Evaluating Hot Deck with Propensity Score Matching For the Advance Monthly Retail Trade Survey (MARTS)[1]

Katherine Jenny Thompson, Laura Bechtel, and Nicole Czaplicki
Economic Statistical Methods Division, U.S. Census Bureau

Proceedings of the 2018 Federal Committee on Statistical Methodology (FCSM) Research Conference.

Introduction

Hot deck procedures use reported values (donors) from the current sample to impute for missing values (recipients). Simply put, a recipient unit is matched to a donor unit based on predetermined criteria, and the missing recipient value is replaced with the valid value obtained or derived from the donor unit. Replacement values can be directly-substituted values of a categorical or quantitative value or can be obtained by prorating an available value from the recipient with a rate or ratio obtained from the donor. The latter approach is especially useful when the expected quantity varies greatly by unit size. In any case, the explicit control over donor eligibility and selection is required to obtain plausible imputed values. This control usually starts with partitioning the sample into disjoint imputation cells (also called adjustment cells and imputation classes) using auxiliary variables that are available for both donors and recipients (Brick and Kalton 1996), before implementing the matching procedure. The categorical variables used to define the imputation cells are either expected to be predictive of the studied outcome variable (e.g., total sales or expenditures for a business) or are predictive of nonresponse (Andridge and Little 2010). Candidate units within the imputation cells are usually subject to other edit (validation) checks to qualify as donors.

Hot deck imputation procedures are often described as "model free" missing data treatments. Certainly, this claim is not entirely true. For example, the variables that define imputation classes implicitly define a nonlinear regression model, where the dependent variable is the probability of responding. When imputation classes are equivalent to sampling strata, there is an implicit assumption that the key outcome variables have the same mean and variance within imputation cell. Likewise, the matching procedures are related to assumed models. For example, random hot deck methods obtain unbiased imputations under a missing-completely-at-random (MCAR) response mechanism or under a missing-at-random (MAR) response mechanism when imputation classes are used. Nearest neighbor imputation assumes that the outcome variables(s) can be predicted by the auxiliary covariate(s) used in the distance function, as do other backwards-forwards selection methods. Similarly, nearest neighbor imputation may be appropriate when the response probability is a function of unit size, common in many business surveys (Thompson and Oliver 2012; Thompson, Oliver and Beck 2015; Thompson and Washington 2013).

In practice, the choice of implemented hot deck method relies on stated or unstated *causal assumptions.* Pearl (2010) distinguishes between associational and causal assumptions as follows:

> "Associational assumptions, even untested, are testable in principle, given sufficiently large sample and sufficiently fine measurements. Causal assumptions, in contrast, cannot be verified even in principle, unless one resorts to experimental control."

Propensity score matching is frequently used in observational or experimental studies to control for causal effects. The objective is to "balance" participants in the treatment and control panels on selected characteristics to eliminate or reduce confounding. This is accomplished by finding similar pairs of units on these characteristics using a predetermined distance measure, then splitting the pairs into separate treatment and control groups so that the study approximates a random experiment. Rosenbaum and Rubin (1983) proposed developing a single propensity score function constructed from the full considered set of characteristics for matching, where the modeled propensity score represents the unit's expected response to the treatment. Rubin and Thomas (1996) recommend including all potential covariates in the propensity score function, even when not significant, especially when multiple outcomes

[1] 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.

are studied and there is no consensus on the (statistical or causal) relationship between a particular covariate and outcome. From a hot deck implementation perspective, it would therefore be preferable to maximize the covariates used in the matching procedure. A single propensity score is certainly convenient but is not necessary for implementation. Many matching algorithms use weighted or unweighted combinations of covariates to pair units. Moreover, there are strong theoretical and practical arguments in the literature against using the single propensity score for matching in the literature, as the score itself is based on an unverifiable model. King and Nielson (2016) argue that propensity score matches are biased towards the implemented model which can be greatly misspecified, proposing instead to implement full blocking in the experimental design a priori. Smith and Todd (2011) had inconsistent results using propensity score matching on income data, even when "robustifying" the procedures by matching on the log-odds ratio rather than propensity scores themselves. Their study provides empirical evidence that demonstrates the importance of the strength of the prediction model used for propensity for the outcome variables and the high sensitivity of propensity score matching to data quality. Reducing a set of matching variables to a single score based on an assumed model can therefore affect the procedure's success.

Alternative hot deck imputation procedures can be tested in an experimental design framework using a repeated measures design, where imputation classes are the blocks, the donor selection procedures are the treatment, and the donated value (or derived imputed value) is the outcome variable. A key objective is to determine the hot deck procedure that best reduces nonresponse bias in the program estimates; surveys might extend this to include preserving statistical associations between outcome variables. In the causal inference framework, the hot deck procedures can control for unit (subject level) effects common to both respondents and nonrespondents in the missing data treatment process. Traditionally, propensity scores are designed to summarize the explanatory variables relationship to the studied outcome variables. However, a response propensity score could be used when there is evidence that covariates predictive of nonresponse are likewise predictive of outcome. For example, the size of a business may be predictive of both the response propensity and the reported value of sales. With hot deck imputation and propensity score modeling, it is good practice to explicitly account for stratification effects.

In this paper, we examine using alternative variations of propensity score matching to obtain hot deck donors to account for late and total unit nonresponse in the Advance Monthly Retail Trade and Food Services Survey (MARTS). The MARTS is a probability sample of companies, subsampled from the Monthly Retail Trade and Food Services Survey (MRTS), whose respondents provide "early" estimates of monthly sales in retail trade industries; the more reliable monthly estimate is provided a few weeks later in the MRTS publication. In the following section, we provide background on the MARTS sample design and estimation procedures, along with a detailed discussion on the specific missing data. This study is part of a larger research project investigating alternative missing data treatments (imputation and estimation procedures) designed to reduce revision difference between the advanced monthly estimates of sales from MARTS and the corresponding preliminary estimates (for the same time period) from MRTS.

Here, we focus on determining an effective propensity matching algorithm for the MARTS nonrespondents, considering greedy and optimal matching procedures. We develop alternative propensity matching approaches, evaluating each donor selection approach empirically on 12 months of historic data from the MARTS and MRTS collections, specifically March 2016 through February 2017. We further evaluate alternative hot deck donor pool criteria on our recommended propensity matching methods, describing the data-availability and quality trade-offs of each proposed criterion. We finish with a few situational observations and recommended next steps, as well as some general remarks.

Background

The MARTS and MRTS are monthly economic indicators published by the U.S. Census Bureau providing estimates of total retail trade across the many industries in the retail trade sector. These estimates serve as inputs into quarterly Gross Domestic Product (GDP) published by the Bureau of Economic Analysis. Estimates from the MARTS are released approximately nine working days after the reference month and provide an early estimate of total monthly sales for certain industries. One month later, the MARTS sales estimate is superseded by the preliminary estimate from the MRTS; the MRTS estimate may be further revised to incorporate data from late reporters. Large revisions between the MARTS estimates and the corresponding MRTS estimates are highly scrutinized, especially when the revision reverses the direction of the seasonally adjusted month-to-month percent change. Consequently, the U.S. Census Bureau is investigating methodological enhancements to the current procedures designed to minimize these

revisions. Czaplicki, Gonzalez, and Bechtel (2018) presents exploratory research into alternative estimators for MARTS. We focus specifically on item imputation.

The MRTS uses a stratified simple random sample without replacement (SRS-WOR) design, subsampled from the Annual Retail Trade Survey. MARTS uses a stratified probability proportional to size sample without replacement (PPS-WOR) design with the unit's MRTS sampling weight as measure of size; the realized MRTS sample is the sampling frame. MARTS is therefore subsampled from a subsample. Companies that exceed a predefined industry cutoff for sales are included in the MARTS sample with a probability of one and are hereafter referred to as certainty units. All other sampled units – whose selection probability is less than one – are referred to as noncertainty units. Sampling weights for the same unit will often differ between the two surveys (MRTS and MARTS). A new MRTS sample is selected approximately every five years. A new MARTS sample is selected approximately every two and a half years with new MARTS samples introduced at the same time as the new MRTS sample and again approximately halfway through the MRTS sample cycle. While there are some large companies that will remain in consecutive MARTS samples, further overlap between consecutive samples is not attempted. For more details about the MARTS design, see https://www.census.gov/retail/marts/how_surveys_are_collected.html; for more details about the MRTS design, see https://www.census.gov/retail/mrts/how_surveys_are_collected.html.

MARTS sample units are asked to provide a response within approximately seven business days, whereas MRTS sample units are given five weeks to provide a response for the same reference period. Figure 1 plots unweighted unit response rates for MARTS from May 2010 through May 2016. A new MRTS sample was introduced in 2012, along with the new MARTS sample. A second MARTS sample was introduced in 2015. The sharp increase in response rates in late 2013 is due the government shutdown of October 2013 which delayed several data releases, thereby giving sampled units more time to respond for those months.

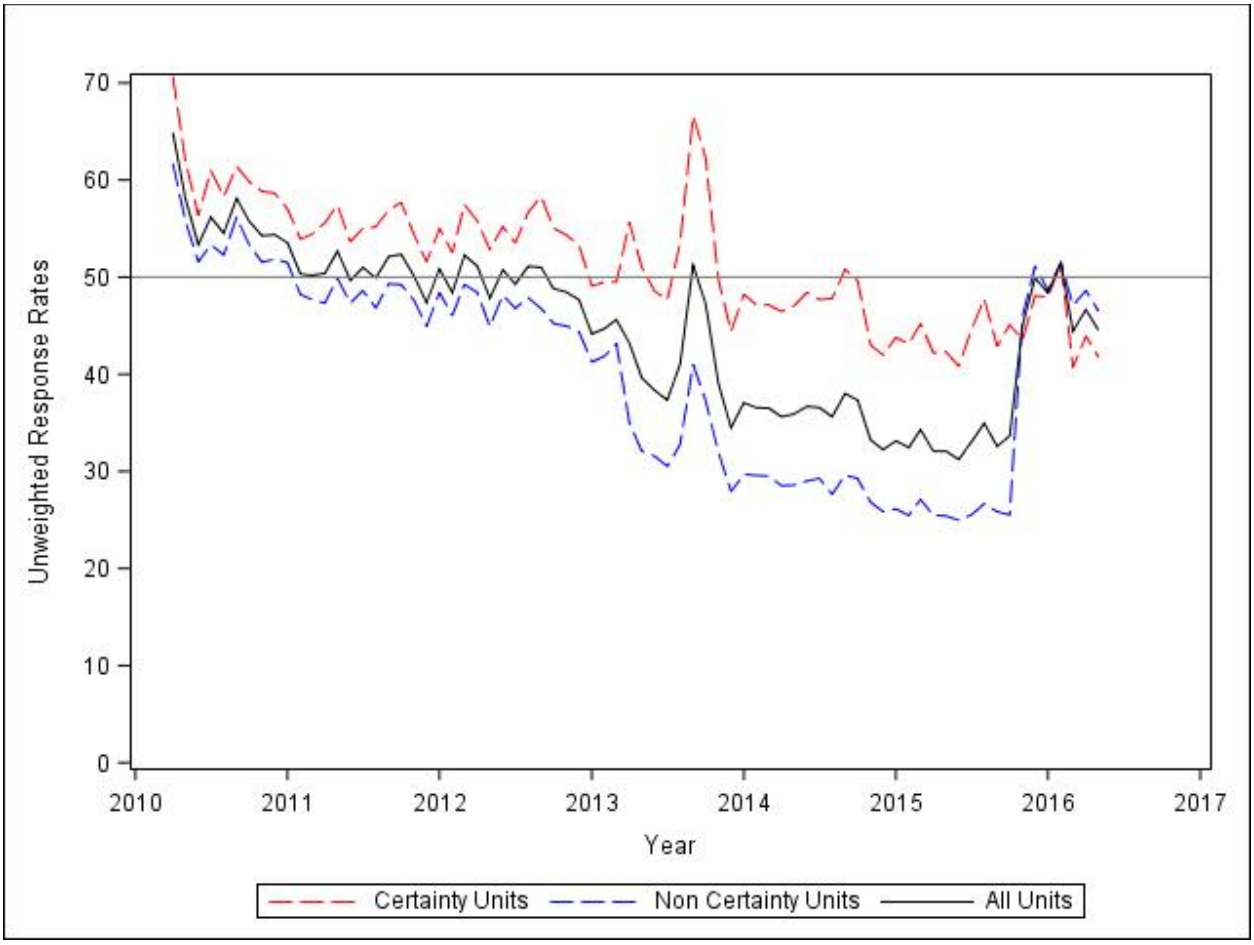

Figure 1: Unweighted unit response rates for MARTS from May 2010 through May 2016 overall and by certainty status

The MARTS respondent sample tends to fatigue over time. Sample fatigue is especially evident in the non-certainty domain. Introducing a new MARTS subsample tends to correct for respondent fatigue; notice the large increase near the end of 2015. In general, certainty units respond to MARTS at a higher rate than the noncertainty units. This is typical of business surveys where large units are more likely to respond than smaller units in part due to the analyst nonresponse follow-up procedures (Thompson and Oliver 2012; Thompson, Oliver, and Beck 2015). However, as many of the MARTS certainty units are in sample "indefinitely," the introduction of a new MARTS subsample does not address their sample fatigue. Unit response rates for MARTS tend to be quite low, especially as the length of time since the sample introduction increases. With response rates as low as seen in the 2014 and 2015 survey years, it is questionable whether the set of MARTS respondents is representative of the intended MARTS sample (Gotway Crawford 2013). Finally, MARTS data are right censored. On occasion, a unit will not respond to MARTS or will provide a late response to MRTS in a reference period, but will provide a timely response to the advance report in the subsequent period. However, MARTS units that do not respond for two or more consecutive months generally drop out of the survey completely; this is also true for the MRTS-only units.

MARTS uses a link relative estimator (Madow and Madow 1978), a synthetic estimator that multiplies a benchmark total from the prior month by an estimate of the month-to-month change (trend) using MARTS sample units with data in both the current and prior month. A unit is considered to be influential if the industry month-to-month change estimate differs "substantively" when the unit is deleted. Under this definition, the certainty MARTS units are often influential, as are several of the largest noncertainty units. Although MARTS does not perform a generalized imputation procedure for all nonrespondents, selected influential nonresponding units' sales values are estimated by subject matter experts using a combination of past company data, calendar effects, and subject matter knowledge. The procedures for obtaining a replacement value may differ by company, industry, and analyst, but do have common elements. An analyst imputed value from MARTS is retained in MRTS unless (1) the unit provides a reported value (late reporter), (2) the analyst-imputed value is visibly different from the estimated industry trend, or (3) or the analyst-imputed value fails an edit test.

Analyst-imputed values in MARTS are reviewed by more than one subject matter expert to promote consistency across the survey. However, the procedure is subjective, and the applications will vary. Replacing these subjective procedures by a more objective and repeatable procedure would be an enhancement to the survey procedures, if the replacement procedure can be shown to help reduce revision error or have better – or equivalent – statistical properties, especially if the replacement procedure(s) were easily automated and not overly computer resource intensive.

Propensity Matching and Hot Deck Imputation Procedures for MARTS

We evaluated two different matching algorithms for pairing donors with recipients, using publicly available SAS software developed by Bergstralh and Kosanke at the Mayo Clinic (http://www.mayo.edu/research/departments-divisions/department-health-sciences-research/division-biomedical-statistics-informatics/software/locally-written-sas-macros): (1) the *gmatch* macro, which implements greedy matching and (2) the *vmatch* macro, which implements optimal matching. These programs allow weighted or unweighted combinations of continuous covariates, where the larger weight indicates more importance in the matching criteria. Both applications allow caliper matching[2] on one or more covariates. Both programs select one or more donors per recipient. The two programs also offer a choice of transformations for the matching variables and two distance measures to match donors to recipients.

There are three key elements to implementing propensity matching:
1. Imputation cell definition (blocks)
2. Matching variables
3. Donor pool requirements (control definitions)

Greedy matching pairs donors to recipients sequentially. Consequently, the sort order of the input file is another factor that might affect the matching outcome. With optimal matching, the recipient selected for a given donor has the closest Euclidean distance over all eligible donors, subject to minimizing the total aggregated distance over all

[2] A maximum allowable distance is specified between match variables.

recipients; sort order is irrelevant. Greedy matches are designed for processing large files with large numbers of variables, where the computing resources required by optimal matching could be prohibitive. Greedy matching can also be advantageous in prioritizing matches, when combined with a directed sort.

With any imputation procedure, it is important to account for the key features of the survey design in the imputation cell definitions to the extent possible, as recommended by Andridge and Little (2009), among others. The usage of the donor value in imputation is another important consideration. The majority of business populations are highly skewed. When the key estimate of interest is a survey total, direct substitution of values from a donor could yield biased estimates. Instead, a common hot deck imputation procedure applies a donor *ratio* to an available recipient value. For example, administrative data could be available for all sampled units. In this case, the selected donor would provide a ratio of the studied outcome variable value to the corresponding administrative data value; the imputed value would be obtained by multiplying the recipient's administrative data value by the donated ratio.

Retail data are highly seasonal. If the imputation cells are designed to group units with similar seasonal patterns, then prorating the recipient value by a donor's current month-to-prior month change can help ensure that the imputed values have similar seasonal effects as reported values, in addition to accounting for differences in unit size. Of course, deriving the imputed value via a donor ratio does restrict both the donor and recipient pools, excluding units with missing or zero prior month values. In the causal inference framework, the outcome variable is the unit-level month-to-month change ratio, with the propensity matching pairing donors (control group) to recipients (treatment group) with similar expected rates of change. Imputation cell definitions should take differing seasonal patterns into account as well as the survey design. Matching variables should "explain" the month-to-month change ratio, in the sense of being highly predictive and having an intuitively understandable interpretation.

The donor pool requirements are intertwined with the imputation cell definitions. One of the strongest arguments for implementing hot deck methods is that they guarantee plausible imputed values (Andridge and Little 2010). If the seasonal patterns differ between imputation cells, then plausibility is only possible when donors are restricted to one imputation cell (i.e. no collapsing, unless the collapsed cells have very similar seasonal patterns). For the MARTS application, the donor pools for MARTS nonrespondents would be drawn from MRTS-only sampled cases and MARTS respondents. Since MRTS-only units have a much later due date than the MARTS units, it is unlikely that there would be sufficient reported data in the current statistical period to use current month-to-prior month change in practice. Instead, the donors would need to provide historic month-to-month change ratios, either from one year ago (accounting for seasonal effects) or from an earlier calendar year with the same calendar effects (number of Mondays, Tuesdays, etc.) as the reference period (accounting for seasonal effects and trading day). The first approach has advantages in terms of timeliness and simplicity and could likely be applied for a large percentage of the sampled MARTS units. However, the second approach requires historic data for the donors from five years ago in our application, an even less recent period. A new sample for MRTS was introduced in December 2012, so that five-year-old historic data values were not available for many of the noncertainty MRTS units. Consequently, we restricted the donor/recipient pool to MRTS certainty units, eliminating MRTS-only noncertainty units from eligibility in the donor pool and MARTS noncertainty units that are also MRTS noncertainty units from the recipient pool.

With MARTS, a natural choice for imputation cell might be the MARTS tabulation industry, which is congruent with the survey's seasonal adjustment procedures (https://www.census.gov/retail/marts/how_surveys_are_collected.html). However, these industry definitions are not necessarily those used in the sample design, which uses a more disaggregated industry definition. Using the 6-digit North American Industry Classification System (NAICS) industry approximates an important feature of the MARTS and MRTS survey designs, without accounting for unit size as done in sampling and estimation. There are other correspondence issues with sampling unit and tabulation unit discussed below that make the choice of tabulation industry as imputation cell less desirable than on the surface. Given the high level of unit nonresponse (and the restriction to MRTS-certainty units), the number of donors in the 6-digit NAICS industries can be very small without incorporating an additional size category. Using the 3-digit NAICS industry as the imputation cell generally sidesteps the small sample size problem but introduces the larger concern discussed above, namely the potential for implausible imputations. Figure 2 illustrates this, presenting side-by-side boxplots of unit-level month-to-month change for MRTS certainty respondent units for the 6-digit NAICS industries within the NAICS 448 category (clothing and clothing accessory stores) using historic data from March 2016. The boxplot on the far left presents the distribution within the aggregated 3-digit industry; the remaining boxplots present the 6-digit level distributions.

Notice that the shape and spread for the 3-digit distribution is quite different from the others, and selection of an unusually large or small donor ratio could lead to an implausibly imputed value in several of the 6-digit NAICS.

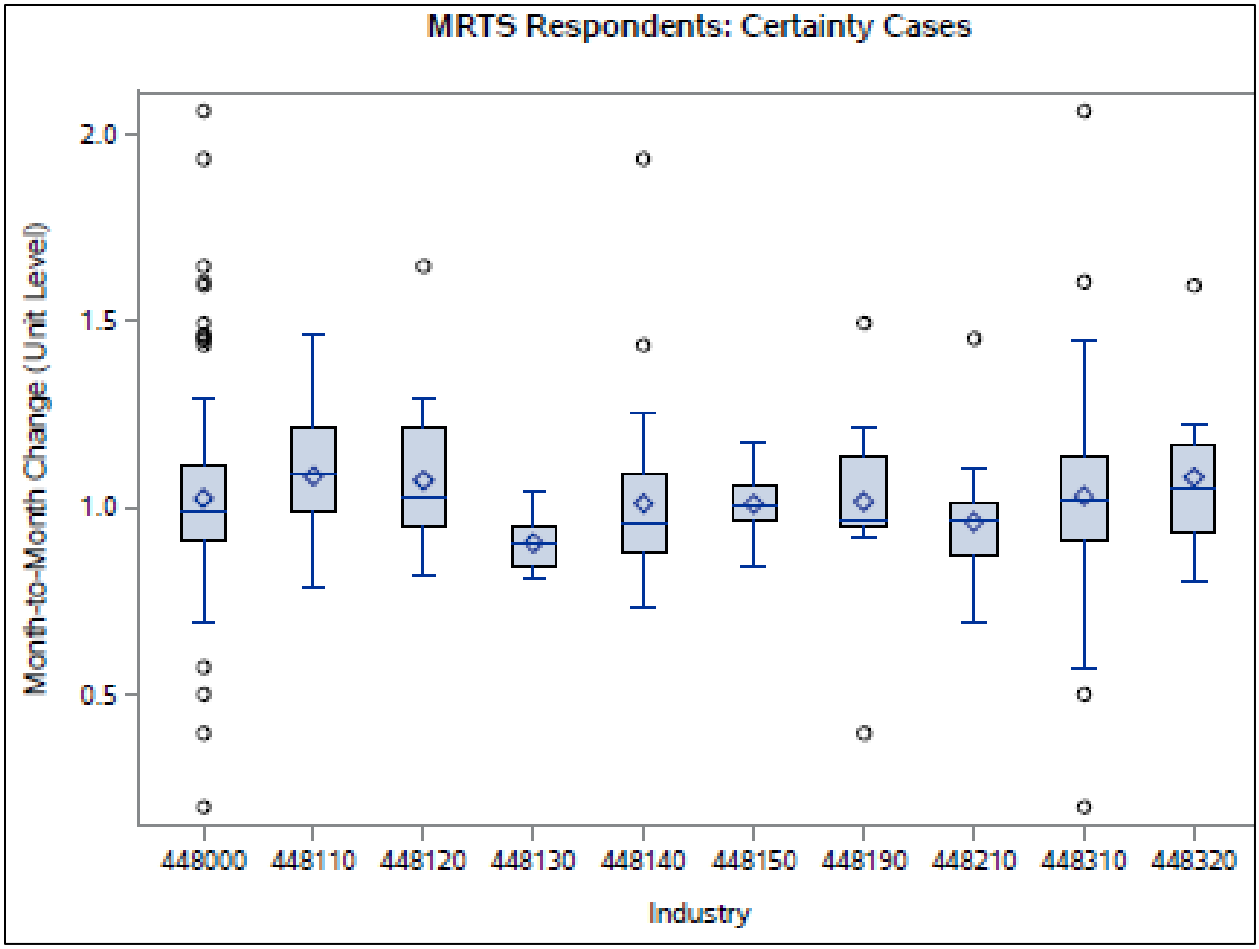

Figure 2: Distributions of Month-to-Month Change in NAICS 448 (Clothing and Clothing Accessory Stores) for MRTS Certainty Respondents in March 2016

The apparent differences in unit level month-to-month change distributions between 3-digit and 6-digit industries are endemic. Ultimately, we decided to use 6-digit industry throughout. Unfortunately, this decision leads to "un-imputable" recipients in some cases, as both the greedy and optimal matching software only use each donor once. This is a limitation of the considered procedures. However, the remaining unit nonresponse will be addressed at the estimation stage via the link-relative estimator.

By and large, we found strong evidence of a causal relationship between unit size and month-to-month change, namely that larger units often exhibit smaller month-to-month change in sales than smaller units within the same 6-digit NAICS. Figure 3 illustrates this, presenting side-by-side boxplots of unit-level month-to-month change MRTS certainty respondent units for the 6-digit NAICS industries within the NAICS 448 category using historic data from March 2016. In these boxplots, MARTS certainty status is a proxy for size, as the largest units are included with certainty in both MARTS and MRTS. The red boxplots present the MARTS certainty unit distribution and the blue boxplots represent the corresponding noncertainty unit distributions in the parent 3-digit NAICS. In six of nine NAICS, the median change value is smaller for the certainty units. In any case, the two sets of distributions are not the same within industry. Accordingly, we decided to incorporate unit size into the matching criteria by using prior month sales as a matching variable.

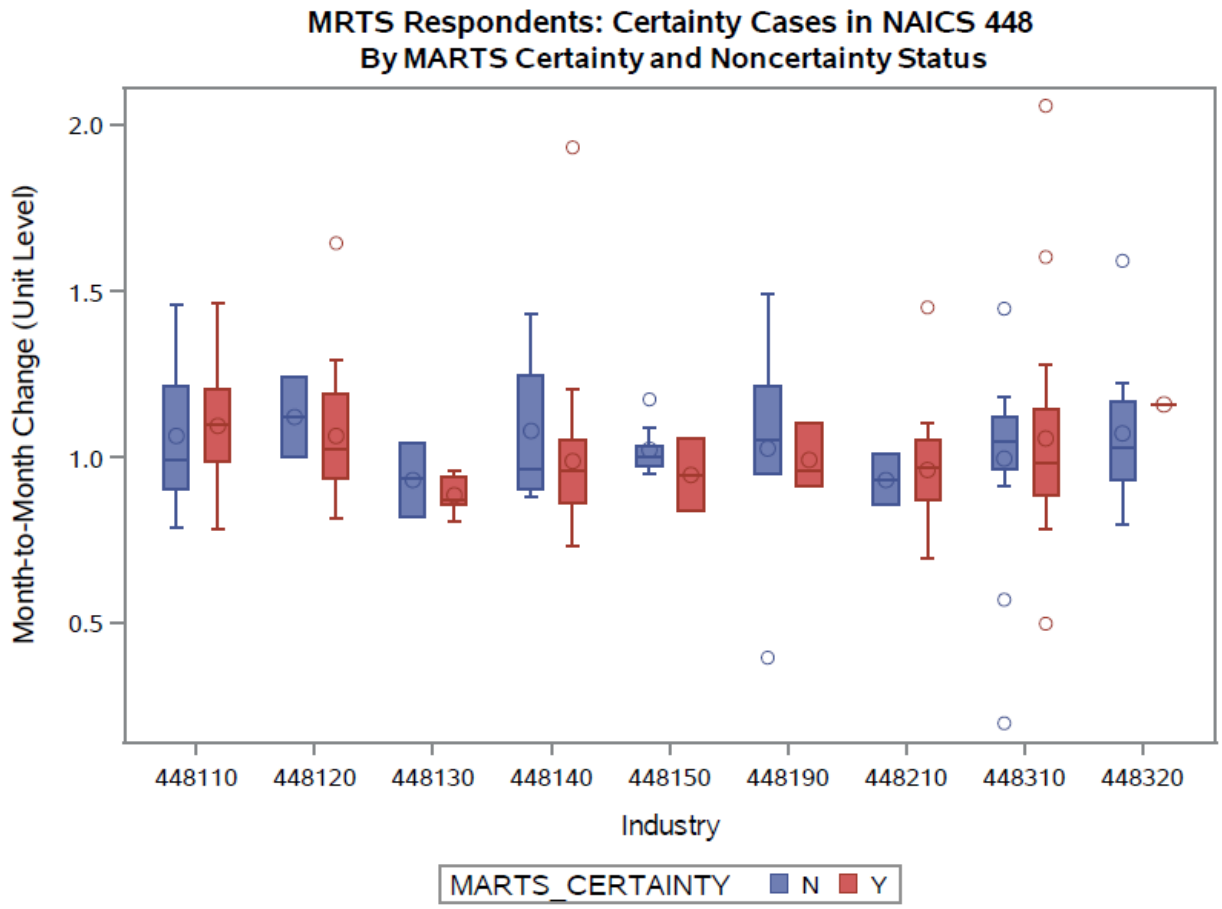

Figure 3: Distributions of Month-to-Month Change in NAICS 448 (Clothing and Clothing Accessory Stores) for MRTS Certainty Respondents by MARTS Certainty Status in March 2016

The survey research literature discusses the potential effects of complex business organizational structure on response (See Thompson, Oliver, and Beck 2015 for a literature review). As in many business surveys, MARTS and MRTS have to distinguish between the survey (sampling) unit, the reporting unit, and the tabulation unit. A *survey unit* is a business selected from the underlying statistical population of similarly constructed units (i.e., from the sampling frame). A *reporting unit* is an entity from which data are collected. Finally, a *tabulation unit* houses the data used in estimation, representing the data at the level used in tabulation.

In a household survey, the reporting unit is usually the sample unit (the sampled address) or is developed from the listed roster of address occupants, and the reporting units are the tabulation units. With a business survey, the three types of unit can differ. The survey unit is defined on the sampling frame, the reporting unit is established by the sampled unit for reporting convenience (of the sampled unit), and the tabulation can be directly derived (from aggregated reporting units) or can be an "artificial" construct used for data storage. In MRTS and MARTS, a sampled company may operate in multiple industries. To ease respondent burden, the reporting unit may provide data on a single form covering all pertinent industries. Upon processing, these response data are allocated ("split out") among tabulation units, using percentage distributions provided by the survey unit or industry models. Figure 4 illustrates the allocation process for a single company that operates in three different industries. For MARTS, the same month-to-month change ratio will be applied to all tabulation units associated with a reporting unit, as the allocation percentages are constant [On a side note, this provides more justification for donating a month-to-month change ratio instead of a value for this survey]. In our analyses, the number of industries associated with the reporting unit is the proxy for organizational complexity and was considered as a matching variable.

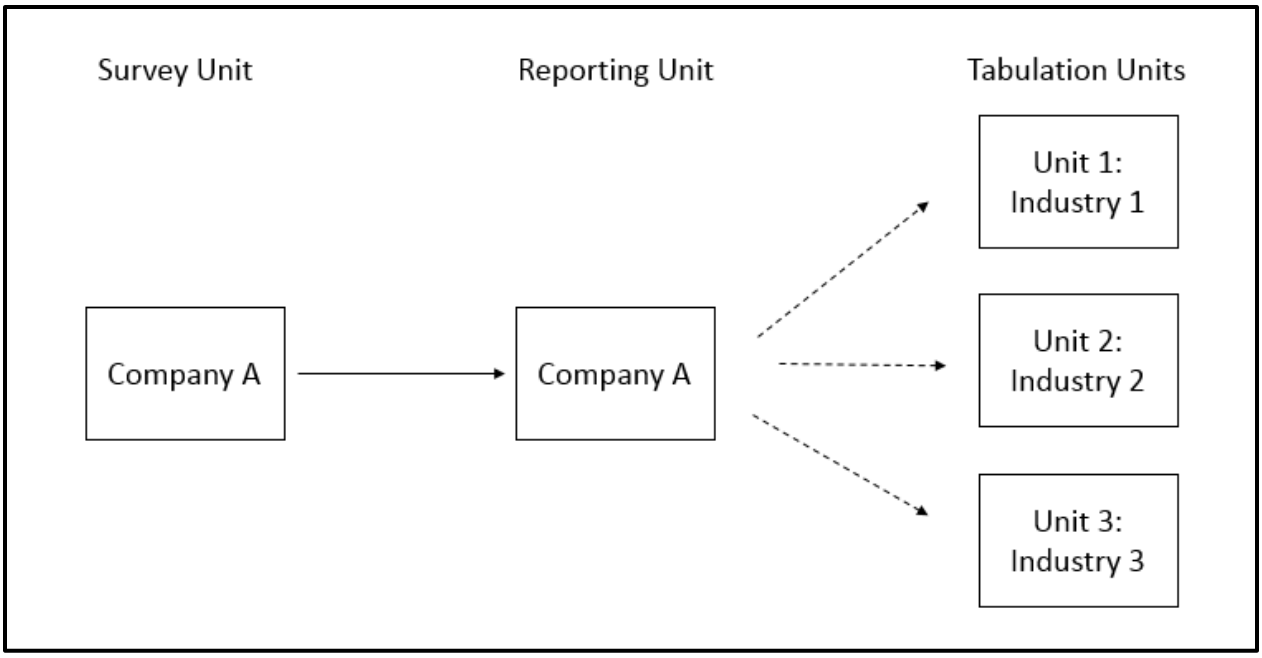

Figure 4: Illustration of Survey Unit, Reporting Unit, and Tabulation Units for MARTS and MRTS

Identifying an informative response propensity model proved to be quite challenging. The association between unit size, length of time in sample, and months since last response and current month response status was consistently strong. Intuitively, we believed that unit drop-out was related to two factors: sample fatigue or business decline. If the former, then we do not anticipate an association or causal relationship between response propensity and month-to-month change in sales. However, a failing business could show a steady decline in sales, yielding very different month-to-month change ratios from other – more typical – businesses in the same industry. We hoped to develop response propensity models for MRTS units by fitting logistic regression models to predict response status or by fitting Cox proportional hazards regression models to the failure-time data (Cox 1975) to develop a survival function score predicting the unit's probability of continuing to participate in MARTS in the current month.

Unfortunately, the fitted logistic regression models[3] using SAS PROC LOGISTIC (SAS/STAT® 9.3 User's Guide 2015) rarely converged. The response pattern for MRTS units is monotonically decreasing as the time in sample increases. However, logistic regression assumes a sinusoidal response function, inappropriate for our data. We considered alternative model fitting approaches to the same data – specifically fitting Binomial, Poisson, and Negative Binomial distribution log and logit link functions using the SAS PROC GENMOD (SAS/STAT® 9.3 User's Guide 2015) – but the models rarely converged. The failure to converge could be due to inappropriate model choice. However, it is more likely due to the weakness of the collective sets of covariates in predicting nonresponse unless the unit is extremely likely to respond (response propensity ≈ 1, generally the largest units who responded in the prior month) or is extremely unlikely to respond (response propensity ≈ 0, generally the smallest units who had not responded in more than one month). Whitehead, Oliver and González (2014) reported similar model-fitting difficulties attempting to fit propensity models to the Quarterly Services Survey data. In the same vein, the omnibus test (all $\beta_I = 0$) was rarely rejected with the proportional hazards regression models fit using the SAS PROC PHREG (SAS/STAT® 9.3 User's Guide 2015).

Table 1 presents the candidate matching variables by matching algorithm. With the greedy matching applications, the first two variable applications represent random hot deck and nearest neighbor hot deck, respectively, with donor usage limits of 1. Random hot deck should be a baseline i.e. any improvements in statistical performance over the baseline is attributable to the purposive match. Nearest neighbor hot deck imputation is recommended when unit

[3] Following Phipps and Toth (2012), we did not incorporate sample design into our models, as we were trying to identify characteristics that were predictive of responding to the survey, not developing models that could be extrapolated to the larger population.

size is predictive of outcome, and represents a simplest case approach to purposive matching. The remaining methods explicitly incorporate our causal assumptions into the matching procedures. For methods three and five, recipients are sorted by descending prior month sales allowing the largest units to get their best matches before the smaller units are matched.

Table 1: Studied Propensity Matching Applications

		Hot deck method	Match Variables	Sort Variables
Greedy	1	Random hot deck	Random number	Random number
	2	Nearest neighbor*	Prior Month Sales	Random number
	3	Propensity	Prior Month Sales	Prior months sales
	4	Propensity	Prior Month Sales and Number of Identified Industries for Reporting Unit	Random number
	5	Propensity	Prior Month Sales and Number of Identified Industries for Reporting Unit	Prior month sales
Optimal	1	Propensity	Prior Month Sales	N/A
	2	Propensity	Prior Month Sales and Number of Identified Industries for Reporting Unit	N/A

*A donor can only be used once.

All applications use Euclidean distances. To facilitate weighting in the matching criteria, the match variables are standardized to have mean 0 and standard deviation of 1 using PROC STANDARD (SAS/STAT® 9.3 User's Guide 2015).

Evaluation Study

We divided our evaluation into two separate phases:

Phase 1: Compare the alternative matching applications listed in Table 1 to find the method(s) that are most effective in selecting donors that yield plausible data

Phase 2: Compare the statistical performance of the recommended matching algorithm from Phase 1 using donated ratios from a year ago (accounting for seasonality) and using donated ratios from five years ago (accounting for seasonality and trading day)

We constructed test decks from 12 months of data from ***respondent*** MRTS certainty units with both current period (x_t) and prior period positive sales values (x_{t-1}). The restriction on the current period value provides a measure of "truth"; the restriction on the second allows imputation in all industries [Note: there are industries where zero values are expected that are ignored in this phase of research]. The first phase of the study uses only the current and prior period values in the statistical period. The second phase of the study further restricts the donor pool by placing the same restrictions on the two sets of historic ratios.

All computations are performed within MARTS tabulation industry. However, to simplify notation, industry code is not referenced in the notation below. The objective of this evaluation is to compare statistical performance of each matching criteria on the available MARTS sample in the studied statistical period. In each industry and statistical period, let n_t represent the number of MARTS sample units that provided valid response data in the current and prior statistical periods and are also MRTS-certainty units. Since the MARTS certainty units are considerably larger than the remaining noncertainty units, we randomly split these cases into donors and recipient groups within 6-digit NAICS in each statistical period. With greedy matching, the results are conditional on the donor and recipient pool; we hoped that the independent splitting by statistical period would lessen the effects of this confounding. The remainder of the donor pool consisted of MRTS-only certainty respondents in the current statistical period. The remainder of the recipient pool consisted of MARTS noncertainty units that were sampled with certainty in MRTS. In summary, there are n_{dt} donors from MARTS and n_{rt} recipients in MARTS ($n_t = n_{dt} + n_{rt}$), with an additional set of donors obtained from the MRTS-only certainty cases. Since the greedy and optimal matching perform one-to-one matches, when there are fewer donors than recipients in an imputation cell, there will be some remaining unimputed cases. Thus, the total number of imputed units in imputation cell is given by $n'_{rt} = \sum_{i=1}^{n_{rt}} I_i$ where I_i is a 0/1 indicator of successful (> 0) imputation for recipient *i*.

We computed the following evaluation statistics within each imputation cell:

Mean Absolute Error (MAE) $$MAE_t^{am} = \sum_{i=1}^{n'_{rt}} \left| x_{t,r(i)} - \left(\frac{x_{t,d(i)}^{am}}{x_{t-1,d(i)}^{am}} \right) x_{t-1,r(i)} \right| / n'_{rt}$$

Unconditional Relative Bias (URB) $$RB_t^{am} = \frac{\sum_{i=1}^{n_{rt}} \left(w_{t,r(i)} \left(\frac{x_{t,d(i)}^{am}}{x_{t-1,d(i)}^{am}} \right) x_{t-1,r(i)} \right) + \sum_{j=1}^{n_{dt}} w_{tj} x_{tj} K_j}{\sum_{r=1}^{n_t} w_{t,} x_t} - 1 = \frac{\hat{x}_t^{am}}{x_t} - 1$$

Conditional Relative Bias (CRB) $$CRB_t^{am} = \frac{\sum_{i=1}^{n_{rt}} \left(w_{t,r(i)} \left(\frac{x_{t,d(i)}^{am}}{x_{t-1,d(i)}^{am}} \right) x_{t-1,r(i)} \cdot I_i \right)}{\sum_{i=1}^{n_{rt}} w_{t,r(i)} x_{t,r(i)} \cdot I_i} - 1 = \frac{\hat{x}_t^{am(R)}}{x_t^R} - 1$$

Where $x_{t,r(i)}$ is the reported (true) value of sales for MARTS recipient *i* at time *t*, $x_{t,d(i)}^{am}$ and $x_{t-1,d(i)}^{am}$ are the values for times *t* and *t-1* for the donor assigned to recipient *i* selected using algorithm *a* and matching variables *m,* x_{tj} is the value of the MARTS unit *j* at time *t*, $w_{t,r(i)}$ is the MARTS sampling weight, and K_j is a 1/0 indicator for a MARTS-donor unit. The unweighted MAE measures the average magnitude of the error per imputed unit and is therefore conditional on obtaining matched values. The unconditional relative bias (URB) measures the overall effect of the imputation error on the tabulated estimates for a given donor selection method. When an imputation cell contains fewer donors than recipients, this measure will be biased. The conditional relative bias (CRB) provides the direction of the imputation bias (if it exists) for the imputed units and gives some indication of magnitude. However, this measure will be extremely sensitive to unit size. For example, the CRB could be close to zero if the imputed value for the largest unit was approximately the same as the true value, even if the remaining imputes are poor. The reverse could also be true i.e. the imputation error could be close to zero for the majority of the smallest units in the imputation cell, but a poorly imputed large unit could yield a very large CRB.

We use a five-percent significance level for all the hypothesis tests discussed in the results presented below.

Evaluation: Phase 1 (Selection of Donors) Results

In this phase of research, we compare the alternative matching applications listed in Table 1 to find the method(s) that are most effective in selecting donors that yield plausible data. We focus on MAE as our primary criterion, using the CRB results to reinforce our initial conclusions. As mentioned above, the CRB is very sensitive to unit size. We do not use the URB in this analysis because the contribution from the un-imputed MARTS donors tends to obscure the contribution from the recipient cases. The URB is more useful in the second phase of this research.

Our first comparison examines differences in accuracy when using one variable versus two (holding the matching algorithm constant), using chi-square tests for independence to substantiate suspected treatment differences. Recall that matching and imputation were performed independently by statistical period and tabulation industry. Thus, in this analysis and the sort comparison described below, the blocks (subjects) are tabulation industry. Our exploratory analyses did not find any evidence for or against a relationship between the number of identified industries for the reporting unit and month-to-month change, but the subject matter experts felt strongly that it should be considered. This contrasts with the consistent evidence of an association between unit size and month-to-month change described above. To perform these tests, we set an indicator variable = 1 when the MAE from the 2-variable match was *smaller* than the corresponding 1-variable match. In all applications, we rejected the null hypothesis of independence, providing evidence that omitting the second variable (number of identified industries for the reporting unit) improved the matching quality on MAE (p-value = 0.0007 for Greedy Match/Random Sort, p-value = 0.0002 for Greedy Match/Ordered Sort, p-value = 0.0314 for Optimal Match). We performed analogous tests with the CRB, with the same results (p-value = 0.0001 for Greedy Match/Random Sort, p-value = 0.0001 for Greedy Match/Ordered Sort, p-value = 0.0246 for Optimal Match). Consequently, we dropped the 2-variable match options from further consideration.

Next, we looked at the effect of sorting with the Greedy Matching (1-variable), again performing chi-square tests for independence. For these analyses, we set an indicator variable = 1 when the MAE (or CRB) from the match obtained using an ordered sort (Greedy match method 3) was *smaller* than the corresponding match obtained using a random sort. With the MAE, the results were less conclusive, with a p-value = 0.0623 and indeed, there were several

industries where the results within the same industry were split between the two methods. With CRB, results differed greatly by industry, and we were unable to reject the null hypothesis of independence (p-value = 0.2146).

This left us with four candidate matching algorithms: (1) random hot deck, (2) nearest neighbor hot deck with a donor limit of 1, (3) greedy matching with units sorted by descending prior month sales with prior month sales as the matching variable, and (4) optimal matching on prior month sales. The evaluation is a complete block design experiment. In our design, the tabulation industries represent the blocks, and the treatments are the matching algorithm. The independent variable (obtained each statistical period) is the evaluation statistic, either MAE or Absolute Value of CRB. 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. 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. We used the more conservative two-way analysis of variance statistic on ranks recommended by Iman and Davenport (1980).

For each evaluation statistic, we ranked the outcome from the four treatments within statistical period and industry (using the mean value of the ranks for ties), aggregated the ranked values within industry, and performed the Friedman Tests within statistical period. Tables 2 and 3 provide the aggregated ranks over industry by statistical period for MAE and CRB, respectively. Here, the omnibus test is of primary interest, where the null hypothesis is that the aggregated ranks for each treatment are all equal. If the null hypothesis is rejected, then pairwise comparisons are appropriate. The final column of each table provides the p-value for the omnibus test.

Table 2: Aggregated Ranks for MAE Comparison with Industries as Subjects,

Statistical Period	Random Hot Deck	Nearest Neighbor	Greedy (Size Sort)	Optimal	P-Value (Friedman)
201603	89	67	76	68	< 0.0001
201604	83	72	74.5	70.5	< 0.0001
201605	87.5	73.5	71.5	67.5	< 0.0001
201606	92	69	69.5	69.5	< 0.0001
201607	89	71.5	68	71.5	< 0.0001
201608	92.5	71.5	59.5	76.5	< 0.0001
201609	108	61.5	66	64.5	< 0.0001
201610	84	80.5	67.5	68	< 0.0001
201611	91.5	67.5	64.5	76.5	< 0.0001
201612	81.5	70	75.5	73	< 0.0001
201701	97.5	75	64	63.5	< 0.0001
201702	89	67.5	76	67.5	< 0.0001
(All)	1084.5	846.5	832.5	836.5	

The significant results shown in Table 2 can be attributed to the poor performance of the Random Hot Deck relative to the other three methods. Optimal matching and greedy matching with units sorted by descending prior month sales yield similar MAE's on average, although there is no consistent pattern within industry and across statistical period.

Table 3: Aggregated Ranks for Absolute Value of CRB Comparison with Industries as Subjects,

Statistical Period	Random Hot Deck	Nearest Neighbor	Greedy (Size Sort)	Optimal	P-Value (Friedman)
201603	71	82	80	67	< 0.0001
201604	90	74	67.5	68.5	< 0.0001
201605	74.5	73.5	70.5	81.5	< 0.0001

201606	86	75	69.5	69.5	< 0.0001
201607	86	66.5	71	76.5	< 0.0001
201608	91.5	60.5	73.5	74.5	< 0.0001
201609	99	70.5	64	66.5	< 0.0001
201610	71	75.5	76.5	77	< 0.0001
201611	71.5	78.5	75.5	74.5	< 0.0001
201612	87.5	69	71.5	72	< 0.0001
201701	94.5	67	64	74.5	< 0.0001
201702	99	73.5	68	59.5	< 0.0001
(All)	1021.5	865.5	851.5	861.5	

Initially, we planned to take the direction of the CRB into account as well as the magnitude. However, negative relative biases were extremely rare, and there was consistent pattern within industry and statistical period by treatment. Consequently, we focus on the magnitude of the CRB. Again, Random Hot Deck has the worst performance. In this case, the other three treatments are about equally effective, even given the earlier caveats about the CRB's sensitivity to unit size.

For both MAE and CRB, the greedy matching with units sorted by size and the optimal matching slightly outperform the nearest neighbor applications. Of the two, there is little compelling evidence to recommend one method over the other for these data. However, greedy matching can give very poor results at the end of the list, when donors may be very different from recipients. This is not a consideration with optimal matching, since the distance is minimized over all recipients. Furthermore, there were software issues with the greedy matching when there were fewer donors than recipients (the software would not execute). In these cases, we "tricked" the application by switching donor and recipient classifications. This bookkeeping was unnecessary with the optimal matching software. For these reasons, we decided to restrict the matching algorithm to optimal matching to compare alternative donor pools.

Evaluation: Phase 2 (Selection of Hot Deck Donor Pool) Results

In practice, insufficient MARTS/MRTS respondent data will be available to use for imputation in the current statistical period. Instead, matched donors will provide historic ratios, either from the same two month period in the prior year (seasonal ratios) or from the most recent year with the same weekday composition by month as the current year (seasonal ratios and trading day effects), which for March 2016-February 2017 is five years earlier. There are excellent arguments for using the second set of donor ratios with retail trade data, which are highly seasonal with consistently significant trading day effects in most industries. However, adding these further restrictions on the donor pools does limit the number of potentially imputed missing observations.

Table 4 presents summary statistics for the donor-to-recipient ratios obtained from our datasets (30 industries). To obtain these measures, we compared counts within industry and statistical period, then averaged the measures within industry (across statistical periods). The summary measures presented in Table 4 use the entire set of averaged ratios (across industries). Of course, these ratios will vary within industry by statistical period and may not be representative of what would be seen in a production system due to our experimental design. However, they do provide some insight into the effects of restricting the donor pool.

Table 4: Summary Statistics on Average Donor to Recipient Ratios and Donor to Donor Ratios in the Study

Ratio	Min.	Q1	Med.	Q3	Max.
Donors (1 Year Ago) to Recipients	0.89	1.69	2.14	3.19	5.58
Donors (5 Years Ago) to Recipients	0.55	0.97	1.38	1.69	2.70
Donors (5 Years Ago) to Donors (1 Year Ago)	0.24	0.50	0.59	0.73	1.00

Using the older ratios greatly reduces the donor pool in many cases. In turn, this will have an effect on the matching, since there are fewer choices. Using the more current ratios (from a year ago) creates more matching options.

Of course, if the trading day effect is as important as the seasonal effects, then it would be unwise to use the larger – and more recent – sets of ratios. We performed chi-square tests for independence to assess the treatment effect

(donor choice) on MAE and on the Absolute Unconditional Relative Bias (URB), again ignoring direction (neither method consistently underestimated). Except for the rare case when there are different numbers of imputed recipients within the same imputation cell – which occurs when one or both of the donor pools contained fewer observations than the intended number of recipients – the denominator for the MAE is the same within imputation cell and industry for each treatment. Using the URB guarantees similar conditions. However, the treatment effects on the URB are somewhat mitigated by the large MARTS donor units' values that are included in the numerator and denominator, making this a less informative measure.

With the MAE, there was a significant treatment effect (p-value = 0.0006): in 17 industries, the 5-year-old donor imputation outperformed the 1-year old donor imputations in the majority of statistical periods; in 11 industries, the reverse was true; and in two industries, the methods tied. With the URB, there was also a significant treatment effect (p-value=0.0475): in 16 industries, the 5-year-old donor imputation outperformed the 1-year old donor imputations in the majority of statistical periods; in nine industries, the reverse was true; and in five industries, the methods tied.

These results are statistically significant. However, there are systematic differences by industry, likely due to the differing trading day effects. Furthermore, the corresponding MAEs within an industry and statistical period often differed by at least 10-percent (i.e. the MAE's were at least 10-percent higher for the "worse" method). Choosing the wrong imputation method could therefore lead to severe overestimation. In the cases where the 5-year old donor ratios were preferable, there really was no "trade-off" between the reduced donor pool size and imputation quality. Unfortunately, when the reverse was true, then the larger donor pool does appear to lead to improved imputation.

Furthermore, the study design that we use is not a realistic proxy for production. We restrict our analysis data to respondents (to provide a "true" value) and randomly split the largest units into donors and recipients. In reality, the donor and recipient composition will not be as congenial. Moreover, optimal matches will be optimized for the available data and may not be as effective as seen here. What's more, for the study period, we were fortunate that the calendar that matched the current-year weekday composition was only five years old. For 2017-2018, the most recent calendar match would be 2006-2007. In this case, the benefits of incorporating trading day effects into the imputation might be offset by changes in the economy beyond seasonality, in addition to the severe limitations on the donor pool. In theory, the imputation method that yields the most accurate predicted values should in turn lead to the estimated MARTS total that is closest to the preliminary MRTS estimate for the same reference period. Failing to find this single method, we recommend testing both on empirical MARTS data, constructing the resultant link relative estimated totals, and performing additional comparisons before making any recommendations.

Finally, another challenge with imputing with historic month-to-month changes, no matter how many years in the past, are the effects of price. Some retail industries, namely gas stations and fuel dealers, have sales levels that are extremely price driven. Prices can shoot up over matter of days or weeks causing large month-to-month changes that are not part of any repeating seasonal pattern. Using these historic month-to-month changes for current month imputation could be problematic.

Conclusion

Hot deck imputation is often lauded for yielding plausible imputed values without depending on an unprovable model, in contrast with other deterministic or stochastic imputation methods such as regression or ratio imputation or sequential regression multiple imputation, to name a few examples. However, hot deck imputation applications implicitly use models, for developing imputation cells and as match variables. Similarly, propensity score matching relies on models. For both, conventional wisdom tends to shy away from parsimonious models. After all, why not include any possible variable that could be predictive of outcome?

In surveys, restricting matching variables can be the wise choice. Often – as was seen here – our intuition is simply not validated by outcome. In our applications, the effectiveness of the propensity score matching actually decreased as variables were added. This decreased effectiveness might be due to a wrongly assumed causal relationship. Equally likely, it could have been the consequence of the more restrictive matching conditions. Regardless of the reason, the recommendation to restrict the set of matching variables seems to defy conventional wisdom.

Or, does it? Traditionally, propensity score matching combines a set of covariates into a single score, the prediction of the collective set of variables to the studied outcome. This presupposes that such a set of covariates can be found. This is the same challenge that often impedes the development of viable response propensity models in business surveys. Demographic surveys often have a suite of available predictors. Business surveys often do not. Hence, the emphasis on data collection, especially for the largest sampled businesses in a survey. Often, there is no substitute for the sampled units.

Of course, this general problem could change, as auxiliary data become more available and "big data" analysis techniques become more accessible. Having more data sources could solve the "lack of predictor" problem. That said, it could also introduce new coverage discrepancies as well as unit discrepancies with their own errors. In the meantime, the simple propensity optimal matching algorithm studied – and ultimately recommended – in this paper provides an easily implemented missing data treatment that relies on a minimal set of assumptions and can (in many cases) provide viable and automated predictions.

Acknowledgments

We thank Brian Dumbacher, Carma Hogue, and Scott Scheleur for the careful review of earlier versions of this manuscript and Andromache Mason for her programming efforts on this project.

References

Andridge, R.and Little, R. (2009). The use of sample weights in hot deck imputation. *Journal of Official Statistics*, 25(1), 21-36.

Andridge, R. and Little, R. (2010). A review of hot deck imputation for survey non-response. *International Statistical Review,* 78 (1), 40-64.

Brick, J.M. and Kalton, G. (1996). Handling missing data in survey research. *Statistical Methods in Medical Research*, 5, 215-238.

Conover, W. (1999). Practical nonparametric statistics. New York: John Wiley.

Cox, D.R. (1975). Partial likelihood. *Biometrika* 62: 269–276. Doi: http:dx.doi.org/10.2307/2335362.

Czaplicki, N., Gonzalez,Y, and Bechtel, L. (2018). Finding an estimator that minimizes revisions in a monthly indicator survey. *Proceedings of the FCSM Research Conference.*

Demsar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*: 7, pp. 1–30.

Friedman, M. (1940). A comparison of alternative tests of significance for the problem of M rankings. *Annals of Mathematical Statistics,* 11, pp. 86–92.

Gotway Crawford, C.A. (2013). Comment. *Journal of Survey Statistics and Methodology*, 1 (2)13, 118–124, https://doi.org/10.1093/jssam/smt013

Iman, R.L. and Davenport. J.M.(1980*).* Approximations of the critical region of the Friedman statistic. *Communications in Statistics*: pp. 571–595.

King, G. and Nielsen, R. (2016). Why propensity scores should not be used for matching. http://gking.harvard.edu/files/gking/files/psnot.pdf.

Madow, L.H. and Madow, W.G. (1978). On link relative estimators. *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Pearl, J. (2010). An introduction to causal inference. *The International Journal of Biostatistics*, *6*(2), http://doi.org/10.2202/1557-4679.1203.

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), 772-794.

Rosenbaum, P.R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41-45.

Rubin, D. B. and Thomas, N. (1996). Matching using estimated propensity scores: relating theory to practice. *Biometrics*, 52(1), 249-264.

"SAS/STAT(R) 9.3 User's Guide". SAS/STAT(R) 9.3 User's Guide. N.p., n.d. Web. 09 Oct. 2015.

Smith, J. and Todd, P.E. (2001). Reconciling conflicting evidence on the performance of propensity-score matching methods. *The American Economic Review*, 91 (2), 112-118.

Thompson, K.J. and Oliver, B. (2012). Response rates in business surveys: going beyond the usual performance measure. *Journal of Official Statistics,* 28, 221–37.

Thompson, K.J. and Oliver, B, and Beck, J. (2015). An analysis of the mixed collection modes for two business surveys conducted by the U.S. Census Bureau. *Public Opinion Quarterly*, 79(3), 769–789, https://doi.org/10.1093/poq/nfv013

Thompson, K.J. and Washington, K.T. (2013). Challenges in the treatment of unit nonresponse for selected business surveys: a case study. *Survey Methods: Insights from the Field*, available at http://surveyinsights.org/?p=2991.

Whitehead, D., Oliver, B., and González, Y. (2014). The use of indicators to assess the quality of business survey returns during data collection. *Proceedings of the Section on Survey Research Methods*, American Statistical Association.