A Comparison of Two CHAID Packages for Modeling Survey Nonresponse

Tien-Huan Lin[1], Ismael Flores Cervantes[1], Carlos Arieira[1], Mike Kwanisai[1]
[1]Westat, 1600 Research Blvd., Rockville, MD 20850

Abstract
When computing survey weights for use in analysis of complex sample survey data, it is common practice to mitigate bias due to unit nonresponse by modeling response propensity and adjusting weights to account for different response propensities. The CHAID (Chi-square Automatic Interactive Detector) algorithm is commonly utilized to produce weighting classes for this purpose. We review two popular software packages that implement the CHAID algorithm: SI-CHAID and HPSPLIT. SI-CHAID is an interactive stand-alone graphical user interface that is easy to manipulate and produces informative graphical images of the decision tree but requires manual intervention and additional effort to incorporate into a code-based environment. HPSPLIT is a SAS code-based procedure. However, manipulation of the tree is less intuitive, and the graphical representation of the tree is less informative than SI-CHAID. We empirically evaluate the two packages in terms of the resulting empirical bias and variance of the weighted estimates using simulations. The simulations account for the complex survey sample design to examine the interchangeability of the two software packages so users can determine the software that best meets the analysis needs of their studies.

Key Words: Survey weighting adjustments, weighting class, nonresponse, CHAID, SAS, HPSPLIT

1. Introduction

In survey research, it is a common practice to attempt to alleviate bias due to unit nonresponse by making weighting adjustments to the design weights that account for the sampled units' unequal selection probabilities. When undertaking this task, researchers are faced with an array of methods and options to choose from to best adjust for nonresponse while minimizing their variance. Brick and Montaquila (2009) discuss several weighting methods for nonresponse adjustment. A method that is much utilized in surveys is the weighting class adjustment method (Lessler & Kalsbeek, 1992). The weighting classes are created either by fitting regression models to predict the response propensity and making cutpoints of the estimated propensity or by utilizing terminal nodes of classification or regression trees (Lohr, Hsu, & Montaquila, 2015). The nonresponse adjustment uses factors computed as the inverse of the weighted response rate in each weighting class (Brick & Kalton, 1996). In each weighting class, the nonresponse adjustment factors are applied to the survey respondents, consequently shifting the weights of the survey nonrespondents appropriately to the respondents, with the aim of reducing the bias from nonresponse (Lin et al., 2017).

Over the past few years, researchers have made progress on the nonresponse weighting class method based on terminal nodes of classification trees fitted to the observed response

status (i.e., respondent and nonrespondent). Toth and Phipps (2014) studied the treatment of survey nonresponse by the use of regression trees. In the same year, Loh (2014) studied 20 or more programs for classification trees and regression trees, reviewing the ideas behind these diverse algorithms. Lohr, Hsu, and Montaquila (2015) compared the estimates of nonresponse adjusted weights from various classification trees and random forest algorithms. Cecere et al. (2020) expanded on Lohr et al. (2015) by including additional tree algorithms such as the SAS procedure HPSPLIT to their comparison while simulating a mail survey with a simple random sample design. Jones et al. (2021) extended the 2020 paper to evaluate the effect of the classification-tree-based methods on the reduction of nonresponse bias simulating a complex survey sample design featuring a two-stage cluster sample.

The weighting class adjustment method models response propensity by identifying auxiliary variables correlated with response propensity alone and produces one set of nonresponse adjusted weights applicable for all analyses of the survey data. It does not utilize survey outcomes in the model fitting. Some researchers have pointed out that nonresponse adjustments should take into account both the probability of response and the survey outcomes in order to reduce bias while controlling for variance (Little & Vartivarian, 2005). Vartivarian and Little (2002), Morral, Gore, and Schell (2014), and Fay and Riddles (2017) have applied this approach to include predictions of the actual survey outcomes in adjusting for nonresponse, instead of modeling on auxiliary variables alone. Lin and Flores Cervantes (2019) compared nonresponse adjusted estimates based on this approach to the popularly utilized weighting class approach. They found little benefits in including predicted survey outcomes in nonresponse adjustment.

This study builds on the work of Lin and Flores Cervantes (2019) and Jones et al. (2021) by targeting two Chi-square Automatic Interactive Detector (CHAID) software packages utilized to produce weighting classes for nonresponse adjustment of survey weights. The two software packages studied in the paper are SI-CHAID, created by Statistical Innovations Inc. (Magidson, 2005), and the SAS procedure HPSPLIT. They are compared empirically through a Monte Carlo simulation study in order to evaluate each package's 'effectiveness of nonresponse bias reduction and the balance between bias and variance of the estimates. Two response mechanisms are assessed: a high response rate and a low response rate. Within each response mechanism setting, two nonresponse patterns are established: missing at random and not missing at random. A brief discussion of these concepts is provided below.

When dealing with unit nonresponse, it is necessary to discuss the potential pattern of missingness in the data. Following the terminology proposed by Rubin (1976) and Little and Rubin (2002), there are three assumptions of nonresponse: missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR). The simplest and strongest assumption is MCAR, where it is assumed that nonresponse is unrelated to any variables in the data. It is the most restrictive assumption and is rarely satisfied in practice, and therefore is out of the scope of this simulation study. The most common assumption from which modern survey statistics are built upon is the MAR. MAR assumes that if covariates are observed for all sampled units, and if missing data occurs only in the outcome variable, the probability to respond depends only on the covariates. Under this assumption, covariates accessible in the data alone are sufficient in mitigating nonresponse bias. The importance of the final assumption, NMAR, has seen a steady increase in recent years due to the decline in response rates. Under this assumption, the probability to respond depends on unobserved data after conditioning on observed data. In

other words, key covariates necessary in reducing nonresponse bias cannot be accessed in the survey data. Since the data necessary for adjusting for nonresponse bias is not available, the existence and level of nonresponse bias is a natural concern.

The rest of this paper is organized as follows. In Section 2, we describe the CHAID algorithm and the two implementations (SI-CHAID and HPSPLIT). Section 3 discusses the simulation setup. Section 4 presents the results. We conclude in Section 5 with a discussion of the results and thoughts for future research.

2. The CHAID Algorithm and the Implementations

The Chi-square Automatic Interaction Detector (Kass, 1980), commonly known as CHAID, is a decision tree technique, based on adjusted significance testing (Bonferroni). Rosenbaum and Rubin (1983) first introduced propensity score methods for analyzing how this method can be used to remove or reduce bias due to unit nonresponse. In survey research, CHAID is commonly utilized to produce weighting classes to reduce nonresponse bias or to identify pools of donors for hot-deck imputations techniques. Using statistical algorithms, the decision trees are split into branches, and the end nodes (or terminal cells) of the trees form the weighting classes for nonresponse adjustment or donor cells for hot-deck imputation. This paper focuses on two implementations for nonresponse bias adjustment, with details of each implementation provided below.

2.1 SI-CHAID

SI-CHAID for Windows is a stand-alone program developed by Statistical Innovations Inc., specifically for the CHAID analyses. It has been utilized in multiple national or multi-national large-scale survey research such as the Medical Expenditure Panel Survey and the Population-Based HIV Impact Assessment surveys, and its effectiveness in mitigating nonresponse bias in survey research has been studied and documented in various publications (see for example, Lin et al. 2017). The disadvantage of this implementation is the fact that all parameter settings are done manually, including the assignment of variable type, which can be labor intensive with large numbers of predictors.

2.2 SAS: HPSPLIT

The HPSPLIT procedure in SAS/STAT® software (2018) is a generalized classification and regression tree package. The procedure offers several options for partitioning criteria; three are commonly used. The first criterion maximizes reduction in node impurity as measured by the Gini index. The second uses entropy information for classification. The third type of criterion is based on a CHAID algorithm, which utilizes chi-square tests to partition the data into trees. The processing of this software is automatic, only requiring specifying and running the procedure in SAS. However, research on its effectiveness in survey nonresponse adjustment is limited and existing literature has not been focused on the CHAID algorithm (see, for example, Cecere et al., 2020).

2.3 Features of Each Package

Variable requirement

SI-CHAID allows for two types of variables: ordinal (with or without missing data) or nominal. The implementation of ordinal variables is a special enhancement of this package (Magidson, 2005) that is not available in other CHAID implementations. It is not friendly to continuous variables, as there is a limit to the number of categories a variable can contain (i.e., 31). If the limit is violated, the package will arbitrarily group the categories into 15 or fewer levels.

HPSPLIT allows for categorical and continuous variables. However, it does not recognize the hierarchical order of ordinal categorical variables.

Missing data
In SI-CHAID, missing categories are combined with the most similar node in terms of percentage of respondents.

Four options are available for the handling of missing data in HPSPLIT: 1) to omit, 2) to create an extra branch for missing data, 3) to collapse with the most popular mode, or 4) to collapse with the most similar node. To maximize the comparableness between the two packages, the last option was applied in the simulation study.

Weights
When a variable is provided in the "weight" parameter, SI-CHAID implements the weighted log-linear modeling (WLM) algorithm by default. The weight variable is treated as a sampling weight and can be any positive value. The use of a weight variable provides unequal treatment to the observations in a data set, whereby an observation is weighted according to the number of population units that it represents in the analysis sample. With complex sampling designs, the WLM algorithm should always be employed.

HPSPLIT: The variable in the weight statement is used as a frequency. If this statement is not included, all observations will have a weight of one (see version 15.1 of the User's Guide; SAS Institute, 2018).

Pruning
SI-CHAID uses maximum iterations and epsilon in conjunction with WLM to set the limit of iterations. Users can also control the level of branches by specifying the depth of the tree and minimum cell size. This implementation allows for users to combine categories of a predictor variable in any way that seems appropriate.

Four pruning options are available for classification trees in HPSPLIT. The default pruning method is Cost Complexity (Breiman et al., 1984, Quinlan, 1987, Zhang & Singer, 2010). Another method is C4.5, which is based on the upper confidence limit for the error rate (Quinlan, 1993). A third method is reduced-error pruning (Quinlan, 1987), which is based on minimizing the error rate in the validation partition at each pruning step and then in the overall subtree sequence. A fourth method is the misclassification pruning, available as one of the options in the "by metric" pruning method (SAS/STAT® 12.3 User's Guide, 2013). It chooses the leaf that has the smallest change in the misclassification rate.

Outputs
When processing the analyses in SI-CHAID, results can be displayed simultaneously in the form of an intuitive tree diagram, as cross-tabulations, and as a gains chart summary. Users can request the displays be output with customized information into a pdf. Users can also request SPSS code or C++ code that can be transformed into SAS code to be integrated into other programming environments.

HPSPLIT outputs a data set with leaf assignments and predicted values for observations. Users can request code and plots to be output as well. The default plot is a high-level tree image with limited node information. Additional plots may be requested: option CVCC produces cross validation plots (default with cost-complexity pruning); PRUNEUNTIL

produces a plot of the metric used to select the final subtree; ROC produces the receiver operating characteristic curve; WHOLETREE produces a plot to visualize the entire tree; and ZOOMEDTREE produces a plot to visualize a portion of the tree.

3. Simulation

3.1 Population and Sample Design

A one-time simple random sample of 200,000 households (excluding group homes) of the 2013-2017 American Community Survey (ACS) Public Use Microdata Sample File (PUMS) was treated as the population for the simulation study. The population frame included 43 variables. Of those variables, 39 were household-level characteristics, while the remaining 4 were person-level characteristics derived by summarizing to the household level the corresponding person-level variables. The 43 predictors included 4 continuous variables and 39 categorical variables. The categorical variables were recoded such that the smallest category contained at least 5 percent of the households in the population.

From this fixed population, repeated samples were selected with a two-stage stratified cluster design, with census region defined as stratum. Within each stratum, primary sampling units (PSUs) were formed using public use microdata areas (PUMAs) or combined PUMAs with each PSU containing at least 300 households. Twenty-five PSUs were sampled from each stratum with probability proportional to size sampling, using number of households as the measure of size. Within each sampled PSU, a simple random sample of 100 households was selected, summing up to a total of 10,000 households in one simulation run. This sample selection is repeated 5,000 times for a single scenario.

Two variables were selected as the outcomes in the simulation study as listed in Table 1. The empirical study compared estimates of means and proportions of these outcome variables.

Table 1: Outcome Estimates

Dependent variable	*Description*	*Type*	*Values*
HINS	Indicator flag for all members in the household to have health insurance coverage. The flag was created and summarized from the person-level health coverage indicator from the ACS person-level file).	Binary	1: yes 0: no
HINC	Household income for the past 12 months	Continuous	

3.2 Response Scenarios

Two response-generating mechanisms were studied in the simulation: a high mechanism (R_{high}) averaging a 70 percent response rate, and a low mechanism (R_{low}) averaging a 30 percent response rate; within each response mechanism, the tree models were altered to create a MAR nonresponse pattern and a NMAR nonresponse pattern.

The response mechanisms were generated in three steps. In the first step, a binary ACS response status (i.e., response vs. nonresponse) was assigned based on a household's survey mode in the ACS PUMS frame. For the high response mechanism, web and mail participants were grouped together to create the "response" category and CATI/CAPI participants were assigned to the "nonresponse" category. For the low-response

mechanism, mail participants were assigned to the "response" category, and web and CATI/CAPI participants were grouped together to create the "nonresponse" category. A generalized linear model was then fit separately to the two versions of ACS response status with all 43 frame variables. The generalized linear model allowed us to identify the covariates significant in predicting ACS response status. In the second step, the two versions of ACS response status were each fit to a logistic regression using the top six most important covariates identified in step 1 along with all potential interactions. Covariates or interactions not statistically significant based on the p-value < .05 criteria were removed from the logistic regressions until all covariates and interactions were statistically significant. In the final step, the two logistic regression models developed in step 2 were applied to the entire frame to compute the synthetic high (R_{high}) and low (R_{low}) response propensities for every sampling unit. The synthetic response propensities generated in this process had almost no correlation with design weights; for R_{high} the correlation to design weights is -0.04, and for R_{low} is -0.03. On the other hand, the synthetic response propensities are virtually not correlated with the outcome estimate HINS (health insurance coverage) but somewhat highly correlated to the outcome estimate HINC (household income). For R_{high} the correlation to HINS was -0.18 and to HINC it was 0.40; for R_{low} the correlation was -0.05 for HINS and -0.32 for HINC. The different level of correlation between the synthetic response propensities, design weights, and outcome estimates will have bearing on the interpretation of simulation results.

Table 2 lists the variables from the 2013-2017 ACS PUMS data identified in the three-step procedure used to generate the two response mechanisms.

Table 2: Variables Used from ACS PUMS for Response Models and Nonresponse Patterns

Variable name	*Variable description*	R_{high}	R_{low}
HHHISP	At least one person in HH is Hispanic	✓	
HHRACE	At least one HH member is not white alone	✓	
INSP	Fire, hazard, flood insurance (yearly amount)	✓	
BROADBND	Broadband (high-speed) Internet service such as cable, fiber optic, or DSL service	✓	
LAPTOP	Laptop or desktop	✓	
WATP	Hot and cold running water	✓	✓
HHBACH	At least one person in HH graduated from college		✓
HISPEED	Broadband (high-speed) Internet service such as cable, fiber optic, or DSL service		✓
FULP	Fuel cost (yearly cost for fuels other than gas and electricity)		✓
SMARTPHONE	Smartphone		✓
R60	Presence of persons 60 years and over in household (unweighted)		✓

To generate the *empirical* response status, in each repeated sample the empirical respondents were drawn using the Poisson sample design, where the selection probability was proportional to the synthetic response propensity (R_{high} or R_{low}).

Within each response mechanism, the tree models used to predict the empirical response propensities were altered to create MAR nonresponse patterns (r_{mar}) and NMAR nonresponse patterns (r_{nmar}). The MAR nonresponse patterns were simply created by

supplying all 43 frame variables, including the six covariates used to develop the underlying response mechanism to the tree models. In contrast, the NMAR nonresponse patterns were created by dropping key covariates in the tree models. For $r_{nmar/high}$, BROADBND and LAPTOP, along with variables highly correlated to the two covariates, were dropped from the pool of potential predictors available to the tree models. For $r_{nmar/low}$, FULP, SMARTPHONE, and R60 and any variables highly correlated to the three covariates were dropped from the pool of potential predictors available to the tree models. The covariates to be withheld from tree modeling were determined based on their correlation to *R*. For the high response mechanism, BROADBND and LAPTOP had the highest overall absolute correlation to the inverse of R_{high}, with the values being 0.68 and 0.65, respectively. For the low response mechanism, the correlation to the inverse of R_{low} were generally low. The highest absolute values were 0.54 and 0.52. However, these variables did not appear on the list of top 25 significant covariates described in the first step of the response mechanism creation and were not used to develop the underlying response mechanism. Within the set of covariates used to develop the response mechanism, FULP showed the highest positive correlation to the inverse of R_{low} at 0.24, and R60 and SMARTPHONE showed the highest negative correlation at -0.30 and -0.29, respectively. These three covariates were further corroborated as they presented the largest Wald Chi-Square values in the logistic regression described in the second step of the response mechanism creation. The low correlation observed for R_{low} in general could potentially lead to unforeseen effects on the simulation.

The four response combinations will be denoted as follows for the remainder of this paper: MAR under high response mechanism ($r_{mar/high}$), NMAR under high response mechanism ($r_{nmar/high}$), MAR under low response mechanism ($r_{mar/low}$), and NMAR under low response mechanism ($r_{nmar/low}$).

The final component to the response scenarios is the correlation between the covariates used to derive the synthetic response mechanisms and the outcomes estimates, as the correlation will also have bearing on the interpretation of simulation results. Table 3 provides this information. For HINS (i.e., household insurance), low correlations are observed for all covariates, regardless of response mechanism. For HINC (i.e., household income), under R_{high}, two outliers are observed in the two covariates that are dropped to create the $R_{nmar/high}$ nonresponse pattern: BROADBND (i.e, -0.36) and LAPTOP (i.e., -0.38). The remaining covariates are virtually not correlated or mildly correlated to HINC. Under R_{low}, the correlation of HINC to the covariates is spread out more evenly compared to R_{high}: relatively high levels of absolute correlation can be observed in HHBACH (i.e, -0.29), HISPEED (i.e., -0.24), and SMARTPHONE (i.e, -0.24). Thus, although SMARTPHONE will be dropped to model $R_{nmar/low}$, the impact on simulation results will be contained.

Table 3: Correlation of Covariates to Outcome Estimates

Variable name	*HINS*		*HINC*	
	R_{high}	R_{low}	R_{high}	R_{low}
HHHISP	-0.18		0.04	
HHRACE	-0.11		0.07	
INSP	0.09		-0.11	
BROADBND	0.05		-0.36	
LAPTOP	0.07		-0.38	
WATP	-0.03	-0.01	0.21	0.17
HHBACH		0.10		-0.29
HISPEED		0.06		-0.24
FULP		-0.03		0.03
SMARTPHONE		-0.01		-0.24
R60		-0.13		-0.03

3.3 Tree Models

Tree models were fit separately to the four sets of 5,000 repeated samples, one set for each response mechanism/nonresponse pattern combination, with either the entire collection of 43 frame variables or the restricted set of variables to predict empirical response propensities using the two CHAID software packages discussed in Section 2. Each software package contains unique sets of parameters to control for tree fitting. Special effort was made to apply global settings among the two packages to minimize subjective differences in result evaluation.

SI-CHAID

The following parameters were set to equal for all trees:

- *Mingrp*: the minimum number of observations in a terminal node was set to 50.
- *Depth*: the maximum level a tree could be grown was set to 5.
- *Iteripf:* 100000. (maximum iterations for the WLM method)
- *Epsipf: 0.000001*. (epsilon limit for the WLM method)
- *Prune:* no pruning was implemented beyond the default settings.

The following factors were varied:

- *Weight:* weight = 1 for all observations or weight = design weight.

All other parameters were set to their default values.

SAS HPSPLIT

The following parameters were set equal for all trees:

- *Minleafsize*: the minimum number of observations in a terminal node was set to 50.
- *Maxdepth*: the maximum level a tree could be grown was set to 5.

The following factors were varied:

- *Weight:* weight = 1 for all observations or weight = design weight.
- *Prune*: misclassification (n <= 100) or reduced error (metric = MISC).

All other parameters were set to their default values.

The fitted empirical response propensity models were then used to compute weighting classes and nonresponse adjustment factors to adjust for design weights.

4. Simulation Results

In this section, we assess the simulation results with three measures. The first two measures include a comparison of the number of weighting classes produced by each tree model and an examination of the set of variables selected by each tree model for the creation of the weighting classes. These two measures are an indication of the homogeneity/heterogeneity of the trees produced by each tree model. The third measure assesses the empirical nonresponse bias and variance by computing the final weighted estimates of mean or proportions adjusted for unbalanced sample selection and nonresponse for the two outcome estimates discussed above and comparing against the true values from the population.

4.1 Number of Weighting Classes

Figure 1 shows the number of weighting classes produced by each tree model among 5,000 repeated samples, separately for the four simulation scenarios. Statistics provided in the table include the lowest, highest, and median number of weighting classes created by each tree model.

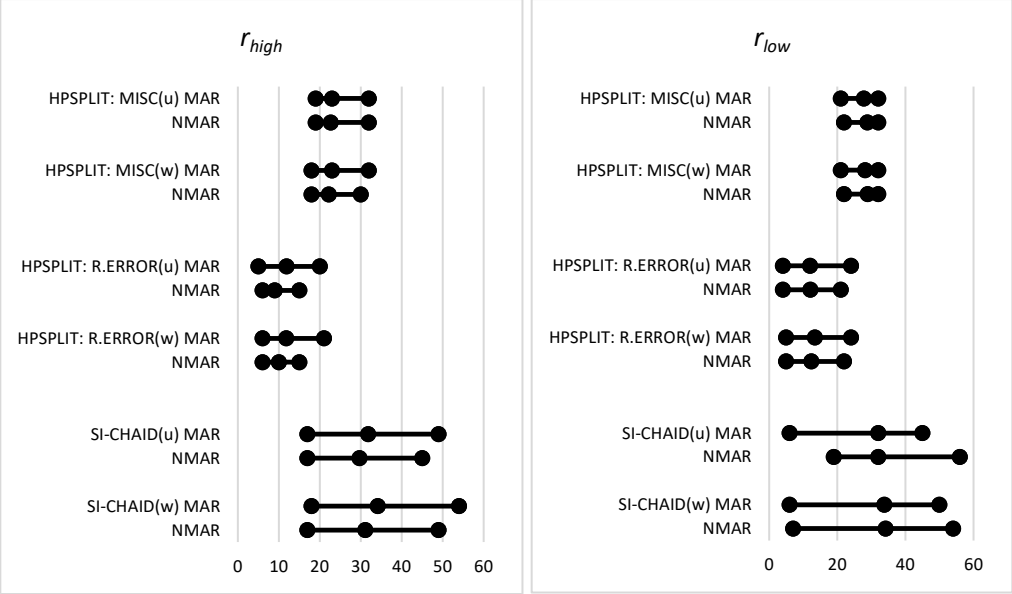

Figure 1: Number of Weighting Classes Produced by Each Tree Model

Overall, the SI-CHAID model produced more weighting classes than the HPSPLIT models regardless of simulation scenario, with the median number of weighting classes around 30-35. Between the two HPSPLIT models, the composition of weighting classes by the misclassification pruning method was closer to the SI-CHAID model, with the median number of weighting classes around 20-30, while the composition of weighting classes by the reduced error pruning method appeared drastically different from the misclassification method as well as SI-CHAID, with the median number of weighting classes around 10.

Among response mechanisms, all three tree models seemed to produce slightly more classes with the low response mechanism. Within response mechanism, nonresponse patterns appeared to have minimal impact on the number of weighting classes produced by each tree model. Not much difference was observed between the unweighted vs. weighted trees for all three models.

4.2 Variables Identified to Be Predictive of Response Propensity

Figure 2 and Figure 3 summarize the top 20 variables identified to be predictive of response propensity by the three tree models over 5,000 repeated samples. Figure 2 provides the summary of the percent of times each of the 20 variables used from ACS PUMS to create Rhigh (i.e., Table 2) were selected in 5,000 repeated samples, separately for rmar/high and rnmar/high, and Figure 3 provides the equivalent for $r_{mar/low}$ and $r_{nmar/low}$.

SI-CHAID is generally consistent and effective in identifying the correct set of covariates predictive of response propensity. In the $r_{mar/high}$ setting, both the unweighted and weighted trees selected all six covariates used to create the underlying response mechanism, and only the six covariates, almost 100 percent of the time. In the $r_{nmar/high}$ setting, both sets of SI-CHAID trees accurately identified the four covariates available for modeling 100 percent of the time (BROADBND and LAPTOP were not supplied to the tree modeling). The same observations can be made for the $r_{mar/low}$ and $r_{nmar/low}$ settings. All six covariate were chosen 100 percent of the time except for FULP under the $r_{mar/low}$ setting, and under the $r_{nmar/low}$ setting the three covariates available for tree modeling were chosen almost 100 percent of the time. The consistency in findings between r_{high} and r_{low} suggests that SI-CHAID performs consistently under different response rates.

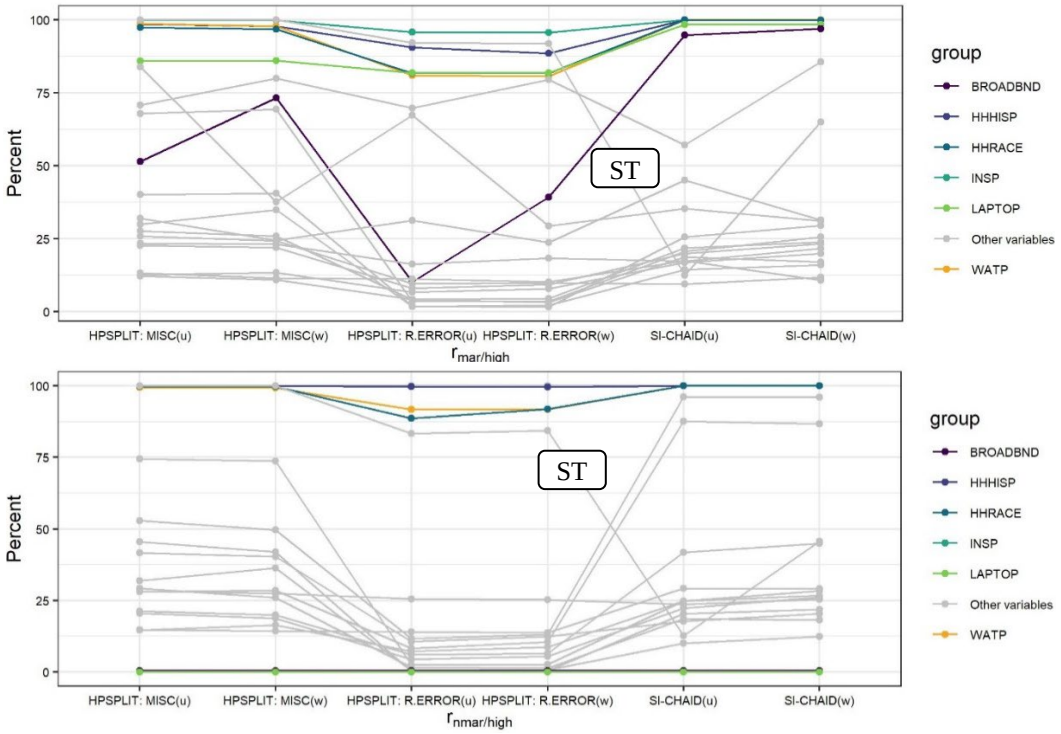

Figure 2: Variable Selection for High Response Mechanism

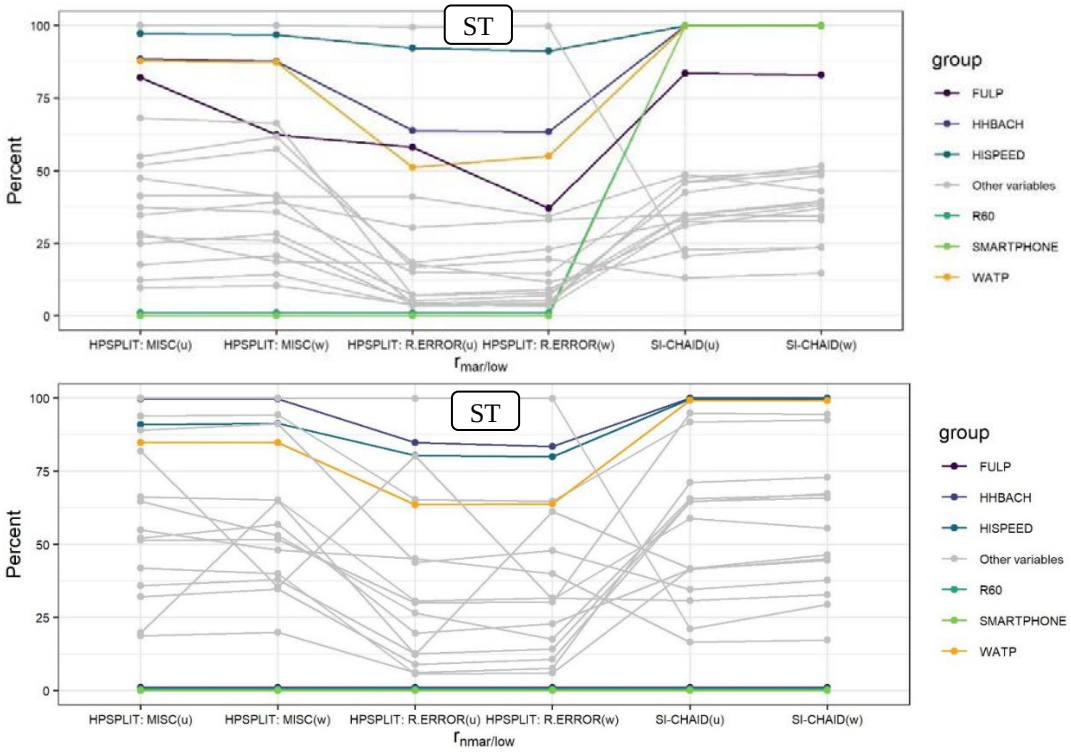

Figure 3: Variable Selection for Low Response Mechanism

HPSPLIT:misclassification appeared to be slightly less effective and contained more inconsistencies compared to SI-CHAID. Under the $r_{mar/high}$ setting, for both the unweighted and weighted trees only four covariates were selected over 90 percent of the time among 5,000 repeated samples. The covariates LAPTOP and BROADBND were selected at relatively low rates. Interestingly, ST (state abbreviation) was selected almost 100 percent of the time for both the unweighted and weighted trees. This variable contains the most number of categories among class variables and is virtually not correlated to response propensity (i.e., -0.04 for the high response rate setting). The $r_{nmar/high}$ setting seemed to inadvertently reduce noise for the tree model, with all covariates being selected at 100 percent, along with the variable ST. Results under the r_{low} settings were even less effective than the r_{high} settings: all covariates were selected at lower rates for both $r_{mar/low}$ and $r_{nmar/low}$, and two covariates (i.e., SMARTPHONE and R60) were never selected under $r_{mar/low}$. Once again, ST was selected at an extremely high rate even though it is virtually not correlated to response propensity (i.e., -0.02), for both $r_{mar/low}$ and $r_{nmar/low}$.

HPSPLIT:reduced error was the least effective among the three tree models in identifying the covariates used to generate the response mechanisms. Similar to HPSPLIT:misclassification, SMARTPHONE and R60 were ignored under $r_{mar/low}$, noise was inadvertently reduced with r_{nmar}, and ST was selected at high rates in most settings.

4.3 Empirical Bias and Variance

In this section, we compare the estimates from each tree model in terms of bias and mean square error.

Ultimately, any nonresponse adjustment is a balancing act between bias and variance. The evaluation tools used in this section are relative bias and relative root mean squared error, with relative bias defined as

$$Relative\ Bias: RB\left(\hat{Y}_E\right)\% = 100 \times B^{-1} \sum_{b=1}^{B} \frac{\hat{Y}_{E,b} - Y}{Y},$$

and relative root mean squared error (*RRMSE*) defined as

$$Relative\ Root\ Mean\ Squared\ Error: RRMSE = \sqrt{\frac{MSE(\hat{Y}_E)}{Y^2}},$$

where $MSE\left(\hat{Y}_E\right) = \frac{\sum_{b=1}^{B}(\hat{Y}_{E,b} - Y)^2}{B}$.

Table 4 and 5 show the $RB(\hat{Y}_E)\%$ and *RRMSE* for the nonresponse adjusted estimates of HINS and HINC, respectively, for each of the tree models. The two tables also include the Horvitz-Thompson (HT) estimator, which is computed as the baseweighted estimate of all sampled units and the baseweighted estimate of respondents (RESP). The former should be unbiased according to the theory, which can be confirmed for both outcome estimates as $RB(\hat{Y}_E)\% \leq .01$ in all settings. The latter provides an indication of the level of empirical nonresponse bias: under r_{high}, $RB(\hat{Y}_E)\% = 9.71$ for HINS and $RB(\hat{Y}_E)\% = 14.73$ for HINC, and under r_{low}, $RB(\hat{Y}_E)\% = 5.84$ for HINS and $RB(\hat{Y}_E)\% = 8.85$ for HINC.

Table 4: Household Insurance (HINS)

	$r_{mar/high}$		$r_{nmar/high}$		$r_{mar/low}$		$r_{nmar/low}$	
Estimates	$RB(\hat{Y}_E)$ *(%)*	*RRMSE (%)*	$RB(\hat{Y}_E)$ *(%)*	*RRMSE (%)*	$RB(\hat{Y}_E)$ *(%)*	*RRMSE (%)*	$RB(\hat{Y}_E)$ *(%)*	*RRMSE (%)*
HT	0.01	<0.01	0.01	<0.01	0.02	<0.01	0.02	<0.01
RESP	9.71	0.01	9.71	0.01	5.84	0.01	5.84	0.01
SI-CHAID (u)	0.38	0.93	0.42	0.93	-0.04	1.12	0.86	1.40
SI-CHAID (w)	0.38	0.93	0.42	0.93	-0.04	1.12	0.86	1.40
HPSPLIT: MISC (u)	0.56	1.07	0.56	1.07	0.09	1.24	0.85	1.46
HPSPLIT: MISC (w)	0.60	1.09	0.54	1.05	0.15	1.25	0.90	1.50
HPSPLIT: R.ERROR (u)	0.80	1.21	0.61	1.03	0.26	1.51	2.12	2.51
HPSPLIT: R.ERROR (w)	0.77	1.18	0.61	1.03	0.41	1.59	2.11	2.50

Table 5: Household Income (HINC)

Estimates	$r_{mar/high}$		$r_{nmar/high}$		$r_{mar/low}$		$r_{nmar/low}$	
	RB($\hat{Y}_E$) (%)	*RRMSE (%)*	*RB($\hat{Y}_E$) (%)*	*RRMSE (%)*	*RB($\hat{Y}_E$) (%)*	*RRMSE (%)*	*RB($\hat{Y}_E$) (%)*	*RRMSE (%)*
HT	0.01	2.51	0.01	2.51	0.01	2.51	0.01	2.51
RESP	14.73	14.84	14.73	14.84	-8.85	9.29	-8.85	9.29
SI-CHAID (u)	0.54	2.66	2.34	3.48	-0.60	3.34	-1.11	3.45
SI-CHAID (w)	0.54	2.66	2.34	3.49	-0.60	3.34	-1.11	3.45
HPSPLIT: MISC (u)	0.95	2.81	2.91	3.93	-1.52	3.71	-1.85	3.82
HPSPLIT: MISC (w)	0.98	2.80	2.87	3.89	-1.40	3.68	-1.78	3.79
HPSPLIT: R.ERROR (u)	1.38	3.02	3.23	4.14	-2.79	4.37	-1.83	3.89
HPSPLIT: R.ERROR (w)	1.34	3.00	3.22	4.13	-2.61	4.23	-1.81	3.89

Overall, all three tree models appeared to be successful in reducing empirical nonresponse bias and minimal empirical differences were observed between unweighted tree models vs. weighted tree models. More detailed evaluation of empirical nonresponse bias examines if the absolute relative bias for each estimate follows two patterns, as stated by the literature: 1) the absolute relative bias under r_{mar} should be close to those of HT (in other words, unbiased); and 2) the absolute relative bias under r_{nmar} should be somewhat higher than those of r_{mar}, but should show improvement from the values by RESP.

1. *The absolute relative bias under r_{mar} should be unbiased*. This can be observed for SI-CHAID, with .04 ≤ absolute $RB(\hat{Y}_E)\%$ ≤ .60. It can be observed in some, but not all, settings of HPSPLIT: misclassification, with.09 ≤ absolute $RB(\hat{Y}_E)\%$ ≤ 1.52; the higher bias tends to occur with the HINC estimates. More departure from this pattern can be observed for HPSPLIT: reduced error. The absolute $RB(\hat{Y}_E)\%$ for HINS estimates are within the range of .26 to .80, which are higher than those observed for SI-CHAID and HPSPLIT: classification but still within reasonable range. However, the absolute $RB(\hat{Y}_E)\%$ for HINC estimates are all above 1.0, and can be as high as 2.79.
2. *The absolute relative bias under r_{nmar} should be somewhat higher than those of r_{mar}, but should show improvement from the values by RESP*. The former part of this statement, "*the absolute relative bias under r_{nmar} should be somewhat higher than those of r_{mar}*," can be largely observed for SI-CHAID and HPSPLIT: misclassification. For SI-CHAID, the absolute $RB(\hat{Y}_E)\%$ of r_{mar} vs. r_{nmar} for HINS were .38 vs. .42 for the high response *mechanism* and .04 vs. .86 for the low response mechanism. For HINC, the rates were .54 vs. 2.34 for the high response mechanism and .60 vs. 1.11 for the low response mechanism. The equivalent numbers for HPSPLIT: misclassifications were HINS|r_{high}: ~.60 vs. ~.60, HINS|r_{low}: ~.10 vs. ~.90, HINC|r_{high}: ~1.0 vs. ~2.9, and HINC|r_{low}: ~1.50 vs. ~1.8. Inconsistent patterns were observed for HSPLIT: reduced error. The trend of HINS|r_{low} and HINC|r_{high} was in agreement with the expectation, with the rates for r_{mar} vs. r_{nmar} being ~0.3 vs. ~2.1 for

HINS|r_{low} and ~1.4 vs. ~3.2 for HINC|r_{high}. On the contrary, the trend of HINS|r_{high} and HINC|r_{low} was in violation of the expectation. The rates for r_{mar} vs. r_{nmar} were ~0.8 vs. ~0.6 for HINS|r_{high} and ~2.7 vs. ~1.8 for HINC|r_{low}. In terms of the latter part of the statement, "*should show improvement from the values by RESP,"* all three tree models were successful. For HINS|r_{high}, over 8 percent of bias was removed (i.e., 9.71% vs. <1.0%) for all tree models regardless of nonresponse pattern. For HINS|r_{low}, over 4 percent of bias was removed (i.e., 5.84 vs. <1.0) regardless of nonresponse pattern except for HPSPLIT: reduced error, $r_{nmar/low}$ where the bias reduction was approximately 3.7 percent (i.e., 5.84% vs. 2.1%). For HINC|r_{high}, the bias reduction ranged from 13.4 to 14.2 percent depending on the tree model under the missing at random nonresponse pattern and from 11.5 to 12.4 percent under the not missing at random nonresponse pattern. For HINC|r_{low}, the bias reduction ranged from 6.0 to 8.2 percent depending on the tree model under the missing at random nonresponse pattern and from 7.0 to 7.7 percent under the not missing at random nonresponse pattern.

Another evaluation that could be of potential interest is the impact of response rate on the three tree models. All three tree models presented similar reactions to the impact of response rate; however, inconsistent patterns were observed among the nonresponse patterns and outcome estimates.

For HINS, all tree models presented the pattern of $r_{mar/high} > r_{mar/low}$ and $r_{nmar/high} < r_{nmar/low}$ in absolute $RB(\hat{Y}_E)\%$. Contrarily, for HINC, for all tree models the pattern was $r_{mar/high} < r_{mar/low}$ and $r_{nmar/high} > r_{nmar/low}$. The interpretation of results can be convoluted since the response mechanisms are confounded by the correlation between outcome estimates and key covariates. In Section 3 we provided the correlation between the response mechanisms and outcome estimates: the correlation of HINS|R_{high} is -0.18 and HINC|R_{high} is 0.40; the correlation of HINS|R_{low} is -0.05 and HINC|R_{low} is -0.32. Table 3 also provided the correlation between the outcome estimates to key covariates. Since HINS has an extremely low correlation to R_{low}, and since one of the key variables for r_{low} (i.e., R60) can potentially be used for indirect modeling of HINS, it is reasonable that the $RB(\hat{Y}_E)\%$ of $r_{mar/low}$ suggests virtually no bias. In the $r_{nmar/low}$ setting, however, since R60 is dropped from modeling, it is reasonable that the values of $RB(\hat{Y}_E)\%$ are higher than the $r_{mar/low}$ counterparts.

Correlation can also help explain the slightly higher bias observed in $r_{mar/high}$. Since HINS has a higher correlation to R_{high}, it is not unreasonable to assume the key covariates available for modeling cannot fully predict for HINC. On the other hand, since the two key covariates dropped for $r_{nmar/high}$ modeling (i.e., BROADBND and LAPTOP) are virtually not correlated to HINS, it is within expectation that not much change in bias is observed. The same analysis can be performed on HINC. HINC has a much higher correlation to both R_{high} and R_{low}; thus, it is expected that the relationship to key covariates would have a stronger impact on bias outcomes. Under r_{high}, the two key covariates highly correlated to HINC were not available for modeling of $r_{nmar/high}$, and therefore the relatively highest increase in bias; under r_{low}, the correlation to HINC was spread among three key covariates: HHBACH, HISPEED, and SMARTPHONE, and therefore although SMARTPHONE was removed for $r_{nmar/low}$ modeling, the impact to bias outcome was limited.

With regard to empirical variance, the *RRMSE* for HINS were generally lower than HINC given that HINS is a binary estimate whereas HINC is a continuous estimate. For both HINS and HINC, and under all response mechanism and nonresponse pattern settings, SI-CHAID appeared to produce the lowest *RRMSE*, followed by HPSPLIT: misclassification,

followed by HPSPLIT: reduced error. The only exception is with HINS, $r_{nmar/high}$, where HPSPLIT: misclassification produced *RRMSE* values that were marginally higher than HPSLIT: reduced error.

In conclusion, SI-CHAID and HPSPLIT: misclassification appeared to be closely in agreement with the literature and showed great effectiveness in mitigating empirical nonresponse bias while limiting empirical variance. HPSPLIT: reduced error was successful in reducing empirical nonresponse bias, though to a slightly lesser degree compared to SI-CHAID and HPSPLIT: misclassification. HPSPLIT: reduced error was also slightly less effective in limiting empirical variance and occasionally departed from the literature in showing moderate biased estimates with r_{mar} as well as inconsistent patterns of bias reduction between r_{mar} and r_{nmar}.

5. Discussion

Using the 2013-2017 ACS PUMS data as a pseudo-population, we investigated the use of two implementations of the CHAID algorithm: SI-CHAID and the SAS procedure HPSPLIT. For HPSPLIT, we included two pruning methods: misclassification and reduced error. Our simulation selected repeated samples drawn from a fixed population with a two-stage stratified cluster design with census region serving as the sampling strata; PSUs (i.e., PUMAs or combination of PUMAs) were selected at the first stage, and addresses (i.e., households) were selected at the second stage. We synthetically generated two response mechanisms, a high response rate and a low response rate, and with each response mechanism we generated a "missing at random" nonresponse pattern and a "not missing at random" nonresponse pattern. Using the ACS PUMS as our fixed population and generating synthetic response mechanisms allowed us to compare between estimates and true population values, and to isolate potential causes for discrepancy.

Our results showed minimal differences for bias and RMSE for the two outcome variables chosen for the simulation study; SI-CHAID may be the most effective in reducing nonresponse bias and restricting the amount of variance associated with bias mitigation, followed by HPSPLIT: misclassification, but the differences were not statistically significant. However, the composition of trees generated by each tree model was quite different. SI-CHAID had an extremely high accuracy rate in identifying the covariates used to create the underlying response mechanisms and also produced the most number of weighting classes. HPSPLIT: misclassification produced comparable results to SI-CHAID, with a slightly lower accuracy rate in variable identification. HPSPLIT: reduced error produced the least desirable results of the three options.

Not much difference was observed between the unweighted and weighted tree models. Although Lohr et al. (2015) suggest that design weights do not provide a benefit when modeling response propensity, we suspect that the lack of improvement from weighted tree models is due to the ignorable nature of our design weights. A further step would be to test these tree models with a sample design that allows for non-ignorable design weights.

Another limitation to our simulation study is the interpretation of comparison between response mechanisms being confounded by the correlation of response propensity to outcome estimates and correlation of key covariates. A different response mechanism designed would be needed to allow for direct response rate comparison.

Acknowledgments

The authors are grateful to David Cantor, Jennifer Kali, and Minsun Riddles for the insightful suggestions.

References

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

Brick, J. M., and Kalton, G. (1996), "Handling Missing Data in Survey Research," *Statistical Methods in Medical Research*, 5, 215-238.

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.

Cecere, W., Lin, T., Kali, J., and Flores Cervantes, I. (2020), "A Comparison of Classification and Regression Tree Methodologies when Modeling Survey Nonresponse," *JSM Proceedings*, Virtual Conference: American Statistical Association.

Fay, R. E., and Riddles, M. K. (2017), "One- Versus Two-Step Approaches to Survey Nonresponse Adjustments." *JSM Proceedings* (pp. 953-964). Baltimore, MD: American Statistical Association.

Jones, M., Cecere, W., Lin, T.-H, Kali, J., and Flores Cervantes, I. (2021), "Modeling Survey Nonresponse under a Cluster Sample Design: Classification and Regression Tree Methodologies Compared." *JSM Proceedings*. Virtual Conference: American Statistical Association.

Kass, G.V. (1980), "An Exploratory Technique for Investigating Large Quantities of Categorical Data," *Applied Statistics*, 29, 119-127. https://doi.org/10.2307/2986296

Lessler, J. T., and Kalsbeek, W.D. (1992), "Nonsampling Errors in Surveys" (1st ed.). New York: John Wiley and Sons.

Lin, T.-H., Weil, N., Flores Cervantes, I., and Saito, S. (2017), "Developing Nonresponse Weighting Adjustments for Population-Based HIV Impact Assessments Surveys in Three African Countries," *JSM Proceedings* (pp. 965-982). Baltimore, MD: American Statistical Association.

Lin, T.-H., and Flores Cervantes, I. (2019), "A modeling approach to compensate for nonresponse and selection bias in surveys?" *JSM Proceedings* (pp. 827-834). Denver, CO: American Statistical Association.

Little, R., and Rubin, D. (2002), *Statistical Analysis with Missing Data*. New York: Wiley.

Little, R.J.A., and Vartivarian, S. (2005), "Does Weighting for Nonresponse Increase the Variance of Survey Means?" *Survey Methodology*, 31, 161-168.

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

Lohr, S., Hsu, V., and Montaquila, J. (2015), "Using Classification and Regression Trees to Model Survey Nonresponse," *Proceedings of the Survey Research Methods Section*, American Statistical Association, 2071-2085.

Magidson, J. (2005), "SI-CHAID 4.0 user's guide," Belmont, MA. Retrieved September 26, 2017, from https://www.statisticalinnovations.com/wp-content/uploads/SICHAIDusersguide.pdf.

Morral, A. R., Gore, K. L., and Schell, T.E. (2014), "Sexual assault and sexual harassment in the U.S. Military: Volume 1. Design of the 2014 RAND Military Workplace Study." Santa Monica, CA: RAND Corporation. Retrieved from www.rand.org/t/RR870zl.

Quinlan, J. R. (1987), "Simplifying Decision Trees," *International Journal of Man-Machine Studies*, 27, 221–234.

Quinlan, J. R. (1993), "C4.5: Programs for Machine Learning," San Francisco: Morgan Kaufmann.

Rosenbaum, P. R., and Rubin, D. B. (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41-55.

Rubin, D. (1976), Inference and missing data. *Biometrika, 63,* 581-590.

SAS Institute, Inc. (2013). SAS/STAT® 12.3 User's Guide. Cary, NC: SAS Institute, Inc.

SAS Institute, Inc. (2018). SAS/STAT® 15.1 User's Guide. Cary, NC: SAS Institute, Inc.

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

Vartivarian, S., and Little, R. J. (2002), "On the Formation of Weighting Adjustment Cells for Unit Nonresponse," The University of Michigan Department of Biostatistics Working Paper Series, Working Paper 10.

Zhang, H., and Singer, B. H. (2010), "Recursive Partitioning and Applications" (2nd ed.), New York: Springer.