Statistica Sinica Preprint No: SS-2017-0225	
Title	Classification and regression trees and forests for incomplete data from sample surveys
Manuscript ID	SS-2017-0225
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202017.0225
Complete List of Authors	Wei-Yin Loh John Eltinge Moon Jung Cho and Yuanzhi Li
Corresponding Author	Wei-Yin Loh
E-mail	loh@stat.wisc.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0225

CLASSIFICATION AND REGRESSION TREES AND FORESTS FOR INCOMPLETE DATA FROM SAMPLE SURVEYS

Wei-Yin Loh, John Eltinge, Moon Jung Cho and Yuanzhi Li

University of Wisconsin-Madison, U.S. Census Bureau and Bureau of Labor Statistics

Abstract:

Analysis of sample survey data often requires adjustments for missing values in the variables of interest. Standard adjustments based on item imputation or on propensity weighting factors rely on the availability of auxiliary variables for both responding and non-responding units. Their application can be challenging when the auxiliary variables are numerous and are themselves subject to incomplete-data problems. This paper shows how classification and regression trees and forests can overcome these difficulties and compares them with likelihood methods in terms of bias and mean squared error. The development centers on a component of income data from the U.S. Consumer Expenditure Survey, which has a relatively high rate of item missingness. Classification trees and forests are used to model the unit-level propensity for item missingness in the income component. Regression trees and forests are used to model the conditional mean of the income component. The methods are then used to estimate the mean of the income component, adjusted for item nonresponse. Thirteen methods for estimating a population mean are compared in simulation experiments. The results show that if the number of auxiliary variables with missing values is not small, or if they have substantial missingness rates, likelihood methods can be impracticable or inapplicable. Tree and forest methods are always applicable, are relatively fast, and have higher efficiency than likelihood methods under real-data situations with incomplete-data patterns similar to that in the abovementioned survey. Their efficiency loss under parametric conditions most favorable to likelihood methods is observed to be between 10–25%.

Key words and phrases: Imputation, item nonresponse, response propensity.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0225

1 Introduction

We consider estimation of a population mean μ of an outcome variable Y based on an incompletely observed simple random sample S without replacement from a finite population. After a review of some existing solutions, we introduce nine new ones based on classification and regression trees and forests (abbreviated to CRTF) and apply them to a data set from the U.S. Consumer Expenditure Survey. In addition, we show, by means of simulations with real and artificial data, that CRTF methods have advantages with respect to (i) bias, (ii) mean squared error, (iii) speed, and (iv) applicability to large numbers of auxiliary predictor variables (denoted by X) subject to non-trivial missing-data rates, although no method is uniformly best. Performance depends on the number of X variables and their incomplete-data patterns, and on the extent to which the observed data and missing-data patterns are consistent with model conditions used implicitly or explicitly by a method.

Given unit i, let w_i denote its sampling weight, with w_i inversely proportional to the probability that unit i is in S. Let S_1 be the subset of S containing the non-missing Y values and let y_i denote the value of Y for unit i in S_1. Let π_i denote the probability that Y is non-missing for unit i and let $\hat{\pi}_i$ be an estimate of π_i. The *inverse probability weighted* (IPW) estimate of μ is

$$\Big(\sum_{i\in S_1}\hat{\pi}_i^{-1}w_i\Big)^{-1}\sum_{i\in S_1}\hat{\pi}_i^{-1}w_iy_i \tag{1.1}$$

(see, e.g., Little (1986) and Seaman and White (2013)). Its bias depends on accurate specification and estimation of a model for π_i (throughout this paper, expectations are evaluated with respect to both the sample design and the nonresponse mechanism). Logistic regression is often used, but it is difficult to apply if there are many missing values in the X variables. The simple weighted mean of the non-missing values (denoted by SIM) given by

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0225

$\bar{y} = (\sum_{i \in S_1} w_i)^{-1} \sum_{i \in S_1} w_i y_i$, is an IPW estimate with constant $\hat{\pi}_i$.

Another approach is *imputation* of the missing Y values. Let $S_2 = S - S_1$ be the subset of S with missing Y. Let $\hat{y}_j$ denote the imputed value of Y for unit j in S_2. Then the *mean imputation* estimate of μ is

$$\Big(\sum_{k \in S} w_k\Big)^{-1} \Big(\sum_{i \in S_1} w_i y_i + \sum_{j \in S_2} w_j \hat{y}_j\Big); \tag{1.2}$$

SIM is a special case with $\hat{y}_j = \bar{y}$ for all $j \in S_2$. If the X variables are completely observed, the $\hat{y}_j$ values may be alternatively obtained by *regression imputation* (Buck (1960)), where a regression model of Y on X is fitted to the observations in S_1 and the $\hat{y}_j$ are predicted from the X values in S_2. If the X variables have missing values as well, the *complete-case* method fits the regression model to the subset of values with complete observations in the X and Y variables. Another solution is *nearest-neighbor imputation*, where $\hat{y}_j$ is the value of y_i in S_1 for which x_i is nearest to x_j. If X has no missing values, this method can yield asymptotically unbiased estimates of functions of means (Chen and Shao (2000)). But this is not directly applicable if the X variables have missing values. *Hot deck* (Little and Rubin (2002)) imputes missing values by random sampling of non-missing values within 'adjustment cells' that are prespecified partitions of the data. A major difficulty is finding suitable partitions.

Yet another method is *maximum likelihood*, which draws random observations from a parametric model fitted to the (X, Y) observations. Assuming that (i) the model is correct, (ii) the X variables are completely observed, and (iii) the Y values are *missing at random* (MAR, the probability that a value is missing does not depend on the value itself, conditional on the non-missing values of the X variables), Rubin (1987) showed that inferences from multiply imputed data are statistically valid for large samples. If there are missing X values,

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0225

the EM algorithm (Dempster et al. (1977)) is often used to estimate the parameters in the model. The AMELIA algorithm (Honaker et al. (2011)) uses multivariate normal likelihoods, with categorical variables converted to dummy 0-1 vectors.

Finally, there is *sequential regression*, a variable-by-variable imputation technique. Initially, all missing values in the X and Y variables are imputed by their means, medians, or modes. Then each variable is regressed in turn on the other variables and missing values are updated with the predicted values. The procedure is continued for several cycles to reduce the effects of the initial imputed values (see, e.g., Raghunathan (2016)). MICE (van Buuren and Groothuis-Oudshoorn (2011)), which stands for *multiple imputation by chained equations*, is one implementation. It uses linear regression for imputation of ordinal variables and polytomous logistic regression for categorical variables. Theoretical arguments for the effectiveness of sequential regression have been given, but they are based on the assumption of a correctly specified linear regression model relationship between the variable being imputed and the covariates (White and Carlin (2010)).

In practice, MICE runs into computational problems when there are many X variables with missing values. Linear regression fails if there is multicollinearity and logistic regression fails if there is quasi-complete separation (Albert and Anderson (1984); Hosmer et al. (2013)). Conversano and Siciliano (2009), Burgette and Reiter (2010), Wallace et al. (2010), and Doove et al. (2014) replace linear and logistic regression with CART (Breiman et al. (1984)) classification and regression trees, respectively, and, after several iterations, obtain multiple imputations by sampling from the observed responses in the terminal nodes of the trees. Vateekul and Sarinnapakorn (2009) first fit a classification tree to each missingness indicator; then at each terminal node of the tree, they fit a regression tree to the observations

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0225

there and use the fitted values from the latter tree for imputation.

Rubin (1987) proposed imputing the missing values multiple times to obtain variance estimates. Although there are many simulation studies on multiple imputation (MI), almost all used normally distributed data and missingness mechanisms defined by logistic regression models (see, e.g., Allison (2000); Schafer and Graham (2002); Carpenter et al. (2007); Burgette and Reiter (2010); White and Carlin (2010)). Little is known about the performance of the methods in actual settings where variables are not normally distributed (e.g., categorical variables) and probabilities of missingness are not determined by logistic regression.

To our knowledge, only three published simulation studies used real data (Ambler et al. (2007); Yu et al. (2007); Andridge and Little (2010)). None had more than 20 X variables and only one had missing values in X. To evaluate the methods under more demanding and most realistic conditions, we use a U.S. Consumer Expenditure (CE) Survey data set as the test-bed. The data contain more than 500 X variables, many with substantial numbers of missing values. We consider nine new CRTF methods based on the CART and GUIDE (Loh (2002, 2009); Loh and Zheng (2013)) algorithms and compare them against AMELIA and MICE on the data as well as on simulated data derived from them. Missing values in the X variables are real, not artificially simulated as MAR. For balance, we also compare the methods under ideal conditions for the parametric methods.

The remainder of the article is organized as follows. Section 2 describes the CE data, Section 3 introduces the methods and Section 4 applies them to the data. Section 5 reports on the simulation experiments and the results, and Section 6 summarizes the conclusions and discusses some potential extensions.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0225

2 2013 CE data

Table 1: Codes and definitions of missing value flag variables

A	valid nonresponse: a response is not anticipated
C	"don't know", refusal or other type of nonresponse
D	valid data value
T	topcoding applied to value

The CE Survey is a longitudinal survey sponsored by the U.S. Bureau of Labor Statistics. It collects information on consumers' expenditures and incomes as well as characteristics of the consumers. We use a subset of the public-use microdata of the 2013 CE Survey. Answers from 25,822 consumer units (CUs), roughly equivalent to households, were obtained on more than 600 questions. Each CU is associated with a sampling weight named FINLWT21 whose values do not vary greatly (coefficient of variation 0.375). Details on the survey may be found in Bureau of Labor Statistics (2016, Chap. 6).

We consider as Y the variable INTRDVX, which is the amount of interest and dividend income received by the CU during the past 12 months. Variables are potentially subject to four types of missingness, as reflected in the "missing value flag variable" codes given in Table 1. The variables are often identified by underscores at the end of their names; e.g, INTRDVX_ is the flag variable associated with INTRDVX. To protect respondent confidentiality, the CE Survey replaces INTRDVX values greater than \$32,000 (called the "upper critical value") with \$98,338 (called the "upper topcode value") and assigns code T to INTRDVX_. In our analysis, we omit CUs with INTRDVX_ codes A and T. This yields a sample size of 4609, of which 1771 CUs have INTRDVX_ = C, i.e., INTRDVX is missing. Topcoding is discussed in Section 6 below.

There are 587 neither constant nor completely missing X variables that may be used to

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0225

estimate the population mean of Y. About 20% of these X variables have missing values; 67 of them have more than 95% missing values. No CU has complete responses on all 587 variables, which include both variables defined at the CU level (e.g., housing tenure) and variables defined for geographical areas. The latter variable type includes STATE, REGION and PSU. The STATE identifier is subject to confidentiality restrictions in some cases. PSUs are generally counties or small clusters of counties. In this data set, each "A" size PSU is a cluster with a population of over 2.7 million people, and is self-representing (i.e., selected with probability one) under the CE design. Only "A" size PSUs are identified publicly and other PSU labels are coded as missing for confidentiality reasons. This paper treats PSU membership as a fixed effect, instead of a random effect, in the models for missing value imputation.

3 Estimation methods

Thirteen methods are compared on their ability to estimate the population mean μ of INTRDVX accurately and quickly. All except the first three are new and all except the first four are based on CRTF.

SIM. This is the "simple" weighted mean of the non-missing Y values with weights proportional to FINLWT21. It yields an estimate of \$1900 for the 2013 CE data.

MICE. This is the R version (van Buuren and Groothuis-Oudshoorn (2011)) of MICE with default options, including five multiple imputations. Multicollinearity in linear regression and quasi-complete separation in logistic regression can severely limit the number of variables it can employ. After much trial and error, 19 X variables were found to work with the software: AGE_REF, BATHRMQ, BEDROOMQ, BLS_URBN, BUILDING, CUTENURE, EARN-

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0225

COMP, EDUC_REF, ETOTA, FAM_TYPE, MARITAL1, NO_EARNR, NUM_AUTO, OCCUCOD1, REF_RACE, REGION, ROOMSQ, SEX_REF, and ST_HOUS. Their definitions are given in the Supplementary Material. Only 5 of the 19 variables have missing values and only 1 (OCCUCOD1) has a substantial number of missing values (1697).

AME. This is AMELIA (Honaker et al. (2011)) with default parameters except that the empirical prior level is set at 5. The prior shrinks the covariances of the data but keeps the means and variances the same. The level of 5 was suggested by AMELIA author M. Blackwell for dealing with many missing values, small sample sizes, large correlations, and categorical variables with many levels.

AIPW. This is the IPW method that uses AMELIA to impute missing X variables and then uses logistic regression to estimate missing propensities.

GCT. This is an IPW method that uses a GUIDE classification tree (Loh (2009)) to estimate π_i, the probability that INTRDVX is non-missing (i.e., INTRDVX_ = D). Figure 1 shows the tree constructed using all 587 X variables (definitions of the variables are given in the Supplementary Material). The estimate of π_i in each terminal node is the proportion of non-missing INTRDVX values in the node. Substituting these estimates in (1.1) yields an estimate of \$1942 for μ. The terminal nodes can also serve as adjustment cells for conditional mean imputation in (1.2).

RCT. This is GCT with the CART algorithm—specifically RPART (Therneau et al. (2015))—instead of GUIDE. Section 4.1 reviews the key differences between CART and GUIDE.

GCF. This is another IPW method where, instead of using a single tree to estimate π_i, it uses a GUIDE classification forest (Loh (2014)), which is an ensemble of 500 unpruned classification trees with each constructed from a bootstrap sample of the data. GUIDE

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0225

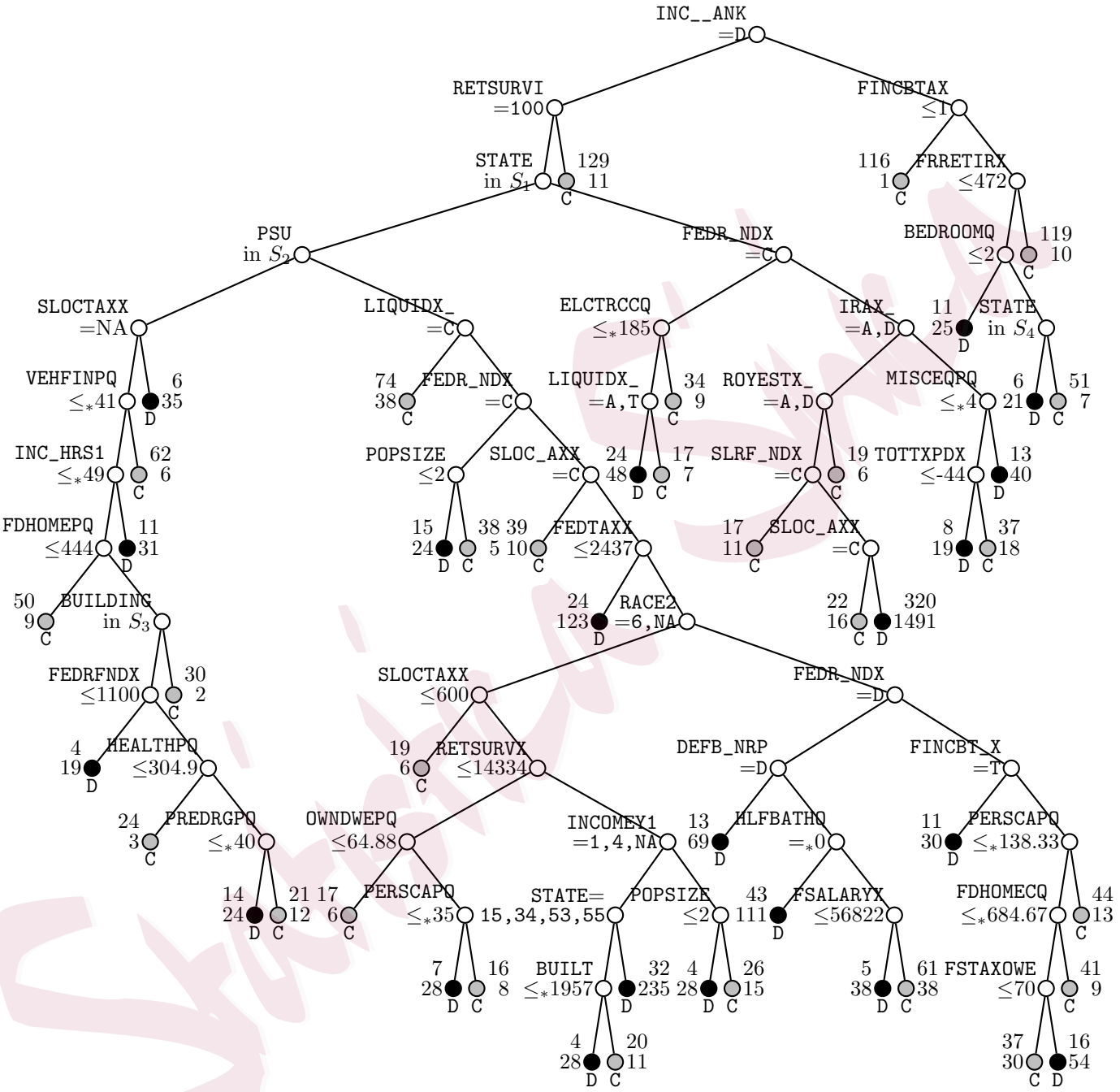

Figure 1: GUIDE classification tree for predicting `INTRDVX_`. At each split, an observation goes to the left branch if and only if the condition is satisfied. The symbol '$\leq_*$' stands for '$\leq$ or missing'. Set S_1 = {10, 12, 15, 17, 22, 25, 26, 34, 36, 39, 42, 45, 47, 53, 55, 8}. Set S_2 = {1102, 1109, 1110, 1423}. Set S_3 = {1, 11, 7, 8}. Set S_4 = {18, 2, 22, 24, 27, 42, 47, 48, 51, 8}. Predicted classes printed below terminal nodes; sample sizes for `INTRDVX_` = `C` and `D`, respectively, beside nodes.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0225

forest is similar to Random forest (Breiman (2001)) except for the former using the unbiased split selection method of GUIDE. The R implementation of Random forest (Liaw and Wiener (2002)) is not directly applicable here because it does not allow variables with missing values and categorical variables with more than 32 levels.

GRT. This is a conditional mean imputation method that uses a GUIDE piecewise-constant weighted least-squares regression tree (Loh (2002)) to impute missing INTRDVX values, with weights given by the variable FINLWT21. Figure 2 shows the result based on all 587 X variables. It is constructed from the 2838 CUs with non-missing values in INTRDVX. Using the mean INTRDVX values in the terminal nodes as $\hat{y}_j$ in (1.2) with w_i = FINLWT21 yields an estimated μ of \$1919.

RRT. This is a version of GRT where RPART is used instead of GUIDE.

GRF. This is an alternative to GRT that uses a GUIDE regression forest (Loh (2012, 2014)) of 500 unpruned regression trees to perform conditional mean imputation.

GMICE. Burgette and Reiter (2010) and Wallace et al. (2010) used the R package `tree` (Ripley (2016)) and RPART, respectively, to implement CART-MICE. Neither is applicable here because their execution times grow exponentially with the the number of categorical levels if the response variable is also categorical with more than two levels. GMICE is CART-MICE with GUIDE in place of CART with ten iterations and a single imputation.

DRT and DRF. If parametric models are employed to model the missing propensity and the mean of the response variable (such as logistic regression and linear regression, respectively), an estimate is said to be "doubly robust" if it is consistent as long as at least one of the two models is correctly specified. A standard method to achieve this is to fit a linear regression model to the response variable and then add an IPW estimate constructed from

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0225

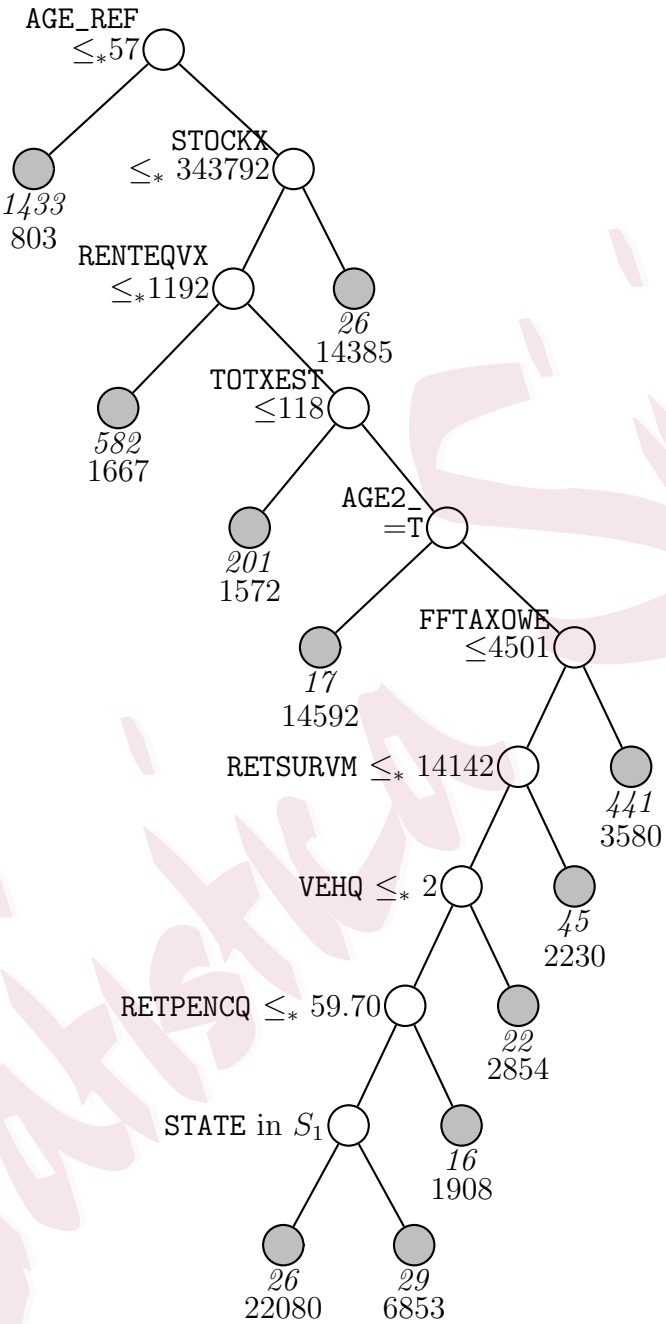

Figure 2: GUIDE regression tree for predicting `INTRDVX`. At each split, an observation goes to the left branch if and only if the condition is satisfied. The symbol '$\leq_*$' stands for '$\leq$ or missing'. For splits on categorical variables, values not present in the training sample go to the right. Set $S_1 = \{12, 13, 17, 29, 36, 48, 53, 6, 8, \text{NA}\}$. Sample size (*in italics*) and mean of `INTRDVX` printed below nodes.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0225

its residuals (Robins and Rotnitzky (1995); Little and An (2004); Kang and Schafer (2007)). Although the methods studied here are not specifically based on any parametric models, we can nonetheless use the double robustness approach to construct new estimators, as follows. Let $\hat{y}_i$ be the predicted value of y_i from GRT or GRF and let $\hat{\epsilon}_i = y_i - \hat{y}_i$ if $i \in S_1$ (i.e., if y_i is observed). The resulting doubly robust estimate, denoted by DRT and DRF, respectively, is

$$\Big(\sum_{i\in S} w_t\Big)^{-1} \sum_{i\in S} w_i \hat{y}_i + \Big(\sum_{i\in S_1} \hat{\pi}_i^{-1} w_i\Big)^{-1} \sum_{i\in S_1} \hat{\pi}_i^{-1} w_i \hat{\epsilon}_i.$$

4 Application to CE data

The tree for predicting INTRDVX_ in Figure 1 first splits on INC_RANK (flag variable for percent income rank). CUs with INC_RANK = D are sent to the left node where they are split on RETSURVI (imputation descriptor variable for the amount of retirement, survivor or disability pensions in past 12 months). The 140 CUs with RETSURVI $\neq$ 100 are sent to the right terminal node, where 129 and 11 have INTRDVX_ = C and D, respectively. Thus the probability p that INTRDVX is non-missing in this node is estimated by 11/140. Let t denote a terminal node of the tree and $p(t)$ the proportion of non-missing INTRDVX in t. Let w_i denote the sampling weight FINLWT21 for unit i in t. Setting $\hat{\pi}_i = p(t)$ for all CUs in t in (1.1) gives the IPW (GCT) estimate of \$1942. The terminal nodes can serve as adjustment cells for hot deck imputation.

The tree can also be used to impute missing values in INTRDVX. Specifically, CUs with missing INTRDVX in a terminal node can be imputed with the mean INTRDVX (weighted by FINLWT21) of the CUs with observed responses in the node. Using the node weighted mean as $\hat{y}_j$ in (1.2) yields the same estimate of \$1942.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0225

A number of missing-data flag variables (FEDR_NDX, INC_ANK, IRAX_, LIQUIDX_, ROYESTX_, SLOC_AXX, and SLRF_NDX) appear in Figure 1, showing that missing-data flag variables for asset and income variables are important predictors of the missingness propensity of INTRDVX. This illustrates the flexible way in which tree methods can explore the use of both observed values and related missing-variable flags as predictors in a response propensity model. It is impossible for logistic regression to do this given the large numbers of predictor variables and their missing values.

A complement to the classification tree for the flag variable INTRDVX_ is the regression tree for variable INTRDVX in Figure 2. Unlike classification models, the regression model uses only the subset of 2838 CUs with non-missing INTRDVX. It splits first on AGE_REF (age of reference person). If AGE_REF ≤ 57 or missing, the CU goes to the left terminal node (after pruning) which contains 1433 CUs with a weighted mean INTRDVX of \$803. Otherwise, if AGE_REF is non-missing and > 57, the tree splits on STOCKX (value of all directly-held stocks, bonds, and mutual funds). The 26 CUs there with STOCKX $>$ \$343,792 go to the right terminal node where they have a weighted mean INTRDVX of \$14,385. Using these weighted terminal node means for $\hat{y}$ in (1.2) gives the GRT estimate of \$1919.

4.1 Differences between CART and GUIDE

At each node, CART searches for the best split on each X variable and then selects the "variable-split set" pair that most reduces node impurity. If X is ordinal with m distinct values, CART evaluates $(m-1)$ splits of the form "$X \leq x_0$", with x_0 being a midpoint between consecutively ordered values. If X is categorical with m levels, CART searches through $(2^{m-1}-1)$ splits of the form "$X \in A$" to find the subset A that yields the best

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0225

split on X. Thus the number of splits to be evaluated increases exponentially with m if X is categorical. For example, STATE has $2^{38} - 1 \approx 2 \times 10^{11}$ splits because it has 39 values. Other variables with many levels include HHID (household identifier), PSU (primary sampling unit), OCCUCOD1 (occupation of reference person), and OCCUCOD2 (spouse occupation), with 46, 21, 15, and 15 levels, respectively. Consequently, CART is known to be biased toward selecting variables that allow more splits (Loh and Shih (1997); Kim and Loh (2001)). GUIDE uses a two-step approximate solution that avoids the bias. It first selects an X variable to split the node by means of chi-squared tests of association with Y. Then it finds the best split on the selected X. See Loh (2002, 2009, 2014) for the additional steps GUIDE uses to reduce the search space for categorical variables with large m.

Another major difference between GUIDE and CART is how they deal with missing values in the X variables. If X has missing values, GUIDE creates a "missing" level to use in the chi-squared tests for variable selection as well as for split set selection. This allows every observation to be used. CART, on the other hand, uses only observations with non-missing values in (X, Y) to find the best split on each X. Then it uses a system of surrogate splits on other X variables to send observations with missing values through the split. This approach is biased toward selecting variables with more missing values (Kim and Loh (2001)). There is some evidence that the GUIDE approach yields higher average classification accuracy than CART's (as implemented in RPART); see Loh (2009).

4.2 Results

To allow comparison with MICE, we applied the 13 methods to 3 nested sets of X variables. The smallest is the set of 19 variables mentioned in Sec. 3 for which MICE does not fail.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0225

They are far from ideal because only 5 variables among them have missing values (and only 1 has a substantial number of missing values). The second set consists of 52 X variables, obtained by combining the above 19 with the top 20 X variables determined by the GUIDE importance ranking method (Loh (2012); Loh et al. (2015)) for predicting INTRDVX_ and INTRDVX, respectively. The third is the full set of 587 variables.

SIM estimates the population mean of INTRDVX by \$1900 for all three sets because it does not use any auxiliary variables. Estimates for the other methods are shown in the top half of Table 2 and graphed in Figure 3. Every method works on the set of 19 variables but MICE fails for the other two sets. The estimates range from a low of \$1674 (GCF, 52 variables) to a high of \$2184 (AME, 52 variables). In Figure 3, SIM is indicated by a solid line and its value plus and minus one standard error of \$146 (estimated using balanced repeated replicate weights) by dotted lines. A large majority of the estimates lie within one standard error of SIM and all are within two standard errors. Qualitatively similar results (reported in the Supplementary Material) were obtained with the 2014 CE data.

Table 2 also gives the computation times on a Linux computer with a 2.4 GHz AMD Opteron 16-core processor and 64 GB memory. For 19 predictors, MICE is the slowest, taking 430 sec. for five imputations—more than 4000 times slower than RCT, which is fastest at 0.1 sec. AME is the next slowest at 139 sec. for the same 19 predictors. For 52 predictors AME took 18 hours, and for 587 predictors it was terminated after executing for 6 months.

5 Simulation experiments

Simulation experiments were carried out to examine the bias and mean squared error of the methods. To avoid issues due to sampling weights, equal-probability sampling was used

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0225

Table 2: Estimates (weighted by FINLWT21) and computation times of mean INTRDVX for 3 sets of predictor variables for the 2013 CE data. Results for AIPW, AME and MICE are based on 5 multiple imputations. AIPW and AME did not yield any results for 587 variables after more than 6 months.

	19 variables		52 variables		587 variables	
	Est.	Sec.	Est.	Sec.	Est.	Sec.
AIPW	2055	122	1900	72029	-	-
AME	2088	139	2184	111068	-	-
GCT	1949	9	1960	17	1942	283
GCF	1731	97	1674	171	1750	2669
GRT	2042	8	2011	18	1919	273
GRF	2007	226	1997	344	1948	2457
DRT	2016	17	2051	35	1990	565
DRF	1975	235	2039	361	1960	2739
GMICE	2094	57	2005	434	2002	76874
MICE	2031	430	Fail	-	Fail	-
RCT	1900	1	1898	1	1920	19
RRT	1940	1	1940	1	1930	7
SIM	1900	-	1900	-	1900	-

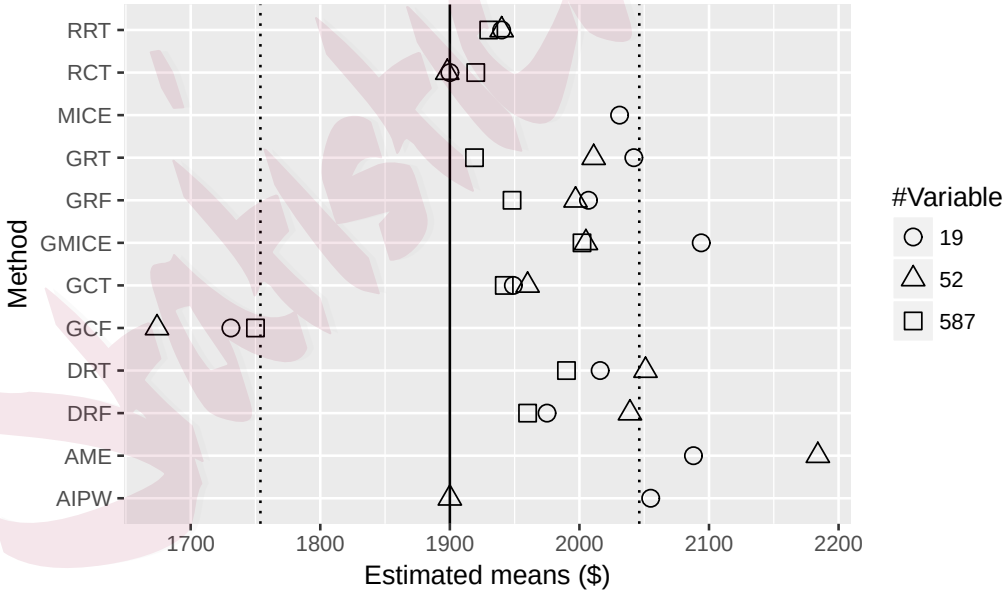

Figure 3: Estimates of mean INTRDVX (weighted by FINLWT21) for 2013 CE data using 19, 52 and 587 predictor variables. The solid line marks the value of the SIM estimate of $1900 and the dotted lines mark SIM plus and minus one standard error of $146 (calculated from balanced repeated replication weights).

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0225

throughout.

5.1 Experimental design

Studies that try to simulate real conditions usually start with a data set $\mathcal{D}$ and then generate artificial populations in two steps:

Step I. Impute the missing Y values in $\mathcal{D}$ and treat the resulting data set as a finite population $\mathcal{P}_1$ so that the mean μ of Y is known.

Step II. Generate a population $\mathcal{P}_2$ from $\mathcal{P}_1$ by making some X and Y values in $\mathcal{P}_1$ missing.

To make $\mathcal{P}_2$ as similar to $\mathcal{D}$ as possible, the way these two steps are carried out is important. Often, MICE is used in Step I and a logistic regression propensity model is fitted to the missing value flag variable in $\mathcal{D}$ to estimate the probability that the variable itself is missing in Step II. These choices have three undesirable consequences.

1. Use of MICE in Step I limits the number of variables with missing values to no more than a few, due to problems with multicollinearity and quasi-complete separation. It is inapplicable to the CE data without prior variable selection. Besides, to impute a Y variable, MICE necessarily imputes all X variables with linear and logistic regression models. This imposes parametric relationships among the X variables in $\mathcal{P}_1$ that do not exist in $\mathcal{D}$. Therefore, the simulated data often do not resemble the data.

2. In Step II, logistic regression propensity models cannot be constructed from X variables with missing values. Either the missing values must be imputed first or the propensity models must be built from subsets of variables or subsets of data. Neither solution is desirable. Imputing the X variables (e.g., with MICE) distorts the data and building propensity models from subsets of data requires the artificial assumption that the X variables are MAR. Some

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0225

studies solve this problem by using only completely observed X variables for propensity modeling (Burgette and Reiter (2010)), but this is artificial too because the probability of a missing value in Y often depends on X variables with missing values—see, for example, the classification tree model in Figure 1, where many predictors of missingness in INTRDVX are missing value flags of other variables.

3. Linear and logistic regression models with prespecified fixed sets of predictor variables are inapplicable if the number of variables exceeds the sample size.

To overcome these difficulties, we started with $\mathcal{D}$ the 2013 CE data consisting of 4609 units and Y = INTRDVX. We skipped Step I by defining $\mathcal{P}_1$ as the subset of 2838 units in $\mathcal{D}$ with observed INTRDVX values. This allows the mean μ of Y to be computed without imputation. In Step II, we fit a GUIDE *classification forest* to the 4609 units and 587 predictor variables in $\mathcal{D}$, using the missing value flag associated with Y as dependent variable. Then we used the fitted model to estimate the probability that Y is missing for each unit in $\mathcal{P}_1$. Finally, we generated $\mathcal{P}_2$ from $\mathcal{P}_1$ by making each value of Y in $\mathcal{P}_1$ independently missing according to these probabilities.

We used the same subsets of 19, 52, and 587 X variables as in Section 4.2 in the simulations. Each trial consisted of generating a simulated population $\mathcal{P}_2$ as described and drawing a simple random sample without replacement from it. Each method was applied to the sample to impute the missing INTRDVX values (for conditional mean imputation) or to estimate π_i (for IPW) and then finally to estimate μ. We used sampling fractions of 5%, 10%, and 25% but report only the results for 10% sampling because the other results are qualitatively similar. Importantly, in our simulations Y is MAR only when all 587 X variables are included. When 19 and 52 variables are used, Y is not MAR. Therefore our

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0225

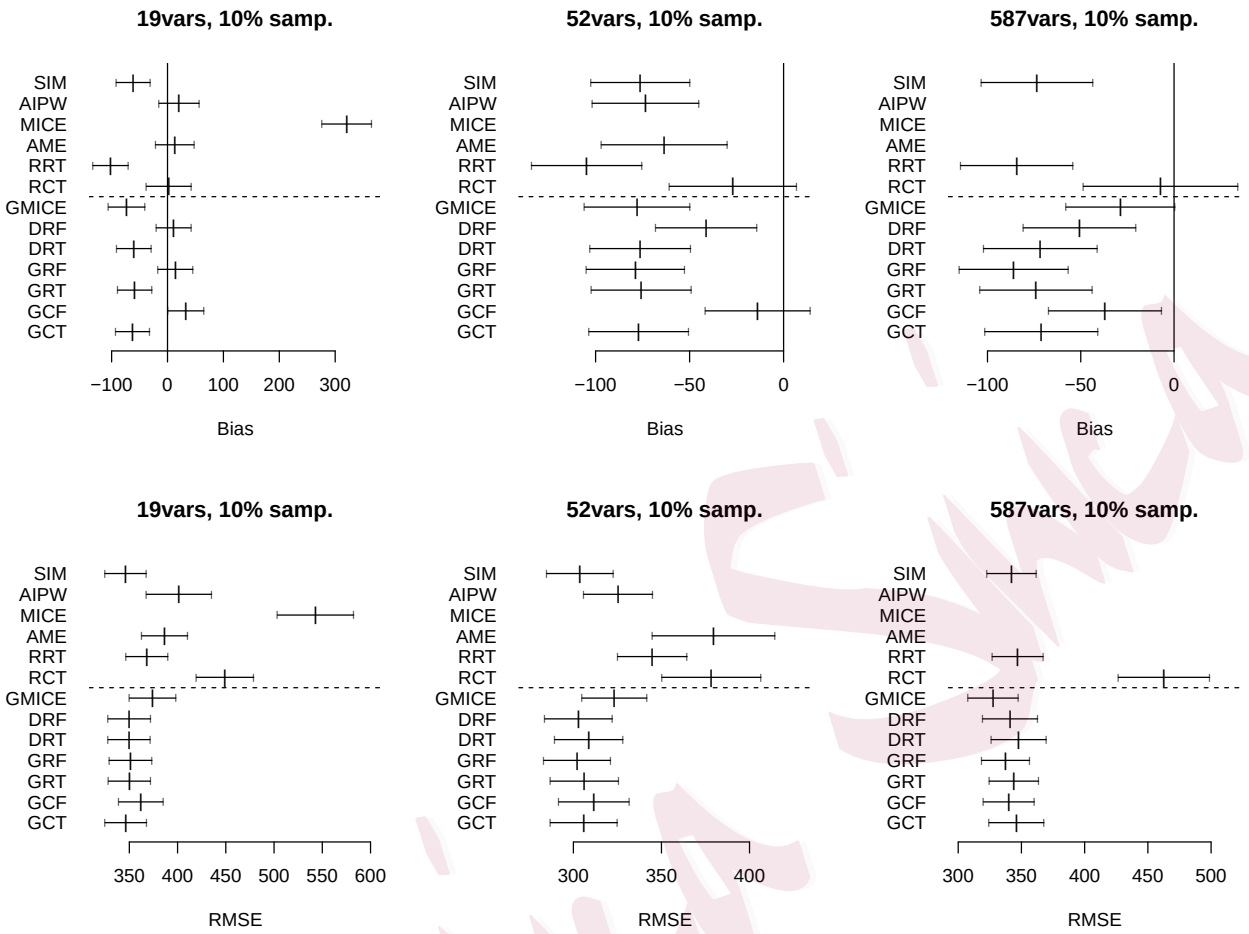

Figure 4: Bias and RMSE (with 2-SE bars) for populations generated from 19, 52 and 587 X variables. MICE failed for 52 and 587 variables. Computations for AME and AIPW were aborted after 6 months for 587 variables. Dashed lines separate GUIDE from non-GUIDE methods.

study allows the methods to be evaluated in a fully realistic situation where the X variables are not MAR and where Y is either MAR or not.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0225

5.2 Simulation results

5.2.1 Data models

We used $M = 500$ simulation trials. For trial m, let μ_m denote the population mean of Y and $\hat{\mu}_m$ denote its estimate. The estimated bias and root mean squared error (RMSE) are

$$\begin{aligned} \text{Bias} &= M^{-1} \sum_m (\hat{\mu}_m - \mu_m) \\ \text{RMSE} &= \left\{ M^{-1} \sum_m (\hat{\mu}_m - \mu_m)^2 \right\}^{1/2} . \end{aligned}$$

Figure 4 displays the bias and RMSE (with 2 simulation standard error bars) of the methods. Again, MICE failed for 52 and 587 variables and AIPW and AME did not finish after running for 6 months. The following observations are apparent from the plots.

1. For the 19-variable situation, MICE is the worst by far, in both bias and RMSE. This seems to contradict conventional wisdom (White and Carlin (2010)), but the X and Y variables here are not MAR (Y is MAR with respect to 587 variables, not 19). On the other hand, only 1 of the 19 X variables has a substantial number of missing values. We conclude that MICE is more dependent on the MAR assumption than the other methods.

2. For the same 19-variable situation, AIPW and AME have low biases and RMSEs. But for 52 variables, neither is unbiased and AME is tied with RCT for having the largest RMSE.

3. Among tree methods, RCT has the largest RMSE and RRT the largest bias. Thus tree algorithms are not alike in performance. The variable selection bias of CART is likely a contributing factor.

4. GUIDE-based methods tend to perform relatively well in terms of bias and RMSE, although the long computation times of GMICE (Table 2) puts it at a practical disadvantage. GCF and DRF seem to have the lowest bias.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0225

5.2.2 Parametric models and correlated predictors

To see how the methods compare under conditions ideal for AIPW, AME and MICE, we repeated the simulation experiments with parametric models. The simulation population was based on the subset of the CE data with completely observed values in the above 19-variable set. The Y variable was generated from a linear model with normal errors and its values were made missing according to a logistic regression model. Five percent of the values in each of variables AGE_REF, BLS_URBN, EDUC_REF, NO_EARNR, NUM_AUTO, and ETOTA were made MCAR. Equal-probability samples without replacement were drawn from the population at the 10% sampling rate.

Two linear models were employed. The first was trivial ($Y = \epsilon$), where the X variables had no effect. The results show that the biases of all methods were within two simulation standard errors of 0. The RMSE of AME was the largest, being about 20% larger than the lowest value (from RRT); MICE and RCT were tied for second largest. The second linear model was non-trivial, with regression coefficients estimated from the data. In this case, only AIPW, AME and MICE had biases within two simulation standard errors of 0; they also have the lowest RMSE. SIM, GCT and RRT were the worst, in both bias and RMSE. The other methods had RMSE ranging between 10–30% that of MICE. Details about the simulation and results are in the Supplementary Material.

Because AIPW, AME and MICE are based on linear and logistic regression, a final simulation experiment was performed to see the effect of correlation among the predictor variables. Four transportation variables in the CE data with correlations greater than 0.9850 were selected as X variables. The data were simulated the same way (i.e., linear model for Y with logistic missing propensity and MCAR for X) with the X variables restricted to these

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0225

four. The results, given in the Supplementary Material, show that only AME, DRF, GCT, RCT, RRT, and SIM had biases within two SEs of 0; and GCF, GCT, RCT, RRT, and SIM had the largest RMSEs.

6 Conclusion

6.1 Summary of results

We introduced several techniques to use classification and regression tree methods for mean estimation of incomplete data. Some employ regression trees to estimate conditional means in adjustment cells defined by the nodes of the trees. Others employ classification trees to estimate the propensity for item missingness, for use in inverse probability weighting. We applied these methods to fit models to the variable INTRDVX (interest and dividend income) and its missingness indicator from the U.S. Consumer Expenditure Survey. Of special importance was the fact that several potential predictor variables were themselves subject to relatively high item-missingness rates.

In addition, using this U.S. Consumer Expenditure Survey data set as a test bed, we performed several simulation experiments to compare the methods with AIPW, AME and MICE. A major feature of the experimental design is the novelty of ensuring that the predictor variables are naturally missing, i.e., not constrained to be MAR, in the simulation populations. The results demonstrate that CRTF methods have the following desirable properties that make them deserving of serious consideration for analysis of incomplete data.

1. They are often competitive with AIPW and AME, and superior to MICE in terms of bias and mean squared error for mean estimation. One reason is the nonparametric nature

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0225

of CRTF models. Another is the ability of the models to avoid the need for imputation of predictor variables, thereby preventing propagation of imputation errors. In contrast, AIPW, AME and MICE are based on normality and multivariate linear model assumptions that are seldom satisfied in real and complex data and they all require imputation of predictor and outcome variables.

2. Even under conditions ideal for AIPW, AME and MICE (namely, logistic missing propensity and linear regression mean models), the CRTF methods compare quite favorably. If the mean model is not constant, CRTF methods other than GCT and RRT have RMSEs that are 10–30% larger than those of AME and MICE. On the other hand, if the mean model is constant, CRTF methods except for RCT tend to have lower RMSEs than AME and MICE. Finally, if the predictor variables are highly correlated, AIPW and MICE are biased but AIPW, AME and MICE have the lowest RMSEs. Among CRTF methods, DRF, DRT, GRF, and GRT are the best performers under these parametric conditions, with efficiency losses in RMSE in the 0–30% range.

3. CRTF methods have no sample size limitations. In contrast, parametric methods typically require the sample size to be substantially larger than the number of parameters.

4. CRTF methods are not hindered or crippled by multicollinearity or quasi-complete separation. In fact, collinearity is often used to advantage in tree algorithms; e.g., CART surrogate splits (Breiman et al. (1984)) and GUIDE linear splits (Loh (2009)).

5. For cases that involve large numbers of candidate predictors, CRTF methods can be orders of magnitude faster compared to traditional methods. Because the speed advantage increases nonlinearly with number of variables, it can be invaluable for imputation of multiple data sets, bootstrapping, and other variance estimation techniques in large surveys.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0225

Little and Vartivarian (2003) extended the ideas of Little (1986) to outline two strategies for reducing the number of adjustment cells for nonresponse: (a) choosing cells that are homogeneous with respect to the probability of response, and (b) choosing cells that are homogeneous with respect to the outcome variable. They observed that weighting based on either of these methods of grouping removes nonresponse bias in estimating population means. Classification and regression trees do this naturally. For example, the terminal nodes of the classification tree in Figure 1 constitute cells that are homogeneous with respect to the estimated probability of response, and the regression tree in Figure 2 gives cells that are homogeneous with respect to the outcome variable.

6.2 Potential extensions

The work here can be extended in several directions. First, the current work used design information in only limited ways, including use of the labels of certain self-representing primary sampling units as potential categorical predictors; use of survey weights in weighted least squares estimation of regression trees and forests; and use of balanced repeated replication to compute a standard error for the SIM estimator in Figure 3. It would be useful to consider further design-adjusted versions of the procedures proposed here. This includes the use of weights in the growth of individual classification trees or forests. Of special interest would be evaluation of the extent to which a given procedure may be sensitive to specified patterns of heterogeneity in the weights. For example, all CRTF algorithms use approximations; the properties of the resulting procedures can be sensitive to the extent to which a given data set is consistent with the approximations used for that procedure; and some weight-heterogeneity patterns may exacerbate that sensitivity. In addition, several analyses

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0225

here used some predictor variables that are equal to membership indicators for certain large primary sample units. It would be of interest to extend previous literature on the use of stratum and PSU labels in regression to the current case.

Second, development of appropriate variance estimators would provide important tools for use in pruning of trees, and for inference related to CRTF-based estimators of population- and subpopulation-level means and related quantiles, or other population parameters. This would require CRTF-related extensions of standard theorems on the properties of replication-based variance estimators under complex sample designs; and may also require complex-sample-design extensions of post-selection inference approaches developed previously for simple random samples (Loh et al. (2016)). Of special importance would be conditioning arguments arising from the fact that the structure of a given tree is data-driven and not determined a priori.

Pending rigorous theoretical development in those areas, one could consider ad hoc procedures based on standard methods of balanced repeated replication (BRR) developed for the U.S. Consumer Expenditure Survey and other large-scale complex surveys. For general background on BRR methods, see, e.g., Krewski and Rao (1981), Shao (1996), Wolter (2007) and references cited therein. For discussion of BRR methods applied to CE, see Bureau of Labor Statistics (2016). In particular, for a population mean estimator $\hat{\mu}$ considered in the current paper, one could compute that estimator separately for the full sample and for each of the CE half-sample replicates, and then use customary BRR procedures to compute the resulting replication-based variance estimator. In addition, one could consider ad hoc estimation of the variance of $\hat{\mu}$ based on bootstrap methods, with the bootstrap draws aligned appropriately with the salient features of the complex sample design. For

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0225

example, in CE applications based on public-use datasets, one could consider resampling of pseudo-PSUs within pseudo-strata, with membership of individual consumer units within the pseudo-PSUs and pseudo-strata determined by the structure of the 44 BRR replicates. Finally, one could consider similar approaches to variance estimation for the mean or other estimator associated with a particular node of a tree. In the latter case, however, due to the data-driven nature of tree construction, the calculation and interpretation of numerical results would be conditional on a specific tree.

Third, the current paper focused on estimation of the means of the "interest and dividend" variable, and a related missing-data flag, for the U.S. Consumer Expenditure Survey. Some users of CE data, however, are interested in carrying out econometric analyses based on, for example, regression, generalized linear models and more complex hierarchical models. For those situations, one may need to impute simultaneously a substantial number of missing income and expenditure variables for a given consumer unit. For such cases, evaluation criteria for the properties of the resulting imputation procedure may be more complex, as discussed in, e.g., Rubin (1996).

Finally, the current analysis focused only on the subpopulation of units that did not have INTRDVX values above the topcoding threshold of $32,000. It would be of interest to explore the use of classification trees to model the joint and conditional probabilities that a consumer unit (a) has a true INTRDVX value below the $32,000 threshold and (b) provides a non-missing INTRDVX item response. In addition, it would be of interest to explore the use of regression trees or forests to analyze public-use data sets that replace large item responses with a single reported value. For example, the CE public-use data set replaced INTRDVX values above $32,000 with the single value $98,338, and a related multiply-imputed public-

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0225

use data set replaced the above-threshold INTRDVX values with the same value $55,632 for each of the multiple imputations.

Supplementary Materials

The online supplementary materials give definitions of the predictor variables, population mean estimates from the 2014 CE data, and results of simulation experiments with parametric models.

Acknowledgments

W.-Y. Loh's research was supported in part by an ASA/NSF/BLS research fellowship and NSF grant DMS-1305725. Work on this paper was carried out while J. Eltinge was at the U.S. Bureau of Labor Statistics. The authors thank Steve Henderson, Geoff Paulin and Adam Safir of the U.S. Bureau of Labor Statistics for very productive discussions and help on the U.S. Consumer Expenditure Survey; and thank the Editor and two referees for very helpful comments on an earlier version of this paper. They also thank Matthew Blackwell for assistance with the AMELIA software. This report is released to inform interested parties of ongoing research and to encourage discussion. The views expressed on statistical, methodological, technical, operational or policy issues are those of the authors and not necessarily those of the U.S. Bureau of Labor Statistics nor of the U.S. Census Bureau.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0225

References

Albert, A. and J. A. Anderson (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika 71*, 1–10.

Allison, P. D. (2000). Multiple imputation for missing data. a cautionary tale. *Sociological Methods and Research 28*, 301–309.

Ambler, G., R. Z. Omar, and P. Royston (2007). A comparison of imputation techniques for handling missing predictor values in a risk model with a binary outcome. *Statistical Methods in Medical Research 16*, 277–298.

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

Breiman, L. (2001). Random forests. *Machine Learning 45*, 5–32.

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

Buck, S. F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society, Series B 22*, 302–306.

Bureau of Labor Statistics (2016). *Handbook of Methods, Consumer Expenditures and Income*. U.S. Department of Labor. `https://www.bls.gov/opub/hom/cex/pdf/cex.pdf`.

Burgette, L. F. and J. P. Reiter (2010). Multiple imputation for missing data via sequential regression trees. *American Journal of Epidemiology 172*, 1070–1076.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0225

Carpenter, J. R., M. G. Kenward, and I. R. White (2007). Sensitivity analysis after multiple imputation under missing at random: a weighting approach. *Statistical Methods in Medical Research 16*, 259–275.

Chen, J. and J. Shao (2000). Nearest neighbor imputation for survey data. *Journal of Official Statistics 16*, 113–131.

Conversano, C. and R. Siciliano (2009). Incremental tree-based missing data imputation with lexicographic ordering. *Journal of Classification 26*, 361–379.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Ser. B 39*, 1–38.

Doove, L., S. V. Buuren, and E. Dusseldorp (2014). Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational Statistics & Data Analysis 72*, 92–104.

Honaker, J., G. King, and M. Blackwell (2011). Amelia II: A program for missing data. *Journal of Statistical Software 45*, 1–47.

Hosmer, D. W., S. Lemeshow, and R. X. Sturdivant (2013). *Applied Logistic Regression* (3rd ed.). Hoboken, NJ: Wiley.

Kang, J. D. Y. and J. L. Schafer (2007). Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data (with discussion). *Statistical Science 22*, 523–580.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0225

Kim, H. and W.-Y. Loh (2001). Classification trees with unbiased multiway splits. *Journal of the American Statistical Association 96*, 589–604.

Krewski, D. and J. N. K. Rao (1981). Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods. *The Annals of Statistics 9*, 1010–1019.

Liaw, A. and M. Wiener (2002). Classification and regression by randomforest. *R News 2*, 18–22.

Little, R. J. A. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review 54*, 139–157.

Little, R. J. A. and H. An (2004). Robust likelihood-based analysis of multivariate data with missing values. *Statistica Sinica 14*, 949–968.

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

Little, R. J. A. and S. Vartivarian (2003). On weighting the rates in non-response weights. *Statistics in Medicine 22*, 1589–1599.

Loh, W.-Y. (2002). Regression trees with unbiased variable selection and interaction detection. *Statistica Sinica 12*, 361–386.

Loh, W.-Y. (2009). Improving the precision of classification trees. *Annals of Applied Statistics 3*, 1710–1737.

Loh, W.-Y. (2012). Variable selection for classification and regression in large p, small n problems. In A. Barbour, H. P. Chan, and D. Siegmund (Eds.), *Probability Approximations*

and Beyond, Volume 205 of *Lecture Notes in Statistics—Proceedings*, New York, pp. 133–157. Springer.

Loh, W.-Y. (2014). Fifty years of classification and regression trees (with discussion). *International Statistical Review 34*, 329–370.

Loh, W.-Y., H. Fu, M. Man, V. Champion, and M. Yu (2016). Identification of subgroups with differential treatment effects for longitudinal and multiresponse variables. *Statistics in Medicine 35*, 4837–4855.

Loh, W.-Y., X. He, and M. Man (2015). A regression tree approach to identifying subgroups with differential treatment effects. *Statistics in Medicine 34*, 1818–1833.

Loh, W.-Y. and Y.-S. Shih (1997). Split selection methods for classification trees. *Statistica Sinica 7*, 815–840.

Loh, W.-Y. and W. Zheng (2013). Regression trees for longitudinal and multiresponse data. *Annals of Applied Statistics 7*, 495–522.

Raghunathan, T. E. (2004). What do we do with missing data? Some options for analysis of incomplete data. *Annual Review of Public Health 25*, 99–117.

Raghunathan, T. E. (2016). *Missing Data Analysis in Practice*. Boca Raton, FL: CRC Press.

Ripley, B. (2016). *tree: Classification and Regression Trees*. R package version 1.0-37.

Robins, J. M. and A. Rotnitzky (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association 90*, 122–129.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York, NY: Wiley.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association 91*, 473–489.

Schafer, J. L. and J. W. Graham (2002). Missing data: our view of the state of the art. *Psychological Methods 7*, 147–177.

Seaman, S. R. and I. R. White (2013). Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research 22*, 278–295.

Shao, J. (1996). Invited discussion paper resampling methods in sample surveys. *Statistics 27*(3-4), 203–237.

Therneau, T., B. Atkinson, and B. Ripley (2015). *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1-10.

van Buuren, S. and K. Groothuis-Oudshoorn (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software 45*, 1–67.

Vateekul, P. and K. Sarinnapakorn (2009). Tree-based approach to missing data imputation. In *IEEE International Conference on Data Mining Workshops*, pp. 70–75.

Wallace, M. L., S. J. Anderson, and S. Mazumdar (2010). A stochastic multiple imputation algorithm for missing covariate data in tree-structured survival analysis. *Statistics in Medicine 29*, 3004–3016.

White, I. R. and J. B. Carlin (2010). Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statistics in Medicine 29*, 2920–2931.

Wolter, K. M. (2007). *Introduction to Varaince Estimation*. New York, NY: Springer.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0225

Yu, L.-M., A. Burton, and O. Rivero-Arias (2007). Evaluation of software for multiple imputation of semi-continuous data. *Statistical Methods in Medical Research 16*, 243–258.

Department of Statistics, University of Wisconsin, Madison, WI 53706

E-mail: loh@stat.wisc.edu and yuanzhi.li@wisc.edu

Research and Methodology Directorate, U.S. Census Bureau, Washington, DC 20233

E-mail: John.L.Eltinge@census.gov

Office of Survey Methods Research, U.S. Bureau of Labor Statistics, Washington, DC 20212

E-mail: Cho.Moon@bls.gov