

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

# Propensity Score Estimation and Optimal Sampling Design when Integrating Probability Samples with Non-probability Data

by Anders Holmberg, Lyndon Ang, Robert Clark and Bronwyn Loong

Release date: September 8, 2025



Statistics  
Canada Statistique  
Canada

Canada

# Propensity Score Estimation and Optimal Sampling Design when Integrating Probability Samples with Non-probability Data<sup>1</sup>

Anders Holmberg, Lyndon Ang, Robert Clark and Bronwyn Loong<sup>2</sup>

## Abstract

Although non-probability data sources are not new to official statistics, a revived interest in the topic has emerged from pressures due to falling survey response rates, increasing data collection costs and a desire to take advantage of new data source opportunities from the ongoing societal digitalisation. Due to the exclusion of certain segments of the target population, inference derived solely from a non-probability data source is likely to result in bias. This work approaches the challenge of addressing the bias by integrating non-probability data with reference probability samples. The focus will be on methods to model the propensity of inclusion in the non-probability dataset with the help of the accompanying reference sample, with the modelled propensities then applied in an inverse probability weighting approach to produce population estimates. The reference sample is sometimes assumed as given. In this presentation however, an objective of finding an optimal strategy will be pursued that is, the combination of a data integration-based estimator and sample design for the reference probability sample. Recent work is discussed in which advantage is taken of the good unit identification possibilities in business surveys to study an estimator based on propensities and derive optimal (unequal) selection probabilities for the reference sample.

Key Words: Data integration; Selection bias; Optimal selection probabilities; Big data; Inverse probability weighting.

## 1. Introduction

### 1.1 Background and context

One of the goals of national statistical organisations (NSOs) is to provide trusted official statistics about the “economic, demographic, social and environmental situation” for informed decision making (United Nations Statistics Division 2014). The data used by NSOs may be drawn from all types of sources, depending on the required levels of quality, timeliness, cost, and respondent burden.

The ongoing digitalisation of society has spawned an increasing amount of data collected about individuals, entities and the environment. These new data sources include “big data”, for example from sensors, satellites, and administrative systems, which carry potential for producing more efficient statistics. Some examples from the Australian Bureau of Statistics (ABS) include: employee payroll data submitted by employers to the Australian Taxation Office (ATO), used to help produce wages and jobs statistics (Australian Bureau of Statistics, 2024b, 2024d), bank transactions data to assist in producing statistics on household spending (Australian Bureau of Statistics 2024c), and agricultural levies data to produce statistics on agricultural production (Australian Bureau of Statistics 2024a).

Big data is usually generated for non-statistical purposes with data acquisition instruments outside the control of the statistical offices. Consequently, the data may suffer from selection bias where specific groups in the target population are less likely to be included in the dataset. Although they may be large, when these datasets are used directly without

---

<sup>1</sup> The views expressed in this paper are those of the authors and do not necessarily represent the views of the Australian Bureau of Statistics. Where quoted or used, they should be attributed clearly to the authors.

<sup>2</sup> Anders Holmberg, Australian Bureau of Statistics (ABS), Belconnen, ACT, 2617, Australia (anders.holmberg@abs.gov.au); Lyndon Ang, ABS and Australian National University, Canberra, ACT, 2600, Australia; Robert Clark, Australian National University, Canberra, ACT, 2600, Australia; Bronwyn Loong, Australian National University, Canberra, 2600, ACT, Australia

adjustment to estimate population parameters such as totals and means, biased estimates are likely to be produced (see, for example, Meng 2018).

In this paper we discuss quasi-randomization approaches to correct selection bias in nonprobability data, and in particular big data. These methods assume that the missingness in the non-probability data is governed by an unknown but discoverable selection mechanism. The idea is to use available auxiliary information from the population (see for example, Burakauskaitė & Čiginas 2023) or from a supplementary reference probability sample (see for example, Chen et al. 2020 and Wang et al. 2021) to estimate propensities of selection into the non-probability dataset. The inverse of these propensities are then used as weights to “weight up” the non-probability data and estimate finite population quantities.

The reference samples are generally assumed to be already drawn; these may be existing surveys undertaken for a different purpose, but which collect the relevant auxiliary information. We pose the question “Is it sensible to design the reference sample more efficiently for the non-probability data?”

In some scenarios, it may be possible to identify which units in the reference sample are also present on the non-probability dataset. This would occur, for example, if a common unit identifier exists and is available on both the non-probability dataset and the reference sample. We explore possibilities for taking advantage of the ability to link reference and non-probability sample units during the sample design and estimation stages.

The rest of the paper is structured as follows. We outline the basic setup for the paper in Section 2. In Section 3 we discuss the optimal allocation of sample as part of the design of the reference sample. Section 4 outlines a propensity score estimator which takes advantage of the ability to identify which units in the reference sample are also on the non-probability dataset. Section 5 examines the performance of this estimator coupled with the design ideas in Section 3 through a simulation study. Section 6 provides some concluding remarks.

## 2. Basic Setup

Let  $U$  be a finite population of size  $N$ . For each unit  $i$  in the population we have values  $(\mathbf{x}_i, y_i, \mathbf{z}_i)^T$  for a data item of interest  $y$ , some auxiliary data items  $\mathbf{x}$ , and some additional variables  $\mathbf{z}$ . In this paper, we are interested in estimating the population mean  $\bar{Y} = \mu_y = N^{-1} \sum_{i=1}^N y_i$ .

Suppose we have a probability sample  $A$  (which we will call the “reference” sample) of size  $n_A$  drawn from the population.  $A$  collects information for  $\mathbf{x}$  but not  $y$ .<sup>3</sup> Define  $\pi_i^A = P(i \in A|U)$  to be the inclusion probability for unit  $i$  being in the probability sample, and  $d_i^A = 1/\pi_i^A$  as the survey design weight for  $i \in A$ . The  $\pi_i^A$  and  $d_i^A$  values are known from the probability sample design and may depend on some of the  $\mathbf{z}$  variables, for example via stratification of the population.

We also have a non-probability sample  $B$ .  $B$  contains information for  $\mathbf{x}$  as well as  $y$ . Let  $\delta_i = I(i \in B)$  be an indicator variable for unit  $i$  being included in  $B$ . The non-probability sample size is  $N_B = \sum_{i=1}^N \delta_i$ . In contrast to  $\pi_i^A$ , the inclusion probabilities  $\pi_i^B = P(\delta_i = 1|U)$  are unknown and need to be estimated. We define  $C = U \setminus B$  to be the units in the population not included in  $B$ . There may or may not be overlap between the two samples  $A$  and  $B$ .

We make some assumptions regarding the selection mechanism for units being included in  $B$ , to facilitate forming inferences using those datasets; see for example Chen et al. (2020) and Yang et al. (2021). These assumptions are:

**A1 Ignorability:** Conditional on the set of covariates  $\mathbf{x}_i$ ,  $\delta_i$  and  $y_i$  are independent.

**A2 Positivity:** Conditional on  $\mathbf{x}_i$ ,  $P(\delta_i = 1|\mathbf{x}_i) > 0, \forall i \in U$ .

**A3 Independence:** Conditional on  $\mathbf{x}_i$  and  $\mathbf{x}_j$ ,  $\delta_i \perp \delta_j, i \neq j, \forall i, j \in U$ .

---

<sup>3</sup> The ABS have agricultural (satellite) and accounting data applications where it is unfeasible to collect the same  $y$  as in the non-probability data source.

Ignorability implies that  $P(\delta_i = 1|\mathbf{x}, y) = P(\delta_i = 1|\mathbf{x})$ . In other words, selection into  $B$  is ignorable conditional on the covariates  $\mathbf{x}$ . This assumption is similar to the Missing-At-Random (MAR) scenario of Rubin (1976). Andridge et al. (2019) and Boonstra et al. (2021) refer to this type of selection process as Selection At Random (SAR).

The positivity assumption states that conditional on  $\mathbf{x}$  every unit in the population has a non-zero chance of inclusion in  $B$ . This may not always hold - for example, in a big dataset generated by a social media platform, only those persons who are members of the platform have a chance of inclusion in the big dataset.

In addition to the above, we assume we can link units in  $B$  to either  $A$  or the population frame. This assumption may be satisfied if there exists a common unit identifier available on  $B$  and the population frame, or if it is possible to probabilistically link the records using a common set of linking variables. On occasion, we will relax this assumption so that it is sufficient to identify which units in  $A$  are also in  $B$ . This relaxed assumption may be met by including a question in the survey instrument for  $A$  which asks whether the respondent is also on the dataset  $B$ .

We wish to estimate the unknown  $\pi_i^B$ , which in this paper are assumed to follow a parametric model so that  $\pi_i^B = \pi(\mathbf{x}_i, \boldsymbol{\theta}_0)$ , where  $\boldsymbol{\theta}_0$  are the true values of the unknown model parameters. Once final values  $\hat{\boldsymbol{\theta}}$  have been obtained,  $\hat{\pi}_i^B$  can be calculated and fed into an Inverse Probability Weighted (IPW) estimator such as the Hájek-like estimator (Chen et al. 2020)

$$\hat{\mu}_{IPW} = \frac{1}{N_B} \sum_{i \in B} \frac{y_i}{\hat{\pi}_i^B} \quad (1)$$

Estimates of variance can be calculated for  $\hat{\mu}_{IPW}$ , for example by using the asymptotic framework developed by Chen et al. (2020).

### 3. Optimal Design for the Reference Sample

#### 3.1 Re-Visioning the Reference Sample as a Supplement to Address Selection Bias

The usual purpose of a probability survey is to collect information about specific topic(s) of interest to the survey statistician (see for example Chapter 1 of Cochran 1977). It will generally be designed to provide efficient estimates for (a subset of) key variables interest, and in our context, these will be separate to the variables collected on the non-probability dataset.

We suggest that there may be opportunity to efficiently design the reference sample  $A$  to specifically achieve an optimal result (lowest cost for a given variance, or vice versa) for the estimate  $\hat{\mu}_{IPW}$  produced using  $B$ . Some scenarios when this might be attractive include:

- An existing survey  $A$  collects certain information about a population which a new, large but incomplete administrative dataset also includes. To reduce respondent burden, we can stop collecting  $y$  in  $A$  and instead rely on  $B$ . The sample  $A$  can then be re-purposed to provide the characteristic information  $\mathbf{x}$  to be used in addressing selection bias.
- We would like to take advantage of a newly available data source  $B$  but do not have sufficient auxiliary information for  $\mathbf{x}$  to help in correcting selection bias. In this case, we might choose to develop a probability survey  $A$  to assist in collection of  $\mathbf{x}$ .

A reasonable question is: “If we are designing the reference survey to enable statistics to be produced from  $y$ , why not just collect  $y$  in the survey?”. Indeed if only a small number of extra data items need to be collected then the additional cost of doing so could be minimal. However, if any additional data items are so substantial that significant increases in costs and respondent burden are introduced, then there is a savings trade-off to be made. Sometimes some variables are also practically unfeasible to collect in  $A$  because of impossible collection design, cost to collect the data items and/or response burden.

The savings increase further if  $A$  can be used to help address selection bias for more than one non-probability dataset. This leads to the potential for the NSO to reduce the number of surveys they run. Before, the NSO may have conducted a suite of surveys, each of which collects data on a particular set of topics. Now, the NSO would rely more heavily on a suite of non-probability data, supplemented by auxiliary information from a ‘‘Population Profile Survey’’  $A$  to help address selection bias issues.

### 3.2 Optimal Sample Design using $\hat{\boldsymbol{x}}$

The variance of the IPW estimator  $\hat{\mu}_{IPW}$  generally depends on several factors. This includes the model that is used for  $\pi_i^B$  (for example, a logistic model), and the sampling scheme used for  $A$  (for example, Poisson sampling). For a given variance expression for the IPW estimator, it is possible to derive optimal selection probabilities. These probabilities will tend to require knowledge of  $\boldsymbol{x}_i$  for  $i \in U$ , and the optimal probabilities will be proportional to a function of the  $\boldsymbol{x}$  variables. It may not always be the case that we have information for all  $\boldsymbol{x}$ . We may only have information available at the population level for some variables  $\boldsymbol{z}$ , for example frame variables available for the sample design.

How might we implement an optimal sample design in practice when  $\boldsymbol{x}$  may not be available for the whole population, or only partially available? We adopt the proposal of Clark & Steel (2022) to minimise the anticipated variance of  $\hat{\mu}_{IPW}$ , conditional on the available information  $\boldsymbol{z}$ . In practice, the following steps may be followed:

1. Use available records  $i$  which have both  $\boldsymbol{x}_i$  and  $\boldsymbol{z}_i$  to impute  $\hat{\boldsymbol{x}}_i$  for the rest of the population
2. Produce  $\hat{\pi}_i^{A,opt}$  using  $\boldsymbol{x}_i$  and, where needed, the imputed  $\hat{\boldsymbol{x}}_i$
3. Repeat Steps 1 and 2  $m$  times
4. Calculate the mean of the  $m$  values of  $\hat{\pi}_i^{A,opt}$  and use that in the sample design

One approach for creating the estimated values would be to apply the  $k$  Nearest Neighbour (kNN) method to mass impute values  $\hat{\boldsymbol{x}}_i$  where they are missing, using  $\boldsymbol{z}_i$  as the distance measure. The idea of mass imputation using kNN has been explored previously, see for example Rivers (2007) and Yang et al. (2021). An alternative which may be less computationally intensive would be to use hot deck imputation, where the hot deck classes are devised based on  $\boldsymbol{z}$ .

Data for the imputation may be obtained from an existing survey  $A$  which collects  $\boldsymbol{x}_i$ , or from a pilot study. Additionally, if we can link records in  $B$  to the population frame, then the linked dataset will contain both  $\boldsymbol{x}_i$  and  $\boldsymbol{z}_i$ ,  $i \in B$ . The linked dataset may then supply  $\boldsymbol{x}_i$  for  $i \in B$ , and imputed values  $\hat{\boldsymbol{x}}_i$  for  $i \in U \setminus B$  can be obtained from the existing survey or pilot study.

## 4. Overlap-Excluded Propensity Score Estimation

Recall that we wish to estimate unknown propensities of selection into a non-probability dataset  $\pi_i^B$ , and these propensities are assumed to i) satisfy SAR missingness, and ii) follow a parametric model with parameters  $\boldsymbol{\theta}_0$  which need to be estimated. The population likelihood function of  $\pi_i^B = \pi(\boldsymbol{x}_i; \boldsymbol{\theta}_0)$  is

$$L(\boldsymbol{\theta}) = \prod_{i \in U} (\pi_i^B)^{\delta_i} (1 - \pi_i^B)^{(1 - \delta_i)} \quad (2)$$

and the corresponding log-likelihood function is

$$l(\boldsymbol{\theta}) = \sum_{i \in U} \delta_i \log \pi_i^B + \sum_{i \in U} (1 - \delta_i) \log(1 - \pi_i^B) \quad (3)$$

The two components in (3) can be estimated in various ways. We assume that it is possible to identify membership in  $B$  for each unit in the reference sample  $A$ . Note that the first component in (3) can be calculated directly using  $B$ . The second component can be estimated using  $A \setminus B$ , that is, by the sample in  $A$  that are not also in  $B$ . The resulting pseudo-likelihood is given by

$$l_{OE}(\boldsymbol{\theta}) = \sum_{i \in B} \log \pi_i^B + \sum_{i \in A \setminus B} d_i^A \log(1 - \pi_i^B) \quad (4)$$

In this paper, we term this the Overlap-Excluded (OE) pseudo-likelihood function. Under a logistic model for the propensity scores,  $\pi_i^B = \exp(\mathbf{x}_i^T \boldsymbol{\theta}_0) / (1 + \exp(\mathbf{x}_i^T \boldsymbol{\theta}_0))$ . Then the pseudo-likelihood function (4) becomes

$$l_{OE}(\boldsymbol{\theta}) = \sum_{i \in B} \{\mathbf{x}_i^T \boldsymbol{\theta} - \log(1 + \exp(\mathbf{x}_i^T \boldsymbol{\theta}))\} - \sum_{i \in A \setminus B} d_i^A \log(1 + \exp(\mathbf{x}_i^T \boldsymbol{\theta})) \quad (5)$$

The score function  $S(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} l_{OE}(\boldsymbol{\theta})$  can be derived, and the maximum pseudo-likelihood estimator  $\hat{\boldsymbol{\theta}}$  can be found numerically by applying the Newton-Raphson iterative procedure to solve  $S(\boldsymbol{\theta}) = 0$ .

Further information on the OE approach, including details of a plug-in variance estimator for the resulting IPW estimator (1), can be found in Ang et al. (2024).

## 4.1 Split-Population Design for the Reference Sample

In the previous section, we assumed it was possible to identify which units in  $A$  are also in  $B$ . In this section, we make a stronger assumption that it is possible to identify which units on the *population frame* are not in  $B$ . This could be possible, for example, when a common unit identifier exists on the frame and on  $B$ . If this is the case, we may decide to design the reference sample  $A$  based only on the sub-population  $U \setminus B$ . This approach has similarities to a ‘‘split-population’’ scenario (see Zhang 2019) where  $U$  is divided into an incomplete ‘‘frame’’  $B$  and  $C = U \setminus B$ ,  $B$  is fully sampled and  $A$  is a sample from  $C$ .

Under the split-population design, the pseudo-likelihood (4) changes only in that the second term now sums over all  $A$ , and  $d_i^A$  are weights that sum to  $(N - N_B)$ :

$$l_{SPOE}(\boldsymbol{\theta}) = \sum_{i \in B} \log \pi_i^B + \sum_{i \in A} d_i^A \log(1 - \pi_i^B) \quad (6)$$

The log-likelihood  $l_{SPOE}$  can be maximised in a similar way to  $l_{OE}$ , for example by first assuming a logistic model holds for the propensity scores, and solving the resulting score equation using the Newton-Raphson approach.

The above scenario would be appropriate if the data in  $B$  is of high quality, so it is possible to use it directly to form inferences. When the data in  $B$  has measurement error, or if we suspect that the missingness in  $B$  also depends on  $y$ , we may wish to design the sample  $A$  so that at least some of the sample overlaps with  $B$ . Collecting  $y$  in  $A$  (or at least in the overlapping portion) would be beneficial for forming a model to correct for the measurement error in  $B$ , or enable the inclusion of  $y$  in the propensity model. How best to design the overlap between  $A$  and  $B$  when there are quality issues in  $B$  is left for future research.

## 5. Simulation Study

A simulation study was conducted to examine the performance of the IPW estimates formed using the OE propensity scores. The setup of the simulation is similar to what is used in Chen et al. (2020). In our simulation we consider a finite population of size  $N = 200,000$  that includes a data item of interest  $y$  and auxiliary variables  $\mathbf{x}$ , related by the model

$$y_i = f(\mathbf{x}, \boldsymbol{\beta}) = 2 + x_{1i} + x_{2i} + x_{3i} + x_{4i} + \sigma \epsilon_i \quad (7)$$

with  $x_{1i} = z_{1i}$ ,  $x_{2i} = z_{2i} + 0.3x_{1i}$ ,  $x_{3i} = 0.2(x_{1i} + x_{2i})$ ,  $x_{4i} = z_{4i} + 0.1(x_{1i} + x_{2i} + x_{3i})$ , and  $z_{1i} \sim \text{Bernoulli}(0.5)$ ,  $z_{2i} \sim \text{Uniform}(0, 2)$ ,  $z_{3i} \sim \text{Exponential}(1)$ , and  $z_{4i} \sim \chi^2(4)$ . The error terms  $\epsilon_i$  are independent and identically distributed as  $N(0, 1)$ . The true propensity scores  $\pi_i^B$  are given by

$$\log\left(\frac{\pi_i^B}{1 - \pi_i^B}\right) = \theta_0 + 0.1x_{1i} + 0.2x_{2i} + 0.3x_{3i} + 0.4x_{4i} \quad (8)$$

where the intercept  $\theta_0$  is chosen to ensure the target sample size for  $N_B$  is achieved. Poisson sampling is used to select the sample  $B$ , while the randomised systematic probability proportional to size sampling (PPS) method is used for the probability sample  $A$ . In general, the inclusion probabilities for  $\pi_i^A$  are proportional to  $z_i = c + x_{3i}$ , where  $c$  is chosen to ensure the ratio  $\max(z_i)/\min(z_i) = 10$ . Different sets of simulations were run for three values of  $N_B$  - 2,000, 50,000 and 140,000. The sample size for  $A$  was kept fixed at  $n_A = 5,000$  for all simulation runs.

For each sample size scenario,  $R = 2,000$  simulations were run. The performance of the different estimators was evaluated by calculating the Monte Carlo percentage Relative Bias (%RB) and percentage Relative Root Mean Squared Error (%RRMSE):

$$\%RB = \frac{1}{R} \sum_{r=1}^R \frac{\hat{Y}_r - \bar{Y}}{\bar{Y}} \times 100$$

$$\%RRMSE = \frac{\sqrt{\frac{1}{R} \sum_{r=1}^R (\hat{Y}_r - \bar{Y})^2}}{\bar{Y}} \times 100$$

where  $\hat{Y}_r$  is the estimate of the population mean computed for the  $r$ 'th simulation run and  $\bar{Y} = \mu_y$  is the true population mean.

Results were produced for the following estimators:

- OE - IPW estimates using the proposed Overlap-Excluded propensity score estimator
- Optimal OE - Optimal selection probabilities were produced, assuming all auxiliary information available
- “Anticipated” Optimal OE - Optimal selection probabilities were estimated, assuming all auxiliary information was missing except for  $x_3$
- Split-Population OE – Sample  $A$  is drawn from  $U \setminus B$ , and the OE approach is used
- CLW - The Hájek-like IPW estimator in Chen et al. (2020)
- WVL - The ALP estimator proposed by Wang et al. (2021)

## 5.1 Simulation Results

Table 5.1-1 provides the %RB and %RRMSE values for each of the estimators under the different sample sizes for  $B$ . The proposed OE approaches, along with the CLW estimator, are all unbiased. The WVL estimator is approximately unbiased, although there is a hint of bias for the smallest sample size for  $B$ .

**Table 5.1-1**  
**Comparative performance of estimators**

	$N_B = 2,000$		$N_B = 50,000$		$N_B = 140,000$	
	%RB	%RRMSE	%RB	%RRMSE	%RB	%RRMSE
OE	-0.06	5.80	-0.01	0.77	-0.02	0.32
Optimal OE	0.10	5.70	0.00	<b>0.72</b>	-0.01	0.27
“Anticipated” Optimal OE	-0.01	5.83	-0.02	0.78	0.00	0.29
Split-Population OE	-0.14	5.84	0.00	0.75	-0.01	<b>0.22</b>
CLW	-0.17	6.16	-0.07	1.27	-0.06	0.74
WVL	1.10	<b>5.59</b>	0.47	0.88	0.04	0.53

Note: Bold values indicate the lowest RRMSE for a particular value of  $N_B$

In terms of the RRMSE performance of the estimators, the WVL approach results in the lowest RRMSE when  $N_B$  is very small. However, for the larger sample sizes the OE approaches outperform both it and the CLW estimator. Optimising the sample design for  $A$  reduces the RRMSE compared to when we do not use the optimal selection

probabilities, as would be expected. Designing  $A$  so that it is drawn only from  $U \setminus B$  results in further gains in accuracy - this approach results in the lowest RRMSE for the largest  $N_B$ . It is left to future research to also look to produce optimal selection probabilities for the split-population approach.

It is encouraging to see that our use of “anticipated” optimal selection probabilities can also be effective. In the simulation, for the largest sample size of  $N_B$  we were able to achieve a small reduction in RRMSE when using these estimated optimal probabilities over the un-optimised design for  $A$ .

## 6. Concluding Remarks

Quasi-randomisation approaches have been developed to enable finite population inferences to be formed from non-probability data. These approaches utilise available auxiliary information from the population or from a reference probability sample to estimate a propensity of selection into the non-probability dataset.

In this paper, we have proposed several ideas for more efficiently designing the reference sample in conjunction with an accompanying IPW estimator which uses estimated propensities produced through the OE pseudo-likelihood. Optimal design of the reference sample  $A$  to collect  $\mathbf{x}$  presents an opportunity for the NSO to repurpose their survey program, facilitating the use of multiple non-probability (especially big) datasets as the primary source for statistics, supplemented by a small auxiliary reference sample.

These design ideas were examined empirically in a simulation study. The results indicate that when data for  $\mathbf{x}$  are available on the population frame, using optimal selection probabilities for the design provides benefits in efficiency. When  $\mathbf{x}$  is not fully available on the frame, but  $B$  is large, the use of “anticipated” optimal selection probabilities based on predicted values  $\hat{\mathbf{x}}$  can also lead to gains. When we can link  $B$  to the population frame, it may be desirable to select the sample  $A$  from  $U \setminus B$  only.

The OE propensity score estimator that we have used in conjunction with the sample design ideas in this paper is applicable when we are able to identify which units in  $A$  are also in  $B$ . The estimator yields estimates that are unbiased under the correct working model, and are efficient, especially when  $B$  is large.

## References

- Ang, L., Clark, R., Loong, B. and Holmberg, A. (2024), “Estimating Propensities of Selection for Big Datasets via Data Integration”, presented at the 7<sup>th</sup> International Conference on Establishment Statistics, Glasgow, United Kingdom. Paper available at <https://arxiv.org/abs/2501.04185>.
- Australian Bureau of Statistics (2024a), “Australian Agriculture: Broadacre Crops, 2022-23 financial year | Australian Bureau of Statistics”. Available at <https://www.abs.gov.au/statistics/industry/agriculture/australian-agriculture-broadacrecrops/latest-release>.
- Australian Bureau of Statistics (2024b), “Monthly Employee Earnings Indicator, October 2023 to March 2024 | Australian Bureau of Statistics”. Available at <https://www.abs.gov.au/statistics/labour/earnings-and-working-conditions/monthlyemployee-earnings-indicator/latest-release>.
- Australian Bureau of Statistics (2024c), “Monthly Household Spending Indicator, June 2024 | Australian Bureau of Statistics”. Available at <https://www.abs.gov.au/statistics/economy/finance/monthly-household-spendingindicator/latest-release>.
- Australian Bureau of Statistics (2024d), “Weekly Payroll Jobs, Week ending 15 June 2024 | Australian Bureau of Statistics”. Available at <https://www.abs.gov.au/statistics/labour/jobs/payrolljobs/latest-release>.
- Burakauskaitė, I., and Čiginas, A. (2023), “An Approach to Integrating a Non-Probability Sample in the Population Census”, *Mathematics* 11(8), p. 1782. DOI: <https://doi.org/10.3390/math11081782>.

- Chen, Y., Li, P., and Wu, C. (2020), “Doubly Robust Inference With Nonprobability Survey Samples”, *Journal of the American Statistical Association*, 115(532), pp. 2011–2021. DOI: <https://doi.org/10.1080/01621459.2019.1677241>.
- Clark, R. G., and Steel, D. G. (2022), “Sample design for analysis using high-influence probability sampling”, *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 185(4), pp. 1733–1756. DOI: <https://doi.org/10.1111/rssa.12916>.
- Cochran, W. G. (1977), *Sampling Techniques*, 3rd edition, New York, NY: John Wiley & Sons, Inc.
- Meng, X.-L. (2018), “Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election”, *The Annals of Applied Statistics*, 12(2), pp. 685–726. DOI: <https://doi.org/10.1214/18-AOAS1161SF>.
- Rivers, D. (2007), Sampling for web surveys, in “Joint Statistical Meetings 2007, Section on Survey Research Methods”.
- United Nations Statistics Division (2014), “Fundamental Principles of Official Statistics”, <https://unstats.un.org/unsd/dnss/gp/fundprinciples.aspx>. Accessed: 2024-11.
- Wang, L., Valliant, R. and Li, Y. (2021), “Adjusted logistic propensity weighting methods for population inference using nonprobability volunteer-based epidemiologic cohorts”, *Statistics in Medicine*, 40(24), pp. 5237–5250. DOI: <https://doi.org/10.1002/sim.9122>
- Yang, S., Kim, J.-K. and Hwang, Y. (2021), “Integration of data from probability surveys and big found data for finite population inference using mass imputation”, *Survey Methodology*, 47(1), pp. 29–58. Paper available at <http://www.statcan.gc.ca/pub/12-001-x/2021001/article/00004-eng.htm>.
- Zhang, L.-C. (2019), ‘On valid descriptive inference from non-probability sample’, *Statistical Theory and Related Fields* 3(2), 103–113. DOI: <https://doi.org/10.1080/24754269.2019.1666241>.