Using Classification and Regression Trees to Model Survey Nonresponse

Sharon Lohr[1], Valerie Hsu[2], Jill Montaquila[1]
[1]Westat, 1600 Research Boulevard, Rockville, MD 20850
[2]Promontory Financial Group, 801 7th St. NW, Washington, DC 20006

Abstract

In the computation of survey weights to be used for the analysis of complex sample survey data, an adjustment for nonresponse is often an important step in reducing bias. These adjustments depend upon estimated response propensities, which are traditionally obtained through empirical response rates within weighting classes or through logistic regression modeling. In this paper, we discuss possible benefits of using regression trees and random forests for estimating response propensities in surveys, and describe how these models might be used to reduce nonresponse bias. We review issues for their use with complex surveys such as the effect of survey weights and clustering, pruning criteria, and loss functions, and we explore the sensitivity of results to these conditions.

Key Words: classification trees, random forests, response propensities, survey weights, weighting class adjustments

1. Nonresponse Adjustments in Surveys

Almost all sample surveys have nonresponse, which reduces the sample size and causes concern about potential bias of estimates. Weighting adjustments are commonly used to try to adjust for nonresponse. In these methods, the design weights for responding units are adjusted to account for the nonrespondents. Brick and Montaquila (2009) gave an overview of weighting methods used to adjust for nonresponse.

Suppose that the quantity of interest is the population mean $\bar{Y} = \sum_{i=1}^{N} y_i / N$. Let $\pi_i =$ P(unit i included in sample). Then the design weight for unit i is $d_i = 1/\pi_i$ and, if everyone in the selected sample S responds, the estimated mean $\bar{y} = \sum_{i \in S} d_i y_i / \sum_{i \in S} d_i$ is approximately unbiased for the population mean $\bar{Y}$.

With nonresponse, however, the set of respondents R is a subset of S. Under theoretical models reviewed in Bethlehem (1988, 2002) and Brick (2013), unit i is assumed to have an intrinsic response propensity $\psi_i =$ P(unit i responds to the survey). The estimated population mean using the design weights and the respondents, $\bar{y} = \sum_{i \in R} d_i y_i / \sum_{i \in R} d_i$, is approximately unbiased if $\text{Cov}(y, \psi) = 0$. Thus, under this theory it is desired to form weighting classes such that the covariance is 0 within each class. The covariance equals 0 if the response variable is homogeneous within the class [$Var(y) = 0$], the response propensity is homogeneous within the class [$Var(\psi) = 0$], or the correlation between y and ψ equals 0. Because most surveys have a large number of possible y variables, in this paper we focus on methods for obtaining weighting classes in which the response propensities are homogeneous within the classes, i.e., having $Var(\psi) = 0$.

Unlike the selection probabilities π_i, the response propensities ψ_i are assumed to be unknown. In practice, the response propensities are estimated using a model that predicts the response indicator from covariates known for each unit in the selected sample S. Logistic regression, a tree-based model, or another type of model may be used to obtain estimated response propensities $\hat{\psi}_i$.

The estimated propensities may be used in several different ways for nonresponse bias adjustment. One method treats each observation as its own class and forms a nonresponse-adjusted weight as $w_i = d_i/\hat{\psi}_i$. This fits in with a view of nonresponse as a second phase of sampling (Oh and Scheuren 1983, Särndal and Lundström 2005). This method, however, can produce highly variable weights, thereby increasing the variance of population estimates; in addition, results can be sensitive to model misspecification.

A second method uses cutpoints of the estimated propensities to form weighting classes (Potter et al. 2006). Eltinge and Yansaneh (1997) noted that using equal quantiles of the response propensities for the cutpoints gives some control over the number of sampled units or respondents in each class. Using classes rather than individual weights can reduce variation from the uncertainty in the estimated propensities (Brick 2013).

A third method forms the weighting classes directly from the terminal nodes of classification or regression trees. As noted by Toth and Phipps (2014), tree-based methods do not require the user to specify a parametric model because they automatically select variables and fit interactions if needed. The trees are easy to interpret, and can provide insight into the nonresponse mechanism for adaptive design or for investigating methods to increase the response rate in future surveys.

There has been a great deal of work on various aspects of fitting classification and regression trees to predict response propensities, and we refer the reader to Toth and Phipps (2014) for a comprehensive bibliography of work done through 2014. There has been less research on the effects of different classification and regression tree methods, or of different options that can be used when fitting trees with survey data. In this paper, we perform a simulation study that explores the performance of the nonresponse weighting adjustments for different tree-type models. The study is not exhaustive: Loh (2014) described more than 20 software packages that have been developed for fitting trees and ensembles of trees to data sets, and each of these algorithms has multiple input parameters. The methods used in this paper include representatives of the major types of algorithms, but we do not study all of them. We focus on the effects of using sampling weights, of adjusting for clustering and weighting, and, primarily, on the differences among the algorithms. The methods are evaluated by the accuracy of the propensity estimates and by the mean squared error of the estimated population mean and quartiles of a response variable, y.

2. Simulation Study Population and Design

Data from the 2009-2013 5-year American Community Survey Public Use Microdata Samples (ACS PUMS) were treated as the population for the simulation study, and repeated samples were drawn with different nonresponse-generating mechanisms from this population. Table 1 lists the variables used from the 2009-2013 ACS.

Table 1: Variables Used from ACS

Variable name	*Variable description*
HINCP	Household income
HUPAC	Household presence and age of children; used to derive binary indicator of children in household
TEN	Home tenure; used to derive binary indicator of whether home is rented
DIS	Disability indicator; used to derive binary indicator of whether at least one person in household has a disability
HICOV	Health insurance coverage indicator; used to derive binary indicator of whether every household member has health insurance coverage
WKL	When last worked; used to derive count of the number of household members who worked within the past 12 months
PINCP	Person's total income; used to derive count of the number of income-earning household members
LNGI	Indicator of limited English-speaking households
REGION	Census region
VALP	Property value. This was recoded as a categorical variable with five categories: the first four categories were based on the quartiles of the variable, and the fifth category indicated the household was missing this variable.

The variable HINCP was chosen as the response variable of interest, y. Household income is often subject to missingness in surveys and has a highly skewed distribution. The mean of household income might therefore be expected to behave relatively badly even after nonresponse adjustment. The other variables in Table 1 were assumed to be known for all units in the selected samples. Some of these were used to generate the nonresponse mechanism for different simulation settings, and all of them were used to model response propensities.

For this study, only records with non-missing data for each of the variables listed in Table 1 (with the exception of VALP) were retained in the population. The resulting set of 6,019,599 household-level records from the ACS was treated as the population for the study. Public use microdata areas formed the primary sampling units (PSUs) in generating clustered samples; the research data set contained a total of 3,344 PSUs.

A $2 \times 2 \times 4 \times 2$ factorial design was used to generate nonresponse mechanisms and samples for the study. The factors were:

- Response rate, with levels 50% and 80%.
- Number of PSUs (NUM_PSU) in each sample, with levels 25 and 100.
- Nonresponse generating mechanism. Four models were used:
    1. Missing completely at random (MCAR). Households in the population were selected randomly to be nonrespondents. For this mechanism, we would expect estimates using the unadjusted sampling weights to be unbiased; ideally, the weight adjustment methods would "do no harm."
    2. Missing at random with linear function of covariates (MAR, linear). The model used to generate nonresponse had main effects for tenure, presence of children, and number of income earners in the household.
    3. Missing at random with interaction (MAR, interaction). The model used to generate nonresponse had main effects for tenure and presence of children and an interaction term involving these two variables.

4. Missing not at random (MNAR). The model used to generate nonresponse had main effects for tenure, presence of children, and household income (the y variable). High-income households were less likely to respond.

- PSU-level variability of latent nonresponse (PSU_NR_VARIANCE), with levels 0 and 0.25.

For each of the 32 simulation settings, the response propensity ψ_i was generated for each household i in the population using three steps. First, $z_k \sim N(0, \text{PSU\_NR\_VARIANCE})$ was generated independently for each PSU k in the population. Then, the latent nonresponse variable for household i was defined as $L_i = f(x_i) + z_{k(i)}$, where $f(x_i)$ is the function of the covariates used to generate the particular nonresponse mechanism (MCAR, MAR linear, MAR interaction, or MNAR) and $z_{k(i)}$ is the value of z_k for the PSU containing observation i. Finally, the values of L_i were scaled to give the desired response rate by finding c such that $1 - \frac{1}{N}\sum_{i=1}^{N} \Phi(L_i - c) =$ response rate, where Φ is the cumulative distribution function for a standard normal distribution. The response propensity was calculated as $\psi_i = 1 - \Phi(L_i - c)$. The response indicator for household i in the population, r_i, was generated independently from a Bernoulli distribution with response propensity ψ_i.

Two hundred samples were selected from the population for each simulation setting, using PROC SURVEYSELECT from SAS/STAT® software (SAS Institute 2011). In order to generate clustered, unequal probability samples of households, a simple random sample of NUM_PSUS PSUs was selected, and within each sampled PSU a simple random sample of 100 households was selected. For each sample the design weight d_i of a sampled household was calculated as the inverse of its inclusion probability. The samples were then exported and models predicting response propensities were fit in the R statistical software package (R Core Team 2015), version 3.2.0.

3. Models for Estimating Response Propensities

The core of tree-based methods is recursive partitioning, in which the units in the selected sample S are split into two subsets based on the division of a covariate. These subsets are split further, and the process is continued until the final subsets, called the terminal nodes, meet user-specified sample size or homogeneity conditions. Different tree-based methods use different criteria to determine tree size and the covariates used for splitting. Forest methods grow many different trees using subsamples of the data, then average the predicted values across the trees.

3.1 Recursive Partitioning (rpart)

The R function *rpart* (Therneau and Atkinson 2015) is based on the classification and regression tree methodology described in Breiman et al. (1984). The measures used to choose variables for splitting are based on a node impurity measure I(node), which is often based on the Gini index, the information index, or the residual sum of squares. For each parent node in turn, the chosen variable split maximizes the reduction in impurity, I(parent node) – I(left child) – I(right child), among all possible variable splits. Each split may be partitioned further based on the best available predictor at each level. The splitting process continues until the specified stopping rules are met. When the variable being predicted is dichotomous, as in this application where we are predicting the response indicator, the initial splits for a classification tree are often similar to those for a

regression tree. The regression tree may have more terminal nodes, however, because regression and classification trees use different loss functions: a regression tree typically uses the deviance while a classification tree typically uses the number misclassified.

The function allows the user to specify a number of parameters to control the tree-fitting. The parameter *minbucket*, the minimum number of observations in a terminal node, was set equal to 20 for all trees. We varied the following factors, using two levels for each:

- *Method:* classification (method = "class") or regression (method = "anova") tree.
- *Weight:* weight = 1 for all observations, or weight = design weight. The function *rpart* treats weights as case weights and permits non-integer-valued weights. The control parameters for *rpart* are based on unweighted counts. For example, *minbucket* = 20 requires each terminal node to have at least 20 sampled households, as opposed to requiring the sum of the weights in the terminal node to be 20. The measures for tree-fitting are invariant to the scale of the weights, so that a tree fit to a data set in which each weight is 1 will be the same as the tree fit to the same data set with each weight set to 1,250. With unequal weights, the fits will differ.
- *Pruning:* no or yes. The procedure developed in Breiman et al. (1984) specifies first growing the tree out as far as possible until either the minimum node size criteria are met or no further improvement in node purity is possible. Then, because the tree-growing procedure can result in some splits occurring because of pure noise, the tree is pruned to a smaller size. With no pruning, the tree was grown out as far as possible subject to *minbucket*. With pruning, cross-validation was used to prune the tree as described in Therneau and Atkinson (2015). The number of cross-validations was set equal to 5.
- *Misclassification loss ratio:* (loss for misclassifying a nonrespondent as a respondent)/(loss for misclassifying a respondent as a nonrespondent) = 1 or 2. This factor applies to classification trees only.

All other parameters were set equal to their default values.

After the trees were grown, each terminal node with fewer than 20 respondents was combined with its nearest neighbor with respect to response propensity. This process was repeated until each terminal node had at least 20 respondents, in order to reduce instability from weighting classes with few respondents.

The splitting rules in *rpart* tend to favor continuous covariates and categorical covariates with many categories, simply because these have more possible splits (Breiman et al. 1984, p. 42; Loh 2014). This occurs because of a problem analogous to multiple testing.

3.2 ctree

The R package *party* (Hothorn et al. 2015) fits conditional trees through the function *ctree*. The method assumes that the observations are independent—an assumption that is not typically met for survey data—and determines splits from the results of hypothesis tests. The first step is to test the global null hypothesis of independence between y and the set of potential explanatory variables $x_1, \ldots x_m$. If that hypothesis is not rejected, the splitting stops. Otherwise, the covariate with the highest association with y is selected for splitting the tree. The global null hypothesis is the intersection of the variable-wise tests, and permutation tests are used for each of the component hypotheses (Hothorn et al. 2006).

The conditional tree method eliminates the step of pruning: the hypothesis tests determine when to stop splitting nodes. Hothorn et al. (2006) argued that the conditional tree method avoids the variable selection bias from *rpart*, which is more likely to choose categorical variables with many categories for splitting.

Ctree, in contrast to *rpart*, requires integer-valued weights, and fits the tree pretending that there are d_i observations in the data set with observation i's variables. The scale of the weights has a large effect on the splits used in *ctree.* The tree from running *ctree* with the weight set equal to 1,250 for every observation will have many more splits than if the weight equals 1 for every observation. This occurs because in the hypothesis testing steps with weighted data, *ctree* thinks there are $1{,}250n$ observations in the data set, which leads to many spurious rejections of the null hypotheses used in the tree-splitting decisions. *Ctree*, like *rpart,* has a control parameter *minbucket*, but in *ctree* the option *minbucket* = 20 requires the sum of the weights in the terminal node to be 20. In many surveys that weight sum requirement will be met with one observation in the terminal node.

Properties of hypothesis tests are affected by a complex survey design, and we performed simple modifications of the input parameters to account for the unequal weighting and clustering of the sample. Because *ctree* requires integer weights, it is not practicable to scale the weights so that they sum to the sample size n. Instead, for the purposes of forming nonresponse adjustments, we defined $d_i^c = \text{ceiling}(d_i)$. Most surveys have relatively large weights, and the effects of using d_i^c instead of d_i, for the purpose of estimating propensities and creating weighting classes, were expected to be small. We set the critical value for the test statistic to be $t_{crit} \times wf$, where t_{crit} is the critical value for the nominal α-level test, and the weight factor $wf = \sqrt{\sum_i d_i^c / n}$ bases the hypothesis test on an effective sample size of n.

Clustering and other features of the complex sampling design also affect properties of hypothesis tests. We accounted for clustering by dividing the test statistic by the square root of the design effect for the response indicator r_i. Note that in general, the design effect for the response variable by itself is greater than the design effect for the regression coefficients (Skinner 1989), so the adjustment is expected to be conservative for this simulation study. In terms of the control parameters for *ctree,* we can include a design effect for the effect of unequal weights, clustering, and stratification by using critical value $t_{crit} \times wf \times \sqrt{deff}$.

We varied the following factors for *ctree*, using two levels for each. Only regression trees were fit.

- *Weight:* weight = 1 for all observations, or weight = d_i^c. With unit weights for all observations, *minbucket* was set to 20. With weight = d_i^c, *minbucket* was set equal to $20 \times \max(1, \min(d_i^c))$, and the critical value for the test statistic was defined to be $t_{crit} \times wf$.
- *Clustering adjustment:* no or yes. When yes, the critical value for the test statistic was multiplied by the square root of the design effect for the response indicator.
- *Bonferroni:* no or yes. When yes, a Bonferroni adjustment was used to compensate for the multiple testing in the global null hypothesis.

3.3 Forest Methods

Methods such as *rpart* and *ctree* produce easily interpretable individual trees that group observations with similar propensities together. The cost of that easy interpretation, however, is that predictions may have high variance. Ensembles of decision trees, or forests (Breiman 2001), reduce the variance by growing multiple trees without pruning, and averaging the predictions from the different trees. Each tree is fit to a different subset of the data, using a different subset of the explanatory variables.

Two packages were used to fit forests in R. The first, *randomForest* (Liaw and Wiener 2002, 2014, based on the original Fortran program by Breiman and Cutler 2004), is derived from the method in Breiman (2001). This package uses the recursive partitioning method in Breiman et al. (1984) for the base learning components. Weights are not allowed. The package *cforest,* in the *party* package (Hothorn et al. 2015), implements the forest methodology using conditional trees as the base learning components.

We varied the following factors for the forest methods:

- Method: randomForest (subsequently referred to as rforest) or cforest.
- Number of trees: 30 or 150. Occasionally, the fit with 30 trees failed to produce a response propensity estimate for one or more households; those were then imputed with the mean propensity.
- Weight: (for cforest only) weight = 1 for all observations, or weight = d_i^c.

Note that the forest methods produce an estimated response propensity but do not produce individual trees. Therefore, weighting classes must be formed by grouping based on estimated response propensities.

3.4 Trees with Random Effects

Half of the simulation settings described in Section 2 include randomly generated PSU-to-PSU variability in response propensities. Such a situation can occur in surveys when, for example, PSUs are establishments or governmental entities that may encourage different participation rates. Between-PSU variability can be modeled by including covariates that are related to PSU-level response rates or by including PSU as a categorical covariate. We explored the alternative of including PSU as a categorical predictor. However, because of the large number of PSUs and subsets, this resulted in infeasibly large computation times for use in a simulation study. In addition, as noted above, *rpart* and related methods have a predilection for splitting on categorical variables with large numbers of categories, with the result that most splits were based on subsets of PSUs and predictions were poor. Instead of using PSU as a fixed effect, we adapted an approach from Skinner and D'Arrigo (2011), who used a random effects model with logit link function to estimate clustered response propensities.

Sela and Simonoff (2012) proposed using trees with random effects estimated by the EM algorithm (*RE-EM*) for data with clustering. The tree building is based on the R function *rpart,* and a linear model is assumed for the random effects. The algorithm iterates between steps of estimating the random effects using maximum likelihood and fitting a tree to model the response indicator after subtracting the modeled random effects. This method produces estimated response propensities that are a combination of estimates from terminal nodes and estimated random effects. As with the forest methods, the estimated response propensities must be used directly or through weighting classes.

3.5 Logistic Regression

The final method used for predicting response propensities was logistic regression, fit using R function *glm*. The logistic models predicted response indicator using the main effects (no interactions) of the covariates derived using the variables in Table 1. One factor was varied for logistic regression: the models were fit with and without weights.

Altogether, 29 models were fit to estimate response propensities from each sample: 8 with *rpart* classification, 4 with *rpart* regression, 8 with *ctree*, 2 with *rforest*, 4 with *cforest*, 1 with *RE-EM*, and 2 with logistic regression.

4. Results

4.1 Accuracy of Estimated Propensities

Each model produces an estimated response propensity $\hat{\psi}_{bi}$ for every household i in sample b. These are compared to the true response propensities ψ_i described in Section 2. Let $\widehat{\boldsymbol{\psi}}_b$ denote the vector of estimated response propensities from sample b. We considered the mean squared error $\text{MSE}\left(\widehat{\boldsymbol{\psi}}_b\right) = n^{-1} \sum_i \left(\hat{\psi}_{bi} - \psi_i\right)^2$ and the mean absolute error $\text{MAE}\left(\widehat{\boldsymbol{\psi}}_b\right) = n^{-1} \sum |\hat{\psi}_{bi} - \psi_i|$ of the estimated propensities.

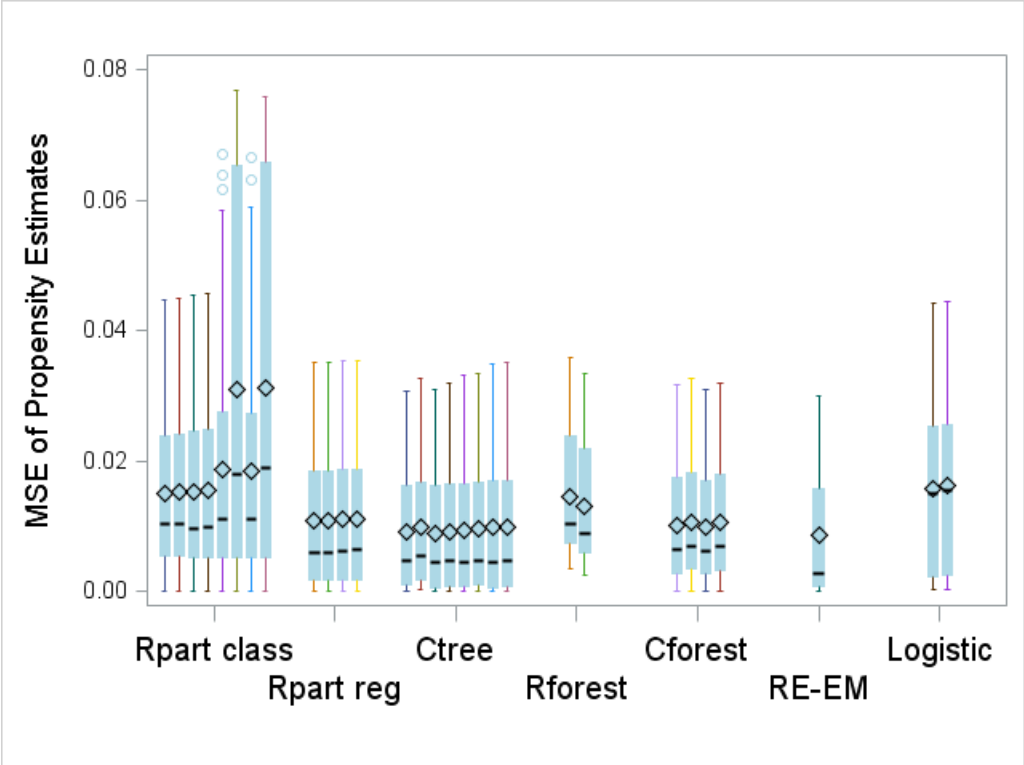

Figure 1: MSE of estimated response propensities. The dark horizontal line in each box represents the median, and the diamond represents the mean. For the *rpart* classification trees, boxes 3, 4, 7, and 8 use weights; boxes 2, 4, 6, and 8 use pruning; and boxes 5, 6, 7, and 8 use unequal misclassification losses. For the *rpart* regression trees, boxes 3 and 4 use weights and boxes 2 and 4 use pruning. For *ctree,* boxes 2, 4, 6, and 8 use weights; boxes 5, 6, 7, and 8 adjust the critical value by a design effect; and boxes 3, 4, 7, and 8 use a Bonferroni adjustment. For *rforest,* box 1 uses 30 trees and box 2 uses 150 trees. For *cforest,* boxes 3 and 4 use 150 trees, and boxes 2 and 4 use weights. For logistic regression, box 2 uses weights.

For the 200 samples in each simulation, we found the mean and standard deviation of the MSE and MAE. Figure 1 shows boxplots of the MSE of the estimated propensities. Each box is constructed from the averages of the MSEs for the 32 sample-generating simulation settings. The patterns from the MAE were similar, and are not shown here. Figure 2 shows the boxplots of accuracy of the propensity estimates separately for the four nonresponse mechanisms used in the simulation.

For *rpart* classification trees, the best results were achieved with no pruning and with equal misclassification costs. However, in general, *rpart* classification trees were the least accurate of the methods.

Figure 1 shows that the *RE-EM* has the lowest mean and median MSE among all of the methods considered. This occurs because it predicts the PSU-to-PSU variability in the simulation runs with PSU_NR_VARIANCE = 0.25; when the nonresponse is missing completely at random as in Figure 2(a), that additional accuracy allows the method to capture the different propensities among PSUs.

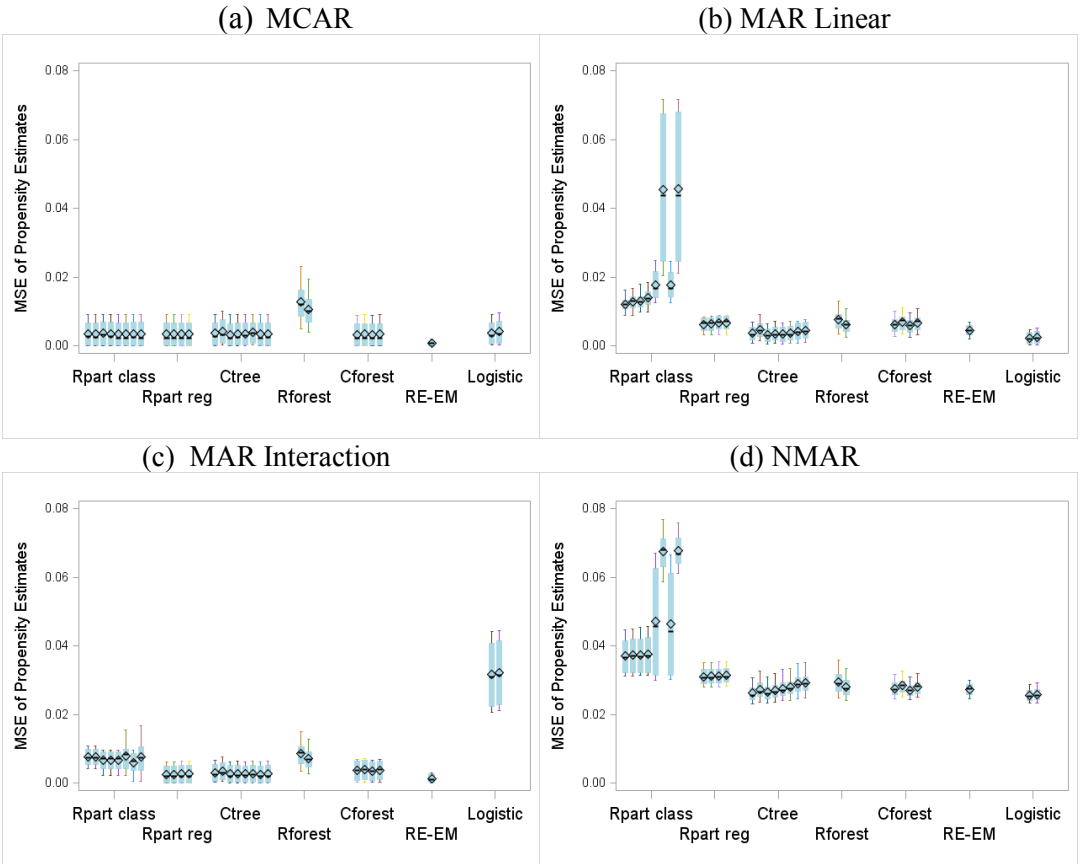

Figure 2: MSE for estimating response propensities, for the four nonresponse mechanisms. The factors associated with the boxes of each tree type are as in Figure 1.

Logistic regression performed very well when the terms that generated the nonresponse were included in the logistic model as in Figure 2(b). As expected, however, it performed poorly when an interaction term used to generate the nonresponse was omitted from the

model as in Figure 2(c). No method performed particularly well for the NMAR mechanism in Figure 2(d).

Using weights in the model-fitting made little difference in the accuracy of the estimated propensities. This was also true for the results in the next section.

4.2 Bias and Mean Squared Error of Response Variable

Ultimately, the test of weighting adjustments is how well they correct for nonresponse bias in variables of interest, while limiting additional variability from unequal weighting. As discussed in the introduction, we studied several types of methods for adjusting weights. For each method, large weighting adjustment factors were trimmed to 10 to avoid excessive weight variation. We compared the methods using the estimated propensities $\hat{\psi}_i$ from each of the 29 models discussed in Section 3. For comparison, we also calculated weight adjustments using the true propensities ψ_i, which are known for this simulation study but of course would be unknown in practice.

Terminal Nodes. The tree models (*rpart* and *ctree*) each produce one tree for each sample. The terminal nodes from this tree form the weighting classes. The terminal node method constructs the weight for respondent i as $d_i \times \min\left[\sum_{j\in S_{k_i}} d_j / \sum_{j\in R_{k_i}} d_j, 10\right]$, where k_i is the terminal node containing respondent i and $S_{k_i}$and R_{k_i} are the sampled units and respondents, respectively, in that terminal node.

Sample Quintiles. Logistic regression, *RE-EM*, and the forest methods do not produce ready-formed weighting classes for nonresponse adjustment, but they produce an estimated response propensity $\hat{\psi}_i$ which can be used to form weighting classes. As suggested by Eltinge and Yansaneh (1997), we formed five weighting classes using the quintiles of the response propensities for the sample. This results in equal numbers of sampled units in each weighting class, and the adjusted weight for respondent i is $d_i \times \min\left[\sum_{j\in S_{q_i}} d_j / \sum_{j\in R_{q_i}} d_j, 10\right]$, where q_i is the quintile containing respondent i and S_{q_i} and $R_{q_i}$are the sampled units and respondents, respectively, in that quintile.

Respondent Quintiles. This is similar to forming weighting classes from the sample quintiles, except that the cutpoints are based on the quintiles of the estimated response propensities for the respondents only. Each weighting class has approximately the same number of respondents. Because the lowest quintile consists of units with low response propensities, the cutpoints for the respondent quintiles method are higher than those for the sample quintiles method.

Divide by Propensity. Each unit is treated as its own weighting class, with weight $d_i \times \min\left(1/\hat{\psi}_i, 10\right)$.

Each of the sets of adjusted weights is used to estimate the mean and the 25th, 50th, and 75th percentiles of HINCP, for each simulated sample. For each statistic t and corresponding population quantity T, we found the mean squared error of t over the 200 samples in each of the 32 simulation settings.

Figure 3 shows the mean of the MSEs for estimating mean household income, across all 32 simulation settings. For simplicity, Figure 3 includes only the "best" parameter settings for each type of model. In general, the settings that gave the most accurate

propensity estimates also gave the smallest MSE for statistics about household income. Therefore, all results displayed in Figure 3 are from models fitted without weights. For *rpart,* no pruning was done and equal misclassification penalties were used. For *ctree,* the critical value was not adjusted for clustering or multiple testing. The forest method results are from the models with 150 trees, although there was very little gain from fitting 150 rather than 30 trees. The first column displays the mean MSE over all 32 simulation settings using the full sample—that is, with no nonresponse—and the respondents with unadjusted weights. The remaining columns show the mean MSE from the divide-by-propensity, sample quintiles, respondent quintiles, and terminal node methods.

Figure 3 shows that for many of the weight adjustment methods, the mean MSE for estimating mean HINCP is higher than when no weight adjustments are made. This occurs largely because of the NMAR nonresponse settings, where the methods do little to reduce the bias but introduce extra weight variation. For the NMAR mechanism, high-income households had a low value for the true response propensity. Consequently, there were few high-income households among the respondents, and the models did not increase the weights sufficiently for those households to counteract the nonresponse bias. For the MAR linear and MAR interaction nonresponse mechanisms, all of the forms of weighting adjustment reduced the MSE for mean HINCP. Note, however, that *ctree* appears to perform well across the different nonresponse mechanisms.

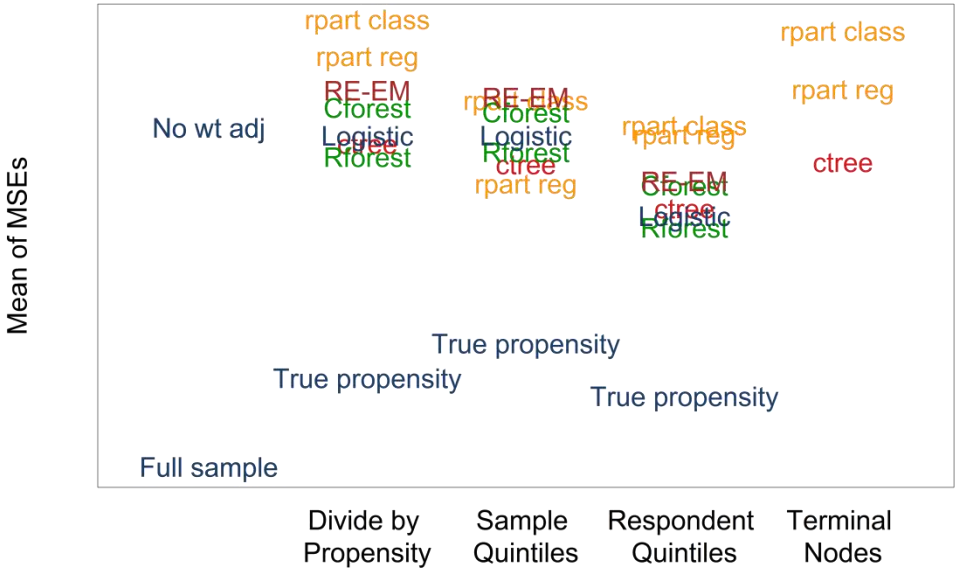

Figure 3: Mean MSE for estimating mean household income.

Figure 4 displays the mean MSE for estimating median household income, which would be expected to be less affected by the nonresponse of high-income households. All of the weight adjustment methods have lower mean MSE than the unadjusted weights. As before, the *rpart* methods tend to have higher MSE than the other methods.

Figure 4 shows an interesting anomaly. One would expect using the true propensities to result in the largest reduction in MSE, but in the fourth column, with weighting classes formed from the quintiles of the respondents, the true propensities resulted in the highest average MSE. This occurred because of the specific MAR linear mechanism used to generate nonresponse as a linear function of children, tenure, and number of income

earners. The MAR linear nonresponse mechanism produced a limited number of discrete values for the true response propensities, and using quintiles of the respondents created an unnatural division in which households with dissimilar propensities were forced into the same class. In fact, none of the weighting adjustments methods based on quintiles of the respondents worked well for the MAR linear mechanism.

It should not be concluded from this anomaly that forming classes using quintiles of the sample works better than forming classes using quintiles of the respondents. The same grouping of dissimilar propensities could occur under a different nonresponse mechanism for the sample quintiles method. The cause of the problem was the algorithm used, which automatically grouped households by quintiles without considering the homogeneity of the resulting classes. This could be remedied by having expert review of the weighting classes. The algorithm could also be improved by using clustering of propensities, rather than strict quintiles, to form weighting classes, and in future research we plan to study clustering methods for forming weighting classes.

The *RE-EM* method gave the most accurate estimates of response propensities, but that superior performance did not carry over to the estimation of mean or median HINCP. The method was most accurate for estimating response propensities because it captured the randomly generated PSU-to-PSU variability. That variability, however, was generated independently of the covariates and of HINCP. Consequently, for these simulations, the improved prediction of ψ_i that was achieved by *RE-EM* did not result in a reduction of $\mathrm{Cov}(\psi, y)$. In situations where the PSU-specific response propensities are related to the y variables, one might expect the *RE-EM* method to result in reduced MSE.

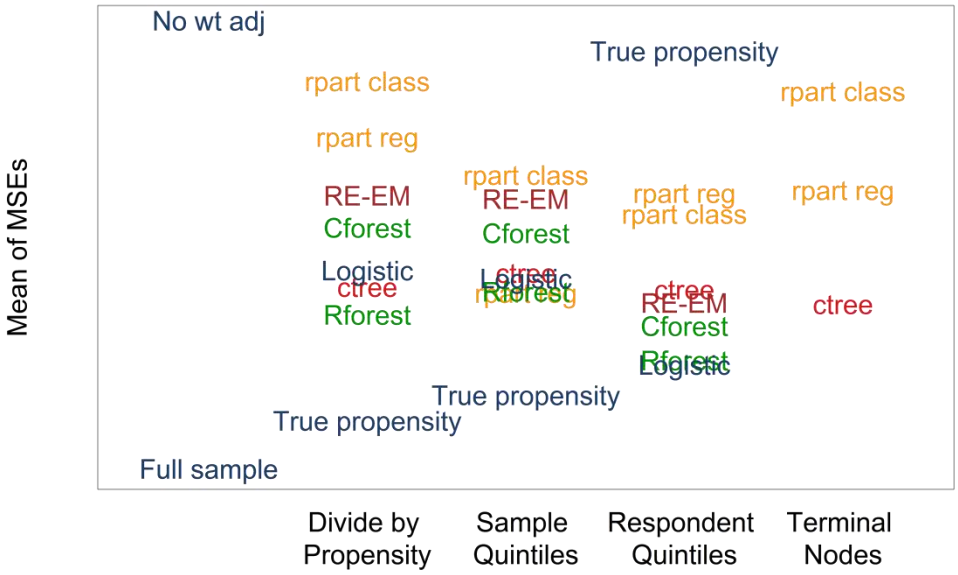

Figure 4: Mean MSE for estimating median household income.

Finally, although not shown here, for the NMAR mechanism, dividing the weight by the true (but unknown) propensity reduced the MSE of median HINCP. Most of the other methods had higher MSE than using no weighting adjustment at all. For this nonresponse mechanism, the covariates did not contain enough information to adjust for the nonresponse. The conditional tree method *ctree* appeared to work better than the other methods, even though it ended up with higher MSE than the unadjusted weights.

5. Conclusions and Future Research Directions

The results in this paper are from one simulation study, and therefore cannot be generalized to all situations. Nevertheless, there are a number of findings that merit further study. First, for MAR nonresponse, most methods reduced the MSE for the estimated mean and quantiles of HINCP. The NMAR mechanism used in this paper was extreme, with some samples having few high-income households that could be used for any type of nonresponse adjustment, and may be atypical of situations likely to be encountered in practice. Our results suggest that any sensible approach to estimating response propensities and adjusting weights is likely to result in improvements for many y variables. However, as the results estimating the mean and median of HINCP with NMAR indicate, it should not be assumed that weight adjustments will reduce bias and MSE for all situations. Typically, the information available about units in S is limited, and the available covariates may be inadequate for modeling the response mechanism. If the classes are poorly formed or if the response propensity estimates are poor, then estimates may still exhibit nonresponse bias; additionally, the MSE may increase because of introduced weight variation.

We found that some settings for tree growing worked better than others, and the same settings were best both for estimating propensities and for estimating the median household income. Overall, as expected, regression trees worked better than classification trees: regression trees focus on predicting the propensity as opposed to simply predicting whether a unit responds.

In these simulations the unadjusted settings—not using weights, pruning, design effects, or Bonferroni adjustments—gave the best performance. We found no benefits of using weights when modeling the response propensities, and recommend that models be fit without weights. However, the effects of pruning and design effect adjustments in the study may be confounded with our use of a minimum terminal node size of 20 for our trees, and collapsing nodes with fewer than 20 respondents. In future research, we plan to vary the minimum number of respondents in a weighting adjustment class. We required a minimum of 20 respondents to reduce the instability from adjusting the weights of few respondents to account for a large number of nonrespondents, but other researchers have explored smaller values for the minimum. Iachan et al. (2014), for example, considered using as few as three respondents per class.

The methods using the original tree algorithm in Breiman et al. (1984)—*rpart* and *rforest*—performed adequately; in fact, *rforest* was among the best methods when we divided the design weight by the propensity estimate. In this study, however, most of the covariates were binary or had only a few categories. In the presence of categorical covariates with many categories, *rpart* performs worse and we do not recommend its use. *Ctree* performed better than *rpart* for almost every simulation setting.

The newer tree models such as *ctree* are worth considering for nonresponse adjustment. *Ctree* performed well for estimating response propensities and better than most other methods for reducing nonresponse bias. More research is needed on how best to use these tree models with data from complex surveys. In particular, research is needed on how to account for the complex survey design in the hypothesis tests used as stopping rules.

Even though *RE-EM* did not work as well for reducing nonresponse bias as some of the other methods, we think that it also merits further study because it had the most accurate

estimates of response propensities when they have PSU-to-PSU variability. In some surveys, that PSU-level variability may be related to the *y* variables. The algorithm performed surprisingly well considering that it was developed to estimate continuous responses as an extension of the linear mixed model. More research is needed on using the method with tree algorithms other than *rpart* and on adapting it for binary responses.

Finally, our investigations indicate that the algorithms help immensely in guiding the formation of weighting classes, but art is needed as well. For many runs, a sampling expert looking at the tree output would be able to suggest improvements that would lead to lower MSE. Expertise is needed to guide the weighting procedures.

Acknowledgments

The authors are grateful to Keith Rust for his helpful comments on this paper, and to Robin Jones for her assistance with SAS® programming. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

References

Bethlehem, J. (1988). "Reduction of Nonresponse Bias Through Regression Estimation." *Journal of Official Statistics,* 4, 251-260.

Bethlehem, J. (2002). "Weighting Nonresponse Adjustments Based on Auxiliary Information." In R. M. Groves, D. Dillman, J. Eltinge, and R. Little (eds.), *Survey Nonresponse.* New York: Wiley, pp. 275-287.

Breiman, L. (2001). "Random Forests." *Machine Learning,* 45, 5-32.

Breiman, L. and Cutler, A. (2004). "Random Forests™." Last accessed May 5, 2015 from http://www.stat.berkeley.edu/~breiman/RandomForests/.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. Belmont, CA: Wadsworth.

Brick, J. M. (2013). "Unit Nonresponse and Weighting Adjustments: A Critical Review." *Journal of Official Statistics,* 29, 329-353.

Brick, J. M. and Montaquila, J. (2009). "Nonresponse and Weighting." In D. Pfeffermann and C.R. Rao (eds.), *Handbook of Statistics, Vol. 29A. Sample Surveys: Design, Methods, and Applications*. Amsterdam: Elsevier, pp. 163-185.

Eltinge, J. L. and Yansaneh, I. S. (1997). "Diagnostics for Formation of Nonresponse Adjustment Cells, With an Application to the U.S. Consumer Expenditure Survey." *Survey Methodology,* 23, 33-40.

Hothorn, T., Hornik, K., and Zeileis, A. (2006). "Unbiased Recursive Partioning: A Conditional Inference Framework." *Journal of Computational and Graphical Statistics,* 15, 651-674.

Hothorn, T., Hornik, K., Strobl, C., and Zeileis, A. (2015). "Party: A Laboratory for Recursive Partytioning." Version 1.0-20. Last accessed May 13, 2015 from http://cran.r-project.org/web/packages/party/.

Iachan, R., Harding, R. L., and Peters, K. (2014). "A Comparison of Weighting Adjustment Methods for Nonresponse." *Proceedings of the Survey Research Methods Section, American Statistical Association,* 3983-3989.

Liaw, A. and Wiener, M. (2002). "Classification and Regression by randomForest." *R News,* 2, 18-22.

Liaw, A. and Wiener, M. (2014). "randomForest: Breiman and Cutler's Random Forests for Classification and Regression." Version 4.6-10. Last accessed May 5, 2015 from http://cran.r-project.org/web/packages/randomForest/index.html.

Loh, W.-Y. (2014). "Fifty Years of Classification and Regression Trees." *International Statistical Review,* 82, 329-348.

Oh, H. L. and Scheuren, F. J. (1983). "Weighting Adjustments for Unit Nonresponse." In W. G. Madow, I. Olkin, and D.B. Rubin (eds.), *Incomplete Data in Sample Surveys, Vol. 2*. New York: Academic Press, pp. 143-184.

Potter, F., Grau, E., Williams, S., Diaz-Tena, N., and Carlson, B. L. (2006). "An Application of Propensity Modeling: Comparing Unweighted and Weighted Logistic Regression Models for Nonresponse Adjustments." *Proceedings of the Survey Research Methods Section, American Statistical Association,* 3555-3560.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. http://www.r-project.org.

SAS Institute, Inc. (2011). *SAS/STAT® 9.3 User's Guide*. Cary, NC: SAS Institute, Inc.

Särndal, C.-E. and Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. Hoboken, NJ: Wiley.

Sela, R. J. and Simonoff, J. S. (2012). "RE-EM Trees: A Data Mining Approach for Longitudinal and Clustered Data." *Machine Learning,* 86, 169-207.

Skinner, C. J. (1989). "Domain Means, Regression, and Multivariate Analysis," in *Analysis of Complex Survey Data*, ed. Skinner, C. J., D. Holt, and T. M. F. Smith, New York: Wiley, pp. 59–88.

Skinner, C. J. and D'Arrigo, J. (2011). "Inverse Probability Weighting for Clustered Nonresponse." *Biometrika,* 98, 953-966.

Therneau, T. M. and Atkinson, E. J. (2015). "An Introduction to Recursive Partitioning using the RPART Routines." Last accessed May 1, 2015 from http://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf

Toth, D. and Phipps, P. (2014). "Regression Tree Models for Analyzing Survey Response." *Proceedings of the Government Statistics Section, American Statistical Association,* 339-351.