Using Classification Trees to Recommend Hot Deck Imputation Methods: A Case Study

Laura T. Bechtel[1], Darcy Steeg Morris, Katherine Jenny Thompson

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. 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 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. An interdisciplinary research team was established to address this, and the outcome was the recommendation of a single (unified) methodology for the treatment of missing product data.

The team conducted a comprehensive evaluation study using empirical and simulated data from the 2012 and 2007 Economic Census. In all of the studied industries, a form of hot deck imputation appeared to be the best compromise of the considered methods. However, the recommended variation (nearest neighbor or random) was split between trade areas. Consequently, the team investigated whether certain properties of establishments or products might be predictive of one method having better statistical properties than another across different subdomains.

Such exploratory analyses readily lend themselves to classification tree analysis. A classification tree starts with a categorical outcome (often binary) and grows branches that represent an explanatory variable. The nodes (splits) at the top of the tree represent the covariates that are most strongly related to predicting outcome. In our application, the outcome was the choice of hot deck method, and the predictors were characteristics of establishment data within the imputation cells. In this paper, we describe how we used classification trees to develop an understanding the underlying causes that led to one method preferable to another and provide some limited guidance for implementation in the 2017 Economic Census and future analyses.

Introduction

The U.S. Census Bureau conducts an Economic Census every five years. 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 sector-specific missing data adjustment practices impossible. Moreover, beginning with the 2017 Economic Census, data collection will be electronic, and the respondents will have greater flexibility in reporting products. NAPCS allows the collection of the same product in different industries. This will allow products to be classified in different industries. Prior to that, each product was only defined within a specified set of industries.

Consequently, an interdisciplinary research team was established to research alternative imputation methods and recommend a single (unified) methodology for the treatment of missing product data. The results of this research are fully reported in Thompson et al (2014), with the experimental design described in Research Background Section. Hot deck imputation produced completed datasets with solid statistical properties in all studied data. However, the

[1] Contact Laura.Bechtel@census.gov, Darcy.Steeg.Morris@census.gov, or Katherine.J.Thompson@census.gov.

recommended hot deck variation (nearest neighbor or random) was split between sectors. For production implementation, subject matter experts require some guidance on selecting a method for a given industry. In practice, little historical data will be available to facilitate this decision-making process besides sampling frame properties (e.g., average weight per sampled unit) and historic imputation cell counts (e.g., donor/recipient ratios, number of sampled units). Taking this into account, the team investigated which of these imputation cells properties might be predictive of improved performance of one hot deck method over the other in a given industry.

Such exploratory analyses readily lend themselves to classification tree analysis. A classification tree starts with a categorical outcome (often binary) and grows branches that represent an explanatory variable. The nodes (splits) at the top of the tree represent the covariates that are most strongly related to predicting outcome. In our application, the outcome is the choice of hot deck method, and the predictors are characteristics of establishment data within the imputation cells. In this paper, we describe how we use classification trees to develop an understanding of the underlying causes that lead to one method being preferable to another and provide some limited guidance for implementation in the 2017 Economic Census and future analyses.

Research Background

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 "product data"). Product data are collected towards the end of the questionnaire and are requested from all establishments included in the Census. The types of products that an establishment is expected to produce or to sell are strongly related to the primary industry in which the establishment operates. The number and frequency of products reported by an establishment in a given industry varies greatly, depending on the sector and the establishment size. However, in a given industry, the majority of the product sales revenue is generally derived from six or fewer products (Ellis and Thompson 2015), although the establishment can provide information on many more products, depending on the industry in which it operates. 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.

For many sectors, the term "Economic Census" is a bit of a misnomer. The majority of industries comprise a small probability subsample of small single-unit establishments, while surveying all multi-unit establishments and the largest single-unit establishments. There are exceptions: prior to the 2017 Economic Census, the manufacturing and mining sectors utilize cut-off sampling, the construction sector selects a probability proportional to size sample of all establishments, and the wholesale trade sector conducts a complete census. In the remaining sectors, the sampling rates tend to be quite high (on average, one-in-five) and never exceed one-in-20. With the exception of the construction sector, all industries construct a complete universe of general statistics values using administrative data. However, product information is collected from only the sampled establishments.

Prior to the 2017 Economic Census, respondents reported their product data on one of more than 400 industry-specific versions of the questionnaires (paper and electronic), with over 8,000 different products reported. Note that many products were rarely reported; product data are characterized by poor item response rates for all but the most frequently reported products. There are additivity constraints, with the reported product dollar values expected to sum to the total receipts reported earlier in the questionnaire. 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 that does not sum to its total receipts (partial product information).

Recall that the research team was charged with finding a single missing data treatment that worked well in all sectors. 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 (RCPTOT), 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 these requirements. The simple ratio (expansion) imputation method currently used in several sectors 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 imputation method as EXP. EXP is easy to implement and preserves the industry reported-product distributions. 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, the team 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. The SRMI method allows the inclusion of additional independent predictors via a parametric imputation model, while the random hot deck (HDR) and nearest neighbor hot deck (HDN) methods are nonparametric. See Andridge and Little (2010) for an excellent overview of the hot deck methodologies; Garcia, Morris, and Diamond (2015) for a discussion of the EXP and SRMI implementation procedure in this study; and Tolliver and Bechtel (2015) for a discussion of the HDN and HDR implementations.

The research team conducted a full factorial experiment to compare the statistical properties of each imputation method over repeated samples. The historic empirical data provided for the study consisted of a set of industries selected by subject matter analysts; these industries do not comprise a representative sample of the Economic Census. All datasets were subject to product data nonresponse. Because the true product values were not available for the nonrespondents, instead of evaluating the four imputation methods on the complete cases, we created *four* pseudo-populations. These pseudo-populations were produced for each industry by applying each considered imputation method to the missing product data, each denoted $POP^{EXP,<SEC>}$, $POP^{HDN,<SEC>}$, $POP^{HDR,<SEC>}$, and $POP^{SRMI,<SEC>}$ where <SEC> refers to the sector, yielding four complete datasets. With the exception of the mining sector, we selected five industries each containing at least two well-represented products to create these pseudo-populations, thus limiting the processing demands; the classification experts had provided only four industries in the mining sector. In the classification tree discussion presented in the following sections, we exclude the $POP^{EXP,<SEC>}$ results because these micro-data are not very variable (by design) and are consequently not particularly realistic, especially since this method will not be used in the 2017 Economic Census. To simulate datasets with missing product data, we randomly induced unit nonresponse in each pseudo-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 our evaluation statistics. To obtain the multiple imputation implicates, we used a slightly modified version of the Approximate Bayesian Bootstrap (ABB) to create data sets with varying sets of respondents (Dong et al. 2014), then applied HDR and HDN to the bootstrapped data. Knutson and Martin (2015) describe these procedures in more detail.

Our focus was to find an imputation method that yields accurate totals and has low variability. We measured these with absolute relative imputation error (ARIE) and the fraction of missing information (FMI), respectively, both obtained using multiple imputation data sets (one data set per replicate with 100 implicates per replicate).

The **absolute relative imputation error (ARIE)** serves as a proxy for measuring the degree of nonresponse bias in the product value tabulations and is defined as

$$ARIE_r^{ip,m} = |(\bar{Y}_r^{ip,m} - Y^{ip})/Y^{ip}|$$

where Y^{ip} is the pseudo-population total of product p in imputation cell i and $\bar{Y}_r^{ipm}$ is the multiply imputed total obtained with imputation method m (HDR or HDN) for all replicates 1 through r. With an unbiased imputation procedure, the ARIE should be near zero.

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. Therefore, while imputation error was certainly important, it could not serve as the sole evaluation criteria. The **fraction of missing information (FMI)** was our measure for evaluating the precision of the product value tabulation. If the imputation method tends to yield consistent distributions, the FMI will be close to zero. If the imputation method performs inconsistently, then the FMI value will approach one, satisfying the desired bounding criteria.

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). Tolliver and Bechtel (2015) found minimal differences in FMI performance between the two hot deck methods on the same products/imputation cell/pseudo-populations, whereas some differences in ARIE performance occurred. Consequently, the classification tree analyses focused on finding factors that "explained" differences in ARIE due to hot deck variation choice.

Hot Deck Methods

Random hot deck (HDR) imputation is the simplest form of hot deck. The donors are selected randomly from the donor pool in the imputation cell shared with the recipient. The assumption is that the missing record is equally as likely to have a value of any of the other observed records in the donor pool. The number of times a donor can be used is predetermined by the statistician. If there are fewer donors than recipients, a donor may need to be used more than once, the imputation cell may need to be collapsed to a higher aggregate level, or an alternative imputation method may need to be used.

With a skewed population, subject matter experts often have reason to believe that an establishment's distribution is strongly correlated with the size of the unit; for example, with product data, one might expect that a large company would report activity of more products than a small establishment. HDR addresses this concern via the formation of imputation cells. Sometimes, sufficiently granular imputation cells cannot be achieved without overly restricting the donor pool.

Nearest neighbor hot deck (HDN) incorporates unit size in the selection of donors without necessarily creating overly restrictive imputation cells. With HDN, the donors are selected via a specified distance function based on auxiliary information. The distance function is used to choose the "most similar" unit when matching a donor with a recipient. Distance measures are calculated for all donors within an imputation cell. The donor with the smallest distance from the recipient is selected as the "most similar" unit and is used for imputation, under the assumption that the missing record is more likely to have a value that is similar to a donor with similar characteristics. Sometimes more than one donor may be closest to the recipient. In this case, the donor is randomly selected from the tied units. When implementing HDN, there is a strong possibility that a donor may be used more than one time if it is the closest donor to more than one recipient. In some implementations, the statistician may choose to limit number of times a donor can be used. If it is decided to limit the number of times a donor is used, at a minimum the statistician must also choose an alternative imputation method that can be used when a particular donor has reached its limit.

HDR is often used in household surveys and HDN is more common in business surveys (Chen and Shao 2001). Both techniques preserve that multivariate relationship since they can be used to impute multiple variables at once from an observed donor, eliminating the chance of imputing an implausible record. To preserve the establishment level multivariate distributions, we use one single donor to provide all product values to a recipient record. A recipient record in this context is a missing record where the sum of product values is not within an acceptable range of the receipt total. For each product p on the donor record, the following value was imputed on the recipient record:

$$Y_p(recipient) = \left(\frac{Y_p\,(donor)}{RCPTOT(donor)}\right) * RCPTOT(recipient).$$

In other words, the donor record replaces all product values on the recipient record even if partial information is reported by the recipient.

Classification Trees for Imputation Method Selection

In all of the studied industries, a form of hot deck imputation appears to be the best compromise of the considered methods. However, the recommended variation is split between sectors. More important, the studied industries are not a probability sample and are not likely to be representative of the larger sectors. Consequently, we investigate

whether we can identify certain properties of establishments or products at the imputation cell level that would lead one method to have smaller imputation error than the other.

Often, classification trees are used to develop imputation cells for a given dataset and imputation method. Note that in our research, the imputation method and the imputation cells have been determined for a given industry. Instead, we are looking for "causes" that can be used to predict whether one hot deck method will outperform another a priori. Specifically, we are looking for characteristics of imputation cell and product combinations (each record representing an aggregated product) to predict the binary outcome:

$$I_{ip} = \begin{cases} 1 \text{ if } ARIE^{ip,HDR} < ARIE^{ip,HDN} \\ 0 \qquad \text{otherwise} \end{cases}$$

where the value of $ARIE^{ip,m}$ is determined using the midpoint of the ARIE range for method m over all replicates for a given product in an imputation cell and sector.

This binary indicator is the outcome variable used to grow the classification tree. Each branch represents an explanatory variable and the splits at the top of the tree represent the covariates that are most strongly related to predicting the outcome. Each entering covariate is included in the model conditioned on the previous splits. In addition to identifying strong explanatory variables, classification trees also automatically identify interactions. Classification trees use forward selection procedures for entering variables, and a stopping rule is usually required or the tree will terminate with one observation at each terminal node.

We use the RPART package in R (Therneau, Atkinson, and Ripley 2013) which implements the Classification and Regression Tree (CART) algorithm. Instead of using a statistical test criteria to terminate the split selection as the tree is grown (e.g., an F-test with pre-specified alpha), Breiman et al. (1984) "recommend pruning the tree." After computing an exhaustive tree, we focus on two pruning approaches: a traditional approach using the one standard error rule and a study-specific approach that guarantees a small tree. The one standard error rule prunes the tree using the cost-complexity parameter (*Cp*) and associated cross-validated errors. Specifically, it chooses the simplest subtree that is no more than one standard error worse than the optimal tree (the one that obtains the smallest cross-validated error). Note that while this method searches for the simplest tree, there is no limit to the number of splits in the resulting pruned tree and thus no guarantee of a reasonable number of splits. Because we are interested in interpreting our tree results as rules to be used in a production setting, we also use a pruning approach that explicitly restricts the tree to a maximum of five splits. That is, we select the subtree from the set of subtrees with no more than five splits that exhibits the smallest cross-validated error.

To evaluate the predictive power of the classification tree, we assess the correct classification rate – the proportion of cases for which the predicted winning method is equal to the observed winning method. Because the differences in imputation error are observed to be small and only marginally significant (Tolliver and Bechtel 2015), we consider the direction of the error to be unimportant. That is, we consider a false positive (HDR predicted to be the winning method when HDN is the observed winner) and a false negative (vice versa) to be equally "bad." For this reason, we focus on overall error via the correct classification rate rather than specific types of error.

Classification and regression trees are often used as a model-building tool to uncover complex interactions among and make precise predictions from a large number of predictors. It is important to note that our motivation for using this machine learning tool is to develop relatively easy and interpretable rules for practical use. Accordingly, we choose a small list of predictors that lend themselves to easy application and make intuitive sense to the subject matter experts that can be obtained from frame data before collection or can be estimated from historic data under the reasonable expectation that the statistics are stable between collection periods. This yields a parsimonious tree with relationships that we expect to hold in future data collection.

Case Study Results

In this section, we present the results from exploratory data analysis and classification tree analysis of the imputation cell/product ARIE estimates. Recall that we translate the ARIE estimates into a zero-one indicator, I_{ip}, and that the

studied data are limited to a small subset of industries focusing on at least two well-represented products. Before we examine the relationship of the predictors with the dependent variable, we study the predictors themselves.

We started with the small pool of possible predictors available from auxiliary information at the imputation cell/product level provided in Table 1. While the data are at the imputation cell/product level, we require that the variables used for production rules be available at the imputation cell level from historical data. To elaborate, in order to implement the criteria specified from the tree, we need (1) adequate historic data available to estimate the variables so that they can be calculated at the time of implementation, (2) variables that are available at the imputation cell level (not imputation cell/product level) because hot deck imputation is designed to impute all products simultaneously for an establishment, not product-by-product, and (3) representative values for the variables included in the analysis (not just the "top" two products in an industry's values). These restrictions eliminate both CORR[2] and RNGPROD from use.

Table 1: List of Available Predictors for Classification Tree

Predictors	Description
CORR	Correlation of products to receipt total
REC	Number of recipients in an imputation cell
DON	Number of donors for a product in an imputation cell
SIZE	Number of donors and recipients in an imputation cell
DRRATIO	Donor (using DON above) to recipient ratio in an imputation cell
CELLPRODS	Number of distinct products reported in an imputation cell
TRADEPRODS	Number of distinct products reported in a trade area (a sector or a combination of sectors)
RNGPROD	Range of product values in an imputation cell
RNGRCPTOT	Range of receipts values at the imputation cell by product level
WGT_F	Percentage of establishments with WGT greater than one at the imputation cell by product level

For the remaining variables, we noted the redundancy in including DON, REC, SIZE and DRRATIO. To make an informed choice about which subset of these variables to use, we look at kernel density plots comparing the distributions; see Figure 2. We find very subtle differences in the distributions for DON, while the differences in the distributions of REC mirror the differences captured by SIZE. Furthermore, we notice that the estimated distribution of DRRATIO is shifted slightly higher for the cases where HDN is the preferred method. Based on these visual and intuitive differences, we retain DRRATIO and SIZE as predictors, dropping DON and REC. Using a similar process, we retain CELLPRODS and drop TRADEPRODS (kernel density plots are available upon request).

[2] Despite it not being practical for use in the classification tree analysis, exploratory data analysis revealed pleasing results about CORR. There is a noticeable difference in the distributions of CORR when separated by winning method, indicating that when there is a strong correlation between a product and the receipts value, HDN is the preferred method.

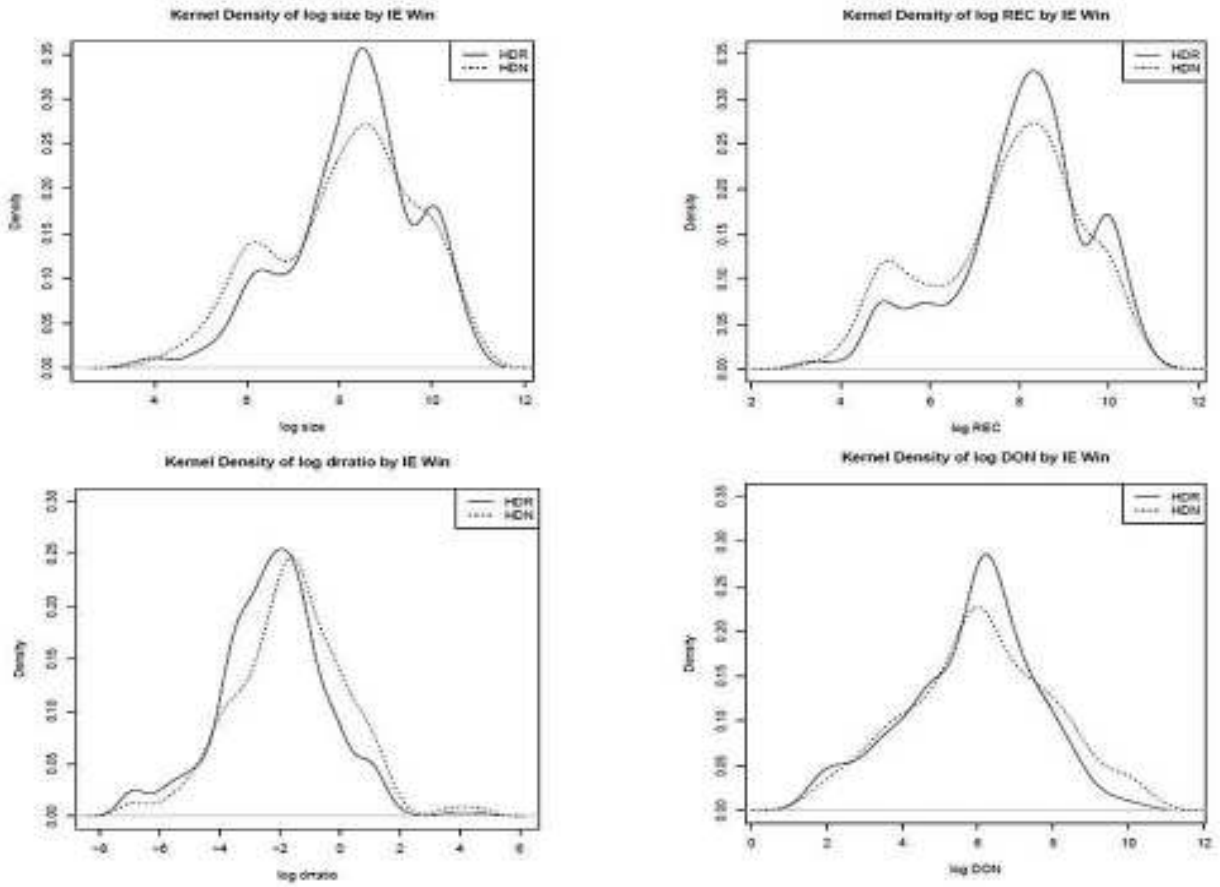

Figure 2: Kernel Density Plots of Select Predictors by IE HDR/HDN Win

The summary statistics for the set of predictors used in the classification tree analysis are given in Table 2.

Table 2: Summary Statistics of Predictors

			Percentile		
Variable	Mean	Std. Dev.	25th	50th	75th
DRRATIO	1.35	8.65	.064	.191	.444
RNGRCPTOT	614,919	2,745,527	41,658	92,493	203,583
WGT_F	42.29	31.63	8.05	43.64	69.98
SIZE	6,655	8,700	1,138	3,767	7,863
CELLPRODS	33.46	20.30	22	31	38

With these predictors, we grow a set of classification trees to assess characteristics of our binary outcome I_{ip}. Given the simulation set-up, we have three different pseudo-populations (HDN, HDR, and SRMI) available for validation. To incorporate this design into our analysis, we grow classification trees on the three pseudo-populations combined and each pseudo-population separately. We then apply each of these four trees to the four universes (one combined and three separate) resulting in 16 tree/pseudo-population combinations. For each of these combinations, we compute the correct classification rate and use this measure to determine if any one tree outperformed the others.

Recall that we implement two pruning approaches: a traditional one standard error rule and a rule that restricts the pruned tree to have no more than five splits. We originally hypothesized that the one standard error rule would result in over-fitting, thus propose the maximum split rule to guarantee a small tree. However, for two of the pseudo-populations (HDR and SRMI) the one standard error rule resulted in null trees (trees without any splits). In the other two pseudo-populations, the one-standard error rule yielded relatively large, hard-to-interpret trees: 18 splits for the combined (ALL) population and seven splits for the HDN pseudo-population. Consequently, we focus our

discussion and conclusions on the trees pruned to have a maximum of 5 splits, an arbitrary number selected because an overfit tree may perform as well, practically, as a tree comprising fewer branches. Fewer splits also have the added benefit of safeguarding against data changes that tend to occur between collection periods. Furthermore, imposing a maximum tree size eases interpretation and comparisons across trees (i.e. a high correct classification rate can be obtained from an over-fit tree, but that does not necessarily make it better).

Table 3 shows the correct classification rates for all combinations of tree, pseudo-population and pruning method. Note that, not surprisingly, we observe the highest correct classification rates for each pseudo-population when the tree is fit and evaluated on the same data (the diagonals of the table). Instead, we will focus on the cross-validated cells. Looking at the results from the maximum split pruning approach, we find that the tree fit from the combined pseudo-populations (ALL) performed the best across all pseudo-populations. Furthermore, the correct classification rates are higher from implementing the ALL trees than from assigning a hot deck method at random based on observed rates. For example, in the HDN pseudo-population, assigning hot deck method via the specifications from the ALL tree achieves a correct classification rate of 0.647. If *all* cases had been selected for imputation using HDN (the "winning" method in this population), then the correct classification rate would be 0.567. In this case, the CC rate obtained from the tree is clearly preferable. Note that the correct classification rates for the ALL tree using the one standard error rule are misleadingly high because the tree is overfit with minimal pruning, which can lead to predictions that are not robust.

Table 3: Correct Classification (CC) Rates

	Tree				Observed	
Pseudo-Population	ALL	HDN	HDR	SRMI	CC Rate	Winner
Pruning: Maximum Number of Splits = 5						
ALL	*.625*	.577	.604	.608	.512	HDN
HDN	**.647**	*.746*	.532	.572	.567	HDN
HDR	**.659**	.497	*.746*	.578	.566	HDR
SRMI	**.570**	.488	.535	*.674*	.535	HDN
Pruning: One Standard Error Rule*						
ALL	*.732*	.595	.581	.512	.512	HDN
HDN	**.751**	*.757*	.503	.567	.567	HDN
HDR	**.728**	.555	*.699*	.434	.566	HDR
SRMI	**.715**	.471	.541	*.535*	.535	HDN

*HDR tree was pruned to the lowest cross-validated error because the one standard error rule resulted in a single node. SRMI tree is null – the lowest cross-validated error is observed with zero splits.

Figure 3 graphically displays the rules determined by the classification tree grown on the combined pseudo-populations (ALL). In this figure, darker terminal nodes denote an HDN "win." The predicted winner is determined to be the method with the empirical majority in the given node. Predicted HDR "win" probabilities are displayed below the HDN/HDR prediction followed by the number of observations that follow that path in the classification tree.

The classification tree results[3] indicate that HDR should *only* be used when the establishments in the imputation cell are observed to report a variety of products, have many more recipients than donors (a donor-to-recipient ratio that is much smaller than one), have large sample sizes within imputation cell (at least 1,000 units in our study data), and contain a not insubstantial sample contribution from noncertainty units (more than ~10% of the weights are greater than 1). Otherwise, HDN yields improved performance in terms of imputation error.

This intuitively makes sense. Nearest neighbor imputation is most effective when the auxiliary variables used in the distance function are highly correlated with the items being imputed. When the establishments report a handful of products (e.g., three of four), this assumption is valid. As the average number of reported products increases, this assumption is less valid. When there are more recipients than donors, then it is less likely that the recipient's

[3] Notice that RNGRCPTOT is not included in the tree indicating that it is not a strong predictor.

"nearest neighbor" is actually close (similar) to the recipient. Large cell sizes combined with relatively low donor to recipient ratios just exacerbates this problem. When there are a lot more recipients than donors, HDR has a less restrictive donor pool available to each recipient than HDN. Finally, the sample composition and the donor selection method should be intertwined. With a high proportion of noncertainty units, it is not unreasonable to assume that donors are somewhat interchangeable, giving an advantage to random selection. With the sample design used in the census, the inverse is actually true, and the donors tend to be quite dissimilar.

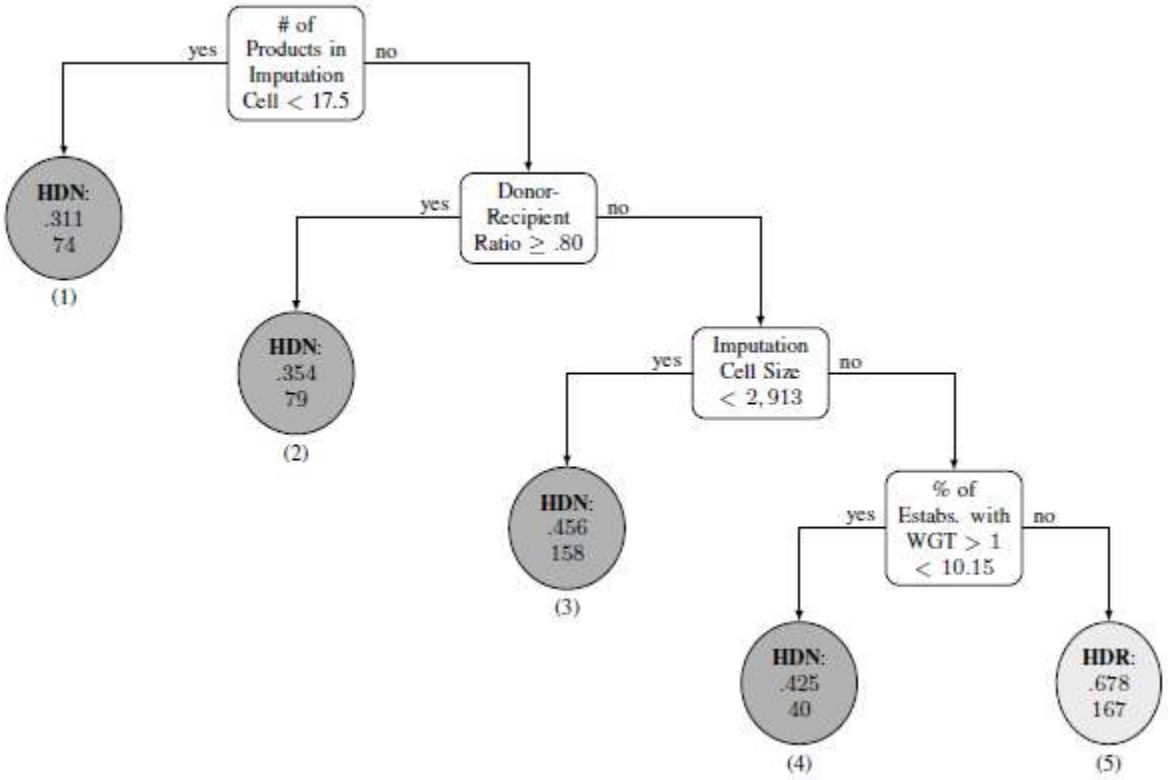

Figure 3: Classification Tree for Imputation Error

One thing to note is that in the actual application of the hot-deck imputation methods, the covariate values indicated in the tree nodes may be very different as our pseudo-population data are not necessarily representative of the entire Economic Census. Therefore, some translation may be needed to match the values of splitting variables with the distribution of the observed data. For this translation, we propose looking at the percentile values associated with the splitting values and using those to find a more appropriate splitting value for the observed data. We find that the splits values in the classification tree correspond to about the 15th, 80th, 45th and 26th percentile for CELLPRODS, DRRATIO, SIZE, and WGT_F, respectively. That said, we recommend retaining the same explanatory variables and the same form of splitting rules for determining the default hot deck variation by industry.

For example, a preliminary analysis of the donor-to-recipients ratios in the production imputation cells for on selected 2012 Economic Census product data obtained by a separate team showed median values above one and below two in the studied sectors; approximately 10-percent of these imputation cells did have less than six donor records and more than six recipient records (tables available upon demand from the authors). Consequently, the distribution of donor-to-recipient ratios in our study data appears to deviate greatly from the full census distribution. In this case, the imputation cell characteristics in the Economic Census are even more conducive to the HDN conditions presented in the tree above than shown in our datasets.

Conclusion

Often, the focus of exploratory research is to determine a viable solution (or viable solutions) to a specific problem. When a single solution presents itself, then the research recommendation easily translates to a production application. However, when two or more equally satisfactory solutions present themselves, then the production application may be determined on other factors, such as cost or timeliness or ease of implementation. If these other practical factors cannot serve as a tiebreaker, then the production application determination might be made on less tangible factors, such as personal preference.

The product imputation research for the Economic Census that motivated this research is such a problem. In the studied datasets, both hot deck methods produced microdata with similar properties in terms of FMI and very few differences in imputation error. Both methods are equally easy to program, run in about the same amount of computing time, require about the same amount of computing space, and can be easily customized to meet survey needs. Nevertheless, there were instances where one method outperformed the other. The processing schedule for the Economic Census is tight, especially for product data, which are reviewed at the end of the collection and processing cycle. There is little time for rework, and it is therefore important to designate an imputation method from the beginning that avoids instances of poor performance and requires extensive rework.

Applying classification trees to the outcome data provided a useful way of categorizing the imputation cell characteristics to develop a single rubric for imputation method choice by industry. Using the intuitive rules derived from the classification tree, we achieved higher correct classification rates than random assignment indicating that we have identified strong criteria to discriminate between HDR and HDN. In part, these rules work because we chose a small list of predictors after performing exploratory data analysis. Indeed, we found that we needed a hypothesis accompanying each predictor to obtain an interpretable tree; earlier (unreported) applications that used a larger list of available covariates yielded classification rules that would likely not apply to the larger census data. The need to specify and interpret covariates before classification is a departure from other practices such as those reported in Phipps and Toth (2012) who used regression trees to develop response propensity models.

With the Economic Census, different reporting behaviors are often exhibited between large and small businesses in the same industry. Although there has been little research on the relationship between product reporting and establishment size, subject matter experts have expressed some reservations about randomly selecting donors without accounting for establishment size in the selection process. The presented results mirrored our subject matter experts' intuition, employing a statistically defensible method to arrive at rules they might have arrived at on their own.

Acknowledgements

The authors thank Carma Hogue, Yves Thibaudeau, and John Ward for their careful review of earlier versions of this manuscript.

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.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). Classification and Regression Trees. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software.

Chen, J. and Shao, J. (2001). "Jackknife Variance Estimation for Nearest-Neighbor Imputation." *Journal of the American Statistical Association*, 96(453), pp. 260-269.

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. (2015). "Exploratory Data Analysis of Economic Census Products: Methods and Results." In *JSM Proceedings*, Section on Survey Research Methods. Alexandria, VA, American Statistical Association.

Garcia, M., Morris, D., and Diamond, L.K. (2015). "Implementation of Ratio Imputation and Sequential Regression Multiple Imputation on Economic Census Products." In *JSM Proceedings*, Section on Survey Research Methods. Alexandria, VA, American Statistical Association.

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

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

Phipps, P. and Toth, D. (2012). Analyzing Establishment Nonresponse Using An Interpretable Regression Tree Model With Linked Administrative Data. *The Annals of Applied Statistics*, 6(2), pp. 772-794.

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.

Therneau, T., Atkinson, B. and Ripley, B. (2013). rpart: Recursive Partitioning. R package version 4.1-3. http://CRAN.R-project.org/package=rpart

Thompson, Katherine J., Liu, X., Bechtel, L., Diamond, Lisa K., Davie Jr., W., Dorsett, F., Ellis, Y., Garcia, M., Kern, J., Knutson, J., Martin, J., Morris, Darcy S., Schuyler, J., Struble, R., Tolliver, K., Ward, J., and Thibadeau, Y. (2014). "Recommendation for Product Line Imputation for 2017 Economic Census: Report from the Product Line Research Team." The Census Bureau internal document, dated 12/9/2014.

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