Journal of the Royal Statistical Society
Statistics in Society
Series A

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

Journal:	*Journal of the Royal Statistical Society: Series A*
Manuscript ID	Draft
Manuscript Type:	Original Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Chambers, Ray; University of Wollongong, School of Mathematics and Applied Statistics; Ranjbar, Setareh Salvati, Nicola; Università di Pisa, Pacini, Barbara ; University of Pisa,

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

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 based on the observed sample. This can be an issue in causal analysis, and double robust methods that protect against misspecification of the sampling process have been the focus of much recent research. In this paper we propose a simple model-based modification to the popular inverse probability weighted estimator of an average treatment effect, and then illustrate its 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; 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.

A further assumption that is often made in this context is the *conditional independence assumption* (CIA)

$$\mathbf{y}_U \perp \mathbf{I}_U \,|\, \mathbf{X}_U$$

where $\mathbf{I}_U$ and $\mathbf{y}_U$ are the population values of I_i and y_i, $\perp$ denotes independence and the conditioning is with respect to known information about the population U which we denote here by a known $N \times p$ matrix $\mathbf{X}_U$. 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.

It is easy to see that if the CIA holds 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. A pure model-based approach is then where it is just the variability in $\mathbf{y}_U$ that drives inference. *Model-assisted* inference is a widely used compromise between design-based inference and pure model-based inference that allows for both sources of variability even under non-informative sampling. This approach is often assumed to provide both the non-parametric robustness of the design-based approach and the parametric efficiency associated with the pure model-based approach. However, this may not be the case, as we shall see.

1.2 Why this paper?

The aim of this paper is to provide an overview of the important issues that arise when using 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, or secondary, data. In this section we also develop a doubly robust estimator for an additive causal effect that behaves similarly to a model-assisted estimator, in that it uses 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, consider the classic design-based version of $\bar{y}_{ws}$. This is the inverse probability weighted (IPW) estimator corresponding to the Hájek (1971) version of the Horvitz and Thompson (1952) estimator for a finite population mean, where $w_{is} = w_{is}^{IPW} = \pi_i^{-1} \Big/ \sum_U \pi_j^{-1} I_j$.

Here $\pi_i = E\left(I_i \middle| \mathbf{X}_U\right)$ is the known sample inclusion probability. Put $\mu_i = E\left(y_i \middle| \mathbf{X}_U\right)$. The IPW estimator $\bar{y}_{ws}^{IPW}$ is consistent for μ under any model for μ_i when the CIA is valid since then $E\left(I_i w_i y_i \middle| \mathbf{X}_U\right) - \mu_i = E\left(I_i w_i (y_i - \mu_i) \middle| \mathbf{X}_U\right) = E\left(I_i w_i \middle| \mathbf{X}_U\right) E\left(y_i - \mu_i \middle| \mathbf{X}_U\right) = 0$. As a consequence, under suitable regularity conditions it follows that

$$E\left(\bar{y}_{ws}^{IPW}\right)-\mu=E\left\{E\left(\frac{\sum_U \pi_i^{-1}I_i y_i}{\sum_U \pi_i^{-1}I_i}-N^{-1}\sum_U y_i\right)\middle|\mathbf{X}_U\right\}\rightarrow E\left\{\frac{E\left(\sum_U \pi_i^{-1}I_i y_i\middle|\mathbf{X}_U\right)}{E\left(\sum_U \pi_i^{-1}I_i\middle|\mathbf{X}_U\right)}-N^{-1}\sum_U \mu_i\right\}=0.$$

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. We return to this issue later.

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 sample weights can be chosen to minimize $Var\left(\bar{y}_{ws}-\bar{y}_U\middle|\mathbf{X}_U\right)$ subject to $E\left(\bar{y}_{ws}-\bar{y}_U\middle|\mathbf{X}_U\right)=\bar{\mu}_{ws}-\bar{\mu}_U=0$. Let $\mathbf{w}_s^{MB}=\left\{w_i^{MB};i\in s\right\}$ denote these model-based weights, with associated estimator $\bar{y}_{ws}^{MB}$. Then by construction,

$$E\left(\bar{y}_{ws}^{MB}\middle|\mathbf{X}_U\right)=\sum_s w_i^{MB}\mu_i=\bar{\mu}_U$$

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 $\mathbf{1}_U$, $\mathbf{e}_U\perp\mathbf{X}_U$, $E\left(\mathbf{e}_U\right)=0$ 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 $\mathbf{w}_s^{MB}=\mathbf{X}_s\left(\mathbf{X}_s^T\mathbf{X}_s\right)^{-1}\bar{\mathbf{x}}_U$. 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. This is accomplished by adding a design-consistent bias correction to the model-based estimator, leading to the estimator

$$\bar{y}_{ws}^{MA} = \bar{y}_{ws}^{MB} + \left(\bar{y}_{ws}^{IPW} - \hat{\bar{\mu}}_{ws}\right) = \hat{\bar{\mu}}_U + \bar{r}_{ws}^{IPW}$$

where $\bar{r}_{ws}^{IPW}$ 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, with mean $\hat{\bar{\mu}}_U$. That is, if the linear model defined in the preceding paragraph is correctly specified then

$$\begin{aligned}
E\left(\bar{y}_{ws}^{MB}\right) &= E\left\{\left(\mathbf{w}_s^{MB}\right)^T \mathbf{y}_s\right\} \\
&= E\left\{E\left(\bar{\mathbf{x}}_U^T\left(\mathbf{X}_U^T diag(\boldsymbol{\pi}_U)\mathbf{X}_U\right)^{-1}\left(\mathbf{X}_U^T diag(\boldsymbol{\pi}_U)\mathbf{y}_U\right)\middle|\mathbf{X}_U\right)\right\} \\
&\rightarrow E\left(\bar{\mathbf{x}}_U^T\beta\right) = \mu
\end{aligned}$$

where $\boldsymbol{\pi}_U$ is the population vector of actual sample inclusion probabilities (which do not necessarily have to be the same as the assumed inclusion probabilities), while

$$\begin{aligned}
E\left(\bar{r}_{ws}^{IPW}\right) &= E\left\{E\left(\bar{y}_{ws}^{IPW} - \bar{\mathbf{x}}_s^T\hat{\beta}\middle|\mathbf{y}_U, \mathbf{X}_U\right)\right\} \\
&\rightarrow E\left\{E\left(\bar{y}_U - \bar{\mathbf{x}}_U^T\left(\mathbf{X}_U^T diag(\boldsymbol{\pi}_U)\mathbf{X}_U\right)^{-1}\mathbf{X}_U^T diag(\boldsymbol{\pi}_U)\mathbf{y}_U\middle|\mathbf{X}_U\right)\right\} \\
&\rightarrow \bar{\mathbf{x}}_U^T\beta - \bar{\mathbf{x}}_U^T\beta = 0
\end{aligned}$$

and so $E\left(\bar{y}_{ws}^{MA}\right) = E\left(\bar{y}_{ws}^{MB}\right) + E\left(\bar{r}_{ws}^{IPW}\right) \rightarrow \mu$ as well.

The convergence behavior indicated above will depend on regularity conditions, chief among which is the CIA. Note that since $\bar{y}_{ws}^{MA} = \bar{y}_{ws}^{IPW} - \left(\hat{\bar{\mu}}_{ws} - \hat{\bar{\mu}}_U\right)$ it follows that if the sample inclusion probabilities used in w_{is}^{IPW} are the same as the actual sample inclusion probabilities π_i then $\bar{y}_{ws}^{MA}$ is the design-consistent IPW estimator $\bar{y}_{ws}^{IPW}$ minus a design-consistent estimator of zero. It follows that $\bar{y}_{ws}^{MA}$ will have the same asymptotic repeated sampling behaviour as

$\bar{y}_{ws}^{IPW}$ irrespective of whether the model for $\mathbf{y}_U|\mathbf{X}_U$ is correctly specified or not. That is, $\bar{y}_{ws}^{MA}$ is design-consistent for $\bar{y}_U$ (and hence μ) even when the assumed model for $\mathbf{y}_U|\mathbf{X}_U$ is incorrectly specified. Alternatively, as we have shown above, $\bar{y}_{ws}^{MA}$ is also model-consistent for μ when the model for $\mathbf{y}_U|\mathbf{X}_U$ is correctly specified, irrespective of whether the sample inclusion probabilities used in w_{is}^{IPW} are correct or not. 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), and 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, Dorfman and Wehrly (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, 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 μ_i and a similar composite value for π_i. 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. We do not pursue this idea further here 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 $\bar{y}_{ws}^{MA}$. Some empirical evidence for this is provided in Chen and Haziza (2019).

3. Chris Skinner's impact on sample weighting methodology

3.1 Contributions to weighting under non-informative sampling

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) investigate how this variability can be reduced while at the same time retaining the desirable design-consistency property of the former two estimators. 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 $\hat{\mu}$ to the estimating equation

$$\sum_U w_{is}^{IPWX} I_i \left(y_i - \hat{\mu} \right) = 0$$

with $w_{is}^{IPWX} = d_i^{IPWX} \Big/ \sum_U d_j^{IPWX} I_j$. Note that with this definition,

$$E\left(\sum_s w_{is}^{IPWX} y_i \right) \rightarrow E\left(\sum_U \pi_i^{-1} q_i I_i y_i \middle| \mathbf{X}_U \right) \Big/ E\left(\sum_U \pi_i^{-1} q_i I_i \middle| \mathbf{X}_U \right) \rightarrow \sum_U q_i \mu_i \Big/ \sum_U q_i \rightarrow \mu .$$

Using a linearization argument, and assuming Poisson sampling of population units, these authors show that the optimal value of value of q_i is $q_i = \left\{ E\left(\pi_i^{-1} \middle| \mathbf{X}_U \right) \right\}^{-1}$, in which case

$$w_{is}^{IPWX} = \pi_i^{-1} \left\{ E\left(\pi_i^{-1} \middle| \mathbf{X}_U \right) \right\}^{-1} \Big/ \sum_U \pi_j^{-1} \left\{ E\left(\pi_j^{-1} \middle| \mathbf{X}_U \right) \right\}^{-1} I_j .$$

There is a subtle but important change in the inference framework used in the preceding development. In particular, the inclusion probabilities $\pi_i = E\left(I_i \middle| \mathbf{X}_U \right)$ are now being treated as unknown random variables rather than as known functions of the values in $\mathbf{X}_U$, putting their values on a par with the values y_i of the response variable. This is important in many practical applications where these probabilities are unknown functions of the values in $\mathbf{X}_U$ and, as is often the case with secondary analysis of survey data, where the values of π_i are only known for the observed sample. Note also that the modified IPW estimator $\bar{y}_{ws}^{IPWX} = \sum_s w_{is}^{IPWX} y_i$ is no longer design-consistent for the population mean $\bar{y}_U$ but instead converges to the q-weighted version of this mean.

3.2 Contributions to weighting under informative sampling

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 $\mathbf{X}_U$ is only partially available. However, it can also be the case that even if $\mathbf{X}_U$ is completely specified, sample inclusion can depend on $\mathbf{y}_U$ as well as $\mathbf{X}_U$. This type of sampling is typically referred to as *informative sampling*. One 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 $\mathbf{X}_U$, but believes that the agency that created the sample did so using information on another, unreleased, variable *Z*. Furthermore, given $\mathbf{X}_U$, the values of *Z* (and hence the realized values of the sample inclusion indicators *I*) and the response variable *Y* 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 informative 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 can now be written $\pi_i = E\left(I_i \middle| \mathbf{y}_U, \mathbf{X}_U\right) = E\left(I_i \middle| y_i, \mathbf{X}_U\right)$. Put

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

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

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

and we have $E\left(\tilde{\pi}_i^{-1} I_i y_i \middle| \mathbf{X}_U\right) = E\left(\tilde{\pi}_i^{-1} \tilde{\pi}_i y_i \middle| \mathbf{X}_U\right) = \mu_i$. 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

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

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 informative 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 $q_i = \left\{E\left(\tilde{\pi}_i^{-1}(y_i - \mu_i)^2 \middle| \mathbf{X}_U\right)\right\}^{-1}$ and modified smoothed IPW weights

$$w_{is}^{SIPWX} = \tilde{\pi}_i^{-1} \left\{E\left(\tilde{\pi}_i^{-1}(y_i - \mu_i)^2 \middle| \mathbf{X}_U\right)\right\}^{-1} \Big/ \sum_U \tilde{\pi}_j^{-1} \left\{E\left(\tilde{\pi}_j^{-1}(y_j - \mu_j)^2 \middle| \mathbf{X}_U\right)\right\}^{-1} I_j .$$

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 informative sampling requires estimation of $E\left(\pi_i^{-1}\middle|y_i, I_i = 1, \mathbf{X}_U\right)$ followed by estimation of $E\left(\tilde{\pi}_i^{-1}(y_i - \mu_i)^2\middle|\mathbf{X}_U\right)$. This can be done by using the sample data to fit an appropriate parametric model to these expectations. In particular, Kim and Skinner (2013) suggest that a model of the form $E\left(\pi_i^{-1}\middle|y_i, I_i = 1, \mathbf{X}_U\right) = 1 + \exp(-\boldsymbol{\phi}_1^T\mathbf{x}_i - \phi_2 y_i)$ will usually be adequate, with the values of $E\left(\tilde{\pi}_i^{-1}(y_i - \mu_i)^2\middle|\mathbf{X}_U\right)$ 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, it is important to note that this approach, as well as its simpler version when the CIA holds, depends crucially on the sample inclusion probabilities being known. In many practical applications of secondary analysis of sample data this is not the case, particularly when there is reason to believe that the sampling was informative. This is the problem that we now address.

4. But what if inclusion probabilities are unknown?

4.1 A brief overview of causal inference using secondary 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 in the experiment population has just two potential outcomes, defined as $Y(0)$ and $Y(1)$ when the unit is a control and when it is treated, respectively. That is, $Y(0)$ denotes the outcome that would be realized for the unit if it were not to be treated and $Y(1)$ denotes the outcome that would be realized for the same unit if it were to be treated. Let **X** denote relevant covariates for the experiment population. A key target of causal inference is the difference

$$D = E\left(Y(1)\middle|\mathbf{X}\right) - E\left(Y(0)\middle|\mathbf{X}\right)$$

between the population expectation of $Y(1)$ and the population expectation of $Y(0)$, which we refer to as the causal effect *D*. Note that *D* is the expected value of the difference between the population averages of *Y*(1) and *Y*(0). Clearly, both these outcomes cannot be observed for the same unit, and so the basic problem with estimating D is in imputing the missing potential outcomes for those units where only one outcome has been observed. That is very similar to the problem of item non-response in survey sampling, except here there are no units with full response.

For the causal effect D to be identifiable, the mechanism through which treatment is assigned to a particular 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:

- *Individualistic assignment*. That is, the probability of a unit being assigned to the treatment only depends on the covariates of that unit and not the values of the covariates for other units. Following Rubin (1980), this condition is sometimes referred to as the Stable Unit Treatment Value Assumption.
- *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). This probability is usually referred to as a propensity score, or just a propensity.
- *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).

Unfortunately, perfectly randomized experiments, although desirable, are not always feasible. Instead, we often have to make do with observational or secondary data when carrying out (or at least trying to carry out) causal inference. In this context, we usually relax the classical assumption that the probability of treatment assignment is known for all population units, and estimate it from the realized values of treatment assignments (which are all assumed known). As noted earlier these estimated probabilities are usually referred to as propensity scores, a convention that we now adopt. The three common strategies used to estimate D are then model-based imputation, where a regression model is used to impute the counterfactuals ($Y(0)$ for a treated unit and $Y(1)$ for a control, or untreated, unit), weighting estimators and matching estimators. Both weighting and matching require knowledge of propensity scores. In this paper we focus on the simplest type of weighting estimator of D, the Inverse Probability

Weighted or IPW estimator. We then combine this estimator with a model for the observed data to obtain an estimator of *D* that has the desirable property of being doubly robust, i.e., it is consistent for *D* provided either the propensity model is valid, or the observed data model is valid.

4.2 Two estimators for *D* that use propensity scores

Let y_i denote the value of a response *Y* for unit *i* in a sample of *n* units taken from a population *U* of *N* units, with y_i equal to y_{1i} if unit *i* is exposed to a treatment, and y_{0i} if not. Put $\boldsymbol{\mu}_i = E\left(y_i \middle| \mathbf{X}_s\right)$ with $\boldsymbol{\mu}_{1i} = E\left(y_{1i} \middle| \mathbf{X}_s\right)$ and $\boldsymbol{\mu}_{0i} = E\left(y_{0i} \middle| \mathbf{X}_s\right)$. Here $\mathbf{X}_s$ denote known sample covariate information, unrelated to treatment exposure, and put $\boldsymbol{\mu} = E\left(\boldsymbol{\mu}_i\right)$. Similarly put $\boldsymbol{\mu}_j = E\left(\boldsymbol{\mu}_{ji}\right)$ for *j* = 0, 1. Furthermore, let $I_i = 1$ denote membership of the treatment subsample (i.e., sample units exposed to the treatment) and $I_i = 0$ denoting membership of the control subsample (i.e., those sample units not exposed to the treatment), with $\boldsymbol{\pi}_i = \Pr\left(I_i = 1 \middle| \mathbf{X}_s\right)$. It immediately follows that $y_i = I_i y_{1i} + (1 - I_i) y_{0i}$. However, as we have already noted, the mechanism underpinning exposure is unknown. If $\boldsymbol{\pi}_i$ is known for all sample units the IPW estimator of $D = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0$, also referred to as the Average Treatment Effect, is

$$\tilde{D}^{IPW} = \sum\nolimits_{i=1}^{n} w_{is}^{\boldsymbol{\pi}} I_i y_i - \sum\nolimits_{i=1}^{n} w_{is}^{1-\boldsymbol{\pi}} (1 - I_i) y_i$$

where $w_{is}^{\boldsymbol{\pi}} = \boldsymbol{\pi}_i^{-1} \Big/ \sum\nolimits_{j=1}^{n} I_j \boldsymbol{\pi}_j^{-1}$ and $w_{is}^{1-\boldsymbol{\pi}} = (1 - \boldsymbol{\pi}_i)^{-1} \Big/ \sum\nolimits_{j=1}^{n} (1 - I_j)(1 - \boldsymbol{\pi}_j)^{-1}$. The more common situation, though, is where $\boldsymbol{\pi}_i$ is unknown but can be modeled as $\boldsymbol{\pi}_i = \boldsymbol{\pi}(\mathbf{x}_i; \Phi)$, where π is a known function. Then $\boldsymbol{\pi}_i$ can be estimated from the sample values of I_i and $\mathbf{x}_i$, leading to the estimator $\hat{\boldsymbol{\pi}}_i = \boldsymbol{\pi}(\mathbf{x}_i; \hat{\Phi})$, where $\hat{\Phi}$ is a vector of estimated parameter values. This leads to the plug-in IPW estimator

$$\hat{D}^{IPW} = \sum\nolimits_{i=1}^{n} w_{is}^{\hat{\boldsymbol{\pi}}} I_i y_i - \sum\nolimits_{i=1}^{n} w_{is}^{1-\hat{\boldsymbol{\pi}}} (1 - I_i) y_i .$$

The estimated probabilities $\hat{\boldsymbol{\pi}}_i$ are the propensity scores, and it is easy to see that if the model for $\boldsymbol{\pi}_i$ is valid and the three basic assumptions listed in Section 4.1 hold when we condition on $\mathbf{X}_s$ then $\hat{D}^{IPW}$ is consistent for *D*.

The estimator $\hat{D}^{IPW}$ does not explicitly control for different covariate distributions between the treatment and control subsamples, assuming instead that these differences cancel out "on average". In many situations, however, 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 $\boldsymbol{\lambda}_i$ then defining the treatment effect for sample unit *i*. Substituting in $\hat{D}^{IPW}$, we see that

$$\hat{D}^{IPW} = \tilde{\boldsymbol{\lambda}} + \tilde{D}_0^{IPW}$$

where $\tilde{\boldsymbol{\lambda}} = \sum_{i=1}^{n} w_{is}^{\hat{\pi}} I_i \boldsymbol{\lambda}_i$ and $\tilde{D}_0^{IPW}$ is the value of $\hat{D}^{IPW}$ when y_i is replaced by y_{0i}. However, there are no treatment effects distinguishing the "treated" units from the "control" units in $\tilde{D}_0^{IPW}$, so this term is purely an estimate of the differential impact of the population covariates on the realized value of $\hat{D}^{IPW}$, something that is asymptotically zero but can be non-zero in any finite sample. It follows that $\tilde{\boldsymbol{\lambda}} = \hat{D}^{IPW} - \tilde{D}_0^{IPW}$ is then a covariate-adjusted estimator of *D*. Note that calculation of $\tilde{\boldsymbol{\lambda}}$ requires estimation of $\boldsymbol{\lambda}_i$, say by $\hat{\boldsymbol{\lambda}}_i$. Since the predicted value of y_{0i} for a treated unit is then $\hat{y}_{0i} = y_i - \hat{\boldsymbol{\lambda}}_i$, it follows that we can estimate $\tilde{D}_0^{IPW}$ by replacing the unobservable values y_{0i} for treated units by $\hat{y}_{0i}$. We denote this estimate by $\hat{D}_0^{IPW}$. Our estimator of $\tilde{\boldsymbol{\lambda}}$ is then

$$\hat{\boldsymbol{\lambda}} = \hat{D}^{IPW} - \hat{D}_0^{IPW}. \tag{1}$$

A simple way of calculating $\hat{D}_0^{IPW}$ is to fit the model $y_i = I_i \boldsymbol{\lambda}_i + m(\mathbf{x}_i; \boldsymbol{\beta}) + e_i$ to the entire sample. Here *m* is a specified function of $\mathbf{x}_i$ (e.g., a linear function), $\boldsymbol{\beta}$ is a vector of fixed effect parameters, and e_i is a suitably specified random effect. We illustrate this approach in the application discussed in the next section.

4.3 $\hat{\boldsymbol{\lambda}}$ is doubly robust

In the previous sub-section, we introduced $\tilde{\boldsymbol{\lambda}}$, a covariate-adjusted estimator of the average treatment effect *D* defined by the difference between the IPW estimator of *D* and the same estimator but with all observed values y_i replaced by control values y_{0i}. Substituting model-based estimates for the treated units values of y_{0i} (which are unobserved) then leads to the estimator $\hat{\boldsymbol{\lambda}}$ defined in (1) above. Note that the first term in $\hat{\boldsymbol{\lambda}}$ is an estimate of *D* while the second term corrects for treatment / control imbalance in the baseline covariates and the error terms of the assumed model. In a randomized trial these two sources of imbalance tend asymptotically to zero. However, they are not necessarily equal to zero in a finite sample situation.

In a causal inference context a doubly robust (DR) estimator is the one that remains consistent if either the model for treatment assignment or the model for the counterfactual outcome (but not necessary both) is correctly specified. Double robustness is often viewed as a desirable property for an estimator since there is usually a good chance that either of these models is incorrectly specified. Rotnitzky, Robins and Scharfstein (1998) proposed the Augmented Inverse Propensity Weighted estimator as a DR estimator in missing data situations and Scharfstein, Rotnitzky and Robins (1999) showed how to construct a DR estimator for causal inference under unconfoundedness. For a more recent discussion of DR see Bang and Robins (2005). Below we show that $\hat{\boldsymbol{\lambda}}$ is a doubly robust estimator of D.

We shall assume 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 that $\hat{y}_{0i} = y_i - I_i\hat{\boldsymbol{\lambda}}_i = m(\mathbf{x}_i;\hat{\boldsymbol{\beta}}) + \hat{e}_i$. Substituting in (1) leads to

$$\begin{aligned}
\hat{\boldsymbol{\lambda}} &= \sum\nolimits_{i=1}^{n} w_{is}^{\hat{\pi}} I_i y_i - \sum\nolimits_{i=1}^{n} w_{is}^{1-\hat{\pi}}(1-I_i)y_i - \sum\nolimits_{i=1}^{n} w_{is}^{\hat{\pi}} I_i \hat{y}_{0i} + \sum\nolimits_{i=1}^{n} w_{is}^{1-\hat{\pi}}(1-I_i)\hat{y}_{0i} \\
&= \sum\nolimits_{i=1}^{n}\left(w_{is}^{\hat{\pi}} I_i - w_{is}^{1-\hat{\pi}}(1-I_i)\right)\left(y_i - \hat{y}_{0i}\right) \\
&= \sum\nolimits_{i=1}^{n}\left(w_{is}^{\hat{\pi}} I_i - w_{is}^{1-\hat{\pi}}(1-I_i)\right)\left(I_i\boldsymbol{\lambda}_i + m(\mathbf{x}_i;\boldsymbol{\beta}) + e_i - \hat{y}_{0i}\right) \\
&= \sum\nolimits_{i=1}^{n} w_{is}^{\hat{\pi}} I_i\boldsymbol{\lambda}_i + \sum\nolimits_{i=1}^{n}\left(w_{is}^{\hat{\pi}} I_i - w_{is}^{1-\hat{\pi}}(1-I_i)\right)\left(m(\mathbf{x}_i;\boldsymbol{\beta}) + e_i - \hat{y}_{0i}\right) \\
&= \sum\nolimits_{i=1}^{n} w_{is}^{\hat{\pi}} I_i\boldsymbol{\lambda}_i + \sum\nolimits_{i=1}^{n}\left(w_{is}^{\hat{\pi}} I_i - w_{is}^{1-\hat{\pi}}(1-I_i)\right)\left(y_{0i} - \hat{y}_{0i}\right).
\end{aligned}$$

That is,

$$\hat{\boldsymbol{\lambda}} = \tilde{\boldsymbol{\lambda}} + \sum\nolimits_{i=1}^{n}\left[w_{is}^{\hat{\pi}} I_i - w_{is}^{1-\hat{\pi}}(1-I_i)\right]R_i \tag{2}$$

where $R_i = y_{0i} - \hat{y}_{0i}$ is the population residual for y_{0i}. Unconfoundedness implies that these residuals will have expectation zero if the model for the control or untreated outcome is valid, in which case the summation on the right hand side of (2) above will have zero expectation irrespective of whether the model for treatment assignment is valid or not. Alternatively, if the model for the treatment assignment is valid then unconfoundedness again implies that the distribution of the control model residuals will be the same in both the treated and untreated parts of the population of interest, in which case the summation on the right hand side of (2) corresponds to the difference of two unbiased estimators of the same expected value, and so has expectation zero irrespective of whether the control model is correctly specified or not. That is, $\hat{\lambda}$ has the same expectation as $\tilde{\lambda}$ in both these situations and so is a doubly robust estimator of *D*.

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 because of operation of 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 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 $\pi(\mathbf{x}) = \Pr\left(I = 1 \middle| \mathbf{X} = \mathbf{x}\right) = E(I|\mathbf{X} = \mathbf{x})$, where $\mathbf{X}$ denotes a vector of covariate measurements such that it is reasonable to assume that Y and I are conditionally independent given $\mathbf{X}$. 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, mainly between 2015 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 allocatable 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. However, it can also be argued that a link could exist between prevailing meteorological conditions (and hence rainfall over the mountains) and operation of the Muscat radiosonde. 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.

5.2 Estimation of average target effects via propensity weighting

Let i index gauge and j index day. We define the average target effect D 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 D 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 (temp.dry.1);
- First principal component of average relative humidity (relh.1);
- First principal component of average ground level air pressure (pres.1).

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**	**ChiSquare**	**Prob>ChiSq**
Intercept	-0.753	0.225	11.166	0.001
total.totals	0.016	0.005	10.499	0.001
temp.dry.1	-0.172	0.040	18.332	0.000
relh.1	-0.110	0.025	19.839	0.000
pres.1	-0.115	0.024	22.718	0.000

Propensity score weighted average actual rainfall based on the 2176 target gauge-day values was 4.853mm, with a corresponding weighted average value for LogRain of 0.554. In comparison, weighted average actual rainfall based the 1992 control gauge-day values was 4.640mm, 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 an estimate $\hat{D}^{IPW}$ = 0.293mm (with outlier removed) for Y = actual rainfall and $\hat{D}^{IPW}$ = 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 1. 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 2. 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 X1, which uses the data from all 488 days.

Figure 1: Scatterplot showing daily numbers of target and control gauge-day observations with actual rainfall, Hajar Mountains trial 2013-2018.

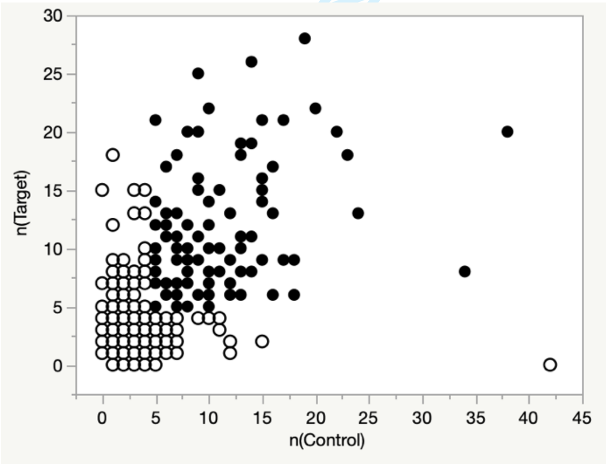
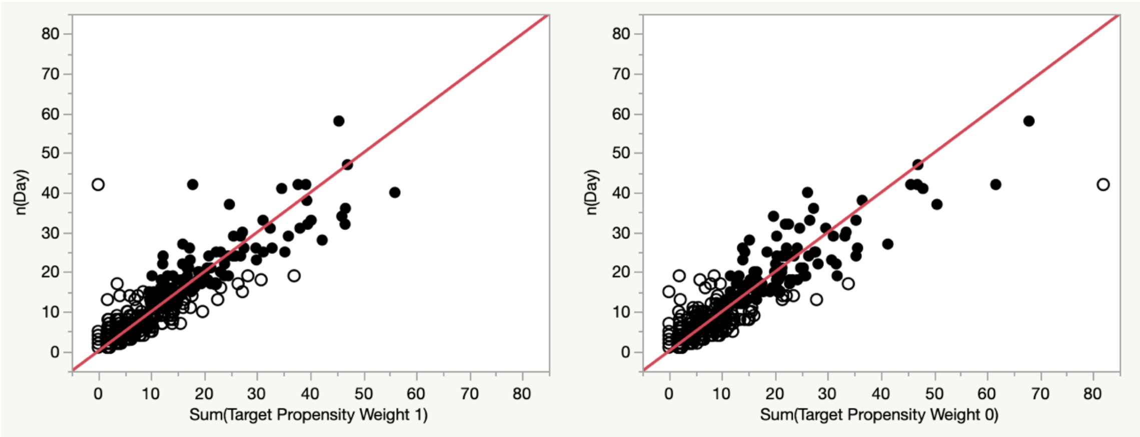


Figure 2: Daily numbers of downwind gauge-days with actual rainfall (y axis) vs. daily sums of propensity scores for target gauge-days (x-axis, left) and control gauge-days (x-axis, right). Plots restricted to days when both target and control rainfall observed. Line is the identity fit.



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 daily $\hat{D}^{IPW}$ values reduce to average target rainfall minus average control rainfall on the day. Averages of these values 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 the lower variability in daily values of $\hat{D}^{IPW}$ on "Good Data" days, when rainfall is more widespread. This implies correlation between daily $\hat{D}^{IPW}$ values 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 counterfactuals defined by 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 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, and 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, with 2015 as 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. This 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 set out in Table 3, with the distribution of predicted Day effects generated by this model shown in Figure 3.

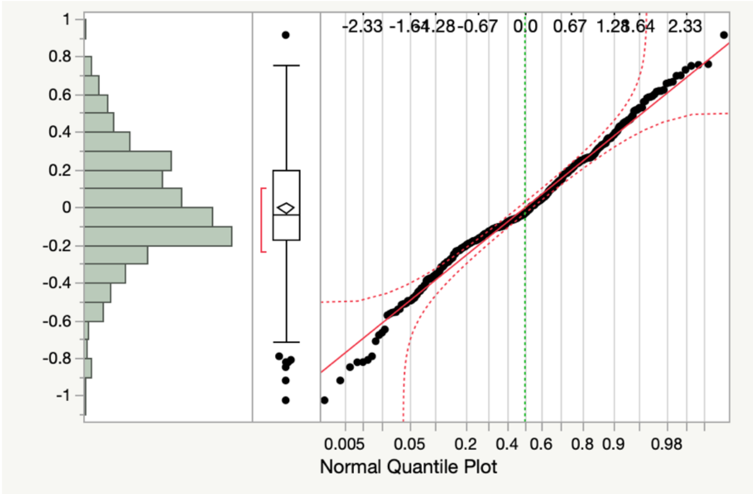
Table 2: Fitted parameter values with estimated standard errors and associate p-values for the downwind LogRain model. Significant p-values are displayed with an asterisk.

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.077	0.121	0.633	0.5269
y2013	0.406	0.113	3.581	0.0004*
y2014	0.336	0.106	3.153	0.0018*
y2016	0.259	0.115	2.241	0.0257*
y2017	0.092	0.119	0.774	0.4397
y2018	0.041	0.146	0.277	0.7821
Gauge Elevation 1	-0.200	0.164	-1.217	0.2237
Gauge Elevation 2	-0.096	0.071	-1.357	0.1748
Upwind LogRain	0.945	0.059	15.983	<.0001*
Target H01	0.481	0.247	1.946	0.0518
Target H02	0.840	0.293	2.867	0.0042*
Target H03	0.241	0.092	2.613	0.0090*
Target H04	-0.114	0.089	-1.283	0.1996
Target H05	0.499	0.132	3.788	0.0002*
Target H06	-0.136	0.149	-0.916	0.3598
Target H07	0.336	0.188	1.785	0.0743
Target H08	0.131	0.129	1.023	0.3063
Target H09	0.711	0.307	2.319	0.0204*
Target H10	0.196	0.170	1.149	0.2504
Gauge Elevation 1*Target H01	-0.488	0.363	-1.342	0.1797
Gauge Elevation 1*Target H02	-1.272	0.469	-2.714	0.0067*
Gauge Elevation 2*Target H01	-0.163	0.159	-1.026	0.3051
Gauge Elevation 2*Target H02	-0.458	0.172	-2.667	0.0077*

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

Source	Var Comp	Std Error	95% Lower	95% Upper	Pct of Total
Day	0.225	0.032	0.162	0.289	10.836
Residual	1.853	0.043	1.772	1.940	89.164
Total	2.078	0.050	1.984	2.179	100.000

Figure 3: Distribution of predicted day effects generated by the downwind model for gauge-day values of LogRain. The usual assumption of Gaussian random effects seems reasonable.



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 D between the target average value of LogRain and the control average value of this variable. In particular 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 these propensity scores, using the doubly robust estimator (2), allows us to at least achieve some measure of protection against this scenario.

We start by writing the model for LogRain set out in Tables 2 and 3 as

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

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 H0b; 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. It immediately follows that for target gauge-day observations we can write $\hat{y}_{0ij} = \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}} + \hat{u}_i + \hat{e}_{ij}$, where $\hat{e}_{ij} = y_{ij} - \mathbf{z}_{ij}^T \hat{\boldsymbol{\theta}} - \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}} - \hat{u}_i$. That is, $\hat{y}_{0ij} = y_{ij} - \mathbf{z}_{ij}^T \hat{\boldsymbol{\theta}}$ and so

$$\begin{aligned}\hat{D}_0^{IPW} &= \sum_{ij} w_{ijs}^{\hat{\pi}} I_{ij} \hat{y}_{0ij} - \sum_{ij} w_{ijs}^{1-\hat{\pi}} (1 - I_{ij}) y_{0ij} \\ &= \sum_{ij} w_{ijs}^{\hat{\pi}} I_{ij} \left(y_{ij} - \mathbf{z}_{ij}^T \hat{\boldsymbol{\theta}} \right) - \sum_{ij} w_{ijs}^{1-\hat{\pi}} (1 - I_{ij}) y_{ij} \\ &= \hat{D}^{IPW} - \sum_{ij} w_{ijs}^{\hat{\pi}} \mathbf{z}_{ij}^T \hat{\boldsymbol{\theta}}.\end{aligned}$$

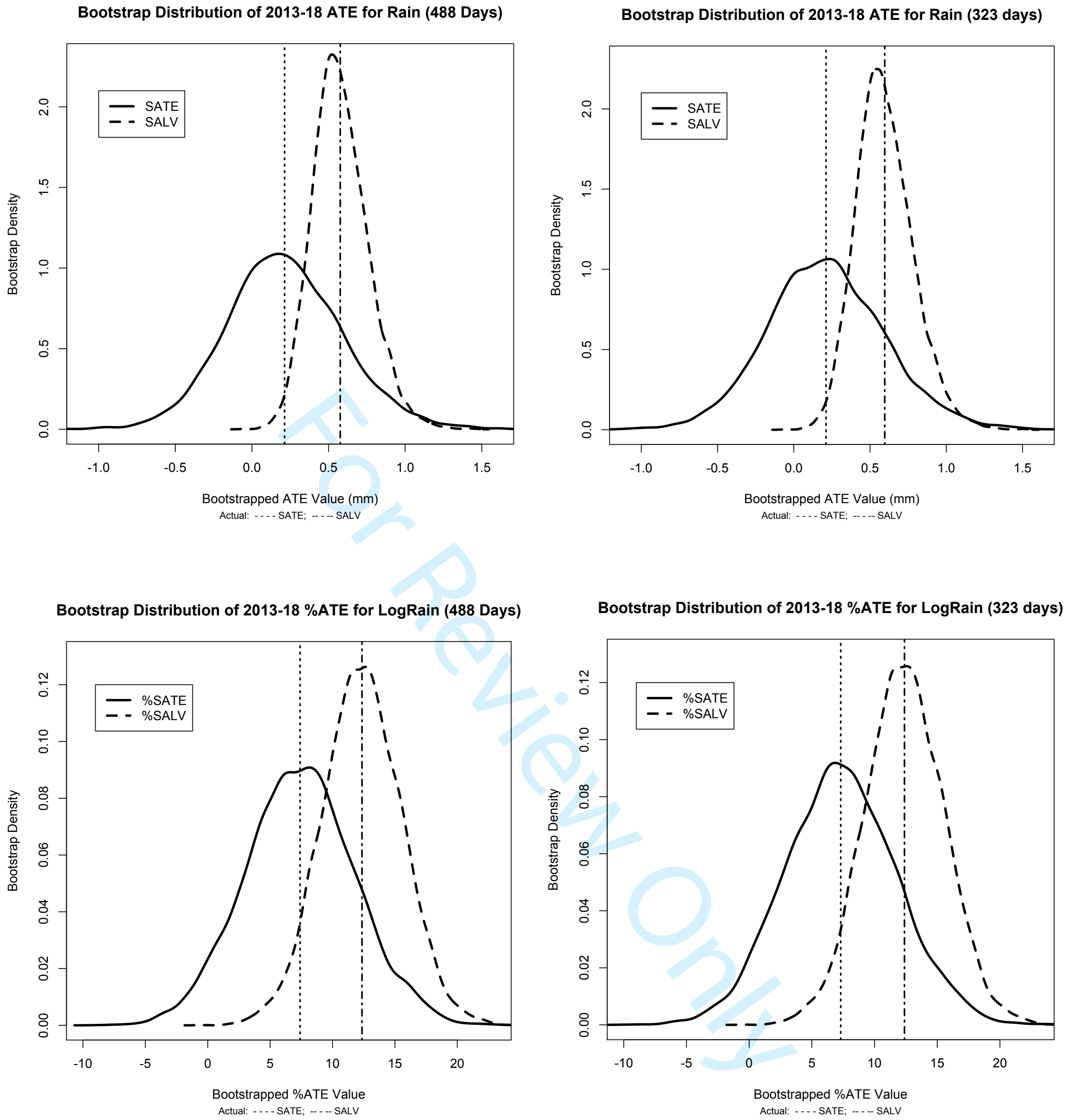
Our estimate of D is therefore $\hat{\boldsymbol{\lambda}} = \sum_{ij} w_{ijs}^{\hat{\pi}} \mathbf{z}_{ij}^T \hat{\boldsymbol{\theta}}$, the propensity score weighted average of the sum of the model-based estimates of the "target effects" defined in Table 2. Let r_{ij} denote actual rainfall (Rain) on a gauge-day. We can then extend the definitions of $\hat{D}_0^{IPW}$ and $\hat{\boldsymbol{\lambda}}$ to gauge-day values of Rain by writing $\hat{r}_{0ij} = \exp\left(y_{ij} - \mathbf{z}_{ij}^T \hat{\theta} \right)$ for this variable.

For notational simplicity we use SATE (Sample Average Treatment Effect), SANE (Sample Average Null Effect) and SALV (Sample Average Lambda Value) to denote $\hat{D}^{IPW}$, $\hat{D}_0^{IPW}$ and $\hat{\boldsymbol{\lambda}}$ respectively in what follows. Following Chambers *et. al*. (2021), we use a two level semi-parametric block bootstrap to calculate standard errors and associated p-values for these estimates, using a total of 10,000 bootstrap samples. As noted in the previous sub-section, there was downwind rainfall data recorded on 488 days of the trial. However, both control and target rainfall data were obtained on only 323 of these days. Results based on this more limited sample are expected to better control for daily meteorological variation and so are presented separately in Table 4, with Figure 4 showing the associated bootstrap distributions associated with the SATE and SALV.

Table 4: Estimated values of SATE, SANE and SALV for Rain and for LogRain for all 488 days when downwind rain was recorded in the Hajar Mountains trial, as well as for the 323 days when both target and control rainfall was recorded. Block bootstrap standard errors and one-sided bootstrap p-values (bootstrap probability of no effect or negative effect) are shown in parentheses below estimated SATE and SALV values.

Sample Days	**SATE(Rain)**	**SANE(Rain)**	**SALV(Rain)**
488	0.2130	-0.3633	0.5763
	(0.3789, 0.2857)	(0.3095)	(0.1806, 0.0002)
323	0.2114	-0.3850	0.5964
	(0.3890, 0.2937)	(0.3186)	(0.1866, 0.0002)
Sample Days	**SATE(LogRain)**	**SANE(LogRain)**	**SALV(LogRain)**
488	0.0740	-0.0496	0.1236
	(0.0436, 0.0449)	(0.0301)	(0.0316, 0.0001)
323	0.0730	-0.0508	0.1238
	(0.0442, 0.0459)	(0.0310)	(0.0316, 0.0001)

Figure 4: Two level semi-parametric block bootstrap distributions for SATE and SALV for Rain (top row) and for LogRain (bottom row).

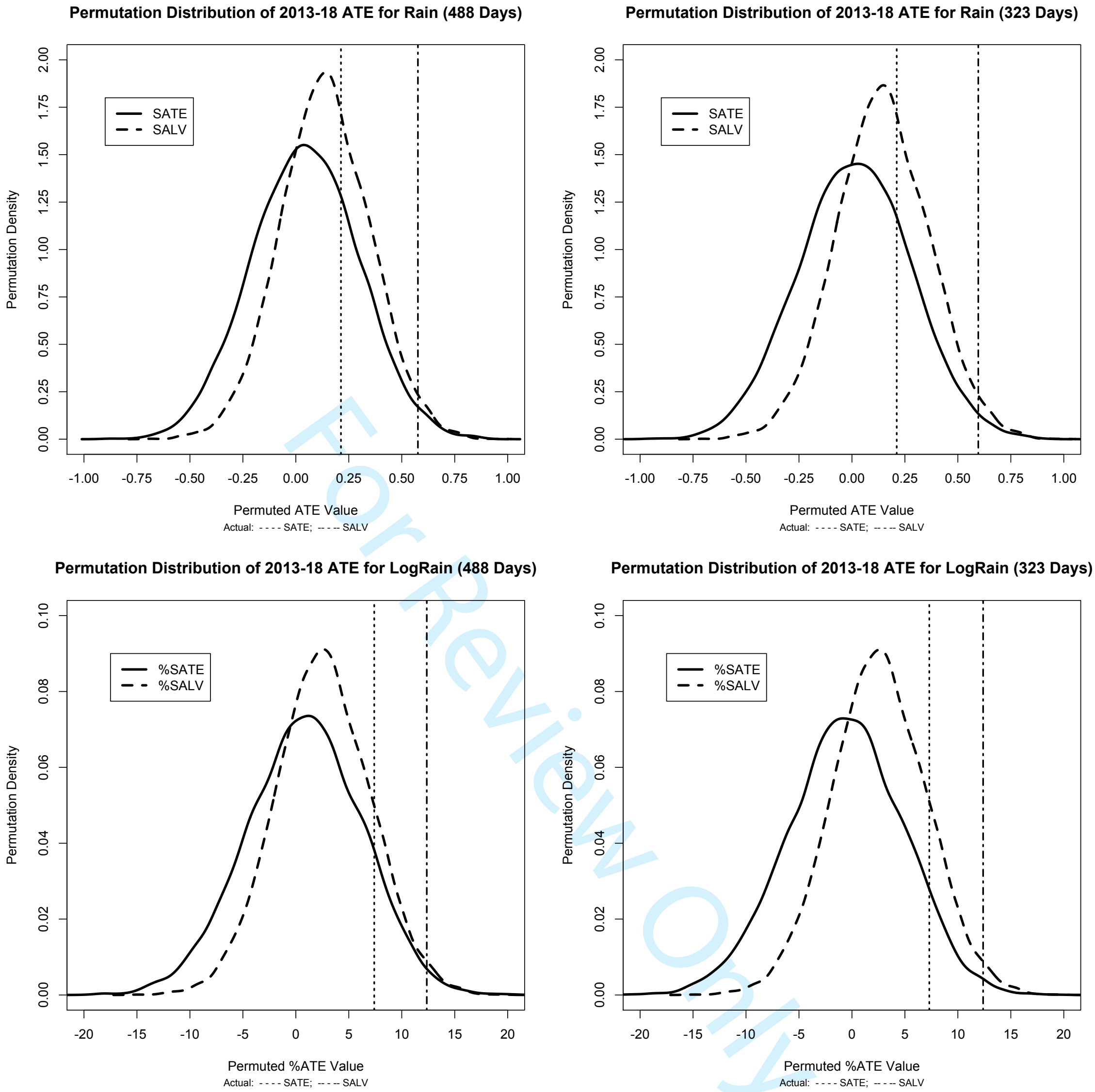


A randomization analysis of the significance of the effect of ionizer operation on positive rainfall as measured by the different values of SATE and SALV was also carried out. This was done by independently randomly permuting the operating states of each ionizer each day (while maintaining the requirement that there were an equal number of operating and non-operating ionizers each day). This was also done 10,000 times. The permutation p-value was then calculated as the proportion of permuted SATE (SALV) values that were greater than the observed SATE (SALV) value. These p-values are set out in Table 5 below, with the randomization distributions for the permuted SATE (SALV) values shown in Figure 5.

Table 5: Permutation p-values of SATE, SANE and SALV for Rain and for LogRain for all 488 days when downwind rain was recorded in the Hajar Mountains trial, as well as for the 323 days when both target and control rainfall was recorded. These p-values were computed as the proportion of permuted values of SATE and SANE greater than the observed values of these estimates. Ionizer operating states over 2013-2018 were randomly permuted a total of 10,000 times in order to generate these permuted values.

Sample Days	p-value SATE (Rain)	p-value SALV (Rain)
488	0.2539	0.0209
323	0.2277	0.0213
Sample Days	**p-value SATE (LogRain)**	**p-value SALV (LogRain)**
488	0.1105	0.0152
323	0.0708	0.0153

Figure 5: Randomization distributions for SATE and SALV for Rain (top row) and for LogRain (bottom row) generated by randomly permuting ionizer operating states over 2013-2018. Actual observed values for SATE and SANE are shown as vertical lines in the plots.



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 ground-breaking 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 not misspecified. 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 D between the expected values of a "treated" response and an "untreated" response. Here we presented the standard inverse probability weighted (IPW) estimator $\hat{D}^{IPW}$ for D, as well as a new estimator (2) that, as in the model-assisted approach, modifies $\hat{D}^{IPW}$ for differences in covariate distributions between treated and untreated sample units. We also showed that (2) is double robust.

Finally, in Section 5 we provide a real life application of causal inference using (2) 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 (Chambers *et al*., 2021) indicated that these devices led to an increase of around 15-18 per cent in rainfall over the trial. The double robust estimator (2) was then 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 0.5763mm per gauge per day of rainfall for target gauge-days (see the SALV(Rain) entry in Table 4 for all 488 days when downwind rainfall was recorded). This can be compared with the propensity score weighted average for control gauge-days, which was 4.560mm after removal of an extreme outlier. That is, the estimated causal effect of operating the ionizers was to increase control, or "natural", rainfall by 0.5763/4.560 or around 12.6 per cent. This is somewhat lower but still consistent with the estimates 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 highly 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.

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.

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.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.

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.

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.

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.

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.

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.

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.