A Tree-Based Dual-Frame Estimation Approach for Combining Probability and Non-probability Samples

Chien-Min Huang and F. Jay Breidt
NORC at the University of Chicago, 55 East Monroe Street, Chicago, IL 60603

Abstract
We consider a dual-frame estimation technique for combining data from two sources, in which a probability sample representative of the population is combined with a non-probability sample with undercoverage and unknown inclusion probabilities. All covariates and responses of interest are observed on both the probability and the non-probability sample. The dual-frame estimator requires probability sampling weights for the non-probability sample and non-probability participation propensities for both samples. We estimate the propensities non-parametrically using a tree-structured model, which we also use to impute probability sampling weights. We assess the properties of the tree-based dual-frame estimator analytically. We also apply the technique to a simulation experiment built from a real study (Culture and Community in a Time of Crisis), evaluating its behavior with respect to estimation bias, effective sample size ratio, and confidence interval coverage.

Key Words: Nonparametric estimation, propensity model, variance estimation

1. Introduction

1.1 Combining probability and non-probability samples

A common problem in modern survey practice is the combination of a probability sample, which is often expensive and small but unbiased, with a non-probability sample, which is often cheap and large but biased. Typical features of a non-probability "sample" include incomplete and unknown coverage, because not all eligible population members may have a chance to be included, and unknown inclusion probabilities, because the participation mechanism is often not probabilistic and in any case is unknown. In the probability sample, we might also have some sparsity or undercoverage due to the incomplete frame. Natural goals are then to make maximal use of the non-probability sample to reduce variance and cost while minimizing any bias from the non-probability sample due to selection effects.

There is a growing literature in combining probability and non-probability samples. Wu (2022) gives a comprehensive review of making statistical inference for non-probability samples. Kim (2022) summarizes the techniques for combining data when the variable of interest is only observed in the non-probability sample. Ganesh et al. (2017) uses bivariate Fay-Herriot models for probability and non-probability direct estimates (explicitly modeling bias in the non-probability estimate) to develop raking targets for some key response variables on small domains as part of an overall calibration strategy. The mass imputation approach (Chen et al., 2020; Kim et al., 2021) is a model-based method that regresses response variables on covariates in the non-probability sample, then predicts ("imputes") the missing response variable using the observed covariates for every element in the probability sample. The sample matching approach (Rivers, 2007; Yang et al., 2018; Chen et al., 2020) assigns sampling weights from the probability sample to the non-probability sample by comparing covariates available in both samples using distance measures. The inverse weighting or quasi-randomization approach (Kim and Wang, 2019; Chen et al., 2020; Elliott and Valliant, 2017) estimates the propensities for the non-probability sample

by combining the probability and non-probability samples based on the missing at random assumption. Chu and Beaumont (2019) uses the regression tree to estimate propensity scores. The terminal nodes of the classification tree are regarded as homogeneous selection propensity classes. Doubly-robust estimators combine an estimated propensity and a regression model for the response on covariates (Chen et al., 2020; Kim and Wang, 2019; Valliant, 2020). The estimator is consistent if either model is correctly specified. Rafei et al. (2020) use nonparametric Bayesian Additive Regression Trees (BART) in the inverse weighting approach to protect against model misspecification. Rafei et al. (2022) extend BART to the doubly-robust method. Bayesian methods on supplementing small sample size of probability samples with non-probability samples are discussed in Nandram and Rao (2021), Sakshaug et al. (2019), Wiśniowski et al. (2020), and Salvatore et al. (2024). The information from the non-probability sample is incorporated in the prior distribution, and the inference is based on the posterior distribution.

The majority of the literature is built with the presumption of parametric models, conditional on the covariates selected in the model. Because parametric methods often suffer when a misspecified model is used, and because variable selection could be hard in a production environment, we propose a tree-based dual-frame approach that uses a non-parametric tree-structured model to model the propensities of the non-probability samples. The models address non-linear data patterns and are robust to model misspecification. The approach can seamlessly handle multiple scenarios of combining probability and non-probability samples.

1.2 Overview

We introduce notation for the dual-frame setting in Section 2.1. The dual-frame estimators are introduced in Section 2.2. Section 2.3 and Section 2.4 outline the tree-based approach. Variance estimators are described in Section 2.5. The Culture and Community in a Time of Crisis (CCTC) simulation platform is described in Section 3.1 and results comparing the dual-frame estimators for different sizes of non-probability samples are compared to probability sampling only in Section 3.2. A brief discussion follows in Section 4.

2. Methods

2.1 Notation

Consider a finite population $U = \{1, 2, \ldots, N\}$ from which two samples are selected: a probability sample $A \subset U$ with known, positive inclusion probabilities $\pi_k^A > 0$ for $k \in A$ and a non-probability sample $B \subset U$ with unknown selection probabilities. Let $A_k = 1$ if $k \in A$ and $A_k = 0$ otherwise, and $B_k = 1$ if $k \in B$ and $B_k = 0$ otherwise. Let n_A denote the A sample size and n_B denote the B sample size. Assume that $U = \cup_{h=1}^{H} L_h$, where the disjoint leaves L_h, $h = \{1, 2, \ldots, H\}$ partition the finite population. Assume that covariates $\boldsymbol{x}_k$ and responses y_k are available for $k \in A \cup B$.

We consider the scenario where the non-probability sample is selected from the complement of the probability sample or deduplicated from the B sample, so that $B \subset U \setminus A$ and $A \cap B = \emptyset$. This can occur, for example, if the B sample is obtained via respondent-driven sampling using randomly-selected seeds A (Huang and Breidt, 2023), or if the B sample is obtained via judgment sampling by the same field crew conducting A.

We assume Poisson sampling for the B sample, selected from the complement of the A sample. Then $\{B_k\}_{k \in U \setminus A}$ are independent random variables, with participation propensity (success probability)

$$\rho(\boldsymbol{x}_k) = \mathrm{P}\left[B_k = 1 \mid k \notin A\right].$$

The first-order inclusion probability for the B sample is then

$$\pi_k^B = \mathrm{P}\left[k \in B \mid k \in A\right]\mathrm{P}\left[k \in A\right] + \mathrm{P}\left[k \in B \mid k \notin A\right]\mathrm{P}\left[k \notin A\right] = 0 + \rho(\boldsymbol{x}_k)(1 - \pi_k^A).$$

In some cases, π_k^A might be known for all $k \in A \cup B$, but in practice it will often be unknown for $k \in B$. Further, the participation propensity $\rho(\boldsymbol{x}_k)$ is unknown and might be zero.

2.2 Dual-frame estimator

We apply a dual-frame approach (see, for example, Singh and Mecatti (2011)) to combine the information from the probability and non-probability samples for estimation of $\sum_{k\in U} y_k$. Dual-frame options for the two samples use either a *combined* sample of probability and non-probability elements, or keeps them *separate*. Separate dual-frame estimators typically rely on choice of some compositing parameter to use in a convex combination of the probability sample estimate and the non-probability sample estimate.

The combined dual-frame approach uses the combined sample $s = A \cup B$. Elements can enter the sample via two paths, probability or non-probability, with first-order inclusion probability $\pi_k = \mathrm{P}\left[k \in A\right] + \mathrm{P}\left[k \in B\right] - \mathrm{P}\left[k \in A \cap B\right] = \pi_k^A + (1 - \pi_k^A)\rho(\boldsymbol{x}_k)$. If π_k were known for all $k \in A \cup B$, then the unbiased combined estimator for $\sum_{k\in U} y_k$ would be

$$\widehat{T}_y = \sum_{k\in A\cup B} \frac{y_k}{\pi_k^A + \rho(\boldsymbol{x}_k)(1 - \pi_k^A)}. \tag{1}$$

There are several advantages of the combined estimator rather than the separate estimator. The first one is we do not need to choose a parameter in the convex combination of probability and non-probability samples. Second, the weights of the combined estimator (1) are stable by construction, since

$$1 \leq \frac{1}{\pi_k^A + \rho(\boldsymbol{x}_k)(1 - \pi_k^A)} \leq \frac{1}{\pi_k^A},$$

which are bounded above by $1/\pi_k^A$ even if $\rho(\boldsymbol{x}_k) = 0$.

The combined estimator requires estimates $\widehat{\rho}(\boldsymbol{x}_k)$ of the participation propensities, not only for the non-probability sample (required for all quasi-randomization approaches), but also for the probability sample. Further, the combined estimator may require imputation of $\{\pi_k^A\}_{k\in B}$, since these will often be unknown in practice for the non-probability sample. We address both issues with a tree-based approach.

2.3 Tree-based approach

We fit a weighted classification tree to the binary outcome of success $B_k = 1$ or failure $B_k = 0$ using the vector of covariates $\boldsymbol{x}_k$. We use the design weights $1/\pi_k^A$ for the probability sample and weights of one for the non-probability sample. The interpretation of the weights of one is that in the absence of other information, the non-probability sample elements represent only themselves.

Our tree-based approach assumes that the leaves of the tree are homogeneous with respect to participation propensity. In this setting, the log-likelihood for one leaf, L_h, would be

$$\begin{aligned}
&\sum_{k\in L_h} (1 - A_k)B_k \ln\{\rho(\boldsymbol{x}_k)\} + \sum_{k\in L_h} (1 - A_k)(1 - B_k)\ln\{1 - \rho(\boldsymbol{x}_k)\} \\
&= \sum_{k\in L_h} (1 - A_k)B_k \ln(\rho_h) + \sum_{k\in L_h} (1 - A_k)(1 - B_k)\ln(1 - \rho_h),
\end{aligned} \tag{2}$$

where $\rho_h = \rho(\boldsymbol{x}_k)$ is assumed to be constant for $k \in L_h$. The first term is fully observed but the second term cannot be computed because it involves covariates for the complement of the combined sample, which are not observed. Accordingly, we replace the second term of (2) by a conditionally unbiased (given the non-probability sample) estimator of its expectation,

$$\sum_{k \in L_h} (1 - A_k) B_k \ln(\rho_h) + \sum_{k \in L_h} \left\{ \frac{A_k}{\pi_k^A} (1 - \pi_k^A) \right\} (1 - B_k) \ln(1 - \rho_h). \tag{3}$$

By maximizing (3), we then estimate ρ_h for leaf h as

$$\widehat{\rho}_h = \frac{\sum_{k \in L_h} B_k}{\sum_{k \in L_h} B_k + \sum_{k \in L_h} (1/\pi_k^A - 1) A_k}, \quad k \in L_h, \tag{4}$$

where we have used $(1 - A_k)B_k = B_k$ and $A_k(1 - B_k) = A_k$. The score of (3) with constant ρ_h within leaves is unbiased for 0,

$$\mathrm{E}\left[\frac{\partial}{\partial \rho_h} L(\rho_h) \right] = \sum_{k \in L_h} (1 - \pi_k^A) + \sum_{k \in L_h} \left\{ \frac{\pi_k^A}{\pi_k^A} \right\} (1 - \pi_k^A)(1 - \rho_h) \frac{-1}{1 - \rho_h} = 0.$$

Since leaves are expected to be homogeneous, we also use the leaf average design probability,

$$\tilde{\pi}_k^A = \frac{\sum_{k \in L_h} \pi_k^A A_k}{\sum_{k \in L_h} A_k},$$

to impute for non-probability elements, $k \in B$.

With the imputed $\tilde{\pi}_k^A$ and the estimated $\widehat{\rho}_h$, the initial combined weights are then

$$c_k = \frac{A_k}{\pi_k^A + \widehat{\rho}_h (1 - \pi_k^A)} + \frac{B_k}{\tilde{\pi}_k^A + \widehat{\rho}_h (1 - \tilde{\pi}_k^A)}.$$

We ratio-adjust these initial combined weights to reflect the representation of the population in the probability sample. A leaf with probability elements represents

$$\widehat{\alpha}_h = \sum_{k \in L_h} \frac{1}{\pi_k^A} A_k$$

such elements in the population. If there are also $\sum_{k \in L_h} B_k > 0$ non-probability elements in the leaf, these elements add detail to the leaf but do not add any information that would change its population representation. Hence, we ratio-adjust the initial combined weights for each leaf to match the original probability sample weights,

$$\omega_k = \left(\sum_{k \in L_h} \frac{1}{\pi_k^A} A_k \right) \frac{c_k}{\sum_{k \in L_h} c_k}. \tag{5}$$

Two special cases of the ratio-adjusted weights are of interest. First, the leaf may contain only elements from A and none from B, in which case $\widehat{\rho}_h = 0$ from (4). In this case, the probability elements in a pure probability leaf have initial combined weights equal to their original probability weights,

$$\frac{1}{\pi_k^A + \widehat{\rho}_h (1 - \pi_k^A)} = \frac{1}{\pi_k^A + 0(1 - \pi_k^A)} = \frac{1}{\pi_k^A}.$$

Further, ratio-adjustment from (5) also has no impact on the original probability weights.

Second, the leaf may contain only elements from B and none from A. Because there is no reference probability sample within the leaf, we do not know what these non-probability elements represent and have no means within the sample to adjust the weights or impute the unknown π_k^A. Accordingly, we impute $\tilde{\pi}_k^A = 0$ and use (4) to obtain $\widehat{\rho}_h = 1$. Hence, the non-probability elements in a pure non-probability leaf have initial combined weights equal to

$$\frac{1}{\tilde{\pi}_k^A + \widehat{\rho}_h(1 - \tilde{\pi}_k^A)} = \frac{1}{0 + 1(1-0)} = 1.$$

Because the leaf has no probability representation, there is no probability estimate for ratio-adjustment, and the final weights for a pure non-probability leaf are equal to one.

The final combined estimator for the total is then

$$\widehat{T}_{y,\text{com}} = \sum_{h \in H} \sum_{k \in (A \cup B) \cap L_h} \widehat{\alpha}_h \frac{c_k}{\sum_{k \in L_h} c_k} y_k = \sum_{h \in H} \sum_{k \in (A \cup B) \cap L_h} y_k \omega_k,$$

and the combined estimator for a ratio is

$$\widehat{R}_{yz,\text{com}} = \frac{\sum_{h \in H} \sum_{k \in (A \cup B) \cap L_h} y_k \omega_k}{\sum_{h \in H} \sum_{k \in (A \cup B) \cap L_h} z_k \omega_k}. \tag{6}$$

2.4 Tree model leaf size

To use a tree-based approach, we often need to determine an appropriate leaf size. A large leaf size with many elements would reduce the homogeneity within leaves, potentially increasing the bias but reducing the variance of within-leaf estimates. On the other hand, a small leaf size with few elements will increase homogeneity, reducing bias but potentially increasing the variance.

We determine the choice of leaf size by evaluating a bias-variance score. We compute the bias score from a set of J key variables at each leaf size ℓ,

$$\text{bias}_\ell = \frac{1}{J} \sum_{j=1}^{J} \left(\frac{\widehat{\theta}_{A \cup B, \ell, j} - \widehat{\theta}_{A,j}}{\widehat{\theta}_{A,j}} \right)^2,$$

where $\widehat{\theta}_{A \cup B, \ell, j}$ is the ratio-adjusted combined estimate for key variable j at leaf size ℓ and $\widehat{\theta}_{A,j}$ is the probability-only estimate for key variable j.

We compute the variance score using the weighting effect. Letting $\{\omega_{k,\ell}\}_{k \in A \cup B}$ denote the ratio-adjusted weights at leaf size ℓ, the variance score is

$$\text{var}_\ell = \text{mean}\left(\omega_{k,\ell}^2\right) / \left(\text{mean}\left(\omega_{k,\ell}\right)\right)^2.$$

We normalize each set of scores to $[0, 1]$ and add them together to get the final bias-variance score:

$$\text{score}_\ell = \frac{\text{bias}_\ell - \min \text{bias}}{\max \text{bias} - \min \text{bias}} + \frac{\text{var}_\ell - \min \text{var}}{\max \text{var} - \min \text{var}}. \tag{7}$$

The leaf size is then chosen to minimize the score_ℓ.

2.5 Variance estimation

Variance estimation for the tree-based dual-frame estimator is challenging. We have considered various approximations and describe here an approach based on estimating equations, treating the estimator as if it were based on a parametric model with a fixed number ($2H$) of parameters. We use a sandwich-type variance estimation formula for estimators obtained as the solution to the joint estimating equations for $\widehat{\theta}_h$ and $\widehat{\rho}_h$ in each

leaf node. Let $\boldsymbol{\eta} = (\theta_1, \theta_2, \ldots, \theta_H, \rho_1, \rho_2, \ldots, \rho_H)^\top$ be the vector of parameters, where $\theta_h = \sum_{k \in U \cap L_h} y_k$ is the population total for leaf h and ρ_h is the participation propensity for the B sample for that leaf. The estimator $\widehat{\boldsymbol{\eta}} = (\widehat{\theta}_1, \widehat{\theta}_2, \ldots, \widehat{\theta}_H, \widehat{\rho}_1, \widehat{\rho}_2, \ldots, \widehat{\rho}_H)^\top$ is the solution to the estimating equations

$$\Phi(\boldsymbol{\eta}) \quad = \quad \begin{bmatrix} J(\theta_1, \rho_1) = \sum_{k \in (A \cup B) \cap L_1} c_k(\widehat{\alpha}_1 y_k - \theta_1) \\ J(\theta_2, \rho_2) = \sum_{k \in (A \cup B) \cap L_2} c_k(\widehat{\alpha}_2 y_k - \theta_2) \\ \vdots \\ J(\theta_H, \rho_H) = \sum_{k \in (A \cup B) \cap L_H} c_k(\widehat{\alpha}_H y_k - \theta_H) \\ G(\rho_1) = \sum_{k \in U \cap L_1} B_k/\rho_1 - \widehat{\beta}_1/(1-\rho_1) \\ G(\rho_2) = \sum_{k \in U \cap L_2} B_k/\rho_2 - \widehat{\beta}_2/(1-\rho_2) \\ \vdots \\ G(\rho_H) = \sum_{k \in U \cap L_H} B_k/\rho_H - \widehat{\beta}_H/(1-\rho_H) \end{bmatrix} = \mathbf{0}, \tag{8}$$

where $c_k = 1/\left\{\pi_k^A + (1-\pi_k^A)\rho_h\right\}$, $\widehat{\alpha}_h = \sum_{k \in A \cap L_h}(1/\pi_k^A)$, and $\widehat{\beta}_h = \sum_{k \in A \cap L_h}(1 - \pi_k^A)/\pi_k^A$. The variance-covariance matrix of $\widehat{\boldsymbol{\eta}}$ has a sandwich form,

$$\text{Var}\left(\widehat{\boldsymbol{\eta}}\right) \simeq \left\{\boldsymbol{\phi}(\boldsymbol{\eta})\right\}^{-1} \text{Var}\left[\Phi\left(\boldsymbol{\eta}\right)\right] \left\{\boldsymbol{\phi}(\boldsymbol{\eta})^\top\right\}^{-1},$$

where $\boldsymbol{\phi}(\boldsymbol{\eta}) = \partial\Phi(\boldsymbol{\eta})/\partial\boldsymbol{\eta}$, and can be estimated by

$$\left\{\boldsymbol{\phi}(\hat{\boldsymbol{\eta}})\right\}^{-1} \widehat{V}\left(\Phi(\hat{\boldsymbol{\eta}})\right) \left\{\boldsymbol{\phi}(\hat{\boldsymbol{\eta}})^{-1}\right\}^\top. \tag{9}$$

The derivative $\phi(\boldsymbol{\eta})$ can be written as

$$\boldsymbol{\phi}(\boldsymbol{\eta}) = \frac{\partial\Phi(\boldsymbol{\eta})}{\partial\boldsymbol{\eta}} = \begin{bmatrix} \frac{\partial J(\boldsymbol{\theta},\boldsymbol{\rho})}{\partial\boldsymbol{\theta}^\top} & \frac{\partial J(\boldsymbol{\theta},\boldsymbol{\rho})}{\partial\boldsymbol{\rho}^\top} \\ \mathbf{0} & \frac{\partial G(\boldsymbol{\rho})}{\partial\boldsymbol{\rho}^\top} \end{bmatrix}, \tag{10}$$

where

$$\frac{\partial J(\boldsymbol{\theta},\boldsymbol{\rho})}{\partial\boldsymbol{\theta}^\top} = \begin{bmatrix} -\sum_{k \in (A \cup B) \cap L_1} c_k & \cdots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \cdots & -\sum_{k \in (A \cup B) \cap L_H} c_k \end{bmatrix}_{(H-\ell_B)\times(H-\ell_B)},$$

$$\frac{\partial J(\boldsymbol{\theta},\boldsymbol{\rho})}{\partial\boldsymbol{\rho}^\top} = \begin{bmatrix} -\sum_k \frac{\left(1-\pi_k^A\right)(\widehat{\alpha}_1 y_k - \theta_1)}{\left(\pi_k^A + (1-\pi_k^A)\rho_1\right)^2} & \cdots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \cdots & -\sum_k \frac{\left(1-\pi_k^A\right)(\widehat{\alpha}_H y_k - \theta_H)}{\left(\pi_k^A + (1-\pi_k^A)\rho_H\right)^2} \end{bmatrix}_{(H-\ell_B)\times(H-\ell_A-\ell_B)},$$

and

$$\frac{\partial G(\boldsymbol{\rho})}{\partial\boldsymbol{\rho}^\top} = \begin{bmatrix} -\frac{n_1^B}{\rho_1^2} - \frac{\widehat{\beta}_1}{(1-\rho_1)^2} & \cdots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \cdots & -\frac{n_H^B}{\rho_H^2} - \frac{\widehat{\beta}_H}{(1-\rho_H)^2} \end{bmatrix}_{(H-\ell_A-\ell_B)\times(H-\ell_A-\ell_B)}.$$

The dimensions on these matrices reflect reductions in the number of estimating equations due to pure probability or pure non-probability leaves. First, a pure non-probability leaf is not representative of the population of interest and has no corresponding θ_h, hence

the corresponding estimating equations $J(\theta_h, \rho_h)$ are removed in (8), leaving $H - \ell_B$ equations, where ℓ_B is the number of pure non-probability leaves.

Further, for a leaf with only elements from A, $\widehat{\rho}_h = 0$ and we treat $\rho_h = 0$ as known. Similarly, for a leaf with only elements from B, $\widehat{\rho}_h = 1$ and we treat $\rho_h = 1$ as known. The corresponding estimating equations $G(\rho_h)$ in (8) are then removed, resulting in $H - \ell_A - \ell_B$ equations, where ℓ_A is the number of pure probability leaves.

The variance-covariance matrix $\text{Var}\left(\Phi(\boldsymbol{\eta})\right)$ can be derived from the iterated variance $\text{E}\left[\text{Var}\left(\Phi(\boldsymbol{\eta})|\, A\right)\right] + \text{Var}\left[\text{E}\left(\Phi(\boldsymbol{\eta})|\, A\right)\right]$. The variance estimator of $\widehat{T}_{y,\text{com}}$ is then the sum of the first $H - \ell_B$ diagonal elements of (9).

The estimated ratio (6) can be approximated by Taylor expansion,

$$\begin{aligned}
\widehat{R}_{yz,\text{com}} &= \frac{\sum_{h\in H}\sum_{k\in(A\cup B)\cap L_h} y_k c_k \widehat{\alpha}_h / \sum_{k\in(A\cup B)\cap L_h} c_k}{\sum_{h\in H}\sum_{k\in(A\cup B)\cap L_h} z_k c_k \widehat{\alpha}_h / \sum_{k\in(A\cup B)\cap L_h} c_k} \\
&= \frac{\sum_{k\in U} y_k}{\sum_{k\in U} z_k} + \sum_{h\in H}\sum_{k\in(A\cup B)\cap L_h} \frac{v_k c_k \widehat{\alpha}_h}{\sum_{k\in(A\cup B)\cap L_h} c_k}
\end{aligned}$$

with

$$v_k = \left\{ \frac{1}{\sum_{k\in U} z_k} \left(y_k - \frac{\sum_{k\in U} y_k}{\sum_{k\in U} z_k} z_k \right) \right\},$$

which can be estimated as $\widehat{v}_k$ by plugging in estimates of the unknown population totals. The variance estimator for the estimated ratio is then obtained by replacing y_k by $\widehat{v}_k$ in (9).

3. Simulation experiment

3.1 Simulation setup

We use the frames and simulated samples described in Benoit-Bryan and Mulrow (2021), which were created to evaluate different estimation approaches that combine probability and non-probability samples. The data used to generate these simulated samples is from a real study called Culture and Community in a Time of Crisis (CCTC), which evaluated behaviors and attitudes during the global COVID-19 pandemic.

In the simulated version, the artificial CCTC population consists of 123,757 records. The probability sampling frame (known to the sampler) consists of 117,276 records. The frame is a subset of the full population, reflecting slight undercoverage. The non-probability sampling "frame" is unknown to the sampler, but in fact is a subset of subpopulation consisting of 74,202 records. Together, the probability and non-probability sampling frames cover the universe. Our goal is to estimate characteristics of the population of interest by combining the probability and non-probability samples.

Stratified probability samples of size 400 are selected from the probability sampling frame and non-probability samples of size 400 and 800 are selected from the non-probability sampling frame. There is no overlap between the probability and non-probability samples. We use 1000 replicate probability and non-probability samples (at each sample size) from the artificial CCTC population.

We combine the samples using the tree-based dual-frame estimation approach. There are many possible demographic covariates for tree modeling in the data. We use all possible covariates and response variables, including education, employment status, income, age, race, region, and metro (binary indicator of metropolitan area). We use the `R` function `rpart` (Therneau and Atkinson, 2023) to fit the weighted tree. To determine the leaf size, we do a grid search across a range of leaf sizes from $(\text{total sample size})^{0.4}$ to $(\text{total sample size})/10$. We choose the leaf size that minimizes score_ℓ from (7).

The estimation methods are evaluated on 20 binary response variables related to a person's behavior or attitude, like attending a live online event, music festival, online exhibition, etc. For each replicate sample, we construct the combined estimator of the population proportions for all 20 variables at each non-probability sample size. The variance estimator for the tree-based dual-frame estimator is constructed by the estimation equations. We compare to a baseline estimator that uses only the probability sample and the original sampling weights.

3.2 Simulation results

Table 1 and Figure 1 summarize the percent relative bias for each of the 20 variables at each total sample size. The estimator using only the probability sample is nearly unbiased, reflecting the small undercoverage error. The combined estimator successfully incorporates the non-probability sample while maintaining small relative bias.

Table 2 and Figure 2 summarize the effective sample size ratio for each of the 20 variables at different total sample sizes. In the ratio, mean square error for the baseline is the numerator and mean square error for the combined estimator is the denominator. If the ratio is greater than 1, there is a benefit of adding non-probability sample to the probability sample. For most variables, there is a gain in efficiency from the addition of the non-probability sample at each non-probability sample size; however, the efficiency is less than linear in the total sample size.

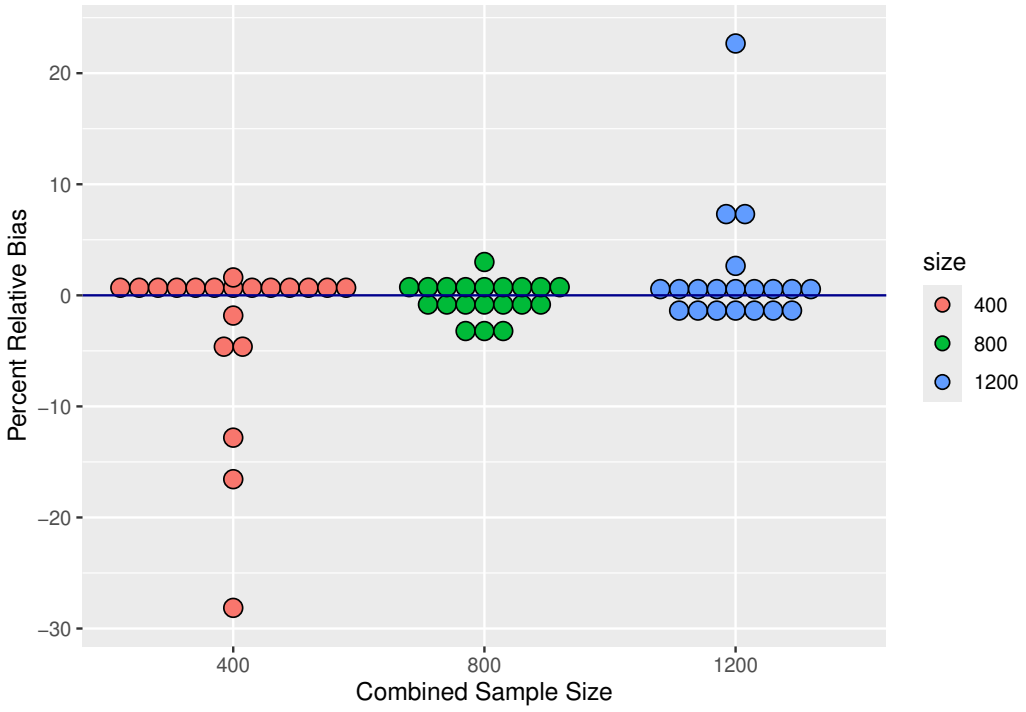

Figure 1: Monte Carlo percent relative bias (based on 1000 replicate samples) for the baseline estimator and the combined estimator (6) for 20 binary variables in the CCTC simulation experiment. Baseline estimator uses only the probability sample (of size 400) and combined estimators use the probability sample plus non-probability samples of size 400 or 800. Each point corresponds to percent relative bias for one estimator type and one binary variable.

Table 1: Monte Carlo percent relative bias (based on 1000 replicate samples) for the baseline estimator and the combined estimator (6) for 20 binary variables in the CCTC simulation experiment. Baseline estimator uses only the probability sample (of size 400) and combined estimator uses the probability sample plus non-probability samples of size 400 or 800.

		Combined	
	Baseline	Estimator	
Total sample size	400	800	1200
Live online event	1.330	1.223	1.125
Past activities	0.594	3.001	2.652
Want more hope	0.047	-0.315	-0.591
Want more fun	-0.023	-0.442	-0.804
Want more in life	0.858	0.102	-0.221
Music festival	0.005	0.048	0.222
Community festival	1.244	0.344	-0.031
Classical music	1.611	1.409	0.931
Online exhibitions	0.825	0.811	0.581
Online kid activities	0.591	0.824	0.675
Q17-Very Unimportant	-12.815	-3.701	7.464
Q17-Unimportant	-5.318	-2.728	-0.695
Q17-Neither	-1.822	-1.665	-1.662
Q17-Important	-0.099	-1.041	-1.703
Q17-Very Important	1.401	1.421	1.342
Q18-Very Unimportant	-28.139	0.012	22.676
Q18-Unimportant	-16.554	-0.400	7.162
Q18-Neither	-3.931	-2.962	-2.153
Q18-Important	0.092	-1.081	-1.901
Q18-Very Important	1.466	1.145	0.981

Table 2: Effective sample size ratios, computed as the ratio of the mean square error of the baseline estimator to the mean square error of the combined estimator (6), for 20 binary variables in the CCTC simulation experiment. Baseline estimator uses only the probability sample (of size 400) and combined estimators use the probability sample plus non-probability samples of size 400 or 800.

	Combined Estimator	
Total sample size	800	1200
Live online event	1.311	1.424
Past activities	0.730	0.663
Want more hope	1.363	1.537
Want more fun	1.342	1.399
Want more in life	1.374	1.615
Music festival	1.333	1.391
Community festival	1.425	1.460
Classical music	1.316	1.578
Online exhibitions	1.261	1.403
Online kid activities	1.237	1.305
Q17-Very Unimportant	1.679	1.956
Q17-Unimportant	1.640	1.976
Q17-Neither	1.299	1.451
Q17-Important	1.212	1.289
Q17-Very Important	1.099	1.129
Q18-Very Unimportant	1.428	1.265
Q18-Unimportant	1.714	1.808
Q18-Neither	1.432	1.638
Q18-Important	1.153	1.191
Q18-Very Important	1.100	1.155

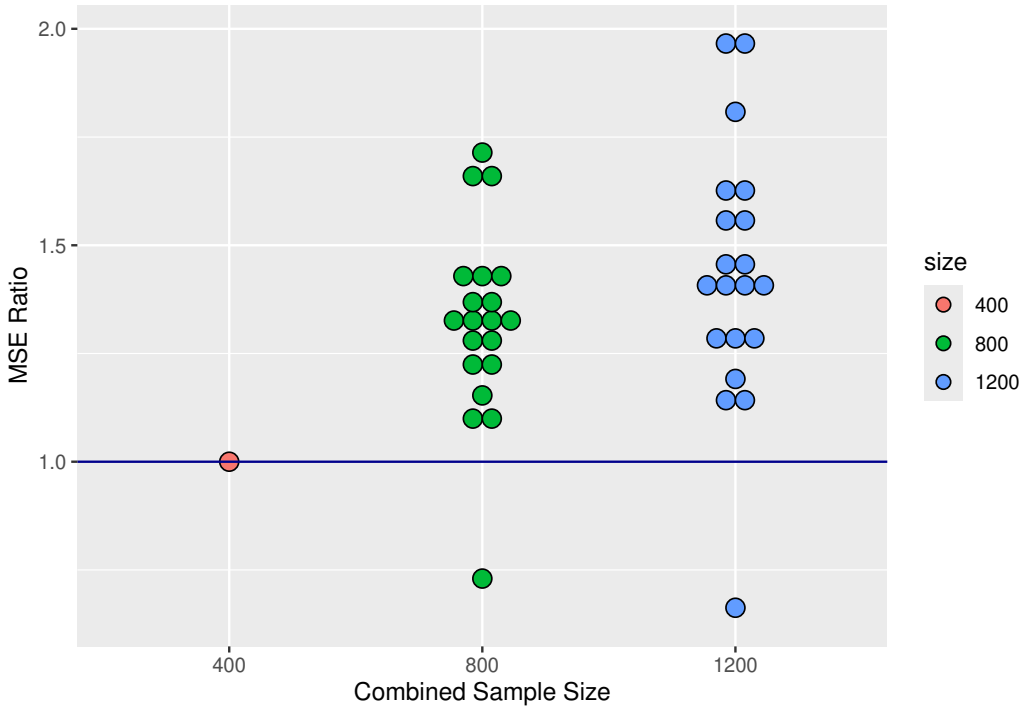

Figure 2: Effective sample size ratios, computed as the ratio of the mean square error of the baseline estimator to the mean square error of the combined estimator (6), for 20 binary variables in the CCTC simulation experiment. Baseline estimator uses only the probability sample (of size 400) and combined estimator uses the probability sample plus non-probability samples of size 400 or 800. Each point corresponds to MSE ratio for one estimator type and one variable.

We estimate the standard deviations for the combined estimator from the estimating equations of the ratio. Figure 3 summarizes the percent relative bias of the estimated standard deviation for each of the 20 variables at each total sample size. The estimators are stable and approximately unbiased across the range of sizes of the non-probability sample although the assumptions does not hold.

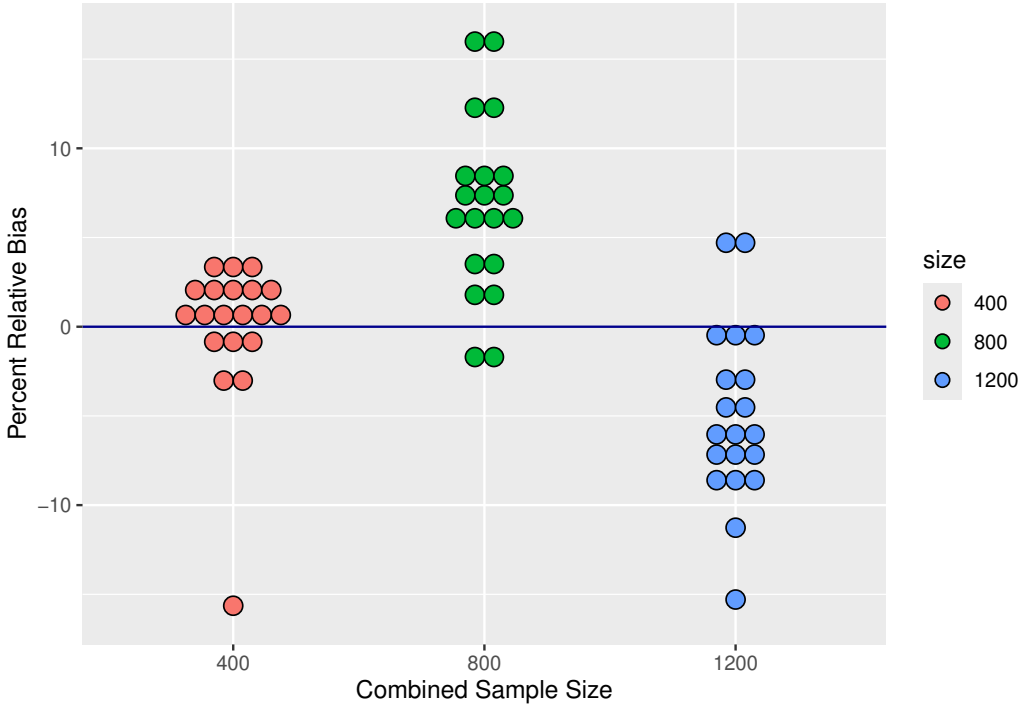

Figure 3: Percent relative bias of standard deviation estimators, computed as 100% times the bias of the estimator divided by the true standard deviation (estimated by Monte Carlo with 1000 replicates), for 20 binary variables in the CCTC simulation experiment. Each point corresponds to percent relative bias for one estimator type and one variable.

Table 3 and Figure 4 summarize the 95% confidence interval coverage for the 20 variables at different total sample sizes for the combined estimators. Almost all confidence intervals have close to the 95% nominal coverage across different total sample sizes. In some cases, like the "Q18-Very Unimportant" variable, the population proportion is very close to zero and confidence intervals for such sparse proportions are known to have poor coverage because the normal approximation is poor.

4. Summary and discussion

We propose tree-based dual-frame estimators for inference when samples include probability and non-probability samples. These estimators are based on nonparametric models and work well in our simulations across a range of variables. The estimator addresses the bias from the non-probability sample and has lower mean square error than the baseline estimator that uses only the probability samples. Our proposed variance estimator, which treats the model as if it were parametric with fixed size H, reduces the bias compared to other alternatives and gives reasonable confidence interval coverage in most cases, at different non-probability sample sizes.

Table 3: Confidence interval coverage (estimated by Monte Carlo with 1000 replicates) for the combined estimator (6) for 20 binary variables in the CCTC simulation experiment. The combined estimator uses the probability sample (of size 400) plus non-probability samples of size 400 or 800. Nominal coverage level is 95%.

		Combined	
	Baseline	Estimator	
Total sample size	400	800	1200
Live online event	95.3	96.7	94.8
Past activities	96.1	90.4	81.1
Want more hope	95.8	97.3	95.6
Want more fun	93.8	96.4	94.0
Want more in life	95.2	96.9	95.9
Music festival	95.7	97.7	95.6
Community festival	93.6	98.3	95.3
Classical music	94.4	97.0	94.9
Online exhibitions	95.1	96.5	94.4
Online kid activities	95.8	96.2	93.4
Q17-Very Unimportant	85.6	91.2	92.1
Q17-Unimportant	90.6	95.4	93.2
Q17-Neither	94.0	95.9	93.9
Q17-Important	95.0	95.8	92.9
Q17-Very Important	93.3	95.2	90.9
Q18-Very Unimportant	65.9	74.9	75.8
Q18-Unimportant	84.3	92.0	92.2
Q18-Neither	92.7	96.0	95.1
Q18-Important	95.1	95.4	92.6
Q18-Very Important	93.3	94.6	93.1

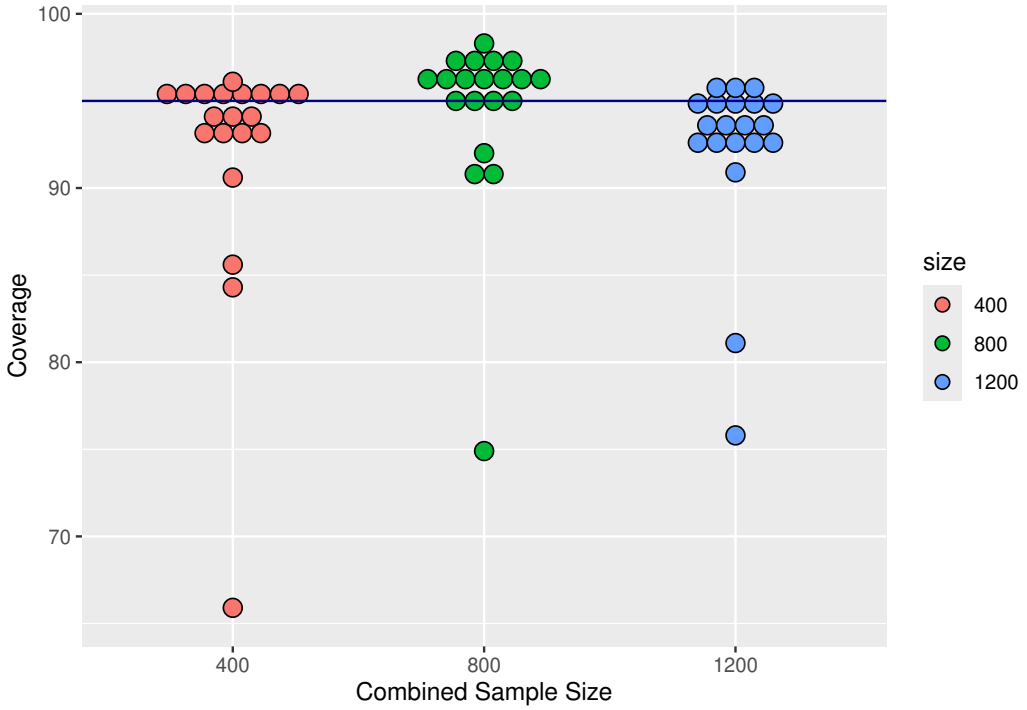

Figure 4: Confidence interval coverage (estimated by Monte Carlo with 1000 replicates), for the baseline estimator and the combined estimator (6) for 20 binary variables in the CCTC simulation experiment. Baseline estimator uses only the probability sample (of size 400) and combined estimator uses the probability sample plus non-probability samples of size 400 or 800. Horizontal reference line is at the nominal level of 95%. Each point corresponds to coverage for one estimator type and one variable.

References

Benoit-Bryan, J. and E. Mulrow (2021). Exploring nonprobability methods with simulations from a common data source: Culture and Community in a Time of Crisis. In *Proceedings of the Section on Survey Research Methods*, Alexandria, VA. American Statistical Association.

Chen, Y., P. Li, and C. Wu (2020). Doubly robust inference with nonprobability survey samples. *Journal of the American Statistical Association 115*(532), 2011–2021.

Chu, K. C. and J.-F. Beaumont (2019). The use of classification trees to reduce selection bias for a non-probability sample with help from a probability sample. In *Proceedings of the Survey Methods Section: SSC Annual Meeting*.

Elliott, M. R. and R. Valliant (2017). Inference for nonprobability samples. *Statistical Science 32*(2), 249–264.

Ganesh, N., V. Pineau, A. Chakraborty, and J. M. Dennis (2017). Combining probability and non-probability samples using small area estimation. In *Proceedings of the Section on Survey Research Methods*, Alexandria, VA, pp. 1657–1667. American Statistical Association.

Huang, C.-M. and F. J. Breidt (2023). A dual-frame approach for estimation with respondent-driven samples. *Metron 81*(1), 65–81.

Kim, J. K. (2022). A gentle introduction to data integration in survey sampling. *85*, 19–29.

Kim, J. K., S. Park, Y. Chen, and C. Wu (2021). Combining non-probability and probability survey samples through mass imputation. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 184*(3), 941–963.

Kim, J. K. and Z. Wang (2019). Sampling techniques for big data analysis. *International Statistical Review 87*, S177–S191.

Nandram, B. and J. N. K. Rao (2021). A bayesian approach for integrating a small probability sample with a non-probability sample. In *Proceedings of the Section on Survey Research Methods*, pp. 1568–1603. American Statistical Association.

Rafei, A., C. A. Flannagan, and M. R. Elliott (2020). Big data for finite population inference: Applying quasi-random approaches to naturalistic driving data using Bayesian additive regression trees. *Journal of Survey Statistics and Methodology 8*(1), 148–180.

Rafei, A., C. A. Flannagan, B. T. West, and M. R. Elliott (2022). Robust Bayesian inference for Big Data: Combining sensor-based records with traditional survey data. *The Annals of Applied Statistics 16*(2), 1038 – 1070.

Rivers, D. (2007). Sampling for web surveys. Paper Prepared for the 2007 Joint Statistical Meetings, Salt Lake City, UT.

Sakshaug, J. W., A. Wiśniowski, D. A. P. Ruiz, and A. G. Blom (2019). Supplementing small probability samples with nonprobability samples: A bayesian approach. *Journal of Official Statistics 35*(3), 653–681.

Salvatore, C., S. Biffignandi, J. W. Sakshaug, A. Wiśniowski, and B. Struminskaya (2024). Bayesian integration of probability and nonprobability samples for logistic regression. *Journal of Survey Statistics and Methodology 12*(2), 458–492.

Singh, A. C. and F. Mecatti (2011). Generalized multiplicity-adjusted Horvitz-Thompson estimation as a unified approach to multiple frame surveys. *Journal of Official Statistics 27*(4), 633–650.

Therneau, T. and B. Atkinson (2023). *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1.23.

Valliant, R. (2020). Comparing alternatives for estimation from nonprobability samples. *Journal of Survey Statistics and Methodology 8*(2), 231–263.

Wiśniowski, A., J. W. Sakshaug, D. A. Perez Ruiz, and A. G. Blom (2020). Integrating probability and nonprobability samples for survey inference. *Journal of Survey Statistics and Methodology 8*(1), 120–147.

Wu, C. (2022). Statistical inference with non-probability survey samples. *Survey Methodology 48*(2), 283–311.

Yang, M., N. Ganesh, E. Mulrow, and V. Pineau (2018). Estimation methods for nonprobability samples with a companion probability sample. In *Proceedings of the Section on Survey Research Methods*, Alexandria, VA, pp. 1715–1723. American Statistical Association.