

## Proceedings of Statistics Canada Symposium 2024: The Future of Official Statistics

### Comparison of Recent Techniques of Combining Probability and Non-probability Samples

by Julie Gershunskaya, Vladislav Beresovsky, and Terrance D. Savitsky

Release date: September 8, 2025



Statistics  
Canada    Statistique  
Canada

Canada

## Comparison of Recent Techniques of Combining Probability and Non-probability Samples

Julie Gershunskaya, Vladislav Beresovsky, and Terrance D. Savitsky<sup>1</sup>

### Abstract

Several recent quasi-randomization methods for inferences from non-probability samples will be compared. The considered techniques are developed under the assumption that the sample selection is governed by an underlying latent random mechanism and that it can be uncovered by combining non-probability survey data with a “reference” probability-based sample, obtained from the same target population. Challenges prompting the development of alternative procedures include (i) non-probability sample participation indicators are available only on the observed sample units and (ii) it is not generally known which units from the underlying population belong to both the non-probability and reference samples. The ways different procedures address these challenges are considered, theoretical properties of the methods are discussed and their comparison is made using simulations.

Key Words: Data combining; Non-probability sample; Reference sample; Participation probabilities; Sample likelihood; Variance estimation.

### 1. Introduction

Recent proliferation of powerful computers and the internet created new opportunities for gathering information. Wealth of data can often be obtained quickly, without a well-designed probability based sample. However, naïve estimates from such data may be substantially biased, even when the sample is large. Proper estimation and inferences from such, non-probability (“convenience”) based, samples require modeling assumptions that account for the sample participation mechanism. Existing methodology can be broadly grouped into model-based, quasi-randomization, and doubly-robust approaches. Elliott and Valliant (2017), Valliant (2020), Beaumont and Rao (2021), Rao (2021), Wu (2022) provide a comprehensive account of this rapidly developing field.

In the current paper, we consider several recent quasi-randomization approaches. The considered techniques are developed under the assumption that the sample selection is governed by an underlying latent random mechanism and that it can be uncovered by combining non-probability survey data with a probability based sample, obtained from the same target population. In this context, the probability based sample is often called “reference” sample. The basic requirement is that both the non-probability and reference samples contain a common set of covariates and the assumption is that the non-probability sample participation probabilities are governed by these covariates and could be estimated from a model.

Recent quasi-randomization approaches available in the literature can be broadly classified into pseudo-likelihood based and sample-based approaches. Beresovsky, Gershunskaya and Savitsky (2024) demonstrated that for important practical situations, where the non-probability sample is large and the probability sample is relatively small, sample-based approaches could be substantially more efficient compared with the pseudo-likelihood based methods.

Chu and Beaumont (2019) and Beaumont et al. (2024) considered the use of classification trees in context of combining probability and non-probability samples. They used the pseudo-likelihood approach of Chen, Li and Wu (2020) in constructing of an objective function for finding the optimal split and for estimation of the non-probability

---

<sup>1</sup>Julie Gershunskaya, U.S. Bureau of Labor Statistics, Washington, DC 20212 ([gershunskaya.julie@bls.gov](mailto:gershunskaya.julie@bls.gov)); Vladislav Beresovsky, U.S. Bureau of Labor Statistics, Washington, DC 20212 ([beresovsky.vladislav@bls.gov](mailto:beresovsky.vladislav@bls.gov)); Terrance D. Savitsky, U.S. Bureau of Labor Statistics, Washington, DC 20212 ([savitsky.terrance@bls.gov](mailto:savitsky.terrance@bls.gov))

sample participation probabilities. We propose to use a sample-based likelihood in place of the pseudo-likelihood as a potentially more efficient alternative. We apply the proposed algorithm in a small simulation study to demonstrate the advantage of the sample-based method over the pseudo-likelihood in the case of a small reference sample.

We start by introducing a setup and formulating basic assumptions in Section 2. In Section 3, we briefly review alternative methods and discuss their asymptotic properties. Application of the methods in a Classification Trees algorithm is described in Section 4. Simulation results are given in Section 5. Section 6 is a summary.

## 2. Setup and notation

Suppose, we are interested in estimation of the finite population mean  $\mu = N^{-1} \sum_{i \in U} y_i$  for units in target population  $U$  of size  $N$  based on available non-probability sample  $S_c$  of size  $n_c$ . Sample  $S_c$  contains variables of interest  $y_i$  and a set of covariates  $\mathbf{x}_i$ ,  $i = 1, \dots, N$ .

Let  $I_{ci}$  be the inclusion indicator of population unit  $i$ , taking on the value of 1 if unit  $i$  is included into  $S_c$ , and 0 if it is in  $U \setminus S_c$ . We assume the existence of an underlying random latent selection mechanism and define *unknown* probabilities  $\pi_{ci} = \pi_c(\mathbf{x}_i) = P\{I_{ci} = 1 | i \in U, \mathbf{x}_i\}$  for units in population  $U$  to be included into non-probability sample  $S_c$ . Inclusion probabilities  $\pi_{ci}$  are also often called “participation” probabilities.

Suppose, probability based sample  $S_r$  of size  $n_r$  is also available from the same population  $U$  and the same set of covariates  $\mathbf{x}_i$  as observed in  $S_c$  is also available on sample  $S_r$ . Probability sample  $S_r$  is selected using a *known* probability survey design with inclusion probabilities  $\pi_{ri} = P\{I_{ri} = 1 | i \in U\}$ , where  $I_{ri}$  is the probability sample inclusion indicator of unit  $i$ .

We assume the following about the selection process:

A1: Each unit in the population has a positive participation probability:  $\pi_{ci} > 0$  for all  $i \in U$ .

A2: Sample  $S_c$  selection mechanism is ignorable given  $\mathbf{x}_i$ :  $P\{I_{ci} = 1 | i \in U, \mathbf{x}_i, y_i\} = P\{I_{ci} = 1 | i \in U, \mathbf{x}_i\}$  for all  $i \in U$ .

A3: Indicators  $I_{ci}$  and  $I_{cj}$  are independent given  $\mathbf{x}_i$  and  $\mathbf{x}_j$  for  $i \neq j$ .

A4: Indicators  $I_{ci}$  and  $I_{ri}$  are independent given  $\mathbf{x}_i$  for any  $i \in U$ .

Assumption A1 states that every unit in the target population has a non-zero chance to participate in the non-probability survey; it is the analog of the complete frame coverage in the probability sample case. Assumption A2 is the Missing At Random (MAR) requirement stating that, after accounting for covariates  $\mathbf{x}_i$ , the non-probability sample participation mechanism is ignorable for study variable  $y_i$ . The mutual independence assumption A3 for indicators  $I_c$  allows us to formulate a likelihood for the indicators as independent variables.

Once we obtain the estimates of sample participation probabilities  $\pi_{ci}$ , we are going to use their inverse values in the

usual form of the Hajek estimator of the target population mean:  $\hat{\mu} = \left( \sum_{i \in S_c} 1 / \hat{\pi}_{ci} \right)^{-1} \sum_{i \in S_c} y_i / \hat{\pi}_{ci}$ , a so called Inverse

Propensity Weighted (IPW) estimator.

### 3. A brief review of available approaches

Chen, Li and Wu (2020) considered a Bernoulli likelihood over the finite population units. They present the log-likelihood as a sum of two terms, where the first term is a total over the observed non-probability sample and the second term is a finite population total. Since the finite population is not available, a *pseudo-likelihood* formulation is employed, where the finite population total is replaced by its probability survey weighted estimate:

$$\hat{\ell}^{CLW}(\boldsymbol{\beta}) = \sum_{i \in \mathcal{S}_c} \log \left[ \frac{\pi_{ci}(\boldsymbol{\beta})}{1 - \pi_{ci}(\boldsymbol{\beta})} \right] + \sum_{i \in \mathcal{S}_r} w_{ri} \log [1 - \pi_{ci}(\boldsymbol{\beta})], \quad (1)$$

where  $\boldsymbol{\beta}$  is the parameter vector in a logistic regression model  $\text{logit}[\pi_{ci}(\boldsymbol{\beta})] = \boldsymbol{\beta}^T \mathbf{x}_i$  and  $w_{ri} = \pi_{ri}^{-1}$ . Estimates are obtained by solving respective pseudo-likelihood based estimating equations.

We refer to the pseudo-likelihood method of Chen, Li and Wu (2020) as “direct” because the Bernoulli likelihood formulated for non-probability sample indicators,  $I_{ci}$ , leads directly to the estimation of non-probability sample participation probabilities of interest,  $\pi_{ci}$ . By contrast, there are several methods considered in the literature where Bernoulli likelihoods are formulated for different indicator variables that are related to the parameters of interest in particular “indirect” ways. These methods include a pseudo-likelihood approach of Wang, Valliant and Li (2021) and sample based approaches of Elliot (2009) and Savitsky et al. (2023). We refer to this group of methods as “indirect” approaches.

Elliot (2009) proposed to consider indicator variable  $I_{zi}$  (in our notation) defined on the combined probability and non-probability samples set  $\mathcal{S}_c \cup \mathcal{S}_r$ , where  $I_{zi} = 1$  for  $i \in \mathcal{S}_c$ , and  $I_{zi} = 0$  for units that *do not* belong to the non-probability sample,  $i \in (\mathcal{S}_c \cup \mathcal{S}_r) \setminus \mathcal{S}_c$ . There is a nonzero chance that units from the finite population may be drawn into both the non-probability and reference samples,  $\mathcal{S}_c \cap \mathcal{S}_r$ . If the identity of units belonging to the sampling overlap is known, it is easy to show (see, for example, Beresovsky (2019), Beaumont et al. (2024), Beresovsky, Gershunskaya and Savitsky (2024)) that the following relationship holds that links the probability of being in the non-probability sample part of the combined set,  $\pi_{zi} = P\{i \in \mathcal{S}_c \mid i \in \mathcal{S}_c \cup \mathcal{S}_r\}$ , and respective samples inclusion probabilities,  $\pi_{ci}$  and  $\pi_{ri}$ :

$$\pi_{zi} = \frac{\pi_{ci}}{\pi_{ci} + \pi_{ri} - \pi_{ci}\pi_{ri}}. \quad (2)$$

The challenge is that it is not generally known which units in the combined set  $\mathcal{S}_c \cup \mathcal{S}_r$  belong to both samples  $\mathcal{S}_c$  and  $\mathcal{S}_r$ . Elliot (2009) assumes that the probability of belonging to the sampling overlap is *negligible* for all units in the finite population. Under this assumption, an *approximate* relationship holds:

$$\pi_{zi} = \frac{\pi_{ci}}{\pi_{ci} + \pi_{ri}}. \quad (3)$$

However, the negligible overlap requirement of the Elliot (2009) method led to a common belief that the combined samples approach is limited to cases where sampling fractions are small. This is unfortunate, because in many practical situations the non-probability sample is large, and the sampling overlap could be significant.

Savitsky et al. (2023) showed that, under a slightly different setup, relationship (3) *holds exactly*, regardless of the samples sizes: instead of the union of two samples, the authors consider a *stack* of the samples,  $\mathcal{S} = \mathcal{S}_c + \mathcal{S}_r$  and, correspondingly, amend the definition of indicator  $I_{zi}$  as taking the value of 1 for all units in non-probability sample

$S_c$ , and  $\mathbf{0}$  for all units in reference sample  $S_r$ ; units belonging to both samples would be included into the stacked sample twice: once with  $I_{zi} = 1$  and once with  $I_{zi} = \mathbf{0}$  (there is no need to know the identity of these units.)

Beresovsky, Gershunskaya and Savitsky (2024) proved that indicators  $I_{zi}, I_{zj}$  for units  $i$  and  $j$  on the stacked set are *nearly* independent, with the correlation as low as  $O(N^{-1})$ , regardless of how large the overlap is, thus motivating the use of the Bernoulli likelihood for independent random variables over the stacked set.

Wang, Valliant and Li (2021) proposed an Adjusted Logistic Propensity (ALP) weighting method where they use an imaginary construct stacking together the non-probability sample (part 1) and the target finite population (part 2). They formulated a Bernoulli likelihood for an indicator variable defined on the proposed stacked set as  $\tilde{I}_{zi} = 1$  for units in part 1 and  $\tilde{I}_{zi} = 0$  for units in part 2, and used a pseudo-likelihood approach by replacing the part involving the finite population units by its probability-based survey weighted estimate:

$$\hat{\ell}^{ALP}(\boldsymbol{\gamma}) = \sum_{i \in S_c} \log[\tilde{\pi}_{zi}(\boldsymbol{\gamma})] + \sum_{i \in S_r} w_{ri} \log[1 - \tilde{\pi}_{zi}(\boldsymbol{\gamma})], \quad (4)$$

where  $\boldsymbol{\gamma}$  is the parameter vector in a logistic regression model  $\text{logit}[\tilde{\pi}_{zi}(\boldsymbol{\gamma})] = \boldsymbol{\gamma}^T \mathbf{x}_i$ .

The relationship between the probability of being in part 1 of the stacked construct and the non-probability sample participation probabilities is

$$\tilde{\pi}_{zi} = \frac{\pi_{ci}}{\pi_{ci} + 1}, \quad (5)$$

which is a particular case of (3), when the whole finite population is available instead of just a sample.

As noted, methods relying on respective identities (2), (3), or (5) are “indirect” in the sense that the respective Bernoulli parameters are probabilities of being in a particular part of the combined set, rather than parameters of interest  $\pi_{ci}$ .

Beresovsky (2019) noted that the estimation of  $\pi_{ci}$  can be performed in a single step if  $\pi_{zi}$  are viewed as a composite function  $\pi_{zi}(\boldsymbol{\pi}_{ci}(\boldsymbol{\beta})) = \pi_{ci}(\boldsymbol{\beta}) / (\pi_{ci}(\boldsymbol{\beta}) + \pi_{ri})$  in log-likelihood

$$\ell^{ILR}(\boldsymbol{\beta}) = \sum_{i \in S_c} \log[\pi_{zi}(\boldsymbol{\pi}_{ci}(\boldsymbol{\beta}))] + \sum_{i \in S_r} \log[1 - \pi_{zi}(\boldsymbol{\pi}_{ci}(\boldsymbol{\beta}))] \quad (6)$$

and called the approach “Implicit Logistic Regression” (ILR) (see also Savitsky et al. (2023) and Beresovsky, Gershunskaya and Savitsky (2024)).

By contrast, Elliott (2009) and Wang, Valliant and Li (2021) apply a two-step procedure: at step 1, they find the estimates of parameters of the standard logistic regression model; at step 2, they use respective identities to obtain estimates of the parameters of interest,  $\pi_{ci}$ . Such a two-step approach, however, may lead to the loss of efficiency and estimates of  $\pi_{ci}$  falling outside of the parameter space.

Beresovsky, Gershunskaya and Savitsky (2024) compared asymptotic properties of the maximum likelihood estimates based on: (1) the “direct” pseudo-likelihood method of Chen, Li and Wu (2020) (CLW), (2) the single-step modification of the “indirect” pseudo-likelihood of Wang, Valliant and Li (2021) (pseudo-ILR, or PILR), and (3) the “indirect” stacked samples based method (ILR).

They found that the three methods are asymptotically equivalent. However, it appears that the ILR approach is less sensitive to the size of the probability sample, relative to the pseudo-likelihood based CLW or PILR approaches.

## 4. Classification Trees

Chu and Beaumont (2019) and Beaumont et al. (2024) adapted Classification and Regression Trees (CART) algorithm in context of combining probability and non-probability based samples. They use the CLW pseudo-likelihood as an objective function for estimation of the probabilities in the left and right branches for a given split and for finding optimal splits. We propose to use the stacked sample-based likelihood for this purpose.

Classification Trees algorithm splits data into homogeneous groups  $g = 1, \dots, G$ , so that all units in a given group have the same probabilities  $\pi_{cg}$ . It works recursively by finding an optimal binary split at a given step for a given covariate, as follows:

- 1) compute participation probabilities in the left and right branches,  $\pi_{cgL}$  and  $\pi_{cgR}$ , on a grid of possible binary splits;
- 2) choose an optimal split based on an objective function.

Under the CLW pseudo-likelihood, the objective function is

$$\hat{I}^{CLW} = - \sum_{g=1}^G \frac{\hat{N}_g}{\hat{N}} \left[ \hat{\pi}_{cg} \log(\hat{\pi}_{cg}) + (1 - \hat{\pi}_{cg}) \log(1 - \hat{\pi}_{cg}) \right] \quad (7)$$

and the estimate of participation probability for the homogeneous group  $g$  is

$$\hat{\pi}_{cg} = \frac{n_{cg}}{\hat{N}_g}, \text{ where } \hat{N}_g = \sum_{i \in g} w_{ri}. \quad (8)$$

For the sample-based (ILR) approach, the objective function is

$$\hat{I}^{ILR} = - \sum_{g=1}^G \left[ \sum_{i \in S_{cg}} \log \hat{\pi}_{zi,g} + \sum_{i \in S_{rg}} \log(1 - \hat{\pi}_{zi,g}) \right], \quad (9)$$

where  $\hat{\pi}_{zi,g} = \hat{\pi}_{cg} / (\pi_{ri} + \hat{\pi}_{cg})$ .

There is no explicit expression for the estimate of  $\pi_{cg}$  under the ILR approach. It can be found as a solution to equation

$$\sum_{i \in S_g} \frac{\pi_{ri}}{\pi_{ri} + \pi_{cg}} = n_{rg}, \quad (10)$$

where  $S_g$  is the part of stacked sample  $S$  belonging to group  $g$ .

Note that although  $\hat{\pi}_{cg}$  in a given homogeneous group is a constant, estimates  $\hat{\pi}_{zi,g}$  in a given group are allowed to vary.

(As noted in Beaumont et al. (2024), estimates of participation probabilities  $\hat{\pi}_{cg}$  for homogeneous groups based on the PILR approach are the same as the CLW based estimates given by (8), although the objective functions based on the CLW and PILR likelihoods differ. We do not include the PILR method in the simulation study of Section 5.)

## 5. Simulations

We performed a small simulation study, where we suppose to already have a grown tree, up to some level  $g$ . We focus on splitting node  $g$  into two parts.

The simulation setup is as follows. Let  $N_g = 1,000$  be the population size at node  $g$ . We generate a single covariate  $x_{ig}$  from the standard normal distribution:  $x_{ig} \sim N(0,1)$ . Let the study variable be  $y_{ig} = 1 + x_{ig} + \delta_{ig}$ , with  $\delta_{ig} \sim N(0,1.5^2)$ .

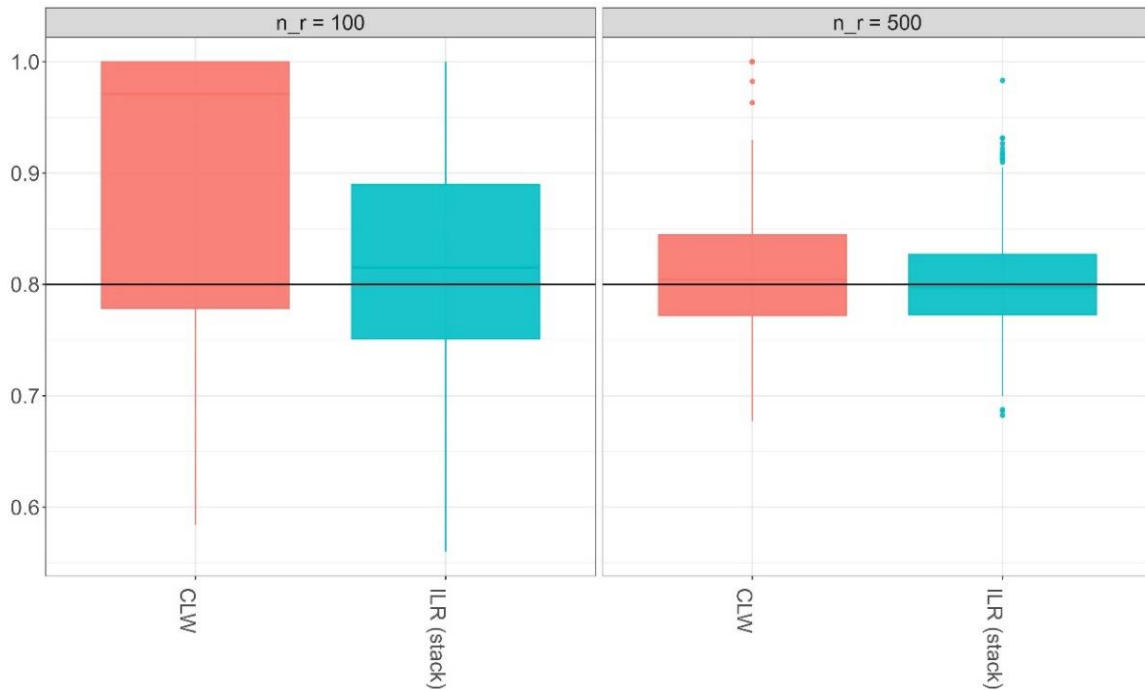
We set true non-probability sample participation probabilities to  $\pi_{cg,L} = 0.80$  for units  $i$  with  $x_{ig} \leq 0$  and  $\pi_{cg,R} = 0.20$  for units  $i$  with  $x_{ig} > 0$ . This setup yields the non-probability sample of approximately  $n_{cg} \approx 500$  units.

We use the probability proportional to size design to generate the probability sample, where inclusion probabilities  $\pi_{rg,i}$  are proportional to  $x_{ig}$ . We consider two simulation scenarios: (1) smaller sample of size  $n_{rg} = 100$  and (2) larger sample of size  $n_{rg} = 500$ .

Results after 1000 simulation runs are presented in Figures 5-1 and 5-2. In Table 5-1, we present results for corresponding estimates of population mean  $\mu$ .

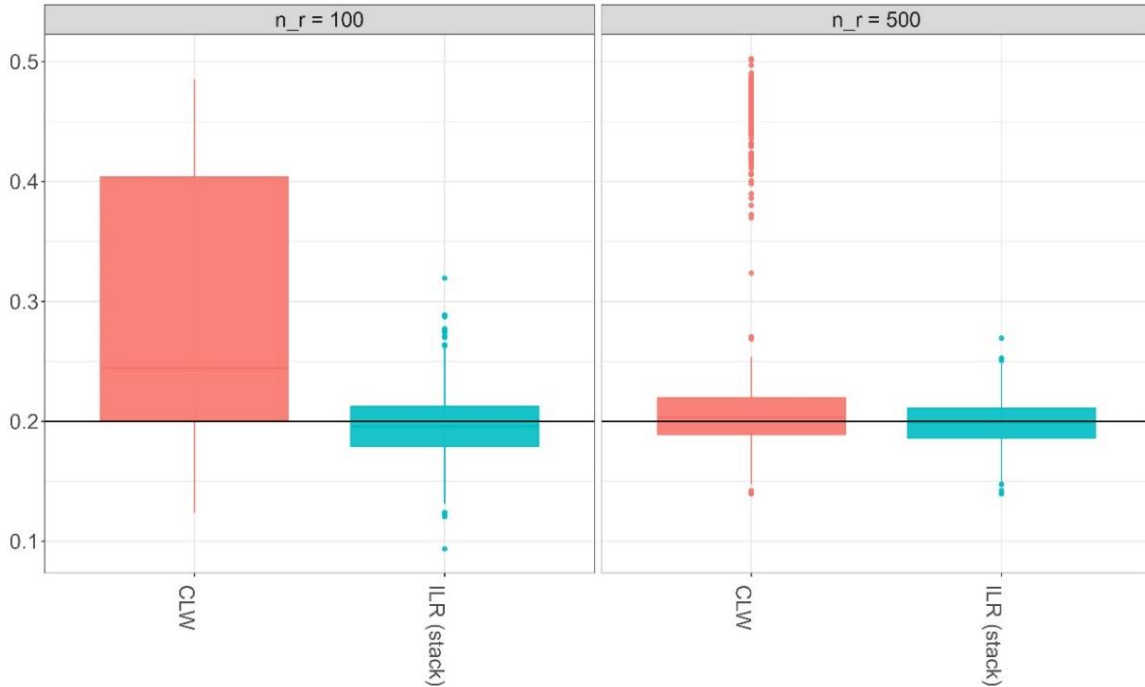
**Figure 5-1**

**Estimates of the "left branch" probabilities, for true  $\pi_{cg,R} = 0.80$ , for smaller ( $n_{rg} = 100$ ) and larger ( $n_{rg} = 500$ ) sample size scenarios**



**Figure 5-2**

Estimates of the "right branch" probabilities, for true  $\pi_{cg,L} = 0.20$ , for smaller ( $n_{rg} = 100$ ) and larger ( $n_{rg} = 500$ ) sample size scenarios



**Table 5-1**

Bias and MSE of estimates of population mean, using unweighted estimator and weights based on CLW and ILR methods

	$n_r = 100$		$n_r = 500$	
	Bias	MSE	Bias	MSE
Unweighted	-0.481	0.234	-0.481	0.234
CLW	-0.128	0.042	-0.047	0.023
ILR	0.012	0.011	0.001	0.007

## 6. Summary

Past research shows that the stacked-samples based approach gives more efficient estimates of participation probabilities as compared to the pseudo-likelihood approaches under many practical scenarios, especially when the probability-based sample is small and there is low overlap in covariates defined domains between the two samples. Therefore, we expected that using this method in conjunction with the regression trees algorithm may be especially beneficial. Our simulations seem to confirm this conjecture.

## References

Beaumont, J.-F. and Rao, J. N. K. (2021), "Pitfalls of making inferences from non-probability samples: Can data integration through probability samples provide remedies", *The Survey Statistician*, 83, pp. 11 – 22.

- 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", *Survey Methodology*, 50, pp. 77–106.
- Beresovsky, V. (2019), "On application of a response propensity model to estimation from web samples", In ResearchGate.
- Beresovsky, V., Gershunskaya, J. and Savitsky, T. D. (2024), "Review of Quasi-Randomization Approaches for Estimation from Non-probability Samples", <https://arxiv.org/abs/2312.05383>.
- Chen, Y., Li, P. and Wu, C. (2020), "Doubly Robust Inference With Nonprobability Survey Samples", *Journal of the American Statistical Association*, 115, pp. 2011-2021.
- Chu, K. and Beaumont, J. F. (2019), "The use of classification trees to reduce selection bias for a nonprobability sample with help from a probability sample", In *Proceedings of the Survey Methods Section, Statistical Society of Canada*.
- Elliott, M. R. (2009), "Combining Data from Probability and Non-Probability Samples Using Pseudo-Weights", *Survey Practice*, 2, pp. 813–845.
- Elliott, M. R. and Valliant, R. (2017), "Inference for Nonprobability Samples", *Statistical Science*, 32, pp. 249 – 264.
- Rao, J. N. K. (2021), "On Making Valid Inferences by Integrating Data from Surveys and Other Sources", *Sankhya*, 83, pp. 242 – 272.
- Savitsky, T. D., Williams, M. R., Gershunskaya, J. and Beresovsky, V. (2023), "Methods for combining probability and nonprobability samples under unknown overlaps", *Statistics in Transition, New Series*, 24, pp. 1 - 34.
- Valliant, R. (2020), "Comparing Alternatives for Estimation from Nonprobability Samples", *Journal of Survey Statistics and Methodology*, 8, pp. 231 – 263.
- Wang, L., Valliant, R. and Li, Y. (2021), "Adjusted logistic propensity weighting methods for population inference using nonprobability volunteer-based epidemiologic cohorts", *Stat Med.*, 40, pp. 5237–5250.
- Wu, C. (2022), "A new criterion for confounder selection", *Survey Methodology*, 48, pp. 283 - 311.