SSC Annual Meeting, May 2019
Proceedings of the Survey Methods Section

THE USE OF CLASSIFICATION TREES TO REDUCE SELECTION BIAS FOR A NON-PROBABILITY SAMPLE WITH HELP FROM A PROBABILITY SAMPLE

Kenneth C.K. Chu[1], Jean-François Beaumont[2]

ABSTRACT

We propose the Tree-based Inverse-Propensity-Weighted (**TrIPW**) estimator, a new population total estimator that primarily uses data from a non-probability sample, but attempts to correct for selection bias by incorporating auxiliary data from an independent probability sample. TrIPW can be regarded as a nonparametric counterpart of the estimator proposed by (Chen et al., 2018). TrIPW executes a variant, which we named **nppCART**, of the CART (Breiman et al., 1984) algorithm to construct a classification tree partitioning the non-probability sample, and uses the auxiliary probability sample to compensate for the missing data that precludes the direct use of CART. The terminal nodes of the resulting tree are regarded as homogeneous selection propensity classes, which in turn yields unit-level selection propensity estimates in the usual way. The reciprocals of these estimates are then used as weights in the TrIPW estimator. The purpose of introducing a nonparametric method such as TrIPW is the expectation that such methods may be more robust against nonlinear relations between the selection mechanism and the auxiliary variables than methods that make explicit (generalized) linear model assumptions. Another advantage of classification trees is that they do not require variable selection or assumptions on the interaction structure among the selected variables. This is particularly useful when the auxiliary variables are categorical. We illustrate the effectiveness and robustness against nonlinearity of TrIPW via two simulation studies.

KEY WORDS: Non-probability samples, Selection bias, Propensity, Classification trees, Probability samples, Non-linearity, Non-parametric methods

RÉSUMÉ

Nous proposons l'estimateur d'arborescence à pondération par l'inverse du score de propension (**TrIPW**), un nouvel estimateur du total de la population qui utilise principalement les données d'un échantillon non probabiliste, mais tente de corriger le biais de sélection en incorporant des données auxiliaires provenant d'un échantillon probabiliste indépendant. TrIPW peut être considéré comme une contrepartie non paramétrique de celle proposée par (Chen et al., 2018). TrIPW exécute une variante, que nous dénommons **nppCART**, de l'algorithme CART (Breiman et al., 1984) pour construire un arbre de classification partitionnant l'échantillon non probabiliste et utilise l'échantillon probabiliste auxiliaire pour compenser les données manquantes empêchant l'utilisation directe de CART. Les nœuds terminaux de l'arbre résultant sont considérés comme des classes homogènes de la propension à la sélection, ce qui donne des estimations de la propension à la sélection au niveau de l'unité de la manière habituelle. Les inverses de ces estimations sont ensuite utilisés comme pondérations dans l'estimateur TrIPW. L'introduction d'une méthode non paramétrique telle que TrIPW a pour objectif de fournir une méthode plus robuste aux relations non linéaires entre le mécanisme de sélection et les variables auxiliaires que les méthodes qui font des hypothèses explicites de modèle linéaire (généralisé). Un autre avantage des arbres de classification est qu'ils ne nécessitent pas de sélection de variable ni d'hypothèses sur la structure d'interaction des variables sélectionnées. Ceci est particulièrement utile lorsque les variables auxiliaires sont catégoriques. Nous illustrons l'efficacité et la robustesse à la non-linéarité de TrIPW via deux études de simulation.

MOTS CLÉS : Échantillons non probabilistes, Biais de sélection, Propension, Arbres de classification, Échantillons probabilistes, Non-linéarité, Méthodes non paramétriques

[1]Kenneth C.K. Chu, Statistics Canada, R.H. Coats Building, 100 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6, kenneth.chu@canada.ca

[2]Jean-François Beaumont, Statistics Canada, R.H. Coats Building, 100 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6, jean-francois.beaumont@canada.ca

1 INTRODUCTION

1.1 Background and motivation

Due to rising costs, declining response rates, continual efforts to reduce response burden and today's relatively easy availability of data from non-probability samples, researchers have begun work on deriving methodologically sound estimators based on non-probability samples.

One of the key obstacles to the utilization of non-probability samples is selection bias. The reader is referred to (Beaumont, 2019) for a detailed review of methods that could mitigate selection bias for non-probability samples. Most of these methods work by incorporating auxiliary information from a "helper" probability sample. They are often model-based or model-assisted, and make explicit assumptions on either the relation between the target variable and auxiliary variables (e.g. linearity) or the relation between the non-probability selection mechanism and auxiliary variables (e.g. a logistic model).

In this article, we propose a novel method that mitigates selection bias for a given non-probability sample by using auxiliary information from a probability sample. It is based on classification trees, a nonparametric method that makes no explicit model assumptions among the target variable, auxiliary variables, and the non-probability selection mechanism except for the usual noninformative selection assumption. Our method can be regarded as a nonparametric counterpart of the one proposed by (Chen et al., 2018), which assumes a logistic model with pre-defined auxiliary variables. Our contribution is the adaptation of the CART algorithm for the automatic selection of auxiliary variables and their interactions.

We implemented our method, and performed related simulation studies, in the R statistical computing language (R Core Team, 2019).

1.2 Description of the objective

Our objective is an archetypal one in survey sampling, namely, to produce a reliable estimate for the population total $T_y = \sum_{i \in U} y_i$ of a population characteristic y (we will also refer to y as the target variable), based on data observed only for units in a certain sample selected from the underlying population U.

We consider the scenario in which the target variable y is observed for units in a non-probability sample $\mathcal{S}_A \subset U$. Traditional methods designed for probability samples therefore do not apply, and we need to correct for the possible selection bias caused by the unknown stochastic selection mechanism that generates $\mathcal{S}_A$. To this end, we assume:

- For each unit in $\mathcal{S}_A$, in addition to the target variable y, a number of auxiliary variables $X_1, \ldots, X_p$ have also been observed. At least some of these auxiliary variables should collectively be strong predictors for the variable of interest, as well as for being selected into $\mathcal{S}_A$.
- $X_1, \ldots, X_p$ have also been observed for each unit in a separate probability sample $\mathcal{S}_B$, which is generated independently from $\mathcal{S}_A$.

See Figure (1.1) for an illustration of this scenario.

Now, for each $i \in U$, let ρ_i be the probability that i is selected into the non-probability sample $\mathcal{S}_A$. If $\rho_i > 0$, for each $i \in U$, and if ρ_i is known for each $i \in \mathcal{S}_A$, then it is straightforward to prove that

$$\widehat{T}_y(\mathcal{S}_A) \;:=\; \sum_{i \in \mathcal{S}_A} \frac{y_i}{\rho_i} \tag{1.1}$$

would define an unbiased estimator for T_y, where the unbiasedness here is with respect to the stochastic selection mechanism that generates the non-probability sample $\mathcal{S}_A$.

Of course, the ρ_i's are unknown in practice. This begs the questions: Can they be estimated, at least for $i \in \mathcal{S}_A$? If so, can we obtain a reliable estimator for T_y by replacing each ρ_i in (1.1) with its corresponding estimate?

X_1	$\cdots$	X_p	y
$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\cdots$	$\cdots$	$\cdots$	$\cdots$

(a) $\mathcal{S}_A$ = non-probability sample

X_1	$\cdots$	X_p	π
$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\cdots$	$\cdots$	$\cdots$	$\cdots$

(b) $\mathcal{S}_B$ = probability sample

Figure (1.1) The objective of TrIPW is to produce a reliable estimate for the population total of the target variable y. (a) The primary data for TrIPW are collected via a non-probability sample $\mathcal{S}_A$; in particular, the target variable y is observed for the units in $\mathcal{S}_A$. (b) The auxiliary data for TrIPW are collected in an independent probability sample $\mathcal{S}_B$. The variables common to both data sets are denoted as $X_1, \ldots, X_p$. π denotes the selection probability of the sampling design that generates the probability sample $\mathcal{S}_B$.

This is reminiscent of the problem of estimating response propensity, which led us to consider classification trees, a class of techniques commonly used for that purpose. More concretely, we would define a binary variable $R : U \longrightarrow \{\, 0, 1\}$ by:

$$R(i) \;=\; R_i \;:=\; \begin{cases} 1, & \text{if } i \in \mathcal{S}_A\,, \\ 0, & \text{if } i \in U \setminus \mathcal{S}_A\,, \end{cases} \tag{1.2}$$

and train a classification tree to predict R, using $X_1, \ldots, X_p$ as predictor variables. Once such a classification tree had been obtained, its terminal nodes would be regarded as homogeneous selection propensity classes, which in turn would yield the unit-level selection propensity estimates needed to replace the ρ_i's in (1.1).

However, it is now clear that there is a fundamental obstacle in our context to applying classification trees for estimating the ρ_i's, namely, the fact that the values of $X_1, \ldots, X_p$ are unknown for the units in $U \setminus \mathcal{S}_A$.

Our proposed method TrIPW compensates for such missing data (the unavailability of data on $X_1, \ldots, X_p$ for units in $U \setminus \mathcal{S}_A$) by incorporating data from the auxiliary probability sample $\mathcal{S}_B$.

2 nppCART & TrIPW ESTIMATOR

2.1 How execution of CART fails when $X_1, \ldots, X_p$ are unknown for $U \setminus \mathcal{S}_A$

Recall that CART (Breiman et al., 1984) is an optimization procedure that executes a greedy search algorithm known as *recursive binary partitioning* in order to minimize (locally) a certain data-dependent objective function defined on binary trees that partition the space of the predictor variables into hyper-rectangles with faces parallel to coordinate hyperplanes. The objective function of CART is known as *tree impurity*. Given a binary tree T, its **tree impurity** $I(T)$ is defined as:

$$I(T) \;:=\; \sum_{l \,\in\, \mathcal{L}(T)} P(X \in l\,) \cdot G(\,l\,)\,,$$

where $\mathcal{L}(T)$ is the set of **terminal nodes** of T, X is the (multivariate) predictor variable, and $G(\,l\,)$ denotes the **Gini index** (see p.104, (Breiman et al., 1984)) of the terminal node $l \in \mathcal{L}(T)$.

For prediction of the binary selection indicator R defined in (1.2) based on $X = (X_1, \ldots, X_p)$, tree impurity

simplifies to:

$$I(T) \;=\; \cdots \;=\; \sum_{l \in \mathcal{L}(T)} \underbrace{\frac{\#(\,l\,)}{\#(U)}}_{P(X \in l)} \times \underbrace{2 \cdot \frac{\#(\,l \cap \{R=1\})}{\#(\,l\,)} \left(1 - \frac{\#(\,l \cap \{R=1\})}{\#(\,l\,)}\right)}_{G(l)\text{, for binary classification}} \tag{2.3}$$

$$= \sum_{l \in \mathcal{L}(T)} \frac{\#(\,l\,)}{\#(U)} \times 2 \cdot \frac{\#(\,l \cap \mathcal{S}_A)}{\#(\,l\,)} \left(1 - \frac{\#(\,l \cap \mathcal{S}_A)}{\#(\,l\,)}\right), \tag{2.4}$$

where $\#(S)$ denotes the number of elements of a given subset $S \subset U$. The execution of CART involves repeated evaluations and comparisons of expressions similar to the summands in (2.4), which reveals precisely why CART cannot be used in practice in our context: The size $\#(\,l\,)$ of a terminal node l is unknown, since

$$\#(\,l\,) \;=\; \#\Big(\,l \cap \{R=1\}\Big) \,+\, \#\Big(\,l \cap \{R=0\}\Big) \;=\; \#\Big(\,l \cap \mathcal{S}_A\Big) \,+\, \#\Big(\,l \cap (U \setminus \mathcal{S}_A)\Big)\,,$$

and the number $\#\Big(\,l \cap (U \setminus \mathcal{S}_A)\Big)$ of units in the population belonging to the terminal node l but not $\mathcal{S}_A$ is unknown in practice.

2.2 nppCART tree-growing algorithm and TrIPW population total estimator

The reason of the failure to execute CART in our context also suggests a remedy: produce somehow an estimate $\widehat{\#}(\,l\,)$ for $\#(\,l\,)$, replace $\#(\,l\,)$ in (2.4) with $\widehat{\#}(\,l\,)$, but otherwise execute recursive binary partitioning as in CART in order to minimize the resulting modified version of tree impurity.

The obvious choice for $\widehat{\#}(\,l\,)$, given the assumed availability of data for units in $\mathcal{S}_B$, is:

$$\widehat{\#}(\,l\,) \;=\; \sum_{j \in\, l \cap \mathcal{S}_B} \frac{1}{\pi_j} \tag{2.5}$$

Our proposed **nppCART** tree-growing algorithm is obtained from CART by using the modified tree impurity as objective function as described above, and by handling certain "boundary scenarios" by tightening stopping criteria and ruling out certain candidate partitions of nodes. (The prefix "npp" in nppCART stands for "non-probability/probability".) These so-called boundary scenarios never occur during the execution of CART; they occur in nppCART due only to the fact that (the hopefully occasional) poor estimates for $\#(\,l\,)$ sometimes lead to certain pathological situations, e.g. division by zero, or estimates for certain probabilities exceeding 1, etc. nppCART does not include a pruning step. In this article, we restrict our attention to the case where $X_1, X_2, \ldots, X_p$ are continuous variables. See Figure (2.1) for the pseudocode of the nppCART tree-growing algorithm.

Once a binary tree has been constructed using the nppCART tree-growing algorithm, its terminal nodes will be regarded as homogeneous propensity classes, and the resulting propensity estimate for each unit $i \in \mathcal{S}_A$ is defined/computed as follows:

$$\widehat{\rho}_i \;:=\; \frac{\#\Big(\,l(i) \cap \mathcal{S}_A\Big)}{\widehat{\#}\Big(\,l(i)\,\Big)}\,, \tag{2.6}$$

where $l(i)$ is the terminal node containing $i \in \mathcal{S}_A$, and $\widehat{\#}(\,l(i)\,)$ is the design-based estimate of its size computed according to (2.5). We can now give the definition of the **TrIPW population total estimator**:

$$\widehat{T}_y^{\,(\mathrm{TrIPW})}(\,\mathcal{S}_A, \mathcal{S}_B\,) \;:=\; \sum_{i \in \mathcal{S}_A} \frac{y_i}{\widehat{\rho}_i}\,. \tag{2.7}$$

3 SIMULATION STUDIES

We illustrate the effectiveness of TrIPW with two simulation studies. The two studies have the same structure but different synthetic populations in order to illustrate the behaviour of TrIPW under different relations among the target variable, the predictor variables, and the non-probability selection mechanism. In each simulation study, the following steps were performed:

- A synthetic population U of size 10,000 of 4-tuples $(X_1, X_2, Y, \rho) \in \mathbb{R}^4$ was generated, where (X_1, X_2) would be treated as the predictor variables, Y the target variable, and ρ the non-probability selection propensity.
- Two hundred Monte Carlo simulations were performed, during each of which a pair, $\mathcal{S}_A$ and $\mathcal{S}_B$, of samples were independently selected from U. $\mathcal{S}_A$ was a Poisson sample, with each unit's selection propensity being its own ρ value. $\mathcal{S}_B$ was a simple random sample without replacement (SRSWOR) with sampling fraction 1/10. nppCART was executed based on the data from $\mathcal{S}_A$ and $\mathcal{S}_B$, with the following choices of parameters: the minimal admissible node size $s_0 = 10$, and the minimal admissible value for the Gini index $g_0 = 0.095$, in order to obtain a binary tree whose terminal nodes would partition $\mathcal{S}_A$, and a non-probability selection propensity estimate for each unit in $\mathcal{S}_A$ according to (2.6). The TrIPW estimate for the population total of Y was then computed according to (2.7).
- For each of the simulations, the intermediate results of the non-probability selection propensity estimates generated by nppCART for units in $\mathcal{S}_A$ were compared against their known values. See Figures (3.1) and (3.3).
- A histogram was generated for the 200 TrIPW population total estimates in order to assess bias and variance. In addition, histograms corresponding to three other methods were also generated for comparison:
  - a calibration estimator, treating $\mathcal{S}_A$ naïvely (and incorrectly) as an SRSWOR, and calibrating to the population size and population totals of X_1 and X_2,
  - the "estimator" (1.1) (note however that this estimator is not implementable in practice as its computation requires knowledge of the true propensity of each unit in $\mathcal{S}_A$),
  - the "naïve" estimator:

$$\widehat{T}_y^{(\text{naïve})}(\,\mathcal{S}_A\,) \;:=\; \frac{\#(\,U\,)}{\#(\,\mathcal{S}_A\,)} \cdot \sum_{i \,\in\, \mathcal{S}_A} y_i \,. \tag{3.8}$$

  See Figures (3.2) and (3.4).

3.1 Simulation study 1

In this study, the target variable Y is a linear function of the predictor variables (X_1, X_2), up to an additive Gaussian noise, while the non-probability selection propensity ρ is a (deterministic) sigmoid function of (X_1, X_2). This is a typical targeted scenario for many model-based methods surveyed in (Beaumont, 2019).

TrIPW demonstrated small relative bias and root mean squared error (RMSE) in this simulation study, but both the calibration estimator and the estimator (1.1) exhibited even smaller relative bias and RMSE. The naïve estimator yielded large values for both performance metrics, as expected. See Figures (3.1) and (3.2).

The very good performance of the calibration estimator, despite its erroneous underlying assumption on $\mathcal{S}_A$, was probably attributable to the strong linear relationship between Y and (X_1, X_2). On the other hand, while (1.1) also gave very good performance, note that it is in fact not implementable in practice as it assumes knowledge of the true propensity of each unit in $\mathcal{S}_A$.

3.2 Simulation study 2

In this study, both the target variable Y and the non-probability selection propensity ρ are in nonlinear relations with the predictor variables (X_1, X_2).

TrIPW demonstrated it could maintain relatively small relative bias and RMSE even under such nonlinearities, while that is no longer the case for the calibration estimator. The estimator (1.1) still showed good performance, though we reiterate that it is not implementable in practice. The "naïve" estimator (3.8) exhibited again large values for both performance metrics, as expected. See Figures (3.3) and (3.4).

The large relative bias and RMSE in this study of the calibration estimator were probably attributable to its erroneous assumption on $\mathcal{S}_A$, and the nonlinear relations between Y (respectively, ρ) and (X_1, X_2).

4 DISCUSSION

4.1 Software

Although the nppCART algorithm was implemented to perform the simulation studies discussed in this article, it was in fact implemented as a reusable and distributable R6 class (Chang, 2019) in the R statistical computing language (R Core Team, 2019). Furthermore, the development of an R package, nppR, featuring nppCART is near completion. Although in this article, we restricted attention to only continuous $X_1, \ldots, X_p$, the nppCART implementation in nppR does handle categorical predictor variables (factors and ordered factors in R).

4.2 Theoretical properties of nppCART and TrIPW

We did not address the theoretical properties of the nppCART tree-growing algorithm nor those of the TrIPW estimator. For instance, it would be desirable to determine conditions on the non-probability selection mechanism, its relation with the predictor variables, and the sampling design of the auxiliary probability sample that would ensure that the terminal nodes of the binary tree produced by nppCART would indeed be good approximations of true homogeneous propensity classes, or that TrIPW would deliver good estimation performance, e.g. consistency and small variance.

4.3 More realistic use-case-specific simulations and experiments with real data

The simulation studies we have conducted have been restricted to low-dimensional synthetic data. It is imperative to examine the effectiveness of TrIPW in more realistic situations, such as high-dimensional data or a mixture of categorical and continuous predictor variables. We plan to conduct computational experiments to assess the performance of nppCART and TrIPW on real data sets, and their potential ultimately to be deployed in production settings.

Figure (2.1)

The nppCART Tree-Growing Algorithm

Input:

- Observed values of $X_1, \ldots, X_p$ and y for units in the non-probability sample $\mathcal{S}_A$ in tabular format; see Figure (1.1a)
- Design selection probability π and observed values of $X_1, \ldots, X_p$ for units in the probability sample $\mathcal{S}_B$ in tabular format; see Figure (1.1b)
- $s_0 \in \mathbb{N}$, minimal admissible node size
- $g_0 > 0$, minimal admissible value for the Gini index

Output:

- A binary tree that partitions $\mathcal{S}_A$ based on the values of $X_1, \ldots, X_p$

Notations:

For a given terminal node l, let

$$\mathcal{D}(\,l\,) \;:=\; \left\{ \;(k,x) \;\middle|\; \begin{array}{c} k \in \{1,2,\ldots,p\}\,, \text{ and } x \in \mathbb{R} \text{ is the mid-point of two distinct} \\ \text{consecutive observed values of } X_k \text{ for units in } \; l \cap \mathcal{S}_A \end{array} \right\}.$$

For a terminal node l with $\mathcal{D}(\,l\,) \neq \varnothing$, and for each $(k,x) \in \mathcal{D}(\,l\,)$, let $l_{\{X_k < x\}}$ and $l_{\{X_k > x\}}$ be the two children nodes of l resulting from splitting l by the variable X_k at the splitting threshold of x, $\widehat{\#}\big(\,l_{\{X_k < x\}}\big)$ and $\widehat{\#}\big(\,l_{\{X_k > x\}}\big)$ be their respective designed-based estimated sizes based on $\mathcal{S}_B$, $\widehat{G}\big(\,l_{\{X_k > x\}}\big)$ and $\widehat{G}\big(\,l_{\{X_k > x\}}\big)$ be their respective Gini indexes, where the $\widehat{\ }$ indicates that design-based estimated counts based on $\mathcal{S}_B$ are used wherever applicable. For a terminal node l with $\mathcal{D}(\,l\,) \neq \varnothing$, define $\Delta : \mathcal{D}(\,l\,) \longrightarrow \mathbb{R}$ as follows: $\Delta(k,x) := \infty$ if any of the following "boundary scenarios" is encountered:

- $\min\Big\{\, \widehat{\#}\big(\,l_{\{X_k < x\}}\big)\,,\, \widehat{\#}\big(\,l_{\{X_k > x\}}\big)\,,\, \widehat{\#}(\,U\,) \,\Big\} \;=\; 0,$
- any design-based estimated probability occurring within $\widehat{G}\big(\,l_{\{X_k > x\}}\big)$ or $\widehat{G}\big(\,l_{\{X_k > x\}}\big)$ strictly exceeds 1, or
- $\min\big\{\, \#\big(\,\mathcal{S}_A \cap l_{\{X_k < x\}}\,\big)\,,\, \#\big(\,\mathcal{S}_A \cap l_{\{X_k > x\}}\,\big)\,,\, \#\big(\,\mathcal{S}_B \cap l_{\{X_k < x\}}\,\big)\,,\, \#\big(\,\mathcal{S}_B \cap l_{\{X_k > x\}}\,\big) \,\big\} \;<\; s_0\,;$

otherwise, $\Delta(k,x)$ is given by:

$$\Delta(k,x) \;:=\; \frac{\widehat{\#}\big(\,l_{\{X_k < x\}}\big)}{\widehat{\#}(\,U\,)} \cdot \widehat{G}\big(\,l_{\{X_k < x\}}\big) \;+\; \frac{\widehat{\#}\big(\,l_{\{X_k > x\}}\big)}{\widehat{\#}(\,U\,)} \cdot \widehat{G}\big(\,l_{\{X_k > x\}}\big)\,.$$

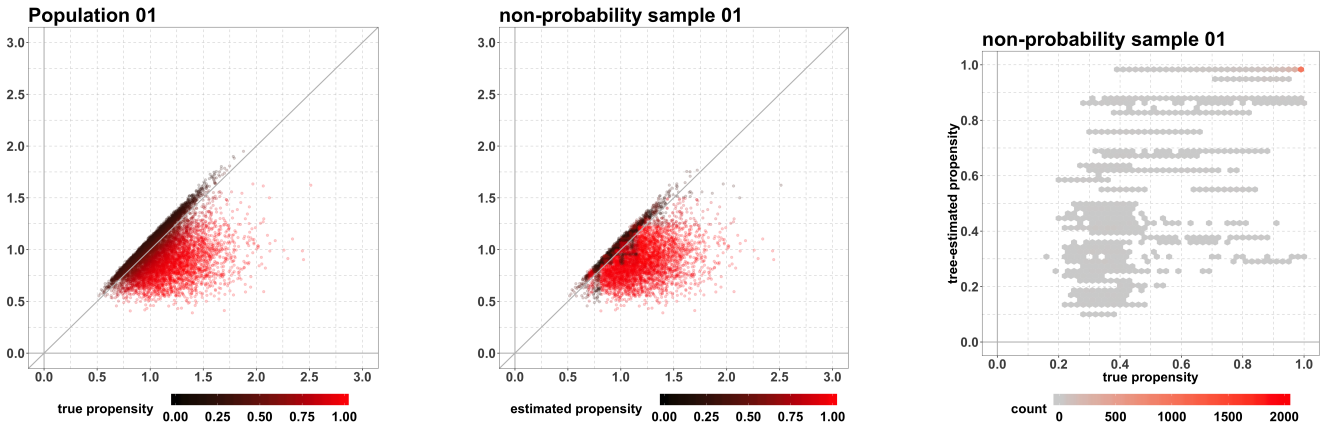
Pseudocode:

1. Define the root node to be all of $\mathcal{S}_A$.
2. For each terminal node l of the current binary tree, check whether the following **Stopping Criterion** is satisfied:

$$\#(\,l \cap \mathcal{S}_A\,) < s_0, \quad \text{or} \quad \widehat{\#}(\,l\,) < \#(\,l \cap \mathcal{S}_A\,), \quad \text{or} \quad \widehat{G}(\,l\,) < g_0, \quad \text{or}$$
$$\mathcal{D}(\,l\,) = \varnothing, \quad \text{or} \quad \min_{(k,x)\,\in\,\mathcal{D}(\,l\,)} \Delta(k,x) = \infty$$

   If so, do nothing further with l (i.e. move on to the next terminal node). Otherwise, fix any one element $(k^*, x^*) \in \operatorname*{argmin}_{(k,x)\,\in\,\mathcal{D}(\,l\,)} \Delta(k,x)$, and replace l with the two children nodes resulting from splitting l by the variable X_{k^*} at the threshold x^*.
3. Repeat Step 2 until each terminal node l satisfies the Stopping Criterion.

Figure (3.1) Simulation Study 1 – Effectiveness of nppCART to reconstruct a logistic propensity



(a) (b) (c)

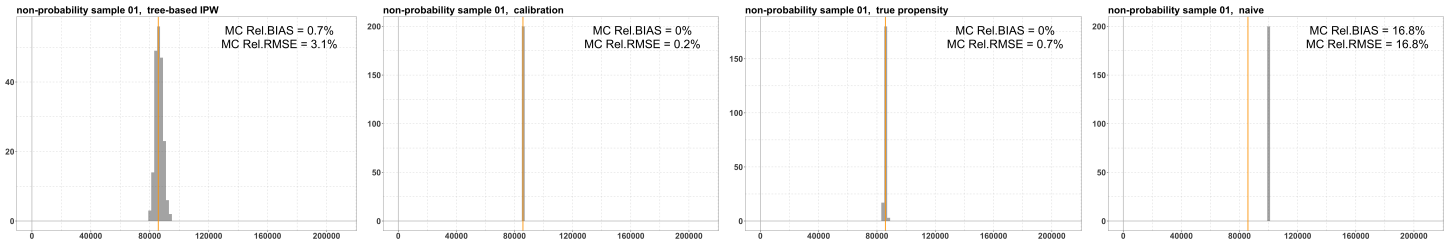
Each unit in the synthetic population is a 4-tuple $(X_1, X_2, Y, \rho) \in \mathbb{R}^4$, generated independently from every other unit. X_1 and X_2 were generated as follows:

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} := \begin{cases} W(Z)\,, & \text{if } W_1 \geq W_2\,, \\ M(\pi/4) \cdot f\Big(M(-\pi/4) \cdot W(Z) \Big)\,, & \text{if } W_1 < W_2\,, \end{cases} \qquad \text{where } W(Z) := \begin{pmatrix} \exp(Z_1/4) \\ \exp(Z_2/4) \end{pmatrix},$$

$Z = (Z_1, Z_2)$, with $Z_1, Z_2 \sim N(0,1)$ being independent standard Gaussian random variables, $M(\theta)$ is the matrix of counterclockwise rotation by the angle θ in the plane, and $f(v_1, v_2) = \Big(v_1\,,\, \frac{3}{10} \cdot \sqrt{v_2/8} \Big)$ is a non-linear compression in the second argument. $Y = 11 + 13 \cdot X_1 - 17 \cdot X_2 + \varepsilon$, where $\varepsilon \sim N(\mu = 0, \sigma = 2)$ is independent of X_1 and X_2. $\rho = (\,1 + \exp[\,-10\,(X_1 - X_2)\,]\,)^{-1}$ is a (deterministic) sigmoid function of X_1 and X_2.

(a) shows the distribution of all population units in the (X_1, X_2)-space, with values of ρ indicated with the red/black gradient. (b) shows only the units in $\mathcal{S}_A$, with the nppCART-generated propensity estimates indicated with the same red/black gradient. (c) is a hexbin plot that compares, for the units in $\mathcal{S}_A$, the true propensities ρ and their respective nppCART-generated estimates.

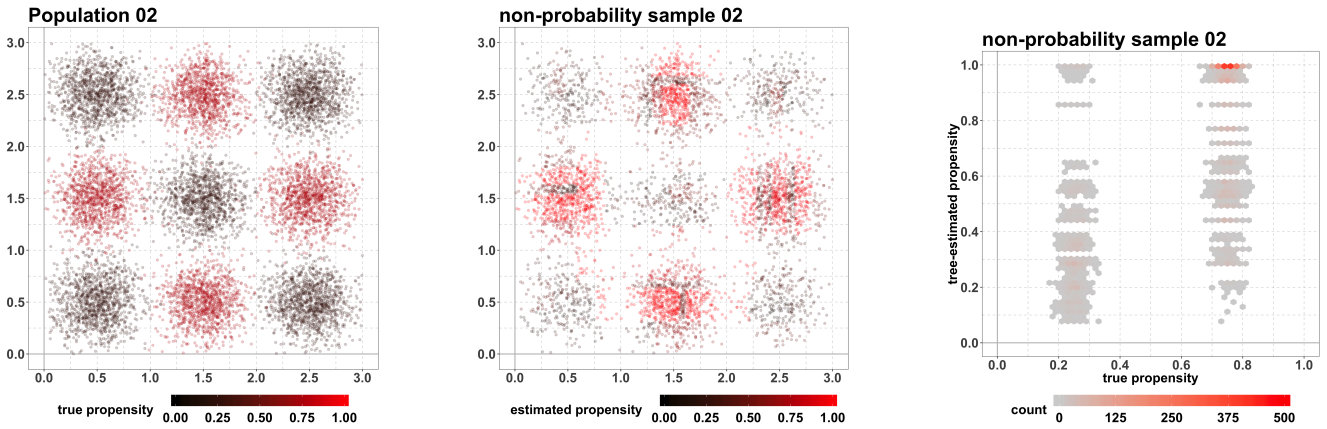
Figure (3.2) Simulation Study 1 – Estimator bias and variance under logistic propensity



(a) (b) (c) (d)

(a) shows the histogram of the 200 resulting TrIPW population total estimates, with the orange vertical line indicating the true value of the population total T_y. The Monte Carlo estimates of the relative bias and root mean squared error (RMSE) are shown near the top right corner. (b), (c) and (d) show respectively the histograms when the calibration estimator, the "estimator" (1.1) and the "naïve" estimator (3.8) were used instead.

Figure (3.3) Simulation Study 2 – Effectiveness of nppCART to reconstruct a nonlinear propensity



(a) (b) (c)

Each unit in the synthetic population is a 4-tuple $(X_1, X_2, Y, \rho) \in \mathbb{R}^4$, generated independently from every other unit, as follows:

$$X_1 \;=\; c_1 + \varepsilon_1\,, \quad X_2 \;=\; c_2 + \varepsilon_2\,, \quad Y \;=\; y_0 + \varepsilon_3^2\,, \quad \rho \;=\; \min\Big\{\,1\,,\,\max\{\,0\,,\,\rho_0 + \varepsilon_4\,\}\,\Big\}\,,$$

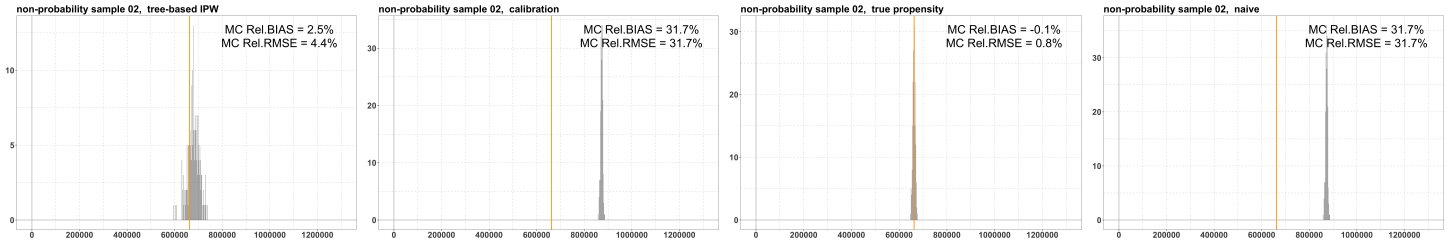
where $c_1, c_2 \sim \text{Uniform}\big(\{\frac{1}{2}, \frac{3}{2}, \frac{5}{2}\}\big)$, $\varepsilon_1, \varepsilon_2 \sim N(\mu = 0, \sigma = 1/5)$, $\varepsilon_3 \sim N(\mu = 0, \sigma = 1)$, $\varepsilon_4 \sim N(\mu = 0, \sigma = 1/40)$, and

$$(y_0, \rho_0) \;=\; \begin{cases} (110, 3/4), & \text{if } c_2 - c_1 \in \{\pm 1\}, \text{ or equivalently, } (c_1, c_2) \in \big\{(\frac{1}{2}, \frac{3}{2}), (\frac{3}{2}, \frac{1}{2}), (\frac{3}{2}, \frac{5}{2}), (\frac{5}{2}, \frac{3}{2})\big\} \\ (\;30, 1/4), & \text{otherwise.} \end{cases}$$

In the above, $c_1, c_2, \varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4$ are all independent. Note the nonlinear relation between Y and (X_1, X_2), as well as the nonlinear relation between ρ and (X_1, X_2).

(a) shows the distribution of all population units in the (X_1, X_2)-space, with values of ρ indicated with the red/black gradient. (b) shows only the units in $\mathcal{S}_A$, with the nppCART-generated propensity estimates indicated with the same red/black gradient. (c) is a hexbin plot that compares, for the units in $\mathcal{S}_A$, the true propensities ρ and their respective nppCART-generated estimates.

Figure (3.4) Simulation Study 2 – Estimator bias and variance under nonlinear propensity



(a) (b) (c) (d)

(a) shows the histogram of the 200 resulting TrIPW population total estimates, with the orange vertical line indicating the true value of the population total T_y. The Monte Carlo estimates of the relative bias and RMSE are shown near the top right corner. (b), (c) and (d) show respectively the histograms when the calibration estimator, the "estimator" (1.1) and the "naïve" estimator (3.8) were used instead.

ACKNOWLEDGEMENTS

We would like to thank Dr. Daniell Toth (Bureau of Labor Statistics, U.S.A.) for helpful discussion on the feasibility of tree-based methods as considered in this article.

DISCLAIMER

The content of this article represents the position of the authors and may not necessarily represent that of Statistics Canada.

REFERENCES

Beaumont, J.-F. (2019). Les enquêtes probabilistes sont-elles vouées à disparaître pour la production de statistiques officielles ? *Survey Methodology*, (Accepted).

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and regression trees*. Wadsworth & Brooks/Cole Advanced Books & Software, 2nd edition.

Chang, W. (2019). *R6: Encapsulated Classes with Reference Semantics*. R package version 2.4.0.

Chen, Y., Li, P., and Wu, C. (2018). Doubly Robust Inference with Non-probability Survey Samples. *arXiv e-prints*, page arXiv:1805.06432.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.