

Catalogue no. 12-001-X  
ISSN 1492-0921

## Survey Methodology

# sCHAID: A tool for constructing nonresponse adjustment cells under a design-based framework

by Jean D. Opsomer and Minsun K. Riddles

Release date: June 30, 2025



Statistics  
Canada Statistique  
Canada

Canada

---

## How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, [www.statcan.gc.ca](http://www.statcan.gc.ca).

You can also contact us by

**Email at** [infostats@statcan.gc.ca](mailto:infostats@statcan.gc.ca)

**Telephone**, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-514-283-9350

## Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, the Agency has developed standards of service which its employees observe in serving its clients. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on [www.statcan.gc.ca](http://www.statcan.gc.ca) under "Contact us" > "[Standards of service to the public.](#)"

## Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Published by authority of the Minister responsible for Statistics Canada

© His Majesty the King in Right of Canada, as represented by the Minister of Industry, 2025

Use of this publication is governed by the Statistics Canada [Open Licence Agreement](#).

**An [HTML version](#) is also available.**

*Cette publication est aussi disponible en français.*

---

# sCHAID: A tool for constructing nonresponse adjustment cells under a design-based framework

Jean D. Opsomer and Minsun K. Riddles<sup>1</sup>

## Abstract

Survey practitioners have increasingly embraced the benefits of modern machine learning techniques, including classification and regression tree algorithms, in the development of nonresponse adjustments. These methods, which do not require a predefined functional relationship between outcomes and predictors, offer a practical means of conducting variable selection and deriving interpretable structures that link response propensity with explanatory variables. However, when applying these algorithms to survey data, it is common to overlook crucial factors like sampling weights, as well as sample design features such as stratification and clustering. To bridge this shortcoming, we propose an extension of the Chi-square Automatic Interaction Detector (CHAID) approach, and we describe the design-based asymptotic properties of the resulting “survey CHAID” (sCHAID) method. To facilitate the practical use of sCHAID, we incorporate a Rao-Scott correction into the splitting criterion, accounting for the survey design. Using data from the U.S. American Community Survey, we illustrate the use of the method and evaluate its performance through comparisons with existing weighted and unweighted algorithms.

**Key Words:** Chi-square test; Recursive partitioning; Response propensity.

## 1. Introduction

Accounting for unit nonresponse is a crucial component of survey estimation and a key step in the creation of survey weights. Most commonly used approaches rely on postulating a response propensity model and creating weights that are inverses of estimated response propensities. The propensity models can be either explicitly stated or only implicitly, through the specification of covariates that are thought to be related to the response propensity. When the models are explicit, they can be either parametric, with the logistic regression model a common choice, or nonparametric, and the estimated response propensities are obtained as regression model predictions. The implicit approach does not specify a model and instead relies on calibration methods to obtain weights that account for the design, the response mechanism and population-level knowledge in a combined adjustment. Because the goal is almost always to create weights that can be applied to all the survey variables, all these approaches rely on the missing-at-random (MAR) assumption. We refer Brick (2013) for an overview of the different frameworks for obtaining response propensity weight adjustments.

Dividing the sample into weighting cells and applying an equal-propensity adjustment within each cell is a common method of obtaining weight adjustments. That method can be justified under multiple explicit models or as post-stratification under the implicit model calibration approach. The *response homogeneity group* (RHG) model (see Särndal, Swensson and Wretman, 1992, Chapter 15.6) assumes that the underlying propensity model consists of cells in which sampling units respond with equal (but unknown) probability.

---

1. Jean D. Opsomer and Minsun K. Riddles, Westat, Inc., 1600 Research Blvd., Rockville, Maryland, 14850, U.S.A. E-mail: jeanopsomer@westat.com.

The cells can be assumed known, for instance by using demographic post-strata as RHG cells, or can be determined based on the sample. The latter is attractive if the number of potential covariates along which to define weighting cells is large relative to the sample size. As will be made more precise in later sections, we will assume here that the responses follow the RHG model, with weighting adjustment cells to be selected by the Chi-square Automatic Interaction Detection (CHAID) algorithm, a popular type of recursive partitioning. Other authors have relied on the RHG model and proposed alternative recursive partitioning methods to create the adjustment cells, see for instance Phipps and Toth (2012).

As an alternative to the RHG model, a parametric functional form such as a logistic regression can be assumed for the response propensity model. However, even in that case, the weight adjustments are often applied as equal adjustments within cells (Little, 1986). This is implemented by dividing the model-predicted propensities into cell and averaging the predictions within each cell to obtain a single weight adjustment. Since these averaged predictions are now estimating an approximation of the true response function unlike in the RHG model, approximation bias issues involving the number and size of the cells need to be addressed when implementing this approach. The use of equal quantiles of the predicted propensities to define the cells is a common approach, see for instance Eltinge and Yansaneh (1997).

CHAID is a recursive partitioning algorithm originally proposed by Kass (1980) and has been used in a wide variety of fields as a classification and data mining method. In the survey context, which is our interest here, it is often used in the construction of weighting cells for nonresponse adjustments. The usual approach consists of fitting a classification tree to the unweighted dataset using CHAID, with the nodes of the resulting tree defining the weighting cells. Following this, the nonresponse weighting adjustments are computed as inverses of the weighted response proportions in each cell. Hence, this method to construct the nonresponse adjustment is implemented as a hybrid of unweighted and weighted approaches. While it is generally found to work well, there is a concern that ignoring the design could lead to inaccurate determination of the weighting cells. Further, its hybrid nature makes it difficult to study or justify theoretically. The goals of this article are to introduce sCHAID, a fully design-based version of CHAID, and to study its asymptotic and practical properties.

The literature on recursive partitioning in the design-based context is limited. The most comparable methodology for recursive partitioning that accounts for the sampling design is that proposed by Toth and Eltinge (2011), which is based on the approach of Gordon and Ohlshen (1980) and is implemented in the R package *rpms* (Toth, 2017). Another recent approach is that of Beaumont, Bosa, Brennan, Charlebois and Chu (2024), who propose a design-based recursive partitioning method based on the Classification and Regression Trees (CART) of Breiman, Friedman, Olshen and Stone (1984) to create selection probability cells for weighting of nonprobability sample data. We will compare our approach to that of Toth and Eltinge (2011) in simulations.

The remainder of the article is structured as follows. In Section 2, we introduce the response propensity model and the proposed estimation approach. Section 3 describes the asymptotic design-based properties of sCHAID. In Section 4, the results of a simulation study comparing sCHAID and CHAID are shown.

## 2. Notation and description of sCHAID

Let  $I_i$  represent the sample membership indicator for  $i \in U$ , with  $U$  denoting the population of size  $N$ . The sample  $s \subset U$  is selected according to the sampling design  $p(s)$ , which determines distribution of the  $I_i$  and in particular the inclusion probabilities  $\pi_i = E(I_i)$ ,  $\pi_{ij} = E(I_i I_j)$ ,  $i, j \in U$ . The response indicators  $R_i$  are independent Bernoulli random variables with  $E(R_i) = p_i$ ,  $i \in U$ . We allow the  $p_i$  to depend on the sampling design but not on the realized sample  $s$ , a common setting for response propensity modeling (see e.g. Shao and Steel, 1999). While the independence of the response indicators is commonly assumed in the development of nonresponse adjustments, it is not always warranted, for instance when there is a notable interviewer effect on unit-level nonresponse in personal-visit or telephone surveys. For simplicity, we will continue to restrict our attention to the independent case.

Under the RHG model,

$$p_i = \sum_{g=1}^G I_{\{i \in U_g^*\}} P_g^*, \quad (2.1)$$

where  $I_{\mathcal{D}} = 1$  if event  $\mathcal{D}$  is true and 0 otherwise, and the non-overlapping response homogeneity groups  $U_g^*$  of size  $N_g^*$  are defined by the intersections of  $K$  categorical (or “categorized”, if originally continuous) auxiliary variables  $X_k, k = 1, \dots, K$ . The portion of the sample  $s$  that falls in  $U_g^*$  is denoted  $s_g^* = s \cap U_g^*$ .

If  $G$  is modest relative to the sample size  $n$ , it would be reasonable to simply use the groups defined by the intersections of the auxiliary variables as adjustment cells. We are interested in the setting in which the number of cells obtained in this manner is too large to be practical, so that choosing a smaller number of cells is of interest to improve the stability of the survey weights. We will use sCHAID for this purpose. The algorithm will be based on a sequence of “sample splits” according to the values of categorical variables defined on the population units, so we first define notation for the relevant subsets here.

Let  $X_{ki}, k = 1, \dots, K$  represent the values of the categorical variables for sampling unit  $i$ , taking values  $1, \dots, L_k$  (possibly after converting categories to integers). For simplicity of notation, we will assume that  $L_k = L$  for all  $k$ , but the case with unequal numbers of values is readily generalized to. For each variable  $X_k$ , we define the subsets  $U_{kl}^*, l = 1, \dots, L$  of size  $N_{kl}^*$ , corresponding to the set of units with  $X_{ki} = l$ . The  $U_{kl}^*$  are non-overlapping for fixed  $k$ . The  $G = L^K$  response homogeneity groups  $U_g^*$  are obtained as intersections of the subsets  $U_{kl}^*$  for  $k = 1, \dots, K$  and  $l = 1, \dots, L$ . We write  $A_{kl}^*$  for the set of indices  $g$  such that  $U_g^* \subseteq U_{kl}^*$ , and we note that  $\sum_{g \in A_{kl}^*} N_g^* = N_{kl}^*$ . In other words,  $A_{kl}^*$  defines a “slice” among the response homogeneity groups corresponding to value  $l$  of categorical variable  $X_k$ . Analogously, we write  $s_{kl}^*$  for the portion of  $s$  that falls in  $U_{kl}^*$ .

The sample splits are chosen by comparing the average response propensity in sample subsets defined by the auxiliary variables. Since the true propensities are unknown, this is implemented in sCHAID (and CHAID) by performing statistical tests based on the estimated response propensities (see below). If  $X_k$  is an ordinal variable (or a categorized continuous variable), we wish to compare the average propensity in  $U_{kl}^*$  with  $l$  less than or equal to a target category, to the average propensity in the  $U_{kl}^*$  with  $l$  larger than that target category. In other words, a split at  $l$  for an ordinal variable  $X_k$  will define two subsets

$$U_{kl} = \cup_{l' \leq l} U_{kl'}^* \quad \text{and} \quad U_{kl^c} = \cup_{l' > l} U_{kl'}^*.$$

Similarly for a purely categorical variable, a split at  $l$  defines two subsets

$$U_{kl} = U_{kl}^* \quad \text{and} \quad U_{kl^c} = \cup_{l' \neq l} U_{kl'}^*.$$

In what follows, we will ignore the distinction between the types of auxiliary variables and use the  $(U_{kl}, U_{kl^c})$  notation when discussing the population subsets being evaluated. Analogously as above, we define the sets of indices  $A_{kl}, A_{kl^c}$ , population subset sizes  $N_{kl}, N_{kl^c}$  and samples  $s_{kl}, s_{kl^c}$  corresponding to  $U_{kl}, U_{kl^c}$ , respectively.

To clarify this population structure, Table 2.1 shows an example with  $K = 2$  ordinal variables, each with  $L = 5$  categories. In this case, the population  $U$  is partitioned into  $G = 25$  subsets  $U_g^*$ , with each  $g$  corresponding to a unique combination of values for  $(X_1, X_2)$  but otherwise arbitrary. We can define the subsets  $U_{1l}^*$  as the columns of the grid in Table 2.1 or equivalently as a set of 5 indices  $A_{1l}^* \subset \{1, \dots, 25\}$ , and similarly, the subsets  $U_{2l}^*$  as the rows of the grid with corresponding index set  $A_{2l}^*$ . For instance,  $A_{12}^* = \{2, 7, 12, 17, 22\}$  is the index set for  $U_{12}^* = \{i : X_{1i} = 2\}$ . Finally, we can define a sample split for  $X_1$  at  $l$  as the subsets  $U_{1l} = \{i : X_{1i} \leq l\}$  and  $U_{1l^c} = \{i : X_{1i} > l\}$ , with index sets  $A_{1l}, A_{1l^c}$  defined as the unions of the corresponding  $A_{1l}^*$ , and similarly for a sample split on  $X_2$ . For instance, the split for  $X_2$  at  $l = 2$  results in the subsets  $U_{22}$  with  $A_{22} = \{1, \dots, 10\}$  and  $U_{22^c}$  with  $A_{22^c} = \{11, \dots, 25\}$ .

**Table 2.1**  
Example of RHG population structure for  $K = 2, L = 5$

(k,l) =	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	
$X_2 \leq 2 \left\{ \begin{array}{l} (2,1) \\ (2,2) \end{array} \right.$	1	2	3	4	5	$\left. \begin{array}{l} \\ \end{array} \right\} U_{22}$
	6	7	8	9	10	
$X_2 > 2 \left\{ \begin{array}{l} (2,3) \\ (2,4) \\ (2,5) \end{array} \right.$	11	12	13	14	15	$\left. \begin{array}{l} \\ \\ \end{array} \right\} U_{22^c}$
	16	17	18	19	20	
	21	22	23	24	25	

Considering a particular  $U_{kl}$ , we define the average response propensities

$$P_{kl} = \frac{\sum_{i \in U_{kl}} p_i}{N_{kl}}, \tag{2.2}$$

which can also be written as a weighted average of the  $P_g^*, g \in A_{kl}$ ,

$$P_{kl} = \frac{\sum_{g \in A_{kl}} N_g^* P_g^*}{N_{kl}}.$$

In the investigation of the asymptotic properties of sCHAID (see Section 3),  $N$  will increase to infinity. Even though population quantities such as  $P_{kl}$  therefore depend on  $N$ , we will not write this explicitly.

We define the Horvitz-Thompson (HT) estimator of  $P_g^*$ , the response probability in  $U_g^*$  as

$$\tilde{R}_g^* = \frac{\sum_{s_g^*} \frac{1}{\pi_i} R_i}{N_g^*}, \quad (2.3)$$

and denote its conditional expectation with respect to the sampling design as

$$R_g^* = \frac{\sum_{U_g^*} R_i}{N_g^*}.$$

The survey estimator of the average response probability in subset  $U_{kl}$  can be defined as either a HT-type,

$$\tilde{P}_{kl} = \frac{\sum_{s_{kl}} \frac{1}{\pi_i} R_i}{N_{kl}} = \frac{\sum_{g \in A_{kl}} N_g^* \tilde{R}_g^*}{N_{kl}}$$

or a Hájek (HA)-type estimator,

$$\hat{P}_{kl} = \frac{\sum_{s_{kl}} \frac{1}{\pi_i} R_i}{\tilde{N}_{kl}} = \frac{\sum_{g \in A_{kl}} N_g^* \tilde{R}_g^*}{\tilde{N}_{kl}} \quad (2.4)$$

with  $\tilde{N}_{kl} = \sum_{g \in A_{kl}} \tilde{N}_g^*$  and  $\tilde{N}_g^* = \sum_{s_g^*} 1 / \pi_i$ . Both are functions of the  $\tilde{R}_g^*$  in (2.3), which will be useful when we study the theoretical properties of the estimators, but is not used explicitly when computing estimates. We will focus on the HA estimator in (2.4) in the selection of the sCHAID splits, because it does not require knowledge of the  $N_{kl}$  and it is generally considered more efficient than the HT estimator.

While CHAID relies on a  $\chi^2$  test for *iid* observations to determine whether  $P_{kl}$  and  $P_{kl^c}$  are different, we replace this by a Wald test using  $\hat{P}_{kl}$  and  $\hat{P}_{kl^c}$ , so that the sampling design can be accounted for. We only consider the Wald test in the theoretical investigation here, but in practice it can be replaced by one of the approximations from Rao and Scott (1981), as implemented in several survey packages. The test statistic is

$$\hat{W}_{kl} = n \frac{(\hat{P}_{kl} - \hat{P}_{kl^c})^2}{\hat{V}_{kl}} \quad (2.5)$$

with  $\hat{V}_{kl}$  a scaled design-based variance estimator to be specified in the next section. The  $p$ -value for the test is

$$\hat{q}_{kl} = \Pr(\chi_1^2 > \hat{W}_{kl}). \quad (2.6)$$

In sCHAID, the test statistic (2.5) will be computed not only on two groups within the overall sample, but also within subsamples obtained during prior iterations of the sCHAID algorithm. Let  $s^0$  represent such a subsample. In that case, the sample groups  $s_{kl}$  and its complement  $s_{kl^c}$  are understood to be within  $s^0$  only, and the test statistic is likewise only computed on  $s^0$ . For simplicity, we will not make explicit which subsamples are used and continue referring to  $s_{kl}$  and  $s_{kl^c}$ .

CHAID constructs the classification tree by performing a set of statistical tests at each step to select and split one of the available tree nodes into smaller nodes. Many variants of the algorithm exist, depending on the significance levels used to determine splits, on the allowable splits, on the algorithm stopping criteria, etc. Most of these can be implemented in the sCHAID context as well. In order to be able to study the proposed method asymptotically, however, we consider a streamlined version of the algorithm here. Specifically, the proposed sCHAID algorithm proceeds to divide the sample into subsamples (“nodes”) as follows:

- Step 1: the sample consists of a single node.
- Step  $r$ : the number of nodes increases from  $r-1$  to  $r$  by performing a binary split on one of the existing nodes, with the splits respecting the ordinal or categorical nature of the auxiliary variables. The eligible splits in step  $r$  consist of all splits that result in further divisions within the nodes, including splits on variables already used at prior steps. The  $p$ -values of the Wald tests  $\widehat{W}_{kl}$  on all eligible splits are computed, and only those with  $p$ -values less than a predetermined value  $\alpha$  are considered further. Among those, the one with the smallest  $p$ -value is selected for splitting. The split consists of classifying the node sample units into two sub-nodes, corresponding to whether they belong to  $U_{kl}$  or  $U_{kl^c}$  for the selected test. If none of the  $p$ -values are less than  $\alpha$ , the algorithm stops.
- Step  $R$ : the algorithm stops when  $R$  nodes have been created, unless it stopped earlier due to lack of eligible splits.

The outcome of the sCHAID algorithm is a set of at most  $R$  nodes, which are uniquely determined by the sequence of splits. These splits, in turn, are determined by the values of the auxiliary variables that correspond to  $\min_{kl} \hat{q}_{kl}$  (or equivalently, to  $\max_{kl} \widehat{W}_{kl}$ ) at each step, where it is understood that splits at later steps are conditional on those performed earlier. We will use the notation  $\widehat{T}_s$  to denote the sCHAID outcome for sample  $s$ , expressible in any of these forms, and develop its theoretical properties in the next section.

### 3. Theoretical results

We will study the properties of sCHAID under a joint design and model framework, using the reverse approach (Shao and Steel, 1999). Under this approach, the respondents are first drawn at the population level as realizations of  $N$  independent Bernoulli random variables with expectation  $p_i, i=1, \dots, N$ , as defined in (2.1). Then, the sample  $s$  is drawn from  $U$  according to the specified sampling design  $p(s)$ , independently from the realized population response status. The finite population is embedded in a sequence indexed by  $N$ , and we let  $N \rightarrow \infty$ . The sampling design  $p(s)$  depends on the population and hence also changes with  $N$ .

In the following assumptions, we state sufficient conditions under which we will prove theoretical results for the proposed sCHAID method.

- A.1.** *The number of response homogeneity groups  $G$  is fixed and there exists  $\lambda_1 > 0$  such that  $N_g^* / N > \lambda_1$  for  $g = 1, \dots, G$ . There exists  $0 < \lambda_2 < \lambda_3 < 1$  such that  $\lambda_2 < P_g^* < \lambda_3$  for  $g = 1, \dots, G$ .*
- A.2.** *For any combination of groups  $A_{kl}, A_{kl^c}$  that are eligible to appear among those evaluated by the sCHAID algorithm, no two values  $(P_{kl} - P_{kl^c})^2 / V_{kl}$  are equal to each other unless  $P_{kl} - P_{kl^c} = 0$ . The number of eligible splits across the whole population is at least  $R$ .*
- A.3.** *The sampling design  $p(s)$  is such that there exists a number  $N_0$  so that, for all  $N > N_0$ ,  $n_g^* \geq 1$  for  $g = 1, \dots, G$  with probability 1. The sample size  $n$  is non-random and satisfies  $n / N \rightarrow \lambda_4$ , with  $\lambda_4 \in (0, 1)$ .*
- A.4.** *For any bounded, non-random variable  $z_i$ , the vector of HT estimators  $\tilde{\mathbf{Z}}^* = (\tilde{z}_1^*, \dots, \tilde{z}_G^*)^T$  with  $\tilde{z}_g^* = \sum_{s_g} \frac{z_i}{\pi_i} / N_g^*$  has the following asymptotic distribution:*

$$\sqrt{n} \left( \tilde{\mathbf{Z}}^* - \mathbf{E}(\tilde{\mathbf{Z}}^*) \right) \Rightarrow \mathcal{N}(\mathbf{0}, \mathbf{V}_z^*)$$
*with respect to the sequence of sampling designs, for some matrix  $\mathbf{V}_z^*$ .*
- A.5.** *The inclusion probabilities satisfy  $\pi_i \geq \lambda_5 > 0$ ,  $\pi_{ij} \geq \lambda_6 > 0$  for all  $i, j \in U$  for some constants  $\lambda_5, \lambda_6$ . Also,*

$$\lim_{N \rightarrow \infty} n \max_{i, j \in U; i \neq j} |\Delta_{ij}| < \infty$$

with  $\Delta_{ij} = \pi_{ij} - \pi_i \pi_j$ , and

$$\lim_{N \rightarrow \infty} \max_{i, j, i', j' \in U; i \neq j, i' \neq j'} E \left( (I_i I_j - \pi_{ij})(I_{i'} I_{j'} - \pi_{i'j'}) \right) = 0.$$

Assumption A1 ensures that the response homogeneity group model in (2.1) remains valid and fixed in the sequence of populations, and that the response probability is bounded away from 0 and 1 everywhere. The former assumption is made to simplify the theory and could be relaxed by allowing  $G$  to grow with  $N$  at a suitably slow rate. The latter assumption is made to avoid degenerate variance components. Assumption A2 is made to avoid ties among the split selections, which could be handled at the cost of more complicated notation and results, and to ensure that the probability that the sCHAID algorithm ends before  $R$  steps due to lack of eligible splits goes to zero. Assumption A3 specifies regularity conditions on the sampling design, preventing the appearance of empty domains for sufficiently large  $N$ , and ensures that the sampling fraction does not become negligible. Assumption A4 states that the sampling design has a Central Limit Theorem for HT estimators. This is a common assumption in asymptotic investigations of survey estimators, because specific designs have different conditions to ensure the normality of their associated estimators, while for others the asymptotic normality is used for inference in practice but has not been formally established in the literature. Finally, the bounds and rate conditions on the inclusion

probabilities of different orders in assumption A5 are made to ensure existence of the HT estimator and the unbiased estimator of its design variance, which is of order  $O(n^{-1})$  and can be consistently estimated.

**Lemma 1.** *Under assumptions A1–A5, the vector  $\tilde{\mathbf{R}}^* = (\tilde{R}_1^*, \dots, \tilde{R}_G^*)^T$  has the following asymptotic distribution with respect to the response mechanism and sequence of sampling designs:*

$$\sqrt{n} (\tilde{\mathbf{R}}^* - \mathbf{P}^*) \Rightarrow \mathcal{N}(\mathbf{0}, \mathbf{V}_R^*) \quad (3.1)$$

with  $\mathbf{P}^* = (P_1^*, \dots, P_G^*)^T$  and  $\mathbf{V}_R^*$  the matrix with element  $gg'$  equal to

$$V_{R,gg'}^* = \frac{n}{N_g^* N_{g'}^*} \sum_{U_g^*} \sum_{U_{g'}^*} \Delta_{ij} \frac{p_i}{\pi_i} \frac{p_j}{\pi_j} + I_{\{g=g'\}} \frac{n}{N_g^{*2}} \sum_{U_g^*} \frac{1}{\pi_i} p_i (1-p_i).$$

The estimator

$$\tilde{V}_{R,gg'}^* = \frac{n}{N_g^* N_{g'}^*} \sum_{s_g^*} \sum_{s_{g'}^*} \frac{\Delta_{ij}}{\pi_{ij}} \frac{R_i}{\pi_i} \frac{R_j}{\pi_j} + I_{\{g=g'\}} \frac{n}{N_g^{*2}} \tilde{N}_g^* \tilde{R}_g^* (1 - \tilde{R}_g^*) \quad (3.2)$$

is consistent for  $V_{R,gg'}^*$  with respect to the response mechanism and sequence of sampling designs.

*Proof of Lemma 1:* We apply Theorem 1.3.6 from Fuller (2009) to obtain the asymptotic distribution under the combination of the response mechanism and sampling design. We first consider the population average  $R_g^* = \sum_{U_g^*} R_i / N_g^*$ , the conditional expectation of  $\tilde{R}_g^*$  with respect to the sampling design. Because  $R_g^*$  is an average of  $N_g^*$  iid Bernoulli variables, it satisfies a Liapounov condition (e.g. Fuller, 1996, Theorem 5.3.2) and is asymptotically normally distributed. Letting  $\mathcal{R}_N = (R_1, \dots, R_N)^T$ , we know that  $E(\tilde{R}_g^* | \mathcal{R}_N) = R_g^*$  and that  $\tilde{R}_g^*$  has a conditional asymptotic normal distribution by A4 with respect to the sequence of sampling designs, which holds for any  $\mathcal{R}_N$ . We are therefore allowed to combine the normal distributions into the overall asymptotic distribution given in the statement of the lemma. The asymptotic variance is obtained by standard conditional moment calculations.

To show that the proposed estimator for the elements of  $\mathbf{V}_R^*$  is consistent, we first consider the first term,

$$\tilde{V}_{R1,gg'}^* = \frac{n}{N_g^* N_{g'}^*} \sum_{s_g^*} \sum_{s_{g'}^*} \frac{\Delta_{ij}}{\pi_{ij}} \frac{R_i}{\pi_i} \frac{R_j}{\pi_j}. \quad (3.3)$$

Conditioning on the realized  $R_i$  in the population, we find

$$E(\tilde{V}_{R1,gg'}^* | \mathcal{R}_N) = \frac{n}{N_g^* N_{g'}^*} \sum_{U_g^*} \sum_{U_{g'}^*} \Delta_{ij} \frac{R_i}{\pi_i} \frac{R_j}{\pi_j}$$

and

$$\begin{aligned} \text{Var}(\tilde{V}_{R1,gg'}^* | \mathcal{R}_N) &= \frac{n^2}{N_g^{*2} N_{g'}^{*2}} \sum_{U_g^*} \sum_{U_{g'}^*} \sum_{U_{g'}^*} \sum_{U_g^*} \Delta_{ij} \Delta_{i'j'} \frac{R_i}{\pi_i} \frac{R_j}{\pi_j} \frac{R_{i'}}{\pi_{i'}} \frac{R_{j'}}{\pi_{j'}} \\ &\quad \times E((I_i I_j - \pi_{ij})(I_{i'} I_{j'} - \pi_{i'j'})). \end{aligned}$$

Hence,

$$E(\tilde{V}_{R1,gg'}^*) = \frac{n}{N_g^* N_{g'}^*} \sum_{U_g^*} \sum_{U_{g'}^*} (\pi_{ij} - \pi_i \pi_j) \frac{p_i}{\pi_i} \frac{p_j}{\pi_j} + I_{\{g=g'\}} \frac{n}{N_g^*} \sum_{U_g^*} \left( \frac{1}{\pi_i} - 1 \right) p_i (1 - p_i).$$

For  $\text{Var}(\tilde{V}_{R1,gg'}^*)$ , we note that  $\text{Var}(\tilde{V}_{R1,gg'}^* | \mathcal{R}_N) \rightarrow 0$  uniformly by applying the bounds on the inclusion probabilities of the different orders in A5 and using  $R_i \leq 1$ . By direct calculation, we also find

$$\text{Var}(E(\tilde{V}_{R1,gg'}^* | \mathcal{R}_N)) = \frac{n^2}{N_g^{*2} N_{g'}^{*2}} \sum_{U_g^*} \sum_{U_{g'}^*} \sum_{U_{g'}^*} \sum_{U_g^*} \Delta_{ij} \Delta_{i'j'} \frac{\text{Cov}(R_i R_j, R_{i'} R_{j'})}{\pi_i \pi_j \pi_{i'} \pi_{j'}},$$

which converges to 0 by the bounds in A5 again and the fact that  $\text{Cov}(R_i R_j, R_{i'} R_{j'}) = 0$  unless  $(i, j)$  overlaps with  $(i', j')$ . Hence,  $\text{Var}(\tilde{V}_{R1,gg'}^*) \rightarrow 0$  and  $\tilde{V}_{R1,gg'}^*$  converges to (3.3).

For second term of  $\tilde{V}_{R,gg'}^*$ , we can show directly that

$$\frac{n}{N_g^{*2}} \tilde{N}_g^* \tilde{R}_g^* (1 - \tilde{R}_g^*) = \frac{n}{N_g^*} P_g^* (1 - P_g^*) + o_p(1) \tag{3.4}$$

by linearization. Combining this with (3.3), the result follows.

**Lemma 2.** Under assumptions A1–A5, the statistic

$$\hat{W}_{kl} = n \frac{(\hat{P}_{kl} - \hat{P}_{kl^c})^2}{\hat{V}_{kl}}$$

with

$$\begin{aligned} \hat{V}_{kl} &= \frac{n}{N_{kl}^2} \sum_{s_{kl}} \sum_{\pi_{ij}} \frac{\Delta_{ij}}{\pi_{ij}} \frac{R_i - P_{kl}}{\pi_i} \frac{R_j - P_{kl}}{\pi_j} \\ &+ \frac{n}{N_{kl^c}^2} \sum_{s_{kl^c}} \sum_{\pi_{ij}} \frac{\Delta_{ij}}{\pi_{ij}} \frac{R_i - P_{kl^c}}{\pi_i} \frac{R_j - P_{kl^c}}{\pi_j} \\ &- \frac{2n}{N_{kl} N_{kl^c}} \sum_{s_{kl}} \sum_{s_{kl^c}} \sum_{\pi_{ij}} \frac{\Delta_{ij}}{\pi_{ij}} \frac{R_i - P_{kl}}{\pi_i} \frac{R_j - P_{kl^c}}{\pi_j} \\ &+ \frac{n}{N_{kl}^2} \sum_{g \in A_{kl}} \tilde{N}_g^* \tilde{R}_g^* (1 - \tilde{R}_g^*) + \frac{n}{N_{kl^c}^2} \sum_{g \in A_{kl^c}} \tilde{N}_g^* \tilde{R}_g^* (1 - \tilde{R}_g^*) \end{aligned}$$

is asymptotically distributed as a non-central  $\chi_1^2$ , with non-centrality parameter  $n\lambda_{kl} = n(P_{kl} - P_{kl^c})^2 / V_{kl}$  and

$$\begin{aligned} V_{kl} &= \frac{n}{N_{kl}^2} \sum_{U_{kl}} \sum_{\Delta_{ij}} \frac{p_i - P_{kl}}{\pi_i} \frac{p_j - P_{kl}}{\pi_j} + \frac{n}{N_{kl^c}^2} \sum_{U_{kl^c}} \sum_{\Delta_{ij}} \frac{p_i - P_{kl^c}}{\pi_i} \frac{p_j - P_{kl^c}}{\pi_j} \\ &- \frac{2n}{N_{kl} N_{kl^c}} \sum_{U_{kl}} \sum_{U_{kl^c}} \sum_{\Delta_{ij}} \frac{p_i - P_{kl}}{\pi_i} \frac{p_j - P_{kl^c}}{\pi_j} \\ &+ \frac{n}{N_{kl}^2} \sum_{U_{kl}} \frac{1}{\pi_i} p_i (1 - p_i) + \frac{n}{N_{kl^c}^2} \sum_{U_{kl^c}} \frac{1}{\pi_i} p_i (1 - p_i). \end{aligned}$$

*Proof of Lemma 2:* Using linearization, we obtain the following approximation

$$\widehat{P}_{kl} - \widehat{P}_{kl^c} - (P_{kl} - P_{kl^c}) = \frac{1}{N_{kl}} \sum_{g=1}^G (N_g^* \widetilde{R}_g^* - P_{kl} \widetilde{N}_g^*) (I_{\{g \in A_{kl}\}} - I_{\{g \in A_{kl^c}\}}) + O_p(n^{-1}). \quad (3.5)$$

The asymptotic normality of  $\widehat{P}_{kl} - \widehat{P}_{kl^c}$  directly follows from A4 and Lemma 1. Starting from the approximation in (3.5), straightforward moment calculations show that the asymptotic distribution of  $\widehat{P}_{kl} - \widehat{P}_{kl^c}$  has expectation  $P_{kl} - P_{kl^c}$ , and variance  $V_{kl}$ . Hence, applying Theorem 5.2.4 of Fuller (1996), we conclude that  $n(\widehat{P}_{kl} - \widehat{P}_{kl^c})^2 / V_{kl}$  has an asymptotic non-central  $\chi_1^2$  distribution, with non-centrality parameter equal to  $n(P_{kl} - P_{kl^c})^2 / V_{kl}$ . The distribution of  $\widehat{W}_{kl}$  follows from Corollary 5.2.6.1 in Fuller (1996) once we replace  $V_{kl}$  by a consistent estimator. Using the same approach as in the proof of Lemma 1, the consistency of  $\widehat{V}_{kl}$  for  $V_{kl}$  can again be shown.

**Theorem 1.** *Under assumptions A1–A5, when evaluating the set of eligible splits at each step of the sCHAID algorithm described in Section 2, the following two statements hold:*

- (i) *For any predetermined  $p$ -value cut-off  $\alpha > 0$ , all splits for which  $P_{kl} - P_{kl^c} \neq 0$  are available for consideration with probability going to 1, and*
- (ii) *By selecting the split with the smallest  $p$ -value  $\hat{q}_{kl}$ , the unique split corresponding to  $\max(P_{kl} - P_{kl^c})^2 / V_{kl}$  will be selected with probability going to 1.*

*Proof of Theorem 1:* To prove (i), we consider the statistic  $\widehat{W}_{kl}$  and associated  $p$ -value  $\hat{q}_{kl}$ . Based on the properties of the non-central  $\chi^2$  distribution, we know that the asymptotic distribution of  $\widehat{W}_{kl}$  has mean  $\lambda_{kl} + 1 = O(n)$  and variance  $2\lambda_{kl} + 4 = O(n)$ , unless  $P_{kl} - P_{kl^c} = 0$ . Hence,  $\widehat{W}_{kl} / n$  converges in probability to a constant and  $\hat{q}_{kl}$  converges in probability to 0 whenever  $P_{kl} - P_{kl^c} \neq 0$ , so that  $\hat{q}_{kl} < \alpha$  with probability going to 1.

For (ii), we consider two possible splits with test statistics  $\widehat{W}_{kl}, \widehat{W}_{kl'}$  and associated  $p$ -value  $\hat{q}_{kl}, \hat{q}_{kl'}$ , with  $P_{kl} - P_{kl^c} \neq 0$  and  $P_{kl'} - P_{kl'^c} \neq 0$ . The test with the smallest  $p$ -value is to be selected for splitting, which we can write as an indicator  $I_{\{\hat{q}_{kl} - \hat{q}_{kl'} < 0\}}$ : if this indicator is 1, then the split on  $X_k = l$  is selected and if this indicator is 0, the split on  $X_k = l'$  is selected. Based on the moments of the non-central  $\chi^2$  again, we have

$$\begin{aligned} I_{\{\hat{q}_{kl} - \hat{q}_{kl'} < 0\}} &= I_{\{\widehat{W}_{kl}/n - \widehat{W}_{kl'}/n > 0\}} \\ &= I_{\{\lambda_{kl} - \lambda_{kl'} > 0\}} + O_p(n^{-1/2}). \end{aligned}$$

This result holds for all pairwise comparisons being evaluated concurrently, so that we conclude that the split with the largest value of  $\lambda_{kl} = n(P_{kl} - P_{kl^c})^2 / V_{kl}$  is selected with probability going to 1.

Theorem 1 describes the asymptotic behavior of the Wald test statistics and the resulting split selection at each step. When these results are considered for the sequence of splits evaluated during the sCHAID algorithm, it is possible to describe the asymptotic behavior of the resulting tree. This is stated in the following corollary, which shows that the algorithm has a well-defined probability limit in the population.

**Corollary 1.** *The sample tree  $\widehat{T}_s$  obtained by the sCHAID algorithm described in Section 2 consists of the sequence of length  $\leq R$  of node splits applied on the sample  $s$ . The sample tree  $\widehat{T}_s$  converges to a population tree  $T_U$  with probability going to 1. The tree  $T_U$  consists of the sequence of length  $R$  of splits performed at the set of values  $X_k = l$  that maximize*

$$\frac{(P_{kl} - P_{kl^c})^2}{V_{kl}}$$

*over all the eligible splits at each step. The tree  $T_U$  is unique under the stated assumptions.*

The variance estimator defined in Lemma 2 is consistent, but its last two terms require the use of the propensity estimators  $\widetilde{R}_g^*$  for the finest level of response homogeneity groups. If  $G$  is large, these estimators might be highly variable or even result in empty cells in practice. This is in contrast to the remaining terms in  $\widehat{V}_{kl}$ , which can all be computed directly for the two sides of the node split being evaluated. Therefore, we propose two alternative approaches to obtain a denominator for the test statistics.

First, the terms involving the  $\widetilde{R}_g^*$  could simply be omitted. This would greatly simplify the computations, because the remaining terms are straightforward to obtain using standard survey software implementations of  $\chi^2$  tests, for instance as provided in the SURVEYFREQ procedure in SAS and the survey package in R. This would be somewhat analogous to only computing the PSU-level variance estimator term in the multi-stage sampling context. As in that context, these terms are expected to represent only a small fraction of the overall variance estimator when the sampling fraction is small. Second, the terms could be replaced by averaging over both sides of the node split, i.e. using

$$\frac{n}{N_{kl}} \widehat{P}_{kl} (1 - \widehat{P}_{kl}) + \frac{n}{N_{kl^c}} \widehat{P}_{kl^c} (1 - \widehat{P}_{kl^c}).$$

The latter option might be preferred if the sampling fraction is non-negligible. We will evaluate both options in the simulations below.

## 4. Simulation

To investigate the performance of the sCHAID algorithm, we performed a simulation study comparing the proposed sCHAID method to traditional CHAID and to rpms (Toth, 2017). All calculations were performed using R statistical software (R Core Team, 2021).

The household-level Public Use Microdata Sample (PUMS) data for the 2017-2021 American Community Survey (ACS) were used as a sampling frame. The frame was stratified by Census Division, the Public Use Microdata Areas (PUMAs) formed the primary sampling units (PSUs) and the secondary sampling units were households (HHs) excluding group quarters. From each of the nine Census Divisions, we draw a simple random sample of 10 PSUs, and a simple random sample of 100 households is selected in each sampled PSU, generating clustered samples of households with unequal probabilities. Table 4.1 presents the number of PUMAs and the average number of households per PUMA by Census Division in the frame.

**Table 4.1****Number of Public Use Microdata Areas and average number of households per Public Use Microdata Area by Census Division, 2017-2021 ACS Microdata Sample**

Census Division	Number of PUMAs	Number of sampled PUMAs	Average number of HHs per PUMA	Number of sampled HHs per PUMA
New England	109	10	2,946	100
Middle Atlantic	310	10	2,790	100
East North Central	339	10	3,026	100
West North Central	159	10	3,008	100
South Atlantic	455	10	2,950	100
East South Central	138	10	2,984	100
West South Central	294	10	2,599	100
Mountain	180	10	2,791	100
Pacific	367	10	2,642	100

Note: ACS = American Community Survey; HHs = households; PUMAs = Public Use Microdata Areas.

Two sets of 1,000 samples of 9,000 households were drawn for the simulation, corresponding to a low and high nonresponse scenario. The base weight of the selected households was set to the product of the ACS household-level weight and the reciprocal of the selection probability. The response status,  $R_i$  was generated for each household ( $i = 1, 2, \dots, 9,000$ ) from a Bernoulli distribution with a probability  $p_i$ , where

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \sum_{j=1}^8 \beta_j x_{j,i}.$$

The eight covariates,  $\mathbf{x} = (x_1, \dots, x_8)^T$  were Census Region (three indicator variables for the Midwest, South, and West regions), building type (one-family house detached home or not), tenure status (owned or not), presence of children in the household, health insurance coverage status, and property value. In order to mimic a realistic response mechanism, the parameters  $\boldsymbol{\beta} = (\beta_0, \dots, \beta_8)^T$  were obtained by fitting a model for early response to the PUMS dataset, with the response indicator set to 1 if the ACS response mode was Web or mail and 0 if the response mode was telephone or in-person (indicating that the respondent did not respond in previous attempts to contact them using the other modes). In the simulation, household  $i$  was considered a respondent if the realized response status  $r_i$  is 1, and a nonrespondent otherwise. The average of response rate across the two sets of 1,000 samples were approximately 74.5% and 31.7%, respectively. The first scenario corresponds to fitting the logistic model to the observed ACS response status described above, and the scenario with the 31.7% response rate was set by modifying the intercept  $\beta_0$ .

When implementing the different recursive partitioning methods, the generated response indicator was the dependent variable and twelve potential covariates were considered. These included the eight covariates in the response model above and four additional covariates that were not used: telephone service status, family type, number of persons in family, and access to internet. For the traditional CHAID method, we implemented two versions: one using the twelve covariates only, and one that adds the design variable (PUMA) and the base weight in the model. The first version is referred to as “CHAID”, while the second is referred to as “CHAID+design info”. The second version may be more comparable to sCHAID as it includes the design information within a model-based framework (Little and Vartivarian, 2003).

In each method, except for setting the p-value to 0.05, we did not adjust any further algorithm parameters to control the behavior of the resulting trees. The second-order (Satterthwaite) Rao-Scott chi-square statistic for two-way tables was employed in sCHAID (Rao and Scott, 1981). This statistic is the default for the “svychisq” function of the “survey” package (Lumley, 2024) in R. We also investigated using the variance estimator  $\hat{V}_{kl}$  from Lemma 2, applying the second simplification option suggested in Section 3 to account for the last two variance terms. Results obtained by adding these last two terms were nearly indistinguishable and are not shown here.

Table 4.2 shows the mean squared prediction error (MSPE) for the response status, the proportion of trees built with all eight predictors used to generate response status, and the proportion of trees built with at least one of the other four covariates that were not used in the response mechanism. Table 4.3 displays the empirical relative bias (RB) and the empirical relative root mean squared error (RRMSE) for the estimates of average household income, an ACS outcome variables. The estimates used the nonresponse adjusted weights, which were computed by dividing the base weight by the estimated response propensities obtained for each of the terminal tree nodes under the RHG model.

**Table 4.2**  
**Mean squared prediction error and proportions of trees built with all eight covariates and trees built with other covariates from 1,000 simulated samples**

Response Rates	Algorithm	MSPE (unweighted)	Proportion of trees built with all eight covariates	Proportion of trees built with other covariates
High 74.5%	CHAID	0.0021	0.99	0.11
	CHAID+design info	0.0022	0.99	0.12
	rpms	0.0043	0.95	0.07
	sCHAID	0.0032	0.99	0.06
Low 31.7%	CHAID	0.0026	0.99	0.12
	CHAID+design info	0.0025	0.99	0.13
	rpms	0.0044	0.96	0.07
	sCHAID	0.0031	0.98	0.07

Note: CHAID = Chi-square Automatic Interaction Detector; sCHAID = survey Chi-square Automatic Interaction Detector; MSPE = mean squared prediction error.

**Table 4.3**  
**Empirical relative bias and the empirical relative root mean squared error for household income by tree algorithm from 1,000 simulated samples**

Response Rates	Algorithm	Relative bias (%)	Relative root mean squared error (%)
High 74.5%	No Adjustment	-3.70	3.98
	CHAID	-0.20	1.68
	CHAID+design info	-0.19	1.70
	rpms	-0.33	1.59
	sCHAID	-0.21	1.50
Low 31.7%	No Adjustment	-6.62	7.57
	CHAID	-1.12	2.75
	CHAID+design info	-1.10	2.77
	rpms	-1.51	2.98
	sCHAID	-0.98	2.24

Note : CHAID = Chi-square Automatic Interaction Detector; sCHAID = survey Chi-square Automatic Interaction Detector.

The results in Table 4.2 show that all four methods performed well in identifying the variables in the response model, although rpms missed at least one variable slightly more often. The two versions of the traditional CHAID had the highest fraction of trees containing extraneous variables, which is not surprising given that they do not account for the reduction in effective sample size due to weighting and clustering in the tree building process. As a result, they tended to overfit more frequently than methods that account for the sampling design. When the estimated trees obtained from all four methods were used to adjust for nonresponse, the results in Table 4.3 show that all were effective in removing most of the bias of the resulting estimators, with only modest differences between them but a slight advantage in precision for the sCHAID-adjusted estimator.

## 5. Discussion

We have proposed sCHAID as an extension of CHAID that is more suitable for forming response adjustment cells in a design-based setting, by incorporating both design weighting and variance features such as clustering and stratification in its partitioning criterion. At the same time, it continues to be based on the general CHAID approach, including the use of  $\chi^2$  tests and  $p$ -values, so that current users of CHAID will be able to readily adapt to this modified recursive partitioning algorithm. We are in the process of developing an easy-to-use R package that integrates with the survey package and existing partitioning packages in R.

We have described sCHAID in the context of constructing nonresponse adjustments. However, recursive partitioning algorithms can also be used for fitting classification and regression trees more generally. The sCHAID algorithm can be used with minor modifications for classification and regression trees for survey data as well. Developing the theory for this would follow the approach of Toth and Eltinge (2011).

## References

- Beaumont, J.-F., Bosa, K. Brennan, A., Charlebois, J. and Chu, K. (2024). [Handling non-probability samples through inverse probability weighting with an application to Statistics Canada's crowdsourcing data](https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2024001/article/00004-eng.pdf). *Survey Methodology*, 50(1), 77-106. Paper available at <https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2024001/article/00004-eng.pdf>.
- Breiman, L., Friedman, J.H., Olshen, R. and Stone, C.J. (1984). *Classification and Regression Trees*. Belmont, CA: Wadsworth.
- Brick, J.M. (2013). Unit nonresponse and weighting adjustments: A critical review. *Journal of Official Statistics*, 29, 329-353.

- Eltinge, J.L., and Yansaneh, I.S. (1997). [Diagnostics for formation of nonresponse adjustment cells, with an application to income nonresponse in the U.S. Consumer Expenditure Survey](https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1997001/article/3103-eng.pdf). *Survey Methodology*, 23(1), 33-40. Paper available at <https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1997001/article/3103-eng.pdf>.
- Fuller, W.A. (1996). *Introduction to Statistical Time Series* (2 ed.). New York: John Wiley & Sons, Inc.
- Fuller, W.A. (2009). *Sampling Statistics*. Hoboken, NJ: John Wiley & Sons, Inc.
- Gordon, L., and Ohlshen, R. (1980). Nonparametric regression from recursive partitioning schemes. *Journal of Multivariate Analysis*, 10, 611-627.
- Kass, G. (1980). An exploratory technique for investigating large quantities of categorical data. *Journal of the Royal Statistical Society, Series C*, 29, 119-127.
- Little, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review*, 54, 139-157.
- Little, R.J.A., and Vartivarian, S. (2003). On weighting the rates in nonresponse weights. *Statistics in Medicine*, 9(22), 1589-1599.
- Lumley, T. (2024). survey: Analysis of complex survey samples. R package version 4.4.
- Phipps, P., and Toth, D. (2012). Analyzing establishment nonresponse using an interpretable regression tree model with linked administrative data. *Annals of Applied Statistics*, 6, 772-794.
- R Core Team (2021). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Rao, J.N.K., and Scott, A. (1981). The analysis of categorical data from complex sample surveys: Chi-squared tests for goodness of fit and independence in two-way tables. *Journal of the American Statistical Association*, 76, 221-230.
- Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- Shao, J., and Steel, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fractions. *Journal of the American Statistical Association*, 94, 254-265.
- Toth, D. (2017). rpms: An R package for modeling survey data with regression trees. US Bureau of Labor Statistics.
- Toth, D., and Eltinge, J.L. (2011). Building consistent regression trees from complex survey data. *Journal of the American Statistical Association*, 106, 1626-1636.