National Institute for Applied Statistics Research Australia

University of Wollongong, Australia

Working Paper

22-21

Weighting, Informativeness and Causal Inference, with an Application to Rainfall Enhancement

Ray Chambers, Setareh Ranjbar,
Nicola Salvati, and Barbara Pacini

National Institute for Applied Statistics Research Australia, University of Wollongong,
Wollongong NSW 2522, Australia Phone +61 2 4221 5076, Fax +61 2 4221 4998.
Email: karink@uow.edu.au

Weighting, Informativeness and Causal Inference, with an Application to Rainfall Enhancement

Ray Chambers
National Institute for Applied Statistical Research Australia
University of Wollongong
ray@uow.edu.au

Setareh Ranjbar
Faculty of Business and Economics
Université de Lausanne
Setareh.Ranjbar@unil.ch

Nicola Salvati
Department of Economics and Management
University of Pisa
nicola.salvati@unipi.it

Barbara Pacini
Department of Political Science
University of Pisa
barbara.pacini@unipi.it

November 2021

Abstract

Sampling is informative when probabilities of sample inclusion depend on unknown variables that are correlated with a response variable of interest. This can be a problem when the sample data analyst only has access to secondary data sources for controlling the impact of the sampling method. When sample inclusion probabilities are available, inverse probability weighting can be used to account for informative sampling in a secondary analysis situation, though usually at the cost of less precise inference. This paper reviews two important research contributions by Chris Skinner that modify these weights to reduce their variability while at the same time retaining consistency of the weighted estimators. In some cases, however, sample inclusion probabilities are not known, and are estimated as propensity scores. This often the situation in causal analysis, and double robust methods that protect against resulting misspecification of the sampling process have been the focus of much recent research. In this paper we propose two model-assisted modifications to the popular inverse propensity score weighted estimator of an average treatment effect, and then illustrate their use in a causal analysis of a rainfall enhancement experiment that was carried out in Oman between 2013 and 2018.

Key Words: Inverse probability weighting; weight modification; propensity scores; double robustness; average treatment effect; model-based analysis; model-assisted estimation; observational data analysis; cloud ionization.

1. Introduction

1.1 A brief background on sample weighting and inference

Weighting is at the core of sampling inference. Virtually every procedure used to make an inference about a population of interest based on data obtained from a sample of population units depends on the statistic $\bar{y}_{ws} = \sum_{i=1}^{N} w_{is} I_i y_i$ being a consistent estimator of the expected value μ of the finite population mean $\bar{y}_U = N^{-1} \sum_{i=1}^{N} y_i$. Here *U* denotes the finite population of interest, *N* is the population size, I_i is a sample inclusion indicator that takes the value 1 if unit *i* is in sample and the value zero otherwise, and y_i is a generic variable value observed for each sample unit and assumed to be observable for any population unit. The set of *n* population units making up the sample is the largest set $\{i \in U : I_i = 1\}$, and is denoted *s*. The sample weights w_{is} are assumed to be known for each sample unit, and are also assumed to be computable for any population unit *i* and any sample *s*. Definition of w_{is} depends to a large extent on the type of inference that one wishes to make about μ. If one replaces μ by $\bar{y}_U$ as the target of inference then inference is said to be *enumerative*, while if μ remains as the target then inference is often referred to as *analytic*. We will be concerned with analytic inference in this paper.

In the case of enumerative inference there are two major approaches. The oldest, first proposed in Neyman (1934), only allows random variation in $\bar{y}_{ws}$ as a consequence of variation in the sample inclusion indicators I_i. That is, the only uncertainty is the outcome of the sampling process. All other finite population measurements, and in particular the values y_i, are considered to be fixed. This is essentially non-parametric inference, typically referred to as *design-based*. Within the last half century however, it has become more common to allow joint variation in both I_i and y_i to underpin inference. This is *model-based* inference, primarily because it is standard to use a stochastic model to describe variability in the population y_i values, with the implicit assumption that variability in the population I_i values is under the control of the sample designer.

Let $\mathbf{I}_U$ and $\mathbf{y}_U$ denote the vectors consisting of the population values of I_i and y_i respectively. The model-based approach implicitly assumes that the distribution of $\mathbf{I}_U$ is a function of population auxiliary information, typically characterized by the values defining a $N \times p$ matrix $\mathbf{X}_U$. Consequently, the *conditional independence assumption* (CIA) is usually made, i.e., $\left(\mathbf{y}_U \,||\, \mathbf{I}_U\right)\big|\mathbf{X}_U$, where || denotes independence. A sampling procedure for which the CIA is valid for some $\mathbf{X}_U$ is commonly referred to as *non-informative sampling*, with the restriction implied by the conditioning on $\mathbf{X}_U$ often ignored. However, as the CIA makes clear, it is this conditioning that is important. Sampling that is non-informative given $\mathbf{X}_U$ may not be so if $\mathbf{X}_U$ is unavailable, or if just a part of it is available. However, if $\mathbf{X}_U$ is known then the realized values $\mathbf{I}_U$ of the sample inclusion indicators are ancillary for inference about μ and so inference can condition on them, i.e., condition on the realized value of the set *s*. In this case it is just the variability in $\mathbf{y}_U$ that underpins inference.

Despite efforts over the last fifty years, e.g., Brewer (1999), design-based and model-based inference cannot be reconciled except in the null case where $\mathbf{X}_U$ provides no information about the variability in $\mathbf{y}_U$ or $\mathbf{I}_U$. However, this is usually not the case, and *model-assisted* inference is then a widely used compromise between design-based inference and model-based inference that allows for both sources of variability. This approach is often assumed to provide both the non-parametric robustness of the design-based approach and the parametric efficiency associated with the model-based approach. However, this may not be the case, as we shall see.

1.2 Why this paper?

This paper aims to provide an overview of the important issues that arise when one uses survey weights for inference, both in the context of Chris Skinner's major contributions in the area and in the context of closely related issues that arise in causal inference. The desirable properties of consistency and double robustness for weighted survey estimators are discussed in the next Section, with Chris's major contributions to improving the efficiency of weighted survey estimates discussed in Section 3. Then in Section 4 we focus on causal inference and the classic problem of estimating a causal effect from observational data. In this section we also develop two doubly robust estimators for an additive causal effect that behave similarly to a model-assisted estimator, in that they use a model to control for bias caused by differences in covariate distributions between treated and untreated groups. In Section 5 we apply the methods developed in Section 4 to a new analysis of a data set collected in a six-year rainfall enhancement trial. Section 6 completes the paper with a more discursive summary of the ideas in it and the results obtained.

2. Consistency and robustness under weighted inference

Chris Skinner firmly believed that model-assisted inference should be the default approach to sample survey inference. His basis for this belief was simple: Defining a statistical model for $\mathbf{y}_U$ given just the sample values $\mathbf{y}_s = \{y_i; i \in s\}$ will almost inevitably result in model misspecification, in the sense that it will not lead to the same model as would be obtained given $\mathbf{y}_U$. On the other hand, the properties of the sample inclusion indicators I_i are known (or at least should be known) to the survey sampler, and these determine whether an estimator of interest is design-consistent, i.e., it converges in probability to its design expectation as the sample size increases. Restricting weights $\mathbf{w}_s = \{w_{is}; i \in s\}$ to use in $\bar{y}_{ws}$ so that this estimator is design-consistent should therefore be a minimum requirement. Modelling assumptions can subsequently be introduced to improve the efficiency of $\bar{y}_{ws}$ assuming that the model holds. However, this efficiency is a secondary consideration.

To illustrate, assume that we know $\pi_i(\mathbf{X}_U) = E\left(I_i \middle| \mathbf{X}_U\right)$ and consider the classic design-based version of $\bar{y}_{ws}$ in this case. This is the inverse probability weighted (IPW) estimator corresponding to the ratio-type Hájek (1971) version of the design-unbiased estimator for $\bar{y}_U$ under without-replacement sampling (Narain, 1951; Horvitz and Thompson, 1952). Here

$$w_{is} = w_{is}^{IPW} = \left(\pi_i(\mathbf{X}_U)\right)^{-1}\left(\sum\nolimits_{j=1}^{N}\left(\pi_j(\mathbf{X}_U)\right)^{-1} I_j\right)^{-1}.$$

Put $\mu_i(\mathbf{X}_U) = E\left(y_i \middle| \mathbf{X}_U\right)$, so $\mu = E\left(\mu_i(\mathbf{X}_U)\right)$. The IPW estimator $\bar{y}_{ws}^{IPW}$ is consistent for μ under any model for μ_i when the CIA is valid and $\pi_i(\mathbf{X}_U)$ is correctly specified since under suitable regularity conditions it then follows

$$\begin{aligned}
E\left(\bar{y}_{ws}^{IPW}\right) - \mu &= E\left\{E\left(\sum\nolimits_{i=1}^{N} w_{is}^{IPW} I_i y_i \middle| \mathbf{X}_U\right) - N^{-1}\sum\nolimits_{i=1}^{N}\mu_i(\mathbf{X}_U)\right\} \\
&\rightarrow E\left\{\frac{E\left(\sum\nolimits_{i=1}^{N}\left(\pi_i(\mathbf{X}_U)\right)^{-1} I_i y_i \middle| \mathbf{X}_U\right)}{E\left(\sum\nolimits_{i=1}^{N}\left(\pi_i(\mathbf{X}_U)\right)^{-1} I_i \middle| \mathbf{X}_U\right)} - N^{-1}\sum\nolimits_{i=1}^{N}\mu_i(\mathbf{X}_U)\right\} \\
&= E\left\{\frac{\sum\nolimits_{i=1}^{N}\left(\pi_i(\mathbf{X}_U)\right)^{-1}\pi_i(\mathbf{X}_U)\mu_i(\mathbf{X}_U)}{\sum\nolimits_{i=1}^{N}\left(\pi_i(\mathbf{X}_U)\right)^{-1}\pi_i(\mathbf{X}_U)} - N^{-1}\sum\nolimits_{i=1}^{N}\mu_i(\mathbf{X}_U)\right\} = 0.
\end{aligned}$$

Unfortunately, as is well known, the IPW estimator can be inefficient. Also, sample inclusion probabilities for sampled units must be known. This is usually not an issue under full response. However, full response is rare, and non-response is usually the case. The probability of sample inclusion then includes the (typically unknown) probability of response. It is also an issue for observational studies where sample inclusion can depend on characteristics of population units that are not captured in $\mathbf{X}_U$, including the value y_i itself.

Improving on the efficiency of the IPW estimator has been the focus of much research over the last fifty years, with most of it is based on the CIA. As we have already noted, the sample inclusion indicators are irrelevant for inference about μ in this case, and so the vector $\mathbf{w}_s$ of efficient model-based sample weights can be chosen to minimize $Var\left(\bar{y}_{ws} - \bar{y}_U \middle| \mathbf{X}_U\right)$ subject to $E\left(\bar{y}_{ws} - \bar{y}_U \middle| \mathbf{X}_U\right) = \bar{\mu}_{ws} - \bar{\mu}_U = 0$. Here $\bar{\mu}_{ws} = \sum_{i=1}^{N} w_{is} I_i \mu_i(\mathbf{X}_U)$ and $\bar{\mu}_U = N^{-1}\sum_{i=1}^{N} \mu_i(\mathbf{X}_U)$. Let $\mathbf{w}_s^{MB} = \left\{ w_{is}^{MB}; i \in s \right\}$ denote these model-based weights, with associated estimator $\bar{y}_{ws}^{MB}$. Then by construction, $E\left(\bar{y}_{ws}^{MB} \middle| \mathbf{X}_U\right) = \bar{\mu}_U$ and so $\bar{y}_{ws}^{MB}$ is model-consistent (but not necessarily design-consistent) for $\mu = E\left(E\left(\bar{y}_U \middle| \mathbf{X}_U\right)\right) = E\left(\bar{\mu}_U\right)$. Note that the final expectation assumes that $\left(\mathbf{y}_U, \mathbf{X}_U\right)$ is a random draw from a conceptual set of finite population values, often referred to as the underlying superpopulation.

To illustrate, suppose that $\mathbf{y}_U = \mathbf{X}_U \beta + \mathbf{e}_U$ where the first column of $\mathbf{X}_U$ is $\mathbf{1}_U$, $\mathbf{e}_U \| \mathbf{X}_U$, $E\left(\mathbf{e}_U\right) = 0$ and $Var\left(\mathbf{e}_U\right) = \sigma^2 diag(\mathbf{1}_U)$. Here $\mathbf{1}_U$ is the *N*-vector with each element equal to 1. Then $\mathbf{w}_s^{MB} = \mathbf{X}_s \left(\mathbf{X}_s^T \mathbf{X}_s\right)^{-1} \bar{\mathbf{x}}_U$, where $\mathbf{X}_s$ denotes the sampled rows of $\mathbf{X}_U$ and $\bar{\mathbf{x}}_U = N^{-1} \mathbf{X}_U^T \mathbf{1}_U$. More sophisticated models (e.g., those with random effects) are discussed in Chapters 13 and 15 of Chambers and Clark (2012).

The adage that all models are wrong applies in survey sampling as much as it does in statistics generally. This concern, echoed in many of the papers that Chris Skinner had a hand in, leads to a compromise between design-based inference and model-based inference that is commonly referred to as model-assisted inference. The basis of this approach, insofar as estimation is concerned, is the idea of using design-based estimation to ensure that a model-based estimator is also design-consistent. Let $\hat{\mu}_i(\mathbf{X}_U)$ denote an unbiased estimator of $\mu_i(\mathbf{X}_U)$ under the assumed model, with associated model-based estimator $\bar{y}_{ws}^{MB}$ of $\bar{y}_U$ that satisfies

$$\bar{y}_{ws}^{MB} = \sum_{i=1}^{N} w_{is}^{MB} I_i y_i = N^{-1}\sum_{i=1}^{N} \hat{\mu}_i(\mathbf{X}_U) = \hat{\bar{\mu}}_U .$$

This condition is satisfied, for example, by efficient model-based weighting under the linear model defined in the preceding paragraph. A model-assisted modification $\bar{y}_{ws}^{MA}$ to $\bar{y}_{ws}^{MB}$ is obtained by adding a design-consistent bias correction to this estimator, i.e.,

$$\bar{y}_{ws}^{MA} = \bar{y}_{ws}^{MB} + \bar{r}_{ws}^{IPW}$$

where $\bar{r}_{ws}^{IPW} = \sum_{i=1}^{N} w_{is}^{IPW} I_i \left(y_i - \hat{\mu}_i(\mathbf{X}_U) \right)$ is the IPW-weighted estimator of the average of the residuals $\mathbf{r}_U = \mathbf{y}_U - \hat{\boldsymbol{\mu}}_U$ and $\hat{\boldsymbol{\mu}}_U$ is the population vector of fitted values under the assumed model. We have seen that $\bar{y}_{ws}^{MB}$ is consistent for μ when $\mu_i(\mathbf{X}_U)$ is correctly specified. Furthermore, if the CIA also holds then $\bar{r}_{ws}^{IPW}$ is model-consistent for zero since $E\left(y_i - \hat{\mu}_i(\mathbf{X}_U) \middle| \mathbf{X}_U \right) = 0$ and so

$$E\left(\bar{r}_{ws}^{IPW}\right) \to E\left\{ \frac{\sum_{i=1}^{N} \left(\pi_i(\mathbf{X}_U)\right)^{-1} E\left(I_i \middle| \mathbf{X}_U\right) E\left(y_i - \hat{\mu}_i(\mathbf{X}_U) \middle| \mathbf{X}_U\right)}{\sum_{i=1}^{N} \left(\pi_i(\mathbf{X}_U)\right)^{-1} E\left(I_i \middle| \mathbf{X}_U\right)} \right\} = 0$$

irrespective of whether $\pi_i(\mathbf{X}_U) = E\left(I_i \middle| \mathbf{X}_U\right)$. That is, under correct model specification and the CIA, $\bar{y}_{ws}^{MA}$ and $\bar{y}_{ws}^{MB}$ are both model-consistent for μ.

Conversely, we can view the definition of $\bar{y}_{ws}^{MA}$ above as adding a model-consistent bias correction to $\bar{y}_{ws}^{IPW}$. That is, we can also write

$$\bar{y}_{ws}^{MA} = \bar{y}_{ws}^{IPW} - \bar{r}_{ws}^{MB}$$

where $\bar{r}_{ws}^{MB} = \sum_{i=1}^{N} w_{is}^{IPW} I_i \hat{\mu}_i(\mathbf{X}_U) - \bar{y}_{ws}^{MB} = \sum_{i=1}^{N} w_{is}^{IPW} I_i \hat{\mu}_i(\mathbf{X}_U) - N^{-1}\sum_{i=1}^{N} \hat{\mu}_i(\mathbf{X}_U)$. Clearly, provided the CIA holds, and $\pi_i(\mathbf{X}_U) = E\left(I_i \middle| \mathbf{X}_U\right)$ is correctly specified, then $\bar{y}_{ws}^{IPW}$ is design-consistent for μ and $\bar{r}_{ws}^{MB}$ is design-consistent for zero irrespective of whether $\hat{\mu}_i(\mathbf{X}_U)$ is model-consistent for $\mu_i(\mathbf{X}_U)$.

This dual property of $\bar{y}_{ws}^{MA}$ is often referred to as *double robustness*. Estimators with a double robustness property have been extensively studied in recent years, see Bang and Robins (2005). In a survey sampling context, Chen *et al.* (2020) have proposed a double robust estimator for the finite population mean given data from a non-probability survey. These methods have been promoted as allowing an analyst to have the best of both worlds – protected against misspecification of the model for $\mathbf{y}_U|\mathbf{X}_U$ if the sample inclusion probabilities are correctly specified, and protected against misspecification of sample inclusion probabilities (as would be the case under sample non-response) if the model for $\mathbf{y}_U|\mathbf{X}_U$ is correctly specified.

Of course, as has been pointed out by many (see Kang and Schafer, 2007), the usual situation is where *both* the sample inclusion probabilities and the model for $\mathbf{y}_U|\mathbf{X}_U$ are incorrectly specified. Because of the ubiquitous nature of non-response, this will still be the case for "well-designed and implemented" surveys. From a pure model-based perspective there appear to be at least two things one can do to protect oneself in this case. The first is to adopt a flexible specification for the model for $\mathbf{y}_U|\mathbf{X}_U$, as in a non-parametric regression specification for μ_i. The second is to replace the IPW weights w_{is}^{IPW} in the bias correction term $\bar{r}_{ws}^{IPW}$ in $\bar{y}_{ws}^{MA}$ by alternative weights that allow for more accurate estimation of the population value of this bias. As Chambers *et al.*, (1993) point out these two strategies lead to the same estimator if the same non-parametric regression-based weighting scheme is used in both. They also point out that the idea of nonparametrically bias correcting a model misspecification bias is essentially an extension of Tukey's idea of "twicing" when fitting a potentially incorrectly specified model.

Other approaches to dealing with model misspecification as well incorrect sample inclusion probabilities that are more in line with the idea of double robustness have also been suggested. For example, Han (2014) and Chen and Haziza (2017) suggest that alternative models for $\mathbf{y}_U|\mathbf{X}_U$ be considered as well as alternative sample inclusion probability specifications, with $\bar{y}_{ws}^{MA}$ then computed based on a suitably averaged fitted value for $\mu_i(\mathbf{X}_U)$ and a similar composite value for $\pi_i(\mathbf{X}_U)$. They show that such a *multiply robust* specification for $\bar{y}_{ws}^{MA}$ can improve on any version of this estimator that uses just one of the alternative models for μ_i and just one of the different sample inclusion probability specifications – provided at least one of these alternatives is correct. This can be useful if different variable selection methods are used in model identification, and these lead to competing model specifications. We do not pursue this idea further beyond noting that in most practical situations it is unlikely that any of the potential alternative specifications will be true, so the utility of this approach will depend on its capacity to reduce the variability of $\bar{y}_{ws}^{MA}$. Here we note that several empirical studies have now shown that despite being inconsistent, multiply robust procedures tend to have good numerical properties, see Han (2014) and Chen and Haziza (2017, 2019).

3. Chris Skinner's impact on sample weighting methodology

3.1 Contributions to weighting under non-informative sampling

In the previous section we used notation that recognized that $\mu_i(\mathbf{X}_U)$ and $\pi_i(\mathbf{X}_U)$ are conditional expectations, and hence functions of the auxiliary information $\mathbf{X}_U$. To avoid notational complexity and improve readability we now make this referencing of $\mathbf{X}_U$ implicit. We also assume that all unit *i* specific expectations that condition on $\mathbf{X}_U$ are purely functions of the vector $\mathbf{x}_i$ defining row *i* of $\mathbf{X}_U$. If one accepts the CIA and that the known values of π_i correctly represent the actual sample inclusion probabilities (as would be the case under controlled probability sampling and full response), then the main issue with both $\bar{y}_{ws}^{IPW}$ and $\bar{y}_{ws}^{MA}$ is their variability compared to $\bar{y}_{ws}^{MB}$. This is basically due to the variability induced by the unit specific "representative" weights π_i^{-1} used in both. In a pioneering paper, Skinner and Mason (2012) showed how this variability can be reduced while at the same time retaining the desirable design-consistency property. In the context of the development so far, their approach corresponds to replacing the π_i^{-1} by modified unit-level weights of the form

$$d_i^{IPWX} = \pi_i^{-1} q_i$$

where q_i is a function of $\mathbf{x}_i$, and is chosen in order to minimize the variance of the solution $\hat{\mu}_i$ to the estimating equation

$$\sum_{i=1}^{N} w_{is}^{IPWX} I_i \left(y_i - \hat{\mu}_i \right) = 0$$

with $w_{is}^{IPWX} = d_i^{IPWX} \Big/ \sum_{j=1}^{N} d_j^{IPWX} I_j$. Note that with this definition, and assuming the CIA,

$$E\left(\sum_{i=1}^{N} w_{is}^{IPWX} I_i y_i \middle| \mathbf{X}_U \right) \to E\left(\sum_{i=1}^{N} \pi_i^{-1} q_i I_i y_i \middle| \mathbf{X}_U \right) \Big/ E\left(\sum_{i=1}^{N} \pi_i^{-1} q_i I_i \middle| \mathbf{X}_U \right) \to \sum_{i=1}^{N} q_i \mu_i \Big/ \sum_{i=1}^{N} q_i .$$

Providing the q_i are $O(1)$, the *q*-weighted mean on the right hand side above will converge to μ. Using a linearization argument, and assuming Poisson sampling of population units, these authors derive the optimal value $q_i = \left\{ E\left(\pi_i^{-1} \right) \right\}^{-1}$, in which case

$$w_{is}^{IPWX} = \pi_i^{-1} \left\{ E\left(\pi_i^{-1} \right) \right\}^{-1} \Big/ \sum_{j=1}^{N} \pi_j^{-1} \left\{ E\left(\pi_j^{-1} \right) \right\}^{-1} I_j .$$

There is a subtle but important change in the inference framework used in the preceding development. In particular, the sample inclusion probability π_i is now being treated as a random variable rather than as a known function of $\mathbf{x}_i$, putting its value on a par with the value y_i of the response variable. This is important in many practical applications where this probability is an unknown function of $\mathbf{x}_i$ and, as is often the case with analysis of observational data, where the value of π_i is only known if unit *i* is part of the observed sample. Note also that although the modified IPW estimator $\bar{y}_{ws}^{IPWX} = \sum_{i=1}^{N} w_{is}^{IPWX} I_i y_i$ is model-consistent for μ, it is not design-consistent for the population mean $\bar{y}_U$ but instead converges to the *q*-weighted version of this mean. In effect, Skinner and Mason (2012) implicitly acknowledge that enforcing strict design-consistency is at odds with efficient weighted inference for μ.

3.2 Contributions to weighting under informative sampling

The development in the previous subsection assumed the CIA holds. But there are situations where the CIA does not hold under conditioning on the available values in $\mathbf{X}_U$. As noted earlier, such cases arise when the conditioning is incomplete, i.e., $\mathbf{X}_U$ is only partially known, or some of its component variables are ignored, as would be the case under variable selection. However, it can also be the case that even if $\mathbf{X}_U$ is completely known, sample inclusion can depend on $\mathbf{y}_U$ as well as $\mathbf{X}_U$. This type of informative sampling is an example of *response-dependent sampling*, with the most obvious example being case-control sampling. Another example where informative sampling is of concern is in the secondary analysis of survey data, where the analyst has access to the sample values of y_i and π_i, as well as the population values of $\mathbf{x}_i$, but believes that the agency that created the sample did so using information on another, unreleased, variable $\mathbf{z}_i$. Furthermore, given $\mathbf{x}_i$, the value of $\mathbf{z}_i$ (and hence the realized value of the sample inclusion indicator I_i) and the response variable y_i are correlated. This clearly violates the CIA.

In a subsequent paper, Kim and Skinner (2013) extended the minimum variance weights concept of Skinner and Mason (2012) to the case of response-dependent sampling, i.e., where the probability of sample inclusion also depends on the value of the response variable of interest. Inverse probability weights can exhibit wide variability in this situation. In order to address this problem, Beaumont (2008) assumes that the sample inclusion probability π_i is known and satisfies $\pi_i = E\left(I_i \middle| y_i, \mathbf{x}_i\right)$. Put

$$\tilde{\pi}_i = E\left(\pi_i \middle| y_i, I_i = 1, \mathbf{x}_i\right) = \left\{E\left(\pi_i^{-1} \middle| y_i, I_i = 1, \mathbf{x}_i\right)\right\}^{-1}$$

where the last equality follows from Pfeffermann and Sverchkov (1999). Then

$$E\left(I_i y_i \middle| \mathbf{x}_i\right) = E\left(E\left(I_i \middle| y_i, \mathbf{x}_i\right) y_i \middle| \mathbf{x}_i\right) = E\left(\pi_i y_i \middle| \mathbf{x}_i\right) = E\left(E\left(\pi_i \middle| y_i, I_i = 1, \mathbf{x}_i\right) y_i \middle| \mathbf{x}_i\right) = E\left(\tilde{\pi}_i y_i \middle| \mathbf{x}_i\right)$$

and we have $E\left(\tilde{\pi}_i^{-1} I_i y_i \middle| \mathbf{x}_i\right) = E\left(\tilde{\pi}_i^{-1} \tilde{\pi}_i y_i \middle| \mathbf{x}_i\right) = \mu_i$. It immediately follows that the IPW estimator based on the smoothed value $\tilde{\pi}_i$ instead of π_i is also consistent for μ. Let $\bar{y}_{ws}^{SIPW}$ denote the IPW estimator based on the smoothed $\tilde{\pi}_i$. Then, since

$$E\left(\sum_{i=1}^{N} \pi_i^{-1} I_i y_i \middle| \mathbf{y}_U, \mathbf{I}_U, \mathbf{X}_U\right) = \sum_{i=1}^{N} E\left(\pi_i^{-1} \middle| y_i, I_i = 1, \mathbf{x}_i\right) I_i y_i = \sum_{i=1}^{N} \tilde{\pi}_i^{-1} I_i y_i$$

we have $\bar{y}_{ws}^{SIPW} \approx E\left(\bar{y}_{ws}^{IPW} \middle| \mathbf{y}_U, \mathbf{I}_U, \mathbf{X}_U\right)$ and hence

$$Var\left(\bar{y}_{ws}^{IPW} \middle| \mathbf{X}_U\right) \geq Var\left(E\left(\bar{y}_{ws}^{IPW} \middle| \mathbf{y}_U, \mathbf{I}_U, \mathbf{X}_U\right) \middle| \mathbf{X}_U\right) \approx Var\left(\bar{y}_{ws}^{SIPW} \middle| \mathbf{X}_U\right).$$

That is, the smoothed version of the IPW estimator will usually be more efficient than the "standard" version of this estimator.

The key contribution of Kim and Skinner (2013) was to improve upon this smoothing approach to weighting under response-dependent sampling by combining it with the optimal weighting approach developed in Skinner and Mason (2012). Using similar approximations to those used in this last reference, including assuming Poisson sampling, they consider a modified smoothed weighting scheme with unit weights $\tilde{\pi}_i^{-1} q_i$ and seek to identify the value of q_i that minimizes the asymptotic variance of the IPW estimator based on these unit weights. This leads to the optimal value $q_i = \left\{E\left(\tilde{\pi}_i^{-1}(y_i - \mu_i)^2\right)\right\}^{-1}$ and modified smoothed IPW weights

$$w_{is}^{SIPWX} = \tilde{\pi}_i^{-1} \left\{E\left(\tilde{\pi}_i^{-1}(y_i - \mu_i)^2\right)\right\}^{-1} \Big/ \sum_{j=1}^{N} \tilde{\pi}_j^{-1} \left\{E\left(\tilde{\pi}_j^{-1}(y_j - \mu_j)^2\right)\right\}^{-1} I_j .$$

Finally, we note that application of both the Beaumont (2008) approach and the Kim and Skinner (2013) approach to computing a more efficient IPW estimator of μ under response-dependent sampling requires estimation of both $\mu_i = E\left(y_i \middle| \mathbf{x}_i\right)$ and $\tilde{\pi}_i = E\left(\pi_i^{-1} \middle| y_i, I_i = 1, \mathbf{x}_i\right)$ followed by estimation of $E\left(\tilde{\pi}_i^{-1}(y_i - \mu_i)^2\right)$. This can be done by using the sample data to fit appropriate parametric models to these expectations. Assuming that the model for μ_i is given, Kim and Skinner (2013) suggest that a model for $\tilde{\pi}_i$ of the form $E\left(\pi_i^{-1}\middle|y_i, I_i = 1, \mathbf{x}_i\right) = 1 + \exp(-\boldsymbol{\phi}_1^T \mathbf{x}_i - \phi_2 y_i)$ will usually be adequate, with the values of $E\left(\tilde{\pi}_i^{-1}(y_i - \mu_i)^2\right)$ then computed by bootstrapping from the sample values $\left\{\left\{1 + \exp(-\hat{\boldsymbol{\phi}}_1^T \mathbf{x}_i - \hat{\phi}_2 y_j)\right\}(y_j - \hat{\mu}_i)^2, j \in s\right\}$. However, use of a parametric sampling model like that considered in Kim and Skinner (2013) can lead to inefficient estimators if this model is misspecified. In contrast, non-parametric machine learning methods such as random forests and boosting should be more robust to model misspecification, and so have therefore become more attractive to National Statistical Offices that now have access to a variety of data sources, potentially containing a large number of observations on a large number of variables (Dagdoud *et al.*, 2020). It is also important to note that the Kim and Skinner approach, as well as its simpler version when the CIA holds, depends crucially on the sample inclusion probabilities π_i being known. In many practical applications of analysis of observational data this is not the case, particularly when there is reason to believe that the sampling was informative. In the next section we address this issue in the context of causal inference, which is an important type of analysis of observational data.

4. But what if inclusion probabilities are unknown?

4.1 A brief overview of causal inference using observational data

Neyman (1923) explicitly defined a framework of potential outcomes with the aim of making causal inferences using the data collected in a *randomized experiment*. The simplest version of this framework is where each unit *i* in the experiment population has two response values, depending on whether or not it is subject to an intervention of interest, and our focus is on answering the causal question: Does the intervention have a significant impact on the associated response? We shall write y_{i0} as the response when unit *i* is not subject to the intervention, and y_{i1} as the response when unit *i* is subject to the intervention. Following standard practice, we refer to a unit not subject to the intervention as a control unit and a unit subject to the intervention as a treated unit. A crucial point to make here is that both response values are *potentially* observable, but the one that is *actually* observed depends on whether the relevant unit is a control unit or a treated unit. An observed data set includes a mix of control units and treated units, depending on how interventions are distributed among the units making up the data set. This distribution, also referred to as the assignment mechanism, can be viewed as the outcome of a process that "samples for potential outcomes", i.e., it determines whether y_{i0} or y_{i1} is observed. Furthermore, unless the data set is the outcome of an experiment where interventions are allocated according to a pre-defined randomized sequence, it is a sampling process where we do not know the sample inclusion probabilities.

As usual, we assume the existence of auxiliary information in the form of a vector of covariates $\mathbf{x}_i$ for unit *i*. A key target of causal inference is the average causal effect

$$\delta = E\left(y_{i1}\right) - E\left(y_{i0}\right) = E\left\{E\left(y_{i1}\middle|\mathbf{x}_i\right) - E\left(y_{i0}\middle|\mathbf{x}_i\right)\right\}.$$

Note that δ is the expected difference between the averages of y_{i1} and y_{i0} over the population corresponding to the expectation operator. Since both outcomes cannot be observed for the same unit, estimation of δ effectively requires us to impute the missing potential outcome for each unit in this population. It also requires that we specify the population underpinning δ. The narrowest specification, and the one that we focus on in the following two subsections, is where this population coincides with the observed sample data. This allows us to replace the definition of δ above with its empirical version

$$\delta_s = n^{-1}\sum\nolimits_{i=1}^{n}\left\{\mu_1(\mathbf{x}_i) - \mu_0(\mathbf{x}_i)\right\} \tag{1}$$

where $\mu_1(\mathbf{x}_i) = E\left(y_{i1}\middle|\mathbf{x}_i\right)$ and $\mu_0(\mathbf{x}_i) = E\left(y_{i0}\middle|\mathbf{x}_i\right)$. This is sometimes referred to as the Sample Average Treatment Effect (SATE). See Imbens and Wooldridge (2009) for a discussion on the use of SATE and similar measures in program evaluation.

More generally, the observed data set may be a sample *s* of size *n* from a larger population *U* of size *N*, in which case δ can be defined as

$$\delta_U = N^{-1}\left[\sum\nolimits_{i\in s}\left\{\mu_1(\mathbf{x}_i) - \mu_0(\mathbf{x}_i)\right\} + \sum\nolimits_{i\notin s}\left\{\mu_1(\mathbf{x}_i) - \mu_0(\mathbf{x}_i)\right\}\right] \tag{2}$$

and the problem of estimating this Population Average Treatment Effect (PATE) is then essentially the problem of predicting the finite population mean of the individual treatment effects (ITE) given by $\delta_i = \mu_1(\mathbf{x}_i) - \mu_0(\mathbf{x}_i)$. Unlike the usual model-based sampling situation, though, these ITE values are not observable, and must be themselves estimated from the sample data. We discuss (2) and this more general situation in Section 6. See Imbens and Rubin (2015) for a general background on causal inference and the estimation of treatment effects.

For the SATE to be identifiable, the decision on whether intervention is carried out for a particular unit, i.e., the treatment is assigned to this unit, needs to be restricted so that the assignment probability is independent of the potential outcomes as well as the values of covariates for other units. This is usually summarized in three basic properties of the assignment mechanism:

1. *Individualistic assignment*. The probability of a unit being assigned to the treatment only depends on the value of the covariate $\mathbf{x}_i$ for that unit and not on the values of the covariates for other units. Following Rubin (1980), this condition is sometimes referred to as the Stable Unit Treatment Value Assumption.
2. *Probabilistic assignment*: This condition is familiar to survey samplers and states that every unit in the population has a probability of being treated that is strictly between zero and one for all units (Rosenbaum and Rubin, 1983). When the population and sample coincide, as with the SATE, this probability only refers to treatment assignment for the sample units.
3. *Unconfounded assignment*: This assumption is essentially the CIA for treatment assignment, in that it states that this assignment is independent of any potential outcomes conditioned either on known covariates or on the propensity scores.

The probabilistic and unconfoundedness properties are essentially the *strong ignorability* assumption of Rosenbaum and Rubin (1983).

There are two main concerns in causal inference based on observational data, The first is the use of an appropriate covariate adjustment method that achieves balance on relevant covariates, while the second is satisfying what Rosenbaum and Rubin (1984) refer to as strongly ignorable treatment assignment. Weighting is often used to address the first concern. Unfortunately, strongly ignorable treatment assignment via randomization, although desirable, is not always feasible. Instead, we must often make do with observational data where the probability of treatment assignment is unknown, and so needs to be estimated from the realized values of these assignments. These estimated probabilities are usually referred to as propensity scores, a convention that we now adopt. Li *et al*. (2018) discuss balancing weighted estimators based on propensity scores and derive the set of overlap propensity score weights that minimize the asymptotic variance of the corresponding estimated average treatment effect. More generally, covariate adjustment can be carried out by combining weighting and regression adjustment in order to achieve double robustness, see Scharfstein *et al*. (1999).

There is a huge literature on propensity score weighting in causal inference and the associated issue of double robustness. In particular, the superiority of double robust estimators compared to propensity score weighting methods for causal inference has been comprehensively investigated, see Lunceford and Davidian (2004). The benefits of taking a double robust approach when estimating an average treatment effect in survival analysis with longitudinal data are discussed in

Yu and van der Laan (2006), while Zhou *et al.* (2019) develop a double robust estimator based on a penalized spline propensity prediction method to impute the missing potential outcomes given time dependent confounders in longitudinal studies. Saarela *et al.* (2016) discuss double robustness in causal inference in from a Bayesian perspective, and propose the use of inverse propensity scores as importance sampling weights in the estimation of the outcome model.

The three common strategies used to estimate the SATE (1) are model-based imputation, where a regression model is used to impute the counterfactuals (y_{i0} for a treated unit and y_{i1} for a control, or untreated, unit), weighting estimators and matching estimators. Both weighting and matching require knowledge of propensity scores. In this paper we discuss three weighting-based estimators of the SATE that use these scores. The first is the widely used Inverse Probability Weighted or IPW estimator. The second and third are model-assisted versions of the IPW, based on differing model assumptions about the specification of the treatment effect. Both have the desirable property of being doubly robust.

4.2 Estimators of the SATE that use propensity scores

Let y_i denote the value of the response for unit *i* in the observational data set of *N* units, with y_i equal to y_{1i} if unit *i* is exposed to a treatment, and y_{0i} if not. Put $\boldsymbol{\mu}_{1s} = n^{-1}\sum_{i=1}^{n}\boldsymbol{\mu}_1(\mathbf{x}_i)$ and $\boldsymbol{\mu}_{0s} = n^{-1}\sum_{i=1}^{n}\boldsymbol{\mu}_0(\mathbf{x}_i)$, so (1) becomes $\delta_s = \mu_{1s} - \mu_{0s}$. Let $I_i = 1$ denote membership of the treatment subsample (i.e., units exposed to the treatment) and $I_i = 0$ denote membership of the control subsample (i.e., units not exposed to the treatment), with $\boldsymbol{\pi}(\mathbf{x}_i) = \Pr\left(I_i = 1 \middle| \mathbf{x}_i\right)$. It immediately follows that $y_i = I_i y_{1i} + (1 - I_i) y_{0i}$. Furthermore, it is easy to see that a consistent IPW-type estimator of μ_{1s} is $\tilde{\boldsymbol{\mu}}_{1s}^{IPW} = \sum_{i=1}^{n} \tilde{w}_{is}^{IPW-1} I_i y_i$ and a similarly consistent IPW-type estimator of μ_{0s} is $\tilde{\boldsymbol{\mu}}_{0s}^{IPW} = \sum_{i=1}^{n} \tilde{w}_{is}^{IPW-0}\left(1 - I_i\right) y_i$, with

$$\tilde{w}_{is}^{IPW-1} = \left(\boldsymbol{\pi}(\mathbf{x}_i)\right)^{-1}\left[\sum_{j=1}^{n}\left(\boldsymbol{\pi}(\mathbf{x}_j)\right)^{-1} I_j\right]^{-1}$$

and

$$\tilde{w}_{is}^{IPW-0} = \left(1 - \boldsymbol{\pi}(\mathbf{x}_i)\right)^{-1}\left[\sum_{j=1}^{n}\left(1 - \boldsymbol{\pi}(\mathbf{x}_j)\right)^{-1}\left(1 - I_j\right)\right]^{-1}.$$

The corresponding IPW-type estimator of (1) is then $\tilde{\boldsymbol{\delta}}_s^{IPW} = \tilde{\boldsymbol{\mu}}_{1s}^{IPW} - \tilde{\boldsymbol{\mu}}_{0s}^{IPW}$. However, as we have already noted, the treatment assignment mechanism underpinning exposure is unknown, and is therefore modeled as $\boldsymbol{\pi}(\mathbf{x}_i;\boldsymbol{\eta})$, where π is a known function and $\boldsymbol{\eta}$ is a vector of unknown parameters. The observed values of I_i and $\mathbf{x}_i$ can be used to estimate $\boldsymbol{\eta}$, leading to propensity scores $\boldsymbol{\pi}(\mathbf{x}_i;\hat{\boldsymbol{\eta}})$, where $\hat{\boldsymbol{\eta}}$ is the vector of estimated parameter values. This leads to the plug-in IPW estimator for (1)

$$\hat{\boldsymbol{\delta}}_s^{IPW} = \sum\nolimits_{i=1}^{n} \hat{w}_{is}^{IPW-1} I_i y_i - \sum\nolimits_{i=1}^{n} \hat{w}_{is}^{IPW-0}\left(1-I_i\right) y_i \qquad (3)$$

where

$$\hat{w}_{is}^{IPW-1} = \left(\boldsymbol{\pi}(\mathbf{x}_i;\hat{\boldsymbol{\eta}})\right)^{-1}\left[\sum\nolimits_{j=1}^{n}\left(\boldsymbol{\pi}(\mathbf{x}_j;\hat{\boldsymbol{\eta}})\right)^{-1} I_j\right]^{-1}$$

and

$$\hat{w}_{is}^{IPW-0} = \left(1-\boldsymbol{\pi}(\mathbf{x}_i;\hat{\boldsymbol{\eta}})\right)^{-1}\left[\sum\nolimits_{j=1}^{n}\left(1-\boldsymbol{\pi}(\mathbf{x}_i;\hat{\boldsymbol{\eta}})\right)^{-1}\left(1-I_j\right)\right]^{-1}.$$

Provided the three basic assignment properties listed in Section 4.1 apply, $\hat{\boldsymbol{\delta}}_s^{IPW}$ is consistent for the SATE δ_s.

The IPW estimator (3) does not explicitly control for differences in the covariate distributions between the treatment and control subsamples, assuming instead that the overlap of the covariate distributions across these subsamples is sufficient to ensure that these differences cancel out. Unfortunately, in many situations these differences account for a significant portion of the variation in the treatment and control response values. A simple way of accounting for these sources of variation is to assume additive treatment effects, that is $y_{1i} = \boldsymbol{\lambda}_i + y_{0i}$, with $\boldsymbol{\lambda}_i$ defining the treatment effect for sample unit *i*. Substituting in $\hat{\boldsymbol{\delta}}_s^{IPW}$, we see that

$$\hat{\boldsymbol{\delta}}_s^{IPW} = \widehat{\boldsymbol{\delta}}_s^{IPW-null} + \sum\nolimits_{i=1}^{n} \hat{w}_{is}^{IPW-1} I_i \boldsymbol{\lambda}_i$$

where $\widehat{\boldsymbol{\delta}}_s^{IPW-null}$ is the value of $\hat{\boldsymbol{\delta}}_s^{IPW}$ when y_i is replaced by $y_i - I_i\boldsymbol{\lambda}_i$. Since there are no treatment effects distinguishing the "treated" units from the "control" units in $\widehat{\boldsymbol{\delta}}_s^{IPW-null}$, we can see that this term is purely an estimate of the differential impact of the population covariates on the realized value of $\hat{\boldsymbol{\delta}}_s^{IPW}$, something that is asymptotically zero under randomized assignment but will usually be non-zero in observational data. It follows that the difference $\hat{\boldsymbol{\delta}}_s^{IPW} - \widehat{\boldsymbol{\delta}}_s^{IPW-null}$ is then a covariate-adjusted estimator of (1). However, calculation of this difference requires estimation of $\boldsymbol{\lambda}_i$, say by

$\hat{\boldsymbol{\lambda}}_i$. Let $\hat{\boldsymbol{\delta}}_s^{IPW-null}$ denote the value of (3) when y_i is replaced by $y_i - I_i\hat{\boldsymbol{\lambda}}_i$. Our covariate-adjusted estimator of (1) is then

$$\hat{\boldsymbol{\delta}}_s^{IPW-L} = \hat{\boldsymbol{\delta}}_s^{IPW} - \hat{\boldsymbol{\delta}}_s^{IPW-null} = \sum\nolimits_{i=1}^{n} \hat{w}_{is}^{IPW-1} I_i \hat{\boldsymbol{\lambda}}_i \,. \tag{4}$$

We illustrate use of (4) in the application discussed in the next section, while in Section 4.2 we show that (4) is a double robust estimator of the SATE under unconfoundedness.

In Section 2 we discussed how one could achieve double robustness, or DR, in a survey-sampling context by adding a design-consistent bias correction to a model-based estimator. This model-assisted approach assumes ICE, which in the context of estimation of the SATE corresponds to unconfoundedness. We now apply this idea to estimation of the SATE. To start, we note that the empirical version of the SATE δ_s is

$$d_s = n^{-1}\sum\nolimits_{i=1}^{n} y_{1i} - n^{-1}\sum\nolimits_{i=1}^{n} y_{0i} \tag{5}$$

so if we view the whole sample as the population of interest and use the treated sample to construct a model-assisted estimate of the first term on the right in (5) and the control sample to construct a similarly model-assisted estimate of the second term on the right in (5) then the difference of these two model-assisted estimates is a model-assisted estimate of d_s and hence of the SATE δ_s:

$$\begin{aligned}\tilde{\delta}_s^{IPW-MA} = &\left(n^{-1}\sum\nolimits_{i=1}^{n} \hat{m}_{1i} + \sum\nolimits_{i=1}^{n} \frac{\pi_i^{-1} I_i}{\sum\nolimits_{j=1}^{n} \pi_j^{-1} I_j} \left(y_i - \hat{m}_{1i} \right) \right) \\ &- \left(n^{-1}\sum\nolimits_{i=1}^{n} \hat{m}_{0i} + \sum\nolimits_{i=1}^{n} \frac{\left(1-\pi_i\right)^{-1}\left(1-I_i\right)}{\sum\nolimits_{j=1}^{n} \left(1-\pi_j\right)^{-1}\left(1-I_j\right)} \left(y_i - \hat{m}_{0i} \right) \right).\end{aligned}$$

Here $\hat{m}_{1i} = m_1(\mathbf{x}_i;\hat{\beta}_1)$ and $\hat{m}_{0i} = m_0(\mathbf{x}_i;\hat{\beta}_0)$ are our model-based estimates of $E\left(y_{1i} \middle| \mathbf{x}_i\right)$ and $E\left(y_{0i} \middle| \mathbf{x}_i\right)$ respectively, based on separate fits to the treated and control sample units. The corresponding propensity score-based version of this model-assisted estimator is therefore

$$\hat{\delta}_s^{IPW-MA} = \left(n^{-1}\sum\nolimits_{i=1}^{n} \hat{m}_{1i} - n^{-1}\sum\nolimits_{i=1}^{n} \hat{m}_{0i} \right) + \sum\nolimits_{i=1}^{n} \hat{w}_{is}^{IPW-1} I_i \left(y_i - \hat{m}_{1i} \right) - \sum\nolimits_{i=1}^{n} \hat{w}_{is}^{IPW-0} \left(1 - I_i\right)\left(y_i - \hat{m}_{0i} \right). \tag{6}$$

Assuming unconfoundedness, (6) is the difference of two DR estimators. Consequently $\hat{\delta}_s^{IPW-MA}$ is DR for the SATE δ_s under the same assumption.

Scharfstein *et al.* (1999) discuss how to construct a DR estimator for causal inference under unconfoundedness (see also Bang and Robins, 2005). Furthermore, the idea for constructing a model-assisted DR correction to the IPW estimator (3) is not new. Robins *et al.* (1994) introduced a similar estimator to (6) that adjusts for differences in covariate distributions between the treated and control samples. This is the Augmented IPW or AIPW estimator, see Rotnitzky *et al.* (1998). Following Lunceford and Davidian (2004) the AIPW is calculated as

$$\hat{\delta}_s^{AIPW} = \sum_{i=1}^n \hat{w}_{is}^{IPW-1}\left\{I_i y_i - \left(I_i - \hat{\pi}_i\right)\hat{m}_{1i}\right\} - \sum_{i=1}^n \hat{w}_{is}^{IPW-0}\left\{\left(1-I_i\right)y_i + \left(I_i - \hat{\pi}_i\right)\hat{m}_{0i}\right\} \qquad (7)$$

which after simplification can be written

$$\hat{\delta}_s^{AIPW} = \hat{\delta}_s^{IPW-MA} + \sum_{i=1}^n \left(\frac{1}{\sum_{j=1}^n \hat{\pi}_j^{-1} I_j} - \frac{1}{n} \right) \hat{m}_{1i} - \sum_{i=1}^n \left(\frac{1}{\sum_{j=1}^n \left(1-\hat{\pi}_j\right)^{-1}\left(1-I_j\right)} - \frac{1}{n} \right) \hat{m}_{0i} .$$

That is, the AIPW is (6) plus two correction terms, each of which is asymptotically zero provided the propensity model is correctly specified. In practice the correction terms will be small since $\sum_{j=1}^n \hat{\pi}_j^{-1} I_j \approx \sum_{j=1}^n \left(1-\hat{\pi}_j\right)^{-1}\left(1-I_j\right) \approx n$, so the AIPW will also be DR.

Beyond the fact that they are calculated using data from the treated and control samples separately, the estimates $\hat{m}_{1i}$ and $\hat{m}_{0i}$ are not constrained in any way above. However, suppose that we proceed as we did in the development leading up to (4), assume an additive treatment effect and so write $y_{1i} = \boldsymbol{\lambda}_i + y_{0i}$. Then

$$\delta_s = n^{-1} \sum_{i=1}^n \lambda_i .$$

For identifiability, we also assume that we have access to a covariate $\mathbf{z}_i$ measuring the amount of exposure to the treatment and write $\lambda_i = f(\mathbf{z}_i;\theta)$. The model

$$y_i = f(\mathbf{z}_i;\theta) + m(\mathbf{x}_i;\beta) + e_i \qquad (8)$$

can be fitted to the data from the entire sample to obtain estimates $\hat{\theta}$ and $\hat{\beta}$. Setting $\hat{\lambda}_i = f(\mathbf{z}_i;\hat{\theta})$ then leads to a model-based estimator of δ_s:

$$\hat{\delta}_s^{MB} = \sum_{i=1}^n \hat{\lambda}_i I_i \Big/ \sum_{i=1}^n I_i . \qquad (9)$$

It turns out that the model-assisted estimator (6) and this model-based estimator (9) are identical when we put $\hat{m}_{0i} = y_i - \hat{\lambda}_i I_i$ and $\hat{m}_{1i} = \hat{\lambda}_i + \hat{m}_{0i} = y_i + (1-I_i)\hat{\lambda}_i$, i.e., we include all non-treatment related sources of variability in our estimate of m_{0i}. To see this, note that (6) in this case is

$$\begin{aligned}\hat{\delta}_s^{IPW-MA} &= \hat{\delta}_s^{MB} + \sum_{i=1}^n \hat{w}_{is}^{IPW-1} I_i\left(y_i - \hat{m}_{1i}\right) - \sum_{i=1}^n \hat{w}_{is}^{IPW-0}\left(1-I_i\right)\left(y_i - \hat{m}_{0i}\right)\\ &= \hat{\delta}_s^{MB} + \sum_{i=1}^n \hat{w}_{is}^{IPW-1} I_i\left(y_i - y_i - (1-I_i)\hat{\lambda}_i\right) - \sum_{i=1}^n \hat{w}_{is}^{IPW-0}\left(1-I_i\right)\left(y_i - y_i + \hat{\lambda}_i I_i\right)\\ &= \hat{\delta}_s^{MB}.\end{aligned}$$

It only remains to note that the model-assisted estimator (4) and the model-based estimator (9) will usually be close. This is not surprising, since the development leading to (4) makes the same additive treatment effects assumption about the relationship between y_{1i} and y_{0i} that underpins (9). As a consequence we expect (4) to be a more efficient model-assisted estimator of δ_s than (6) unless the latter is also based on an additive treatment effects model.

4.3 $\hat{\boldsymbol{\delta}}_s^{IPW-L}$ is DR under unconfounded treatment assignment and additive treatment effects

We assume unconfounded treatment assignment and the general additive treatment effect specification:

$$y_i = I_i\boldsymbol{\lambda}_i + m(\mathbf{x}_i;\boldsymbol{\beta}) + e_i = I_i\boldsymbol{\lambda}_i + y_{0i}$$

so $\boldsymbol{\delta}_s = n^{-1}\sum_{i=1}^n \boldsymbol{\lambda}_i$. We shall also assume that we have unbiased estimators $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ and $\hat{\boldsymbol{\lambda}}_i$ of $\boldsymbol{\lambda}_i$. Put $\hat{y}_{0i} = y_i - I_i\hat{\boldsymbol{\lambda}}_i = m(\mathbf{x}_i;\hat{\boldsymbol{\beta}}) + \hat{e}_i$, where $\hat{e}_i = y_i - I_i\hat{\boldsymbol{\lambda}}_i - m(\mathbf{x}_i;\hat{\boldsymbol{\beta}})$. Substituting in (4) leads to

$$\begin{aligned}
\hat{\boldsymbol{\delta}}_s^{IPW-L} &= \sum_{i=1}^n \hat{w}_{is}^{IPW-1} I_i y_i - \sum_{i=1}^n \hat{w}_{is}^{IPW-0}(1-I_i)y_i - \sum_{i=1}^n \hat{w}_{is}^{IPW-1} I_i\hat{y}_{0i} + \sum_{i=1}^n \hat{w}_{is}^{IPW-0}(1-I_i)\hat{y}_{0i} \\
&= \sum_{i=1}^n \left(\hat{w}_{is}^{IPW-1}I_i - \hat{w}_{is}^{IPW-0}(1-I_i)\right)\left(y_i - \hat{y}_{0i}\right) \\
&= \sum_{i=1}^n \left(\hat{w}_{is}^{IPW-1}I_i - \hat{w}_{is}^{IPW-0}(1-I_i)\right)\left(I_i\boldsymbol{\lambda}_i + m(\mathbf{x}_i;\boldsymbol{\beta}) + e_i - \hat{y}_{0i}\right) \\
&= \sum_{i=1}^n \hat{w}_{is}^{IPW-1} I_i\boldsymbol{\lambda}_i + \sum_{i=1}^n \left(\hat{w}_{is}^{IPW-1}I_i - \hat{w}_{is}^{IPW-0}(1-I_i)\right)\left(m(\mathbf{x}_i;\boldsymbol{\beta}) + e_i - \hat{y}_{0i}\right).
\end{aligned}$$

That is, (4) can be written

$$\hat{\boldsymbol{\delta}}_s^{IPW-L} = \sum_{i=1}^n \hat{w}_{is}^{IPW-1} I_i\boldsymbol{\lambda}_i + \sum_{i=1}^n \hat{w}_{is}^{IPW-1} R_{0i} - \sum_{i=1}^n \hat{w}_{is}^{IPW-0}(1-I_i)R_{0i} \qquad (10)$$

where $R_{0i} = y_{0i} - \hat{y}_{0i}$ is the population residual for y_{0i}. The sample mean of these residuals will be a consistent estimator of zero if the model for the control or untreated outcome is valid, in which case the second and third terms on the right hand side of (10) will also each be consistent estimators of zero provided unconfoundedness applies. Similarly, the first term is consistent for the sample mean $\boldsymbol{\delta}_s$ of the $\boldsymbol{\lambda}_i$ in this case. That is, under unconfoundedness, (10) is consistent for $\boldsymbol{\delta}_s$ provided the model for y_i is correctly specified, irrespective of the validity of the model for treatment assignment. Alternatively, if the model for the treatment assignment is correctly specified then unconfoundedness again implies that the asymptotic distributions of the R_{0i} will be the same in both the treated and untreated parts of the sample, in which case the second and third summations on the right hand side of (10) correspond to the difference of two consistent estimators of the same expected value, and so this difference converges to zero irrespective of whether the control model is valid or not. Furthermore, the first term on the right hand side of (5) then converges to the IPW estimator of $\boldsymbol{\delta}_s$. This estimator is design-unbiased for $\boldsymbol{\delta}_s$ under a correctly specified treatment assignment model. That is, under unconfoundedness and additive treatment effects, (5) is consistent for $\boldsymbol{\delta}_s$ when the model for treatment assignment is correctly specified, irrespective of the validity of the control model. We conclude that (10) is a DR estimator of $\boldsymbol{\delta}_s$ provided the assumptions of unconfounded treatment assignment and additive treatment effects hold true.

5. An application of model-based causal inference: Rainfall enhancement in Oman

5.1 Background

A randomized trial of a ground-based rainfall enhancement technology was carried out in the Hajar Mountains of Oman 2013 – 2018. The hypothetical mechanism for rainfall enhancement using this technology is via downwind transport of natural aerosols that have become ionized following exposure to an operating ionizer, resulting in larger raindrop formation downwind and hence heavier rain than would be the case if the ionizer was not operating. During the trial, ionizers were operated according to a randomized daily operating schedule, subject to equal numbers of deployed ionizers being switched on and switched off each day. However, it is impossible to randomize the exposure of any particular downwind rain gauge to an operating ionizer since this depends on whether the gauge is downwind of the operating ionizer, and the downwind direction changes daily according to prevailing meteorological conditions.

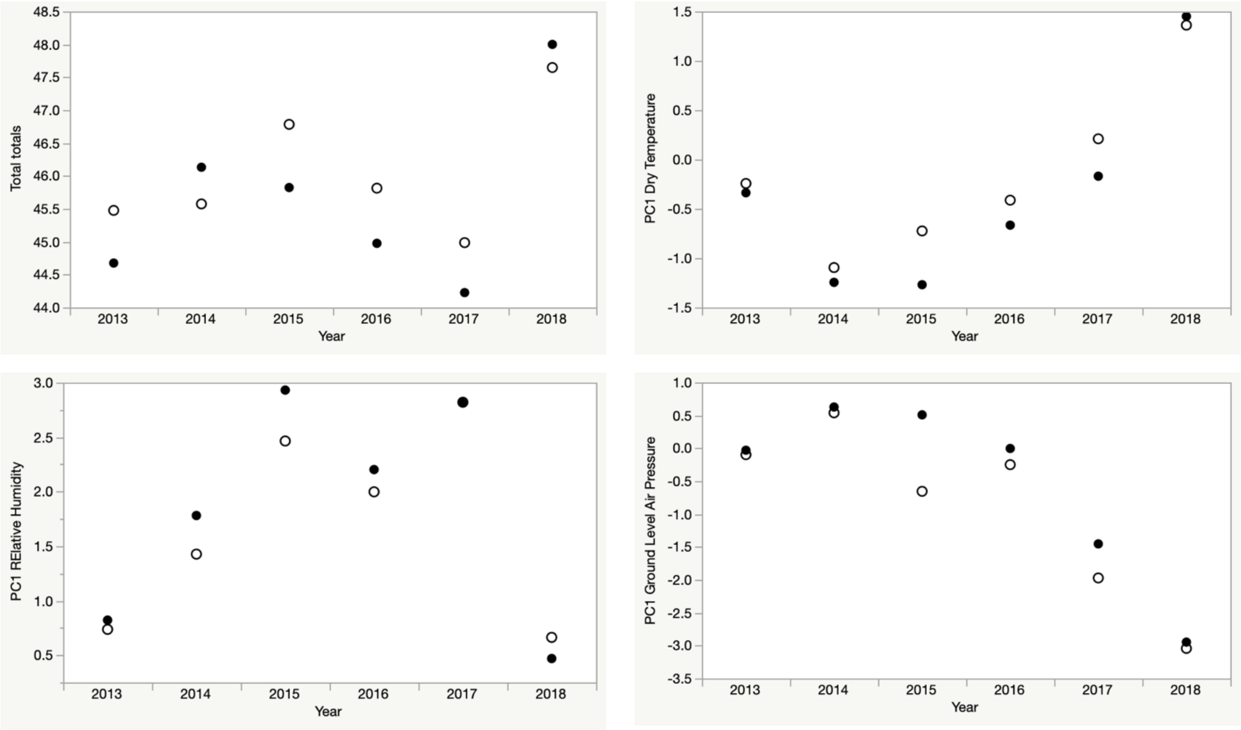
Our aim here is to test the causal hypothesis that exposure to an operating ionizer led to enhanced rainfall in rain gauges that were downwind of installed ionizers in the Hajar Mountains over 2013 – 2018. Our observation units are gauge-days, with a positive rainfall value at a gauge on a day classified as a target value when that gauge is downwind of at least one operating ionizer on the day. Otherwise, it is classified as a control value. From a causal perspective, target values are "treated" values, while control values are "untreated". We consider two types of rainfall measurements, actual rainfall (*Rain*), defined as positive values of rainfall, and the logarithm of actual rainfall (*LogRain*), with the latter of more interest given the huge skewness in the distribution of actual gauge-day rainfall measurements in the Hajar Mountains 2013 – 2018.

Let Y denote either *Rain* or *LogRain* for an actual rainfall gauge-day, and let I denote the zero-one indicator for whether an actual rainfall gauge-day value is a control value ($I = 0$) or a target value (I

= 1). Put $\pi(\mathbf{x}) = \Pr\left(I = 1 \middle| \mathbf{X} = \mathbf{x}\right) = E(I \middle| \mathbf{X} = \mathbf{x})$, where **X** denotes a vector of covariate measurements such that it is reasonable to assume that *Y* and *I* are conditionally independent given **X**. There were *n* = 4168 actual rainfall gauge-day observations spread over 488 days during 2013 – 2018 for the Hajar Mountains trial, and we seek to test the hypothesis that, on average, those observations that were exposed to an operating ionizer (i.e., the target observations) were significantly larger than those that were not (i.e., the control observations). By "on average" here we mean over the 4168 gauge-day values of actual rainfall that were observed 2013 - 2018.

The full duration of the Hajar Mountains trial over 2013 – 2018 was 849 days. However, wind direction data were missing for 109 of these days, all between 2016 and 2018. This was essentially due to problems with the operation of the radiosonde at Muscat International Airport. Since these wind direction data are necessary to determine whether a gauge-day rainfall measurement is downwind or not (and hence can be allocated as either a target or a control value), this meant that the final analysis of the trial data is restricted to the 740 days for which wind direction data were available. Ionizer operations over the entire trial were carried out according to a balanced randomized operating schedule, so a more detailed analysis of the trial reported in Chambers *et. al.* (2021) treats the missing days as missing completely at random, since there seems no obvious reason to link issues with radiosonde operation at Muscat International Airport with ionizer operation in the Hajar Mountains. Here though we proceed more cautiously, noting that it can also be argued that despite the randomized operating schedule for the ionizer mechanisms, this did not necessarily translate into a randomized target vs. control allocation for downwind rain events. In particular, the missing days for operation of the Muscat radiosonde were all in the last three years of the trial: 2016 (18.3% days missing), 2017 (29.4% days missing) and 2018 (43.1% days missing). Furthermore, as can be seen in Figure 1, these years corresponded to both a drying out of the trial region and, in 2018, a switch in the meteorological conditions (and hence rainfall over the mountains) to which target and control gauge-days were exposed. This raised the prospect of a lack of independence between the gauge-day rainfall measurements and their target/control status. Consequently, it becomes important that one also takes account of the possible informativeness of the sample of days when wind directions were available. This is the issue that we address in this section, referring the reader to Chambers *et. al*. (2021) for a more comprehensive model-based analysis of the trial data that ignores this potential source of bias.

Figure 1: Annual average values of key meteorological indicators for target gauge-days (filled circles) and control gauge-days (open circles), Hajar Mountains trial 2013-2018. The variables are defined in Section 5.2.



5.2 Estimation of an average target effect using propensity score weighting

Let *i* index gauge and *j* index day. We define the average target effect δ_s as the difference between the average response values for the 2176 target gauge-days and the 1992 control gauge-days over the trial when actual rainfall was recorded downwind of the installed ionizers. In order to calculate the IPW estimator of this effect we first need to model the propensity score associated with gauge-day ij. This is the estimate of the expected value $\pi(\mathbf{x}_{ij})$ for the binary indicator I_{ij} defined by the target status (target/control) of rainfall on gauge-day ij conditional on a covariate $\mathbf{x}_{ij}$ reflecting observed meteorological conditions on day *j*. We use a logistic specification for $\pi(\mathbf{x}_{ij})$. Standard model searches lead to the specification, with associated estimated parameter values, set out in Table 1. All terms are highly significant and are given by:

- An index for storm development potential (Total.totals);
- First principal component of average dry air temperature (PC1 Dry Temperature);
- First principal component of average relative humidity (PC1 Relative Humidity);
- First principal component of average ground level air pressure (PC1 Ground Level Air Pressure).

Note that principal components were based on daily 10:00 – 20:00 average values computed across the network of automatic weather stations located in the Hajar Mountains.

Table 1: Parameter estimates for fitted propensity score model

Term	Estimate	Std Error	t Ratio
Intercept	0.753	0.225	3.342
Total.totals	-0.016	0.005	-3.240
PC1 Dry Temperature	0.172	0.040	4.282
PC1 Relative Humidity	0.110	0.025	4.454
PC1 Ground Air Pressure	0.115	0.024	4.766

Propensity score weighted average actual rainfall based on the 2176 target gauge-day values is 4.853 mm, with a corresponding weighted average value for *LogRain* of 0.554. In comparison, propensity score weighted average actual rainfall based the 1992 control gauge-day values is 4.640 mm, with a corresponding weighted average value for *LogRain* of 0.480. However, there was a large outlier in the control values of actual gauge-day rainfall. When this value is removed, weighted average actual rainfall for control gauge-days reduces to 4.560mm. These values imply IPW estimates of 0.293 mm (with outlier removed) for $Y = Rain$ and 0.074 (using all values) for $Y = LogRain$.

The sample design for the Hajar Mountains trial was such that on any given day a random half of the installed ionizers were operated, with the remaining half not operated, with the aim of ensuring treatment-control balance in exposure to daily meteorological conditions. Assuming these conditions were uniformly distributed across the trial area, this should have led to the number of target gauge-day observations downwind of the operating ionizers each day being approximately the same as the number of control gauge-day observations that were downwind of the non-operating ionizers. However, spatial variability in rainfall meant that numbers of targets and controls varied significantly from day to day. For the 488 days when rainfall was recorded downwind, 165 days either have no target data, or no control data. And, of the remaining 323 days, only 115 have at least 5 target values and at least 5 control values. These “Good Data” days correspond to solid circles in Figure 2. Furthermore, daily sums of propensity scores for the 323 days when there are data for both targets and controls track daily sample sizes but are very variable. See Figure 3. Finally, we note that refitting the propensity score model just using the data from the "Good Data" days leads to a rather different model specification compared to that shown in Table 1, which uses the data from all 488 days.

Figure 2: Scatterplot showing daily numbers of target and control gauge-day observations with actual rainfall, Hajar Mountains trial 2013-2018. Solid circles are "Good Data" days.

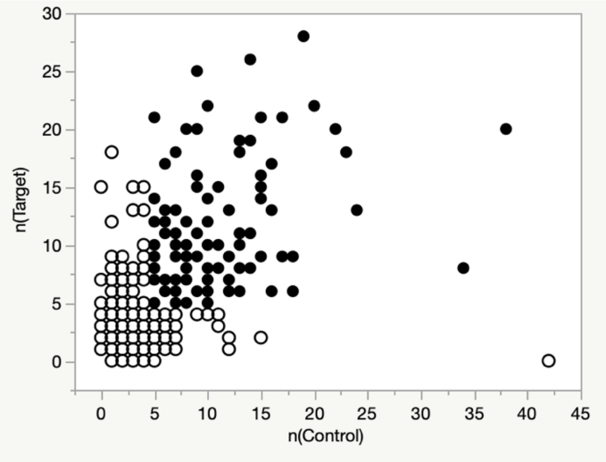
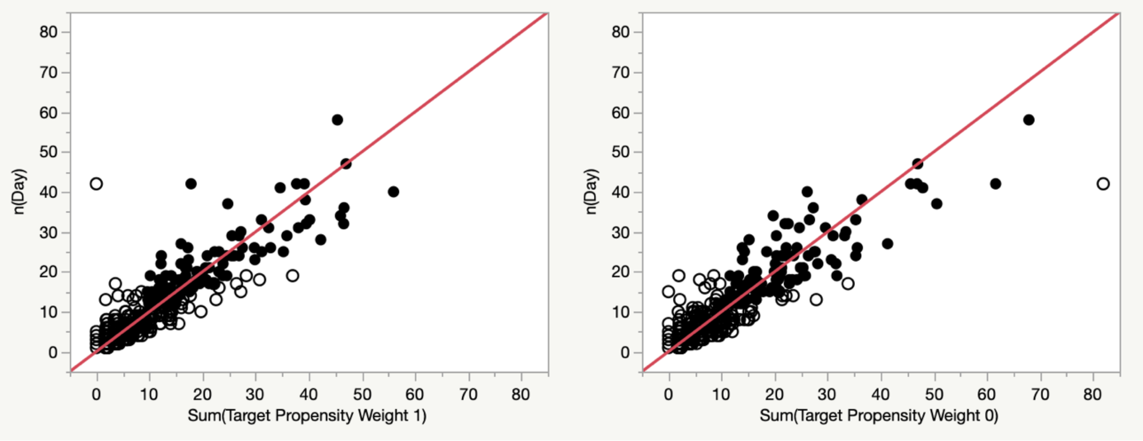


Figure 3: Daily numbers of downwind gauge-days with actual rainfall (vertical axis) vs. daily sums of propensity scores for target gauge-days (horizontal axis, left) and control gauge-days (horizontal axis, right). Plots restricted to days when both target and control rainfall observed. Line is the identity fit. Solid circles are "Good Data" days.



The propensity scores defined by the model set out in Table 1 are constant within a day (since they are a function of daily meteorological measurements), so the IPW estimate is the weighted sum of daily differences equal to average target rainfall minus average control rainfall on the day. Averages of these differences based on data from 323 days with both target and control gauge-day data are positive but not significantly greater than zero for both *Rain* and *LogRain*. When based on data from the 115 “Good Data” days, they are larger and significantly different from zero. This appears to be a consequence of their lower variability on “Good Data” days, when rainfall is more widespread. This implies correlation between these daily differences and meteorological conditions. However, there is no evidence of correlation with the meteorological variables defining the propensity scores, indicating other factors beyond target/control propensity may be present. It is possible that one or more of these factors may be correlated with the response variables (*Rain*, *LogRain*), suggesting that a more complex analysis of the rainfall data collected in the Hajar Mountains trial is necessary.

5.3 Using a random effects model for *LogRain* to control for unobserved sources of variation in rainfall

An alternative model-based approach to estimation of ionizer impact on rainfall enhancement was used in the analysis of the Hajar Mountains trial data described in Chambers *et. al*. (2021). This approach explicitly estimated the the counterfactuals corresponding to control values for target gauge-day observations. A key component of this analysis involved fitting a linear model with random day effects to the downwind *LogRain* values obtained in the trial. This model is specified in Table 2. It depends on daily meteorology via another linear model with random day effects fitted to *LogRain* values from gauges that were upwind of the ionizer sites each day. These upwind gauge-day readings should be unaffected by ionizer operation but should also be strong predictors of "natural" rainfall downwind of the ionizers. Fitted values from this upwind model (denoted Upwind LogRain in Table 2) were therefore used as a measure of expected downwind control rainfall. They were combined in the downwind model for *LogRain* with two elevation measures Gauge Elevation 1 (equal to gauge elevation when this value is 1km or less and is zero otherwise) and Gauge Elevation 2 (equal to gauge elevation when this value is greater than 1km and is zero otherwise) together with indicator variables for the year the data were obtained. The year 2015 is the reference year (these variables are denoted y2013, y2014, y2016, y2017 and y2018 below). Over the course of the trial, there were ten ionizers, denoted H01 – H10, that were operated, with H01 and H02 operated in 2013, H01 – H04 operated in 2014, H01 – H06 operated in 2015, H01 – H08 operated in 2016 and H01 – H10 operated in 2017 and 2018. The downwind model therefore included indicator variables (denoted Target H01 – Target H10 below) for whether the gauge-day observation was a target value for each of these ten ionizers H01 – H10 on the day. The REML estimates of the variance components for the downwind *LogRain* model are shown in Table 3, with the distribution of predicted Day effects generated by this model shown in Figure 4. The usual assumption of Gaussian random effects seems reasonable. Six out of the fourteen target-related effects in the fitted model are clearly significant, in the sense of having t ratios greater than 2. One more (Target H01, t ratio = 1.946) is very close to being classified as significant.

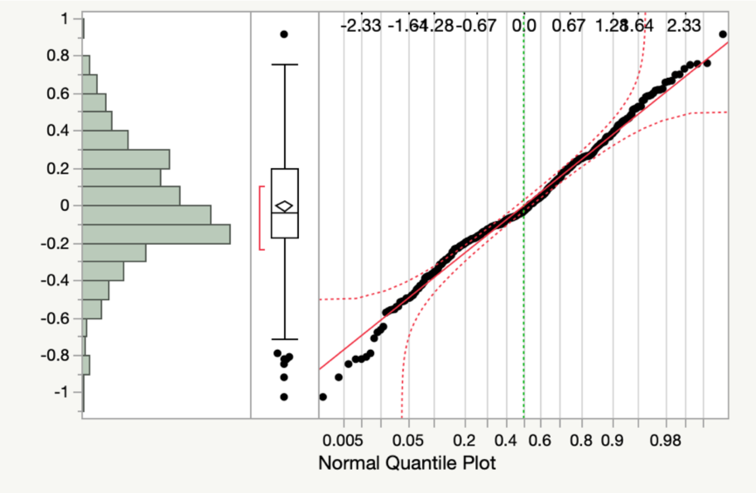
Table 2: Fitted parameter values with estimated standard errors and associated t ratios for the downwind *LogRain* model. Based on 4168 gauge-day observations. Entries for variables with absolute values of t ratios greater than 2 are in bold. Estimated Target Effects are greyed out.

Term	Estimate	Std Error	t Ratio
Intercept	0.077	0.121	0.633
y2013	**0.406**	**0.113**	**3.581**
y2014	**0.336**	**0.106**	**3.153**
y2016	**0.259**	**0.115**	**2.241**
y2017	0.092	0.119	0.774
y2018	0.041	0.146	0.277
Gauge Elevation 1	-0.200	0.164	-1.217
Gauge Elevation 2	-0.096	0.071	-1.357
Upwind LogRain	**0.945**	**0.059**	**15.983**
Target H01	0.481	0.247	1.946
Target H02	**0.840**	**0.293**	**2.867**
Target H03	**0.241**	**0.092**	**2.613**
Target H04	-0.114	0.089	-1.283
Target H05	**0.499**	**0.132**	**3.788**
Target H06	-0.136	0.149	-0.916
Target H07	0.336	0.188	1.785
Target H08	0.131	0.129	1.023
Target H09	**0.711**	**0.307**	**2.319**
Target H10	0.196	0.170	1.149
Gauge Elevation 1*Target H01	-0.488	0.363	-1.342
Gauge Elevation 1*Target H02	**-1.272**	**0.469**	**-2.714**
Gauge Elevation 2*Target H01	-0.163	0.159	-1.026
Gauge Elevation 2*Target H02	**-0.458**	**0.172**	**-2.667**

Table 3: REML variance component estimates for the downwind model for *LogRain*

Source	Var Comp	Pct of Total
Day	0.225	10.836
Residual	1.853	89.164

Figure 4: Distribution of predicted day effects generated by the fitted downwind model for gauge-day values of *LogRain*.



At the end of Section 5.2 we expressed concern that the propensity scores defined by Table 1 are not sufficient to control for the impact of unobserved variables on the difference δ_s between the target average value of *LogRain* and the control average value of this variable. In particular, for the reasons outlined at the end of Section 5.1, it is possible that the achieved target/control allocation in the available data (i.e., for those days when radiosonde operation made it possible to identify a wind direction) may in fact be informative. As a consequence, we now investigate how combining the model for *LogRain* defined by Tables 2 and 3 with the propensity scores defined by the model set out in Table 1, and using the alternative DR estimators (6), (7) and (4), which we now denote by IPW-MA, AIPW and IPW-L respectively, as well as the model-based estimator (9), now denoted MB, allows us to at least achieve some measure of protection against this scenario.

Let y_{ij} denote the value of *LogRain* for gauge *i* and day *j*. We start by writing the model (8) for *LogRain* set out in Tables 2 and 3 in mixed linear model form:

$$y_{ij} = \mathbf{z}_{ij}^T\boldsymbol{\theta} + \mathbf{x}_{ij}^T\boldsymbol{\beta} + u_i + e_{ij}$$

where *i* indexes gauge and *j* indexes day, $\mathbf{z}_{ij}$ is the vector of target indicator variables Target H01 – Target H10 plus the four interaction terms Gauge Elevation *a* * Target H0*b*; *a*, *b* = 1, 2; $\mathbf{x}_{ij}$ is the vector of other fixed effects in the model for *LogRain*, including the intercept, u_i is a random day effect and e_{ij} is the gauge-day residual. Note that $\mathbf{z}_{ij}$ is a zero vector when the gauge-day observation is a control value. Put $\hat{\lambda}_{ij} = \mathbf{z}_{ij}^T\hat{\boldsymbol{\theta}}$. Then IPW-L can be calculated using (4), i.e., the propensity score weighted average of the sum of the greyed out "target effects" defined in Table 2, with MB defined analogously via (9). Note that these estimates relate to $Y = LogRain$. They can be extended to $Y = Rain$ by exponentiating y_{ij}, $\hat{y}_{0ij}$ and $\hat{\lambda}_{ij}$.

In order to compute IPW-MA and AIPW we first need to fit separate mixed linear models to the target and control values of $Y = LogRain$. Estimates for the fixed effects in these models are shown in Table 3, while variance component estimates are shown in Table 4.

Table 3: Fitted parameter values with estimated standard errors and associate p-values for separate *LogRain* models for target and control cases. Entries for variables with absolute values of t ratios greater than 2 are in bold.

Term	Estimate	Std Error	t Ratio
Target model (based on 2176 gauge-day observations over 407 days)			
Intercept	**0.439**	**0.147**	**2.992**
y2013	**0.389**	**0.145**	**2.681**
y2014	**0.454**	**0.138**	**3.286**
y2016	**0.347**	**0.148**	**2.351**
y2017	0.222	0.147	1.509
y2018	0.085	0.185	0.459
Gauge Elevation 1	**-0.694**	**0.198**	**-3.503**
Gauge Elevation 2	**-0.296**	**0.083**	**-3.590**
Upwind LogRain	**0.942**	**0.076**	**12.413**
Control model (based on 1992 gauge-day observations over 404 days)			
Intercept	0.025	0.147	0.172
y2013	**0.491**	**0.146**	**3.372**
y2014	0.248	0.135	1.835
y2016	0.277	0.147	1.884
y2017	0.096	0.153	0.627
y2018	0.097	0.191	0.508
Gauge Elevation 1	-0.066	0.204	-0.325
Gauge Elevation 2	-0.062	0.088	-0.702
Upwind LogRain	**0.903**	**0.079**	**11.459**

Table 4: REML variance component estimates for the target and control models for *LogRain*

Source	Var Comp	Pct of Total
Target model		
Day	0.297	13.983
Residual	1.827	86.017
Control model		
Day	0.281	13.641
Residual	1.779	86.359

Let $\hat{\beta}_1$ and $\hat{\beta}_0$ denote the parameter estimates shown in Table 3, where a subscript of "1" indicates target and "0" indicates control. Then, after setting $\hat{m}_{1ij} = \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}}_1$ and $\hat{m}_{0ij} = \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}}_0$, we calculate IPW-MA and AIPW using (6) and (7) respectively. We again note that these estimates relate to *Y* = *LogRain*, with the corresponding estimates for *Y* = *Rain* obtained using exponentiated versions of $\hat{m}_{1ij}$ and $\hat{m}_{0ij}$. We also note that the estimates for *Rain* could be improved by correcting for back transformation bias, which we ignore here.

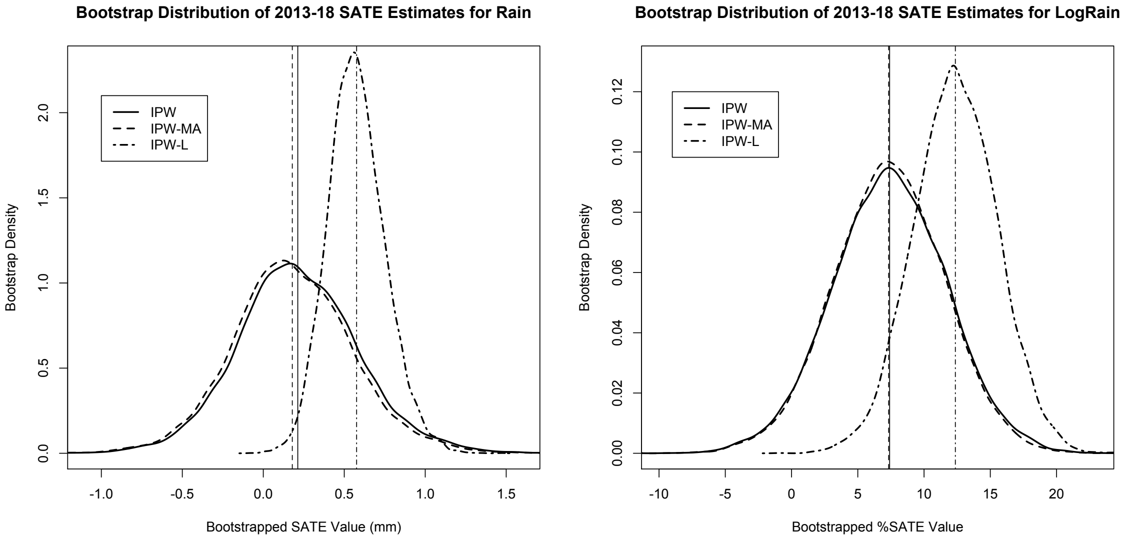
We use a two level semi-parametric block bootstrap (Chambers and Chandra, 2013) to calculate standard errors and associated p-values for these estimates, using a total of 10,000 bootstrap samples. Table 5 shows the estimates for δ_s, as well as their bootstrap standard errors and key points in their bootstrap distributions, for both *Rain* and *LogRain* based on the downwind rainfall recorded on the 488 days of the trial when rain was recorded downwind of the ionizer mechanisms. Figure 5 shows the bootstrap distributions for IPW, IPW-MA and IPW-L. The bootstrap distribution for AIPW is virtually indistinguishable from that for IPW-MA, and the bootstrap distribution for MB is also essentially indistinguishable from that for IPW-L.

It is noteworthy that the model-assisted estimator IPW-MA that only corrects for variability in the expected value of the response (*LogRain* here) due to differences in target and control exposure to meteorological and topographical conditions leads to estimates that that are very close to those obtained using IPW, an estimator that does not control for this variability. In contrast, the model-assisted estimator IPW-L that corrects for all non-treatment related sources of variability leads to substantially greater estimates of δ_s and noticeably less variability. Furthermore, all five propensity score based estimators that we considered in our analysis recorded bootstrap probabilities of less than five per cent for zero or negative rainfall enhancement effects from exposure to ionization operation, and in the case of IPW-L and MB, vanishingly less.

Table 5: Estimated values of δ_s for *Rain* and for *LogRain* for all 488 days when downwind rain was recorded in the Hajar Mountains trial, 2013-2018. Block bootstrap standard errors and one-sided bootstrap p-values (bootstrap probability of no effect or negative effect) are shown as well as key points in the bootstrap distribution.

	IPW	IPW-MA	AIPW	IPW-L	MB
			Rain		
Estimate	0.213	0.179	0.178	0.576	0.591
Std Dev	0.376	0.372	0.372	0.176	0.178
Pr<0	0.280	0.312	0.314	<0.001	<0.001
0.1%	-1.043	-1.082	-1.084	0.070	0.081
0.5%	-0.792	-0.824	-0.825	0.153	0.162
2.5%	-0.506	-0.536	-0.537	0.254	0.270
50%	0.201	0.167	0.166	0.567	0.582
97.5%	0.992	0.946	0.944	0.949	0.970
99.5%	1.296	1.224	1.222	1.087	1.102
99.9%	1.610	1.562	1.560	1.190	1.214
			LogRain		
Estimate	0.074	0.073	0.073	0.124	0.126
Std Dev	0.043	0.042	0.042	0.031	0.031
Pr<0	0.043	0.039	0.039	<0.001	<0.001
0.1%	-0.060	-0.063	-0.064	0.025	0.027
0.5%	-0.039	-0.039	-0.039	0.040	0.042
2.5%	-0.010	-0.009	-0.010	0.062	0.064
50%	0.074	0.074	0.073	0.124	0.126
97.5%	0.158	0.154	0.153	0.184	0.187
99.5%	0.184	0.179	0.178	0.203	0.205
99.9%	0.206	0.201	0.201	0.217	0.219

Figure 5: Two level semi-parametric block bootstrap distributions for IPW, IPW-MA and IPW-L estimates for *Rain* and for *LogRain*. Estimates for *LogRain* are multiplied by 100.

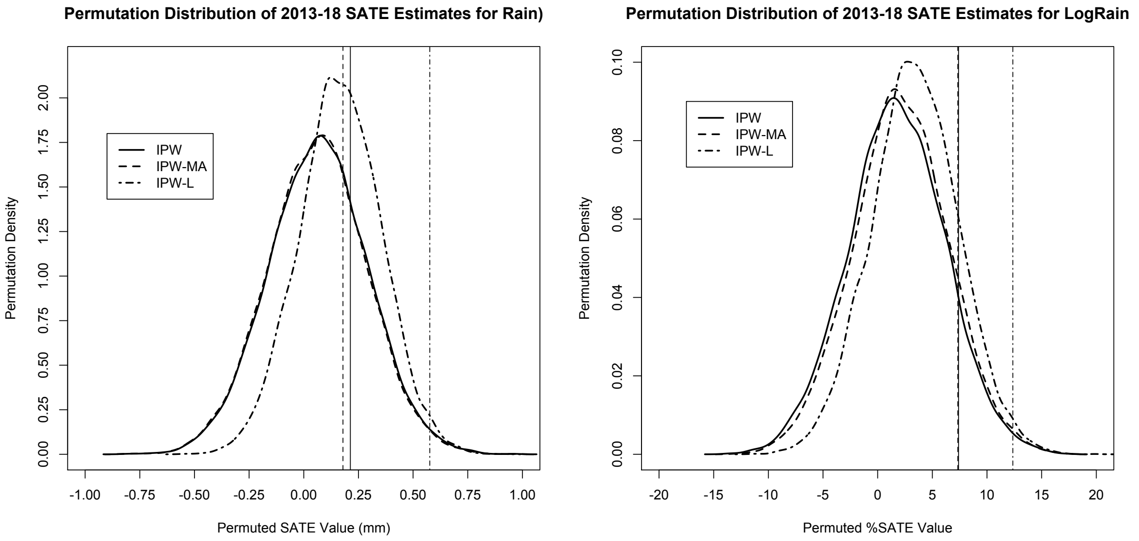


Following Chambers et al. (2021), we also report a randomization analysis of the significance of the effect of ionizer operation on positive rainfall as measured by the different estimators of δ_s shown in Table 5. This was done by independently randomly permuting the operating states of each ionizer each day (while maintaining the requirement that there were an equal number of operating and non-operating ionizers each day). This was done 10,000 times. The permutation p-value for an estimator was then calculated as the proportion of permuted values for that estimator that were greater than the observed estimate value. These p-values are set out in Table 6 below, with the randomization distributions for the permuted values shown in Figure 6. Again, we just show the distributions for IPW, IPW-MA and IPW-L. Note that on both the natural (*LogRain*) and the raw (*Rain*) scales there is a very low probability that our estimated values of IPW-L and MB could have been by chance. This probability is greater for the more variable estimators IPW, IPW-MA and AIPW, but still only around 10 per cent on the natural scale.

Table 6: Permutation p-values for estimators of δ_s for *Rain* and for *LogRain* for the 488 days when downwind rain was recorded in the Hajar Mountains trial, 2013-2018.

Y	IPW	IPW-MA	AIPW	IPW-L	MB
Rain	0.254	0.296	0.299	0.015	0.012
LogRain	0.096	0.115	0.117	0.013	0.011

Figure 6: Permutation distributions for IPW, IPW-MA and IPW-L for Rain and for LogRain generated by randomly permuting ionizer operating states over 2013-2018. Actual values for these estimators are shown as vertical lines in the plots. Values for *LogRain* are multiplied by 100.



6. Summary and a conclusion

As we stated at the beginning of this paper, weighting is at the core of sample survey inference. However, that does not mean that it is always required when analyzing sample survey data. In fact, in situations where the survey variable of interest can be modeled using covariates that include the factors that underpin the sample design, or when the sample design is completely random (a very rare event!), then the survey weights play a much reduced role. Thus, if a strict design-based approach to inference is taken, and so the survey weights are essentially the inverses of the sample inclusion probabilities, then it is easy to see that these weights are ancillary and their use leads to potentially inefficient inference. Much more efficient model-based or model-assisted methods of inference are possible, with the primary distinction between these two approaches being that the first takes the model seriously and consequently leads to more efficient inferences than the second – provided the assumption that the sample inclusion probabilities are ancillary is valid. On the other hand, the second approach is more cautious, allowing for model misspecification by including a design-based bias correction. This insurance comes at a cost, however, with typically reduced efficiency if in fact the model is correctly specified.

This paper has been written for a special issue of Series A of the Journal of the Royal Statistical Society commemorating the research achievements of Fred Smith and Chris Skinner in survey sampling. Both men were giants in the field, and it was Chris Skinner who produced groundbreaking research on survey weighting. Again, as noted earlier, Chris was very definitely a proponent of the model-assisted approach, in that he viewed a model for a survey variable as essentially a working hypothesis and so inevitably a misspecification of reality. However, he also recognized that the insurance premium in terms of loss of efficiency when adopting a model-assisted approach could be high, and so looked for ways to minimize it. This led to two papers with colleagues, Skinner and Mason (2012) and Kim and Skinner (2013), that described methods for stabilizing the variability in model-assisted weights. These are discussed in Section 3, with the latter contribution focusing on the case of informative sampling, i.e., where there is incomplete knowledge of the factors underpinning the sample design, or where the survey response itself is a design factor.

In both of the papers referred to in the previous paragraph, there is an implicit assumption that the sample inclusion probabilities are available. This may be reasonable when the same organization carrying out inference is also responsible for the survey design. However, it is usually not reasonable for secondary analysis, where the analyst and the sample designer may have no contact at all. In this situation heroic assumptions are often made about the non-informativeness of the sample design. One important area of application where this assumption is usually avoided is in causal inference, where explicit models are built for sample inclusion probabilities. A key property of the resulting inference is then its double robustness, where the inference remains valid if either the sample inclusion (or allocation) model is correctly specified or if the assumed model for the survey variable is correctly specified. In fact, it turns out that this is precisely the insurance provided by adopting a model-assisted approach. In Section 4 of this paper we therefore focused on the most straightforward causal inference scenario, where the interest is in estimating the difference δ between the expected values of a "treated" response and an "untreated" response. Here we discussed the standard inverse probability weighted (IPW) estimator $\hat{\delta}_s^{IPW}$ for the sample average δ_s of these differences, as well as two alternative model-assisted estimators, (4) and (6), that modify $\hat{\delta}_s^{IPW}$ in order to control for differences in covariate distributions between treated and untreated sample units. We also show that (4) is double robust.

In Section 5 we provide a realistic application of causal inference based on data obtained in a multi-year randomized experiment investigating the use of ionization devices for rainfall enhancement in the Hajar Mountains of Oman. A model-based analysis of these data that ignored missing wind direction data (Chambers *et al.*, 2021) indicated that these devices led to an increase of around 15-18 per cent in rainfall over the trial. In order to account for this missingness, the estimators (4) and (6) were applied to these data using the same rainfall model specification as in this reference, together with a propensity score model for whether a rainfall gauge was impacted by operation of one or more of these ionization devices on a day. This analysis indicated that, over all days when rainfall was recorded at rain gauges downwind of the devices, there was a highly significant average increase of 0.576 mm per gauge per day of rainfall for target gauge-days when estimated using (4) and a significant increase of 0.179 mm per gauge per day when estimated using (6), with bootstrap variability and permutation test diagnostics indicating the former estimate is preferable. In attribution terms, i.e., as a percentage of expected control rainfall under the same conditions, this corresponds to a 13.9 per cent increase in rainfall when estimated using (4) and 10.4 per cent increase when estimated using (6). These values are somewhat lower but still consistent with the estimated attribution values obtained by the pure model-based analysis reported in Chambers *et al.* (2021). Since both the model-based and double robust methods show enhancement at over 10 per cent for this trial, with both methodologies indicating significant results, it seems reasonable to conclude that the ionization-based rain enhancement technology used in the Hajar Mountain trial did actually lead to increases in rainfall. This has quite significant implications for the use of this technology in other arid areas similar to those that exist in Oman.

Before closing, we comment briefly variance estimation issues for the estimators developed in Section 4 and on making causal inferences about a target population based on sample data. As far as variance estimation is concerned, analytic approximations can clearly be developed. However, the bootstrap method that we use in Section 5 seems a good general-purpose tool for this purpose, albeit one that is computationally intensive. Turning to population causal inference, we note that in this case we are interested in external, rather than internal, validity. Staying with the idea of a single treatment, we can consider the situation where every unit in a population has either been exposed (treatment) or not (control), and we have observational data corresponding to a sample from the population. In this sample we observe treatment/control status as well as values of the response of interest, and our target of inference is the Population Average Treatment Effect (PATE) defined in (2). Estimation of this quantity depends on what information we have about the distribution of treated and control units in the population. The simplest case is where we do not have this information, but we do have sample weights. Typically these will control for both sample inclusion probabilities as well as differences between the sampled and non-sampled population units with respect to auxiliary population information. Let $v_{is}(\mathbf{X}_U)$ denote the sample weight for unit *i* given its auxiliary population information. We can then replace $\tilde{w}_{is}^{IPW-1}$ and $\tilde{w}_{is}^{IPW-0}$ in the IPW development at the start of Section 4.2 by

$$\tilde{w}_{is}^{IPW-1} = v_{is}(\mathbf{X}_U)\left(\pi(\mathbf{x}_i)\right)^{-1}\left[\sum_{j=1}^{n} v_{js}(\mathbf{X}_U)\left(\pi(\mathbf{x}_j)\right)^{-1} I_j\right]^{-1}$$

and

$$\tilde{w}_{is}^{IPW-0} = v_{is}(\mathbf{X}_U)\left(1-\pi(\mathbf{x}_i)\right)^{-1}\left[\sum_{j=1}^{n} v_{js}(\mathbf{X}_U)\left(1-\pi(\mathbf{x}_j)\right)^{-1}\left(1-I_j\right)\right]^{-1}$$

respectively. It is clear that the IPW (3) defined in terms of these "externally valid" weights will be a design-consistent estimator of the PATE. Extension to corresponding externally valid versions of IPW-MA and IPW-L is straightforward. What is not so straightforward, however, is the extension when we do know treatment/control allocation across the population, as would be the case if all individuals in a population are given the choice to take the treatment or not, and these choices are stored on a population register. We are currently researching this problem.

References

Bang, H. and Robins, J.M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, **61**, 962-973.

Beaumont, J.F. (2008). A new approach to weighting and inference in sample surveys. *Biometrika*, **95**, 539-553.

Brewer, K.R.W. (1999). Cosmetic calibration with unequal probability sampling. *Survey Methodology*, **25**, 205-212.

Chambers, R., Beare, S., Peak, S. and Al-Kalbani, M. (2021). Nudging a pseudo-science towards a science - The role of statistics in a rainfall enhancement trial in Oman. Submitted to the *International Statistical Review*.

Chambers, R. and Chandra, H. (2013). A random effect block bootstrap for clustered data. *Journal of Computational and Graphical Statistics*, **22**, 452-470.

Chambers, R.L. and Clark, R.G. (2012). An Introduction to Model-Based Survey Sampling with Applications. Oxford University Press: Oxford.

Chambers, R.L., Dorfman, A.H. and Wehrly, T.E. (1993). Bias robust estimation in finite populations using nonparametric calibration. *Journal of the American Statistical Association* **88**, 268-277.

Chen, S. and Haziza, D. (2017). Multiply robust imputation procedures for the treatment of item non-response in surveys. *Biometrika*, **104**, 439-453.

Chen, S. and Haziza, D. (2019). Multiply robust nonparametric multiple imputation for the treatment of missing data. *Statistica Sinica*, **29**, 2035-2053.

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

Hájek, J. (1971). Comment on "An Essay on the Logical Foundations of Survey Sampling, Part One". Page 236 in The Foundations of Survey Sampling (eds. V.P. Godambe and D.A. Sprott). Holt, Rinehart and Winston.

Han, P. (2014). Multiply robust estimation in regression analysis with missing data. *Journal of the American Statistical Association* **109**, 1159-1173.

Horvitz, D.G. and Thompson, D.J. (1952). A generalization of sampling with unequal probabilities and without replacement. *Journal of the American Statistical Association*, **47**, 663-685.

Imbens, G.W. and J. M. Wooldridge, J.M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, **47**, 5–86.

Imbens, G.W. and Rubin, D.B. (2015). Causal Inference in Statistics, Social and Biomedical Sciences. Cambridge University Press. Cambridge.

Kang, J.D.Y. and Schafer, J.L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, **22**, 523-539

Kim, J.K. and Skinner, C.J. (2013). Weighting in survey analysis under informative sampling. *Biometrika*, **100**, 385-398.

Li, F., Morgan, K.L. and Zaslavsky, A.M. (2018). Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, **113**, 390-400.

Lunceford, J. K., & Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine*, **23**, 2937-2960.

Narain, R.D. (1951). On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, **3**, 169–175.

Neyman, J. (1923). On the application of probability theory to agricultural experiments: Essay on principles, Section 9 of Mater Thesis. Translation republished in *Statistical Science*, **5**, 465-480, 1990.

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, **97**, 558-666.

Pfeffermann, D. and Sverchkov, M.Y. (1999). Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhya B*, **61**, 166-186.

Rosenbaum, P.R. and Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**, 41-55.

Rosenbaum PR, Rubin DB. (1984). Reducing bias in observational studies using subclassication on the propensity score. *Journal of the American Statistical Association*, **79**, 516–524.

Rotnitzky, A., Robins, J.M. and Scharfstein, D.O. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association*, **93**, 1321-1339.

Rubin, D.B. (1980). Comment on ‘Randomization Analysis of experimental data. The Fisher randomization test’, by D. Basu, *Journal of the American Statistical Association*, **75**, 591-593.

Saarela, O., Belzile, L. R., & Stephens, D. A. (2016). A Bayesian view of doubly robust causal inference. *Biometrika*, **103**, 667-681.

Scharfstein, D.O., Rotnitzky, A. and Robins, J.M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, **94**, 1096-1120.

Skinner, C. and Mason, B. (2012). Weighting in the regression analysis of survey data with a cross-national application. *The Canadian Journal of Statistics*, **40**, 697-711.

Yu, Z., & van der Laan, M. (2006). Double robust estimation in longitudinal marginal structural models. *Journal of Statistical Planning and Inference*, **136**, 1061-1089.

Zhou, T., Elliott, M. R., & Little, R. J. (2019). Penalized spline of propensity methods for treatment comparison. *Journal of the American Statistical Association*, **114**, 1-19.