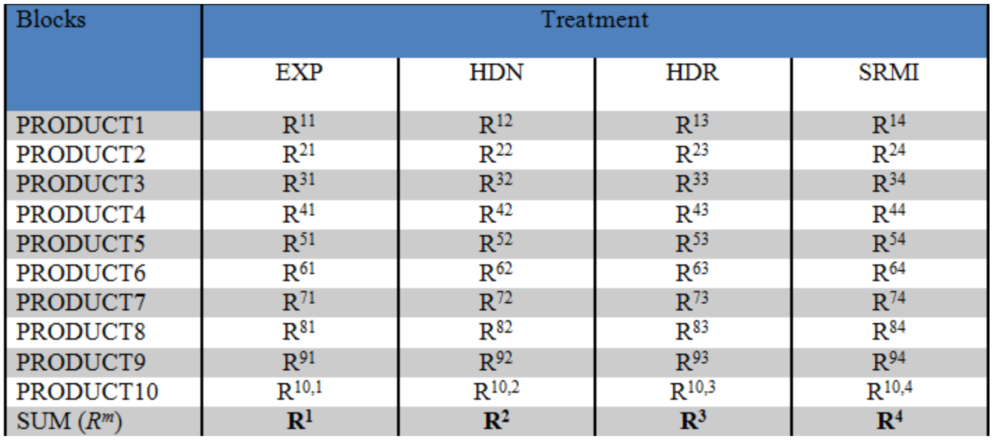
3.2.2. Imputation method selection within trade area population

The simulation study is a complete block design experiment performed independently in each trade area population. In our design, the ten studied products within trade area represent the blocks, and the treatments are the imputation methods (repeated measures on each establishment). Typically, a complete block repeated measures design is analyzed using a two-way analysis of variance (ANOVA). At a minimum, ANOVA assumes that the residuals have the same variances (homoscedasticity), but inferences that use the F-test require that variances are i.i.d. normal.

Instead of making this tenuous assumption, we used the Friedman Test (Friedman, 1940), the two-way ANOVA that uses *rank* as the measure of interest. There are two assumptions for this test: (1) the results between block are approximately independent i.e. the results for one product do not influence the results for the other products; and (2) within block, the observations can be ranked in order of interest. Technically, we may not have complete independence among products collected within the same industry. However, we believed that the number of potentially reported products is large enough within industry to offset the dependence.

Figure 1 shows the experimental set-up, where R^{pm} is the within-product rank assigned within trade-area population for product *p* imputed with method *m*. Demsar (2006) recommends a minimum of five treatments to attain comparable power to the ANOVA test; Conover (1999, Chapter 5.8) does not provide a similar limit on number of treatments or number of blocks, but does note that the power of the tests is directly affected by both.

The omnibus test determines whether all four treatments exhibit the same performance i.e., testing that all treatments have equal average rank ($R^1 = R^2 = R^3 = R^4$). If the omnibus test is not rejected, then each method is assigned a rank of 2.5 within the trade area population, so that the aggregated ranks would equal 10 (the number of studied products within trade area).



Blocks	Treatment			
	EXP	HDN	HDR	SRMI
PRODUCT1	R^{11}	R^{12}	R^{13}	R^{14}
PRODUCT2	R^{21}	R^{22}	R^{23}	R^{24}
PRODUCT3	R^{31}	R^{32}	R^{33}	R^{34}
PRODUCT4	R^{41}	R^{42}	R^{43}	R^{44}
PRODUCT5	R^{51}	R^{52}	R^{53}	R^{54}
PRODUCT6	R^{61}	R^{62}	R^{63}	R^{64}
PRODUCT7	R^{71}	R^{72}	R^{73}	R^{74}
PRODUCT8	R^{81}	R^{82}	R^{83}	R^{84}
PRODUCT9	R^{91}	R^{92}	R^{93}	R^{94}
PRODUCT10	$R^{10,1}$	$R^{10,2}$	$R^{10,3}$	$R^{10,4}$
SUM (R^m)	$\mathbf{R^1}$	$\mathbf{R^2}$	$\mathbf{R^3}$	$\mathbf{R^4}$

Figure 1: Experimental Set-Up for Friedman Test Within Trade Area Population. R^{pm} represents the within-product rank for the p^{th} product and the m^{th} imputation method ($m = 1, \ldots, 4$)**.**

If the omnibus test is rejected, then it is appropriate to perform pairwise comparisons of ranks, adjusted for multiple comparisons. We use the method outlined in Conover (1999, Ch. 5.8), Note that several other options are provided in Demsar (2006). To assign a ranking to each imputation method within population:

- Test all possible pairwise comparisons within the trade area population
- Graph results
  - Each treatment is a node (Summed rank indicated in node)
  - Solid lines connect nodes with significant differences
  - Dotted lines indicate no significant difference
- Rank results
  - Significantly different pairs receive different ranks
  - Items whose pairwise difference is not significant receive the same rank
  - Err on the side of inclusion/conservatism
  - Total rank within population/statistic = 10

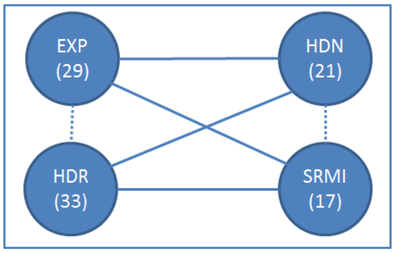


Figure 2: Graphical representation of pairwise comparisons

Figure 2 illustrates this ranking procedure. In this example, the p-value of the omnibus test for differences in IE by treatment is 0.014. Notice that the SRMI and HDN imputation methods have approximately the same summed ranks (not significantly different) and the EXP and HDR imputation methods have approximately the same summed ranks. In this example, the SRMI and HDN methods would each be assigned a rank of 1.5, and the other two methods would be assigned a rank of 3.5.

3.2.3. Imputation method selection between trade area population

The product scoring and the Friedman testing and treatment scoring procedures described in Sections 3.2.1 and 3.2.2 were performed independently in each trade area population. When this was completed, we created summary tables to examine the relative performance of the imputation methods on both statistics within trade area in the considered industries and studied products. Table 1 illustrates the trade area recommendation process, using fictional scores. The H_0 P-value column presents the results of the Friedman omnibus test for differences by treatment within trade area population for the studied statistic (IE or FMI). The other columns present the imputation method's score within trade area population for the studied statistic.

Table 1: Fictional Summary Table for Completed Simulation Study

"Population"	Imputation Error					FMI				
	H_0 P-value	EXP	HDN	HDR	SRMI	H_0 P-value	EXP	HDN	HDR	SRMI
EXP	0.90	2.5	2.5	2.5	2.5	0.01	3.5	1.5	3.5	1.5
HDN	0.50	2.5	2.5	2.5	2.5	0.02	4	1	3	2
HDR	0.09	1	3	2	4	0.05	4	1.5	3	1.5
SRMI	0.08	3	1	2	4	0.03	3.5	2.5	2.5	1.5

The statistician assigned to the trade area would complete the table and present his/her interpretation and recommendation to the assembled team. The original recommendation could therefore be modified after group discussion. In this example, all but the SRMI perform equally well in terms of imputation error, or alternatively, the SRMI method has significantly worse performance in terms of imputation error than the other three methods in this trade area. In terms of FMI, the HDN and SRMI methods perform equally well, and both are improvements over the EXP and HDR methods. Since we are trying to choose a method that does not result in large IE or FMI, we would recommend pursuing the HDN method for this population.

The recommended methods by trade area are provided in Table 2 below.

Table 2: Recommended Imputation Method for Product Lines by Trade Area

Trade Area	Recommended Method
Manufacturing	HDN
Mining	HDN
Retail Trade	HDR
Wholesale Trade	HDR
Services Industries	HDN
Finance, Insurance, and Real Estate (FIRE)	HDR
Transportation, Communication, and Utilities (Utilities)	HDR
Construction Industries	HDN

Knutson and Martin (2015) provide a detailed discussion of the decision-making process and present the final results from the research. Ultimately, the team's "data-based" recommendation was adopted, and implementation is underway.

4. Conclusion

The classic definition of a comedy is a story that begins badly and ends well. By this definition, this report is a comedy. It begins by describing a research project with a narrow timeframe, a large scope, and an inexperienced team of researchers. Being a technical paper, it ignores the softer concerns, such as team morale, misgivings about methods (particularly implementing multiple imputation), and personal preferences in terms of imputation methods. They were there, and they occupied quite a bit of the team leaders' time, especially at the beginning of the project.

The story ends with a recommendation supported by strong and repeatable findings using accepted statistical methods that were ultimately endorsed by the project stakeholders. The journey from chaos to order was consciously mapped out, not happenstance. To address concerns about scope, the study data were narrowed to the most frequently collected products in a small set of industries. Rescaling the size of the problem reduced computation time and increased available time for analysis, although it did impact the study's "representativeness."

Even so, the project would not have been accomplished in a timely manner – and the results would not have been as credible – if the evaluation plan had *not* been carefully mapped out *before* the simulation study was completed. Even with the rescaled size, the sheer number of evaluation statistics available for comparison was staggering. Providing a set of diagnostic measures for consideration before evaluation allowed the design of a simulation study that did not require substantive rework.

Developing a standard method for comparison allowed the partitioning of work between team members, which in turn created a sense of personal ownership along with a collective confidence in the results. Team leaders provided all supporting literature on a flow basis, leading to lively debates. Most important, providing a skeleton evaluation plan to the team, then using the team discussions to "flesh out" the details ensured that the plan was statistically sound, that analyses were consistent, and that potential errors (programming and assumptions) were avoided.

Acknowledgements

The authors thank Carol Caldwell, Carma Hogue, John Burns Murphy, and Thuy Ta Trang Nguyen for their careful review and thoughtful comments on earlier versions of the manuscript. We also thank the collective members of the Product Line Imputation Team (Laura Bechtel, Kaili Diamond, Jeremy Knutson, Jared Martin, William Davie Jr., Darcy Steeg Morris, Fay Dorsett, Jonathan Schuyler, Yukiko Ellis, Robert Struble, Maria Garcia, Kevin Tolliver, John Kern, John Ward, and Yves Thibadeau (consultant)) for their dedicated and excellent work on the project.

References

Andridge, R. and Little, R. (2010). A Review of Hot Deck Imputation for Survey Non-response. *International Statistical Review*, 78 (1), pp 40-64.

Beaumont, J.F. and Bissonnette, J. (2011). Variance Estimation under Composite Imputation: The Methodology Behind SEVANI. *Survey Methodology*, 37, pp. 171-179.

Charlton, J. (2004). Editorial: Evaluating Automatic Edit and Imputation Methods, and the EUREDIT Project. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*: 167(2), pp. 199-207.

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

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

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

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

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

Fuller, W. and Kim, J. (2005). Hot Deck Imputation for the Response Model. *Survey Methodology*, 31(2), pp. 139-149.

Garcia, M., Morris, D., and Diamond, L.K. (forthcoming in 2015). Implementation of Ratio Imputation and Sequential Regression Multiple Imputation on Economic Census Products. *Proceedings of the Section on Survey Research Methods*: American Statistical Association.

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

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

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

Little, R.J.A. and Rubin, D.B. (2002). Statistical Analysis with Missing Data (2nd Edition). New York: Wiley.

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

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

Nordholt, E.S. (1998). Imputation: Methods, Simulation Experiments and Practical Examples. *International Statistical Review*: 66(2), pp. 157-180.

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.* Retrieved from 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.

Sarndal, C.E. and Lundstrom, S. (2005). Estimation in Surveys with Nonresponse. New York: John Wiley and Sons.

Shao, J., and Steel, P. (1999). Variance Estimation For Survey Data with Composite Imputation and nonnegligible Sampling Fractions. *Journal of the American Statistical Association*, 94, pp. 254-265.

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.

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

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

Appendix

Omnibus Test

H_0: $R^1 = R^2 = R^3 = R^4$

H_A: At least one treatment has a different performance from the others

Let $A = \sum_p \sum_m (R^{pm})^2$, the sum of the squares of the (average) ranks

$C = \frac{PM(M+1)^2}{4} = \frac{10 \times 4(4+1)^2}{4}$, the "correction factor" for ties in rank

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

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

Friedman (1940) proposed the T_1 measure; the T_2 is the two-way analysis of variance statistics on ranks recommended by Iman and Davenport (1980). Under H_0, $T_2 \sim F(M\text{-}1,(P\text{-}1)(M\text{-}1)) = F(3,27)$. Reject H_0 if $T_2 > F(3,27,\alpha=0.10)$.

Test for Pairwise Comparisons

At $\alpha = 0.10$, a pair of summary ranks $(R^p, R^{p'})$ is significantly different when

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