Direct and Stable Weight Adjustment in Non-Experimental Studies with Multi-valued Treatments: Analysis of the Impact of an Earthquake on Posttraumatic Stress*

María de los Angeles Resa† José R. Zubizarreta‡

Abstract

In February of 2010, a massive earthquake struck Chile, causing devastation in certain parts of the country, affecting other areas, and leaving territories untouched. Two months after the earthquake, Chile's Ministry of Social Development re-interviewed a representative subsample to its National Socioeconomic Characterization Survey (CASEN), completed two months before the earthquake, thereby creating a prospective longitudinal survey with detailed information of the same individuals before and after the earthquake. In this paper, we use a new weighting method for non-experimental studies with multi-valued treatments to estimate the impact of levels of exposure to the earthquake on posttraumatic stress. Unlike common weighting approaches for multi-valued treatments, this new method does not require explicit modeling of the generalized propensity score and instead focuses on directly balancing the covariates across the multi-valued treatments with weights that have minimum variance. As a result, the weighting estimator is stable and approximately unbiased. Furthermore, the weights are constrained to avoid model extrapolation. We illustrate this new method in a simulation study, both with categorical and continuous treatments. The results show that directly targeting balance instead of explicitly modeling the treatment assignment probabilities, tends to provide the best results in terms of bias and root mean square error. Using this method, we estimate the impact of the intensity of the earthquake on posttraumatic stress. We implement this method in the new package `msbw` for `R`.

KEYWORDS: Causal inference; Inverse probability weights; Propensity score; Observational studies.

*We thank Zach Branson, Ambarish Chattopadhyay, Yige Li, Sherri Rose, and Paul Rosenbaum for helpful conversations and input. This work was supported by grants 1DP2MD012722 from NIH and G-2018-10118 from the Alfred P. Sloan Foundation.

†Department of Statistics, Columbia University, 1255 Amsterdam Avenue, 901 SSW, New York, NY 10027, Email: `maria@stat.columbia.edu`.

‡Department of Health Care Policy and Department of Statistics, Harvard University, 180 Longwood Ave, Office 307-D, Boston, MA 02115, Email: `zubizarreta@hcp.med.harvard.edu`.

1 Introduction

1.1 Impact of the 2010 Chilean earthquake on mental health

The morning of February 27, 2010, an earthquake of magnitude 8.8 struck in the Pacific, 65 miles west of Chile's second largest city, Concepción — the 6$^{\text{th}}$ most severe earthquake since 1900 (USGS 2014). The quake and tsunami caused more than 30 billion dollars in damages and over a million people lost their homes (USGS 2011a). The force of the earthquake moved the city of Concepción 3.04 meters to the west (Pollitz et al. 2011). More than 500 people were crushed, drowned, or perished in fires (USGS 2011a).

Earthquake survivors experience mental conditions similar to those seen after other traumatic experiences such as rape, car accident, crime, combat, and captivity in concentration camps (Wang et al. 2000). In particular, reports of increased posttraumatic stress (PTS) symptoms or disorder (PTSD) have been consistently reported following earthquakes (Sharan et al. 1996; Goenjian et al. 2000; Roussos et al. 2005; Yehuda et al. 2005; Neria et al. 2008). However, a limitation of many studies of the psychological effects of earthquakes and similar disasters is that, because the disaster was not anticipated by the investigators, the studies rely on retrospective recall both of intensity of exposure and of conditions prior to exposure. It is conceivable that a person in distress following an earthquake may recall exposure to the earthquake as more severe than does another person not in distress. Similarly, a person in distress may recall life and conditions before the earthquake in a different way than does a person not in distress. If distress distorts memory, it may also distort associations between current status and recalled exposure adjusting for recalled pre-exposure status. In contrast, this paper uses prospectively recorded conditions prior to the earthquake, and objective geologic measures of ground shaking define exposure.

Natural disasters strike without purpose or deliberate target, but are less equitable than randomized experiments. Lower-income residential areas may be more often found along fault

lines, low-income homes in earthquake-prone areas may be constructed of poorer materials that are less resistant to earthquake damage, and poor health or limited financial resources may limit ability to respond and recover after the earthquake has struck. As seen below, Chileans exposed to intense earth shaking had lower incomes, lower rates of employment and fewer years of education prior to the earthquake than did the control group of Chileans largely spared exposure the earthquake. Using new weighting methods for multi-valued treatments, we compared individuals that experienced different degrees of shaking but that were highly comparable prior to the earthquake in terms of these measured variables.

Chile's earthquake provides a unique perspective because two months before the earthquake, Chile's Ministry of Social Development completed its National Socioeconomic Characterization Survey (CASEN), and two months after the earthquake, they re-interviewed a representative subsample of it (MDS 2011a). It is unusual to have detailed data on the same individuals for a large sample both before and after a major disaster, free of recall bias. To our knowledge, these data has not been used to study impact of levels of exposure to the earthquake.

1.2 A new weighting method for multi-valued treatments

Our analysis illustrates a new weighting method for non-experimental or observational studies with multi-valued treatments. In observational studies with binary treatments, investigators often rely on the propensity score to balance covariates and remove biases due to observed covariates either by matching, weighting, or stratifying. While the propensity score has the attractive property of being the coarsest balancing score (Rosenbaum and Rubin 1983), in practice it will not produce good balance if the propensity score model is misspecified, or if the covariates are sparse relative to the sample size (Zubizarreta et al. 2011). To address these limitations, other matching and weighting methods that directly balance the observed covariates have been proposed (e.g., Hainmueller 2012; Zubizarreta 2012; Diamond

and Sekhon 2013; Imai and Ratkovic 2014; Zubizarreta 2015; Pimentel et al. 2015); however, they are designed for binary treatments.

Generalizing the propensity score to multi-valued treatments is not straighforward and there is no standard way of using it when matching or stratifying. Lopez et al. (2017) provide a review of estimation methods for causal effects with multi-valued treatments and propose the method of vector matching. This method matches subjects with similar values on the complete vector of generalized propensity scores. With multi-valued treatments, an alternative approach to matching is weighting by the inverse predicted probability of receiving treatment (Imbens 2000). However, extreme weights that result from estimated treatment assignment probabilities that are very close to zero or one can generate unstable estimates with large variance (Kang and Schafer 2007). Furthermore, if the probabilities are not correctly modeled, which is more likely when there are more than two treatments, or if the weights are truncated to reduce variability of the estimates, the balancing properties of these weights will not necessarily hold.

To address these problems, different weighting methods have been proposed. Imai and Ratkovic (2014) and Fong et al. (2018) developed the covariate balancing propensity score (CBPS) in which balance conditions are incorporated in the estimation process within a generalized method of moments or empirical likelihood framework. Another approach is the use of generalized boosted models (GBM) to estimate the multi-valued treatment assignment probabilities (McCaffrey et al. 2004, 2013). GBM is a flexible nonparametric model that iteratively fits multiple regression trees and has the advantage that it can incorporate covariate balance into that process by means of a tuning parameter. These methods, however, do not explicitly address the variability and extremeness of the weights in the estimation process.

In this paper, we propose a robust approach to weighting in observational studies with multi-valued treatments. While this approach can be understood within the framework of marginal

structural models (Robins et al. 2000), we note that alternative approaches to estimation with multi-valued treatments are the G-methods by Robins (1986). In this paper, we generalize the stable balancing weights (SBW) of Zubizarreta (2015). The proposed weights address the problems of balance (bias) and stability (variance) in weight estimation without explicitly modeling the treatment assignment probabilities and instead directly balancing the observed covariates with weights of minimum variance. Therefore, the weights do not need to be truncated. Also, since the weights are required to be non-negative, the resulting covariate adjustments are an interpolation grounded to the available study sample as opposed to an extrapolation based on a model that can be misspecified (Mattei and Mealli 2015). See Li and Li (2019) for a related approach.

The remainder of the paper is organized as follows. In Section 2, we establish the setup, notation, and assumptions used throughout the paper. In Section 3, we provide a justification for the proposed approach and describe it subsequently. In Section 4, we present a simulation study that compares the proposed method to other currently available ones. In Section 5, we use the proposed weights to estimate the effect that the 2010 earthquake in Chile on PTSD for varying levels of ground shaking. Finally, in Section 6 we provide concluding remarks and a summary of the work.

2 Setup, notation, and assumptions

For a sample of n independent observations indexed by $i = 1, ..., n$, we let Y_i be the observed outcome, $\boldsymbol{X}_i$ be the observed (pre-treatment) covariates, and $Z_i \in \mathcal{Z}$ be the treatment assignment indicator, where $\mathcal{Z}$ is the set of all possible treatment values. Under the potential outcomes framework (Neyman 1923, 1990; Rubin 1974), we define the set of potential outcomes for each subject as $\mathcal{Y}_i = \{Y_i(z), z \in \mathcal{Z}\}$, where $Y_i(z)$ is the potential outcome that subject i would exhibit if he or she were to be assigned to treatment z (Holland 1986). We only observe

the potential outcome under the treatment z that subject i receives, $Y_i = \sum_{z \in \mathcal{Z}} \mathbb{1}_{\{Z_i = z\}} Y_i(z)$, and the other potential outcomes remain unobserved counterfactuals. For categorical and ordinal treatments, $|\mathcal{Z}| = K$, where $|\mathcal{Z}|$ is the cardinality of the set $\mathcal{Z}$, and so K is the total number of treatment levels. This notation implicitly assumes the stable unit treatment value assumption (SUTVA; Rubin 1980, 1986) holds.

In order to identify causal effects from observed data, we assume positivity and weak ignorability (Imbens 2000); that is,

$$\Pr(Z_i = z \mid \boldsymbol{X}_i = \boldsymbol{x}) > 0 \ \forall \boldsymbol{x} \in \text{supp}(\boldsymbol{X}_i) \ \text{ and } \ Y_i(z) \perp\!\!\!\perp \mathbb{1}_{\{Z_i = z\}} \mid \boldsymbol{X}_i \ \forall z \in \mathcal{Z} \tag{1}$$

respectively. The positivity assumption requires that every subject in the target population has a positive probability of receiving any given treatment. Without this assumption, there could be subjects for which there is no potential outcome information under some treatments, and we would not be able to estimate treatment effects without extrapolation (Mattei and Mealli 2015). Additionally, in practice, subjects with a very low probability of receiving the treatment they received would be very influential in the analysis and the variance of the estimates.

The ignorability assumption states that the treatment assignment is random given the observed covariates. In other words, there are no unmeasured confounders. Unfortunately, this assumption is not directly testable, so subject-matter knowledge is often needed to ensure that $\boldsymbol{X}_i$ satisfies this assumption (see chapters 12 and 21 of Imbens and Rubin 2015 for a discussion of the ignorability assumption).

3 Stable balancing weights for multi-valued treatments

We wish to estimate $\mathrm{E}[Y_i(z)]$, the mean of the potential outcomes $Y_i(z)$ for each possible multi-valued treatment z. For this we write the marginal structural model $\mathrm{E}[Y_i(z)] = g(z; \boldsymbol{\beta})$

(Robins et al. 2000), where g represents the potential outcome mean model

$$g(z;\boldsymbol{\beta}) = \sum_{k=1}^{K} \beta_k \mathbb{1}_{\{z=k\}}. \tag{2}$$

By design, in randomized experiments we have that $Y_i(z) \perp\!\!\!\perp Z_i$ for all z, and therefore, that $\mathrm{E}[Y_i(z)] = \mathrm{E}[Y_i(z) \mid Z_i = z] = \mathrm{E}[Y_i \mid Z_i = z]$. That is to say we can unbiasedly estimate the mean of each potential outcome by the sample mean of each group. In other words, we can estimate the parameters of the causal model (2) by estimating the parameters of the associational model

$$\mathrm{E}[Y_i \mid Z_i = z] = \sum_{k=1}^{K} \gamma_k \mathbb{1}_{\{z=k\}},$$

for example, via ordinary least squares.

In non-randomized or observational studies, when we have enough variables such that the assumption of weak ignorability (1) is plausible, we can still estimate the causal parameters of (2) if we use weights for which the treatment assignment is independent of the observed covariates. In practice, this can be accomplished by balancing the empirical distributions of the observed covariates across all treatments.

As mentioned before, standard weighting approaches can result in unstable effect estimates. Thus, in addition to balancing covariates in order to unbiasedly estimate the causal parameters, it is also desirable that those weights provide the least variable estimator given a level of covariate balance.

The weighted least squares estimator for each parameter in (2) is given by

$$\hat{\beta}_k = \frac{\sum_{i=1}^{n} w_i Y_i \mathbb{1}_{\{Z_i=k\}}}{\sum_{i=1}^{n} w_i \mathbb{1}_{\{Z_i=k\}}}, \qquad k = 1, ..., K,$$

and its corresponding variance is

$$\mathrm{Var}(\hat{\beta}_k) = \sigma^2 \frac{\sum_{i=1}^{n} w_i^2 \mathbb{1}_{\{Z_i=k\}}}{\left(\sum_{i=1}^{n} w_i \mathbb{1}_{\{Z_i=k\}}\right)^2}, \qquad k = 1, ..., K,$$

for an outcome model with homoscedastic variance σ^2. Since the estimated parameters $\hat{\beta}_k$ are independent, the variance of any linear combination of them is just a linear combination of the individual variances using positive loadings. From this we can see that, in order to minimize the variance of an estimator for treatment contrasts, we need to minimize the sum of squares of the weights, for a fixed total sum of weights for each treatment.

Now, how should we balance the covariates in order to unbiasedly estimate the causal parameter of interest? Let $\beta_k^S = \frac{1}{n}\sum_{i=1}^{n} Y_i(k)$ be the (unobserved) average potential outcome in the sample under treatment k. If we have the conditional mean model

$$\mathrm{E}[Y_i(z) \mid \boldsymbol{X}_i] = \sum_{k=1}^{K} h_k(\boldsymbol{X}_i)\mathbb{1}_{\{z=k\}},$$

then, in terms of $\boldsymbol{X}_i$, the parameters in (2) are given by $\beta_k = \mathrm{E}[h_k(\boldsymbol{X}_i)]$. Hence,

$$\begin{aligned}
\left|\mathrm{E}\left[\hat{\beta}_k - \beta_k^S \mid \boldsymbol{X}, Z\right]\right| &= \left|\mathrm{E}\left[\frac{\sum_{i=1}^{n} w_i Y_i \mathbb{1}_{\{Z_i=k\}}}{\sum_{i=1}^{n} w_i \mathbb{1}_{\{Z_i=k\}}} - \frac{1}{n}\sum_{i=1}^{n} Y_i(k) \;\middle|\; \boldsymbol{X}, Z\right]\right| \\
&= \left|\frac{\sum_{i=1}^{n} w_i \,\mathrm{E}[Y_i(k) \mid \boldsymbol{X}_i, \mathbb{1}_{\{Z_i=k\}}]\mathbb{1}_{\{Z_i=k\}}}{\sum_{i=1}^{n} w_i \mathbb{1}_{\{Z_i=k\}}} - \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}[Y_i(k) \mid \boldsymbol{X}_i, \mathbb{1}_{\{Z_i=k\}}]\right| \\
&= \left|\frac{\sum_{i=1}^{n} w_i h_k(\boldsymbol{X}_i)\mathbb{1}_{\{Z_i=k\}}}{\sum_{i=1}^{n} w_i \mathbb{1}_{\{Z_i=k\}}} - \frac{1}{n}\sum_{i=1}^{n} h_k(\boldsymbol{X}_i)\right|.
\end{aligned}$$

This means that the closer we force $\frac{\sum_{i=1}^{n} w_i h_k(\boldsymbol{X}_i)\mathbb{1}_{\{Z_i=k\}}}{\sum_{i=1}^{n} w_i \mathbb{1}_{\{Z_i=k\}}}$ to be to $\frac{1}{n}\sum_{i=1}^{n} h_k(\boldsymbol{X}_i)$, the smaller the bias will be. If we assume $h_k(\boldsymbol{x})$ is linear in each of the covariates, then this simplifies to forcing $\frac{\sum_{i=1}^{n} w_i X_{ip}\mathbb{1}_{\{Z_i=k\}}}{\sum_{i=1}^{n} w_i \mathbb{1}_{\{Z_i=k\}}}$ to be close to $\frac{1}{n}\sum_{i=1}^{n} X_{ip}$ for all $p = 1, ..., P$, where X_{ip} is the p-th observed covariate for subject i.

More specifically, if $h_k(\boldsymbol{x}_i) = \alpha_0^k + \sum_{p=1}^{P} \alpha_p^k x_{ip}$ and we require a maximum imbalance of $\delta_p > 0$, that is, $\left| \frac{\sum_{i=1}^{n} w_i X_{ip}\mathbb{1}_{\{Z_i=k\}}}{\sum_{i=1}^{n} w_i \mathbb{1}_{\{Z_i=k\}}} - \frac{1}{n}\sum_{i=1}^{n} X_{ip} \right| < \delta_p$ for all $p = 1, ..., P$, then the absolute bias of $\hat{\beta}_k$ with respect to β_k^s will be bounded by $\sum_{p=1}^{P} \delta_p \left|\alpha_p^k\right|$,

$$
\begin{aligned}
\left|\mathrm{E}\left[\hat{\beta}_k - \beta_k^S \;\middle|\; Z = z\right]\right| &\leq \mathrm{E}\left[\left|\mathrm{E}\left[\hat{\beta}_k - \beta_k^S \;\middle|\; \boldsymbol{X}, Z\right]\right| \;\middle|\; Z = z\right] \\
&= \mathrm{E}\left[\left| \frac{\sum_{i=1}^{n} w_i \left(\alpha_0^k + \sum_{p=1}^{P} \alpha_p^k X_{ip}\right)\mathbb{1}_{\{Z_i=k\}}}{\sum_{i=1}^{n} w_i \mathbb{1}_{\{Z_i=k\}}} \right.\right. \\
&\qquad \left.\left. -\frac{1}{n}\sum_{i=1}^{n}\left(\alpha_0^k + \sum_{p=1}^{P} \alpha_p^k X_{ip}\right)\right| \;\middle|\; Z = z\right] \\
&= \mathrm{E}\left[\left|\sum_{p=1}^{P} \alpha_p^k \left(\frac{\sum_{i=1}^{n} w_i X_{ip}\mathbb{1}_{\{Z_i=k\}}}{\sum_{i=1}^{n} w_i \mathbb{1}_{\{Z_i=k\}}} - \frac{1}{n}\sum_{i=1}^{n} X_{ip} \right)\right| \;\middle|\; Z = z\right] \\
&\leq \sum_{p=1}^{P} \delta_p |\alpha_p^k|,
\end{aligned}
$$

where conditioning on $Z = z$ means conditioning on the observed treatment assignments in the sample.

It is important to note that this is a bound for the absolute bias with respect to the corre-

sponding sample average potential outcome. The bias with respect to the target population will depend on how well the sample represents that population. To avoid additional sampling bias, we need either to sample $\boldsymbol{X}_i$'s from the marginal distribution of $\boldsymbol{X}_i$ in the target population or know the marginal treatment assignment distribution in the target population.

When the functions h_k are non-linear, we can balance auxiliary covariates in order to approximately balance the entire distribution of the observed covariates and control bias. For example, if the functions h_k are non-linear but additive on the covariates X_p, then the auxiliary covariates can be polynomial splines or indicators for the quantiles of the marginal distribution of X_p. See Zubizarreta (2015) and Resa and Zubizarreta (2016) for a discussion on how to balance non-linear functions of the covariates. See also Wong and Chan (2018), Hirshberg and Wager (2018), Wang and Zubizarreta (2019), and Kallus (2020) for frameworks where the auxiliary covariates are basis functions that span general function spaces for $\mathrm{E}[Y_i(z)|\boldsymbol{X}_i]$.

Having established the conditions we would like the weights to satisfy, we solve the following convex optimization problem in order to find the weights that minimize the variance of the estimator subject to constraints that control bias via covariate balance. We minimize the sum of the squared weight values $\|\boldsymbol{w}\|_2^2$ of the n-dimensional vector of weights $\boldsymbol{w}$, subject to the covariate balance constraints

$$\left| \frac{\sum_{i=1}^{n} w_i X_{ip} \mathbb{1}_{\{Z_i=k\}}}{\sum_{i=1}^{n} w_i \mathbb{1}_{\{Z_i=k\}}} - \frac{1}{n} \sum_{i=1}^{n} X_{ip} \right| \leq \delta_p, \ k = 1..., K, \ p = 1, ..., P,$$

where δ_p is a scalar determined by the researcher that represents the desired level of covariate balance, and a constraint on the sum of the weights

$$\sum_{i=1}^{n} w_i \mathbb{1}_{\{Z_i=k\}} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{Z_i=k\}}, \ k = 1..., K, \tag{3}$$

for non-negative weights, $w_i \geq 0$, $i = 1, ..., n$.

Alternatively, instead of including the bias terms as restrictions, we can incorporate them in the objective function (Athey et al. 2018). Doing this would imply sacrificing covariate balance in favor of ensuring feasibility of the optimization problem. For this we minimize

$$\lambda \|\boldsymbol{w}\|_2^2 + (1-\lambda) \left\| \frac{\sum_{i=1}^{n} w_i X_{ip} \mathbb{1}_{\{Z_i=k\}}}{\sum_{i=1}^{n} w_i \mathbb{1}_{\{Z_i=k\}}} - \frac{1}{n} \sum_{i=1}^{n} X_{ip} \right\|_\infty^2$$

subject to constraint (3) on the sum of positive weights, $w_i \geq 0$, $i = 1, ..., n$. In this formulation, $\lambda \in (0, 1)$ is a tuning parameter that represents the trade-off between bias and variance.

In both formulations, we suggest fixing the sum of weights for each treatment group to be equal to the sample proportion of subjects in each group, giving a total sum of weights equal to one. This is analogous to building stabilized inverse probability weights. Between both formulations, we favor the covariate balance constrained one, as it serves as a check to practical violations to the positivity assumption. In fact, while the unconstrained formulation will always give a solution, with the covariate balance constrained formulation the problem may be infeasible. We view this as helpful information about the data at hand, since it tells us if it is possible at all or not to balance the covariates to the extent defined by δ.

An important question in practice with both the constrained and an unconstrained formulations is how to choose the parameters δ and λ. In this discussion, we focus on the choice of δ since we favor the constrained formulation of the weighting problem and δ is a function of λ. Following Wang and Zubizarreta (2019), solving the constrained formulation is equivalent to doing shrinkage estimation of the inverse propensity score. Wang and Zubizarreta rely on this connection to devise an algorithm for selecting the degree of approximate covariate balance δ. In this algorithm, the covariates X_p (or transformations thereof, $B_q(X_p)$, that

span a more general function space, for $q = 1, ..., Q$) are standardized and δ is constant for all $p = 1, ..., P$ (albeit the algorithm can be generalized, at a computational cost, for variable tolerances δ_{pq}). The intuition behind this algorithm is that the correct (true) inverse propensity score weights will balance the covariates at both the population and the sample levels, and by analogy, they will balance the covariates at both the original sample and bootstrap sample levels. Thus, for a grid of candidate values for δ, Wang and Zubizarreta (2019) select the value of δ that produces the best covariate balance across bootstrap samples. This algorithm is a data-driven procedure for selecting δ. This algorithm is implemented in the `msbw` package for `R` and we use it in the following simulation study.

4 Simulation study

In this section, we evaluate the performance of the proposed weighting method in a simulation study with two settings: one with a categorical treatment, and another with a continuous treatment. Each setting has two scenarios: one with correct model specification, and another with model mispecification. In addition to the proposed weighting method, we consider three other methods that estimate the generalized propensity score: maximum likelihood, CBPS (Fong et al. 2018), and GBM (McCaffrey et al. 2004, 2013). We consider parametric implementations of these and our method.

4.1 Data generating mechanisms

4.1.1 First simulation setting: categorical treatment

The first simulation setting is based on one of the designs in Yang et al. (2016). Here, the treatment variable has three values. There are six covariates in the outcome model as well as in the treatment assignment model. In this design, $\boldsymbol{X_i} = (X_{i1}, X_{i2}, X_{i3}, X_{i4}, X_{i5}, X_{i6})$

is generated as: $X_{i1}, X_{i2}, X_{i3} \sim \mathcal{N}_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\boldsymbol{\mu} = (0,0,0)^\top$, $\sigma_{11} = 2$, $\sigma_{22} = \sigma_{33} = 1$, $\sigma_{12} = \sigma_{21} = 1$, $\sigma_{13} = \sigma_{31} = -1$, and $\sigma_{23} = \sigma_{32} = -0.5$; $X_{i4} \sim \text{Unif}[-3,3]$; $X_{i5} \sim \chi^2_1$; and $X_{i6} \sim \text{Bern}(0.5)$. The treatment variable Z_i is constructed from the treatment indicators $(\mathbb{1}_{\{Z_i=1\}}, \mathbb{1}_{\{Z_i=2\}}, \mathbb{1}_{\{Z_i=3\}}) \sim \text{Multinom}(p(1|\boldsymbol{X_i}), p(2|\boldsymbol{X_i}), p(3|\boldsymbol{X_i}))$, where $p(z|\boldsymbol{X_i}) = \frac{exp(X_i^\top \eta_z)}{\sum_{\zeta=1}^{3} exp(X_i^\top \eta_\zeta)}$, with $\eta_1 = (0,0,0,0,0,0)^\top$, $\eta_2 = 0.7 \times (1,1,1,-1,1,1)^\top$, and $\eta_3 = 0.4 \times (1,1,1,1,1,1)^\top$. The outcome is given by $Y_i(1) \mid \boldsymbol{X_i} = -1.5 + X_{i1} + X_{i2} + X_{i3} + X_{i4} + X_{i5} + X_{i6} + \varepsilon_i$, $Y_i(2) \mid \boldsymbol{X_i} = -3 + 2X_{i1} + 3X_{i2} + X_{i3} + 2X_{i4} + 2X_{i5} + 2X_{i6} + \varepsilon_i$, and $Y_i(3) \mid \boldsymbol{X_i} = 1.5 + 3X_{i1} + 1X_{i2} + 2X_{i3} - X_{i4} - X_{i5} - X_{i6} + \varepsilon_i$, where $\varepsilon_i \sim \mathcal{N}(0,1)$. Under this generating mechanism, the true dose-response function is $\text{E}\left[Y_i(z)\right] = \text{E}\left[\text{E}\left[Y_i(z) \mid \boldsymbol{X_i}\right]\right] = 0$ for all $z = 1,2,3$. Therefore the average effect is zero when comparing any two treatment levels.

To compare the methods' performance under misspecification of the outcome and treatment models, we simulate data under a second scenario. In this scenario, we introduce non-linearities in both models by using transformed covariates $\boldsymbol{X_i'}$, where $X'_{i2} = sign(X_{i2}) \times |X_{i2}|^{\frac{1}{2}}$, $X'_{i5} = \frac{1}{exp(X_{i5})}$, and $X'_{ij} = X_{ij}$ for $j = 1,3,4,6$ as the observed covariates used to obtain the weights. The true treatment effect remains the same. For both scenarios we generate a total of 1000 replications, each one with a sample size of $n = 1500$. In the Online Supplementary Material we present additional results for smaller and larger sample sizes, $n = 750$ and $n = 3000$.

4.1.2 Second simulation setting: continuous treatment

The goal of this second simulation setting is to evaluate the performance of the proposed method when the treatment is continuous. For this, we follow Fong et al. (2018) and consider two of the scenarios in their study: one with treatment and outcome models that are linear in the covariates, and another with models that are both non-linear. In this design, $\boldsymbol{X_i} = (X_{i1}, ..., X_{i10}) \sim \mathcal{N}_{10}(\mathbf{0}, \boldsymbol{\Sigma})$, with $\sigma_{jj} = 1$ for $j = 1, ..., 10$ and $\sigma_{jj'} = 0.2$ for $j \neq j'$. In the first

scenario, the outcome and treatment are given by $Z_i = X_{i1}+X_{i2}+0.2X_{i3}+0.2X_{i4}+0.2X_{i5}+\epsilon_i$, and $Y_i(z) \mid \boldsymbol{X_i} = X_{i2}+0.1X_{i4}+0.1X_{i5}+0.1X_{i6}+z+\varepsilon_i$, where $\epsilon_i \sim \mathcal{N}(0,9)$ and $\varepsilon_i \sim \mathcal{N}(0,25)$. This means that the true dose-response function is $\mathrm{E}\left[Y_i(z)\right] = \mathrm{E}\left[\mathrm{E}\left[Y_i(z) \mid \boldsymbol{X_i}\right]\right] = z$, and thus, the value of the relevant parameter is one.

In the second scenario, $Z_i = (X_{i2}+0.5)^2 + 0.4X_{i3} + 0.4X_{i4} + 0.4X_{i5} + \epsilon_i$, and $Y_i(z) \mid \boldsymbol{X_i} = 2(X_{i2}+0.5)^2 + 0.5X_{i4} + 0.5X_{i5} + 0.6X_{i6} + z + \varepsilon_i$. Here, the true dose-response function is $\mathrm{E}\left[Y_i(z)\right] = \mathrm{E}\left[\mathrm{E}\left[Y_i(z) \mid \boldsymbol{X_i}\right]\right] = 2.5 + z$, so the relevant parameter remains equal to one. Again, we conduct 1000 replications each with sample size $n = 1500$. In the Online Supplementary Material we present additional the results for $n = 750$ and $n = 3000$.

4.2 Estimation methods

The first method we use to estimate the generalized propensity score is GLM. In the categorical treatment setting, we fit a multinomial logistic regression model with the function `multinom` from the `nnet` package for `R`, while in the continuous treatment setting, we fit a linear model with the `lm` function. After estimating the treatment assignment probabilities, we compute the stabilized versions of the weights. The second method is CBPS for which we use the `CBPS` package for `R` (Fong et al. 2018). We obtain the weights directly from the function output. The third method is GBM, which we fit with the function `mnps` from the `R twang` package (Ridgeway et al. 2006). We obtain the corresponding weights by applying the `get.weights` function. For the three methods, we include the covariates as linear terms, which means that the model is correctly specified in the first scenario of each setting and is misspecified in the second scenarios. For comparison, in the second scenarios we also include more adequate transformations of the covariates as indicators for the deciles of their distribution plus indicators for the 5^{th} and 95^{th} percentiles. The inclusion of these covariates improves the specification of the methods. We run all of the above functions using their default values.

For the multi-valued SBW, we obtain the weights by solving the formulation of the optimization problem that includes the covariate balance requirements as constraints with a small value for the tolerance selected by the algorithm in Wang and Zubizarreta (2019). In a similar way to the three other methods, in the first scenarios of both simulation settings, we only require mean balance for the covariates, and in the second scenarios we also consider the previous indicator transformations of the covariates. For this we use the new package `msbw` for `R`.

Since the `twang` package does not include an option for a continuous treatment, and our method is designed for categorical treatments, in the second setting, for both `twang` and the multi-valued SBW, we transform the original continuous treatment variable into a categorical one with ten levels for the treatment deciles. As a measure of comparison of the original bias, we also include a naive estimator where no weights are used.

In the first setting, the estimands are $\tau_{12} = \mathrm{E}[Y_i(2) - Y_i(1)]$, $\tau_{13} = \mathrm{E}[Y_i(3) - Y_i(1)]$, and $\tau_{23} = \mathrm{E}[Y_i(3) - Y_i(2)]$; and in the second setting, they are β_0 and β_1 from the MSM $\mathrm{E}[Y_i(z)] = \beta_0 + \beta_1 z$. We compare the performance of the methods in terms of bias and root mean square error (RMSE), as well as in average length and coverage of 95% confidence intervals obtained using the robust sandwich variance estimator.

4.3 Results

4.3.1 Categorical treatment

For each method, Figure 1 shows the distributions of the estimated treatment effects across simulated data sets, and Table 1, their bias and RMSE. In the figures, the dashed line denotes the true parameter value. In the second plot of Figure 1, we can see that without weighting the estimated difference between treatment 3 and treatment 1, $\hat{\tau}_{13}$, is not badly biased, and none of the methods exacerbate this bias. However, the greater variability of

the GLM weights translates into greater variability of the estimates, resulting in an RMSE twice as large as the one of the unweighted (Unwt) estimator. For the two other treatment effect contrasts, τ_{12} and τ_{23}, GBM improves with respect to the unweighted estimator in terms of bias and RMSE, but not quite as much as the other methods. CBPS exhibits very good performance with its exact specification (see Section 2.2 of Imai and Ratkovic 2014 for details). In this first scenario, the multi-valued SBW provides the lowest absolute bias and RMSE for the three parameters. This translates into confidence intervals with smaller length and near to nominal coverage (Table 2). Albeit being conservative, in Table 2 we see that the CBPS and SBW intervals cover the true parameter value in almost every simulation repetition with the shortest average lengths, albeit being conservative.

Figure 1: Boxplots of the estimated parameters in the first setting, first scenario.

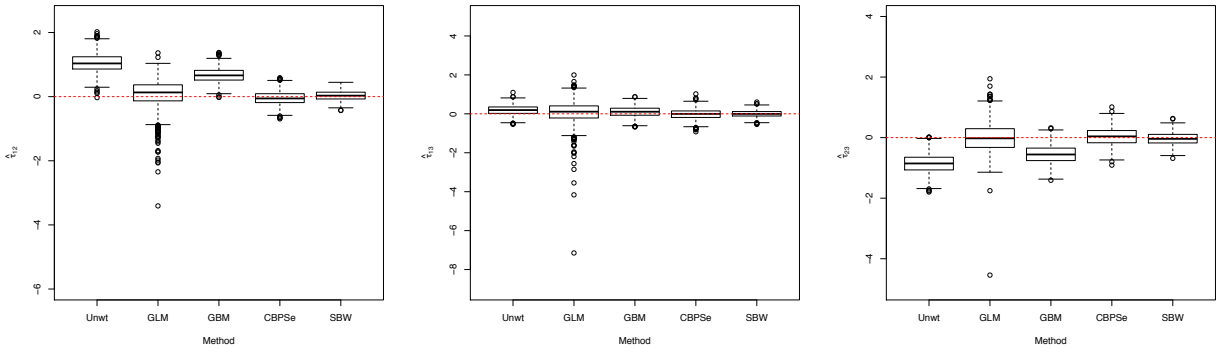


Table 1: Bias and RMSE of the estimated parameters in the first setting, first scenario.

Method	$\hat{\tau}_{12}$		$\hat{\tau}_{13}$		$\hat{\tau}_{23}$	
	Bias	RMSE	Bias	RMSE	Bias	RMSE
Unwt	1.05	1.08	0.19	0.31	-0.86	0.91
GLM	0.06	0.52	0.07	0.61	0.00	0.50
GBM	0.66	0.70	0.11	0.29	-0.56	0.63
CBPS	-0.05	0.22	-0.02	0.26	0.03	0.30
SBW	0.04	0.15	0.00	0.18	-0.03	0.21

Figure 2 and tables 3 and 5 show the results for the second scenario with misspecification. As expected, here all the methods produce biased effect estimates. The comparative performance of the methods is similar to the one of the previous scenario, except for the relative improvement of GBM. To some extent, the flexibility of GBM makes it more robust to model

Table 2: Coverage (%) and average length of 95% confidence intervals using the robust sandwich variance estimator in the first setting, first scenario.

Method	τ_{12}		τ_{13}		τ_{23}	
	Coverage	Length	Coverage	Length	Coverage	Length
Unwt	4.50	1.17	88.30	0.98	21.40	1.24
GLM	98.50	2.06	97.70	2.19	98.30	2.26
GBM	60.20	1.45	98.90	1.36	77.10	1.57
CBPS	100.00	2.03	100.00	2.08	100.00	2.24
SBW	100.00	1.61	100.00	1.45	100.00	1.66

misspecification, providing practically the same results than when the correct, original covariates are used. Nonetheless, the bias and RMSE for this method is still the largest for $\hat{\tau}_{12}$ and $\hat{\tau}_{23}$, which are the ones with larger bias on the unweighted sample. The other methods perform similarly relative to each other, with CBPS and SBW producing the best results in terms of bias, RMSE, and coverage. Figure 3 and tables 4 and 6 shows the results when more accurate transformations of the covariates are used. As a consequence, the biases and RMSEs in Figure 2 and Table 3 are reduced, and the coverage of the confidence intervals in Table 5 is increased.

Figure 2: Boxplots of the estimated parameters in the first setting, second scenario.

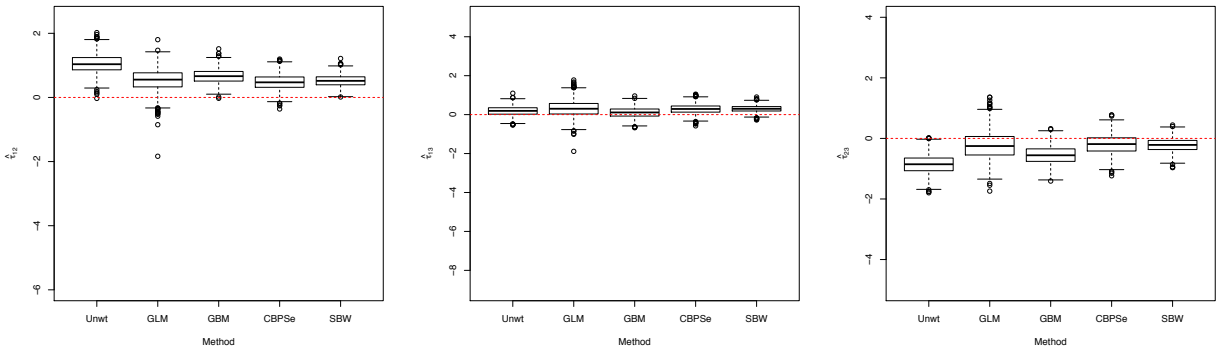


Figure 3: Boxplots of the estimated parameters in the first setting, second scenario. Here, all the methods include more accurate transformations of the observed covariates. As a result, the biases in Figure 2 are reduced.

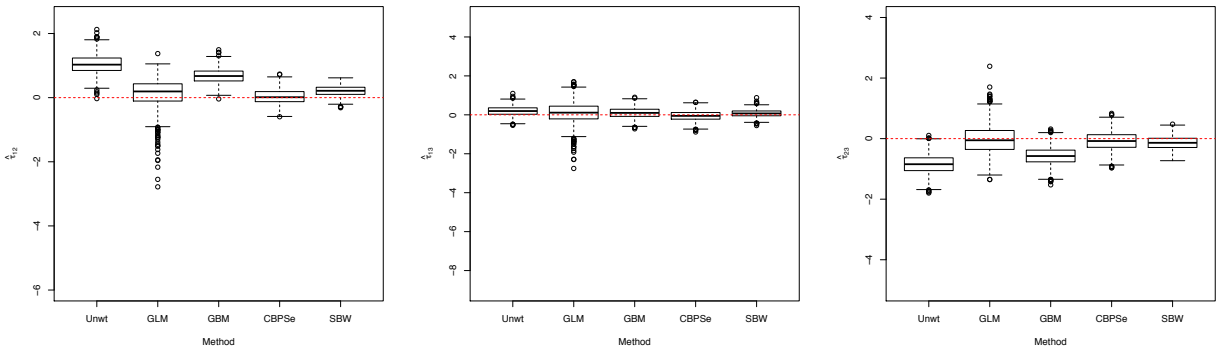


Table 3: Bias and RMSE of the estimated parameters in the first setting, second scenario.

Method	$\hat{\tau}_{12}$		$\hat{\tau}_{13}$		$\hat{\tau}_{23}$	
	Bias	RMSE	Bias	RMSE	Bias	RMSE
Unwt	1.05	1.08	0.19	0.31	-0.86	0.91
GLM	0.54	0.64	0.32	0.53	-0.22	0.51
GBM	0.66	0.70	0.11	0.28	-0.56	0.63
CBPS	0.48	0.54	0.28	0.38	-0.20	0.37
SBW	0.52	0.55	0.30	0.34	-0.22	0.32

Table 4: Bias and RMSE of the estimated parameters in the first setting, second scenario. Here, all the methods include more accurate transformations of the observed covariates. As a result, the biases and RMSEs in Table 3 are reduced.

Method	$\hat{\tau}_{12}$		$\hat{\tau}_{13}$		$\hat{\tau}_{23}$	
	Bias	RMSE	Bias	RMSE	Bias	RMSE
Unwt	1.04	1.08	0.19	0.31	-0.85	0.91
GLM	0.13	0.48	0.10	0.54	-0.03	0.49
GBM	0.68	0.71	0.10	0.28	-0.57	0.64
CBPS	0.03	0.23	-0.06	0.26	-0.08	0.32
SBW	0.21	0.26	0.07	0.20	-0.14	0.25

4.3.2 Continuous treatment

The results for the setting with a continuous treatment in the correctly specified scenario appear in Figure 4 and tables 7 and 8. The spirit of the results is similar to the one in the previous setting. While the performance of the methods is comparable in terms of RMSE, in terms of bias the best performance is by CBPS and SBW. Interval coverage is close to the

Table 5: Coverage (%) and average length of 95% confidence intervals using the robust sandwich variance estimator in the first setting, second scenario.

Method	τ_{12}		τ_{13}		τ_{23}	
	Coverage	Length	Coverage	Length	Coverage	Length
Unwt	4.50	1.17	88.30	0.98	21.40	1.24
GLM	84.60	1.87	96.30	1.92	96.50	2.20
GBM	60.40	1.45	98.70	1.35	77.10	1.57
CBPS	96.90	1.89	99.40	1.88	99.60	2.19
SBW	93.70	1.61	98.50	1.35	99.70	1.70

Table 6: Coverage (%) and average length of 95% confidence intervals using the robust sandwich variance estimator in the first setting, second scenario. Here, all the methods include more accurate transformations of the observed covariates. As a result, the coverage of the confidence intervals in Table 5 is increased.

Method	τ_{12}		τ_{13}		τ_{23}	
	Coverage	Length	Coverage	Length	Coverage	Length
Unwt	5.20	1.17	88.80	0.98	22.70	1.24
GLM	98.10	2.09	98.10	2.21	98.10	2.30
GBM	57.20	1.44	98.30	1.34	75.00	1.57
CBPS	100.00	2.09	100.00	1.96	100.00	2.26
SBW	100.00	1.76	99.80	1.56	100.00	1.74

nominal 95% with all methods except for GBM.

Figure 4: Boxplots of $\hat{\beta}_1$ in the second setting, first scenario.

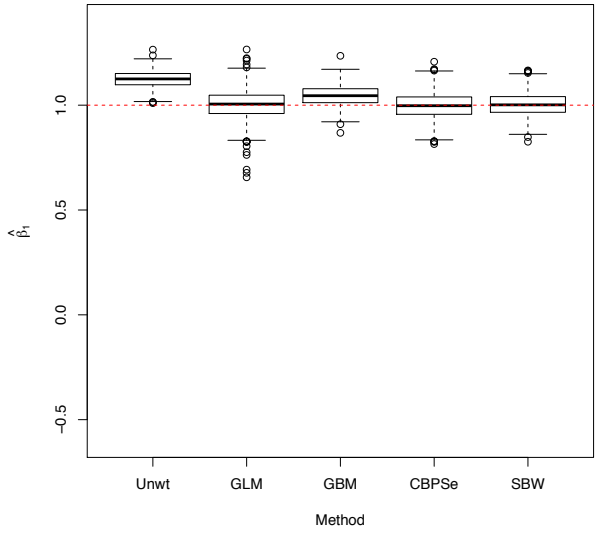


Table 7: Bias and RMSE of the estimated parameters of the MSM in the second setting, first scenario.

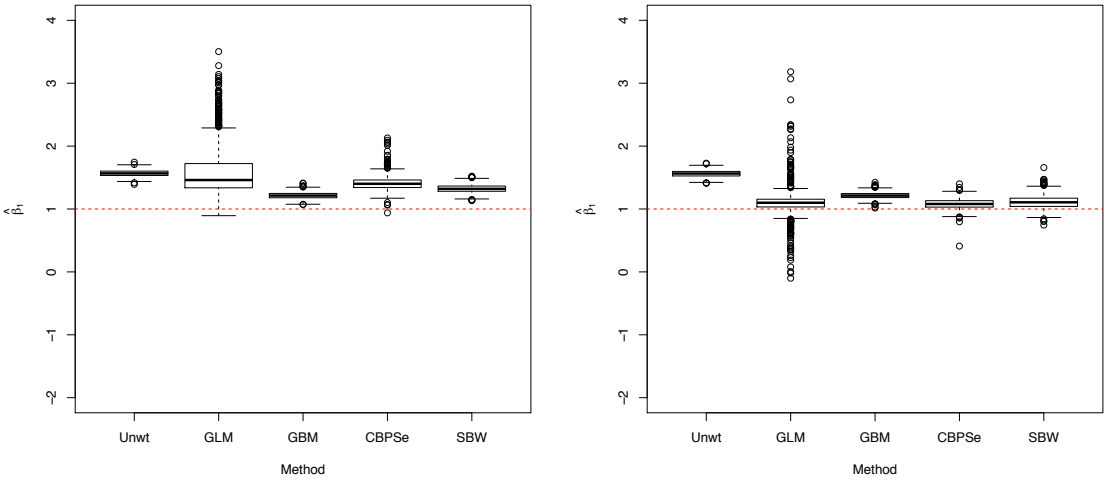
Method	$\hat{\beta}_0$		$\hat{\beta}_1$	
	Bias	RMSE	Bias	RMSE
Unwt	0.00	0.13	0.12	0.13
GLM	0.01	0.18	0.00	0.07
GBM	0.01	0.16	0.05	0.07
CBPS	0.01	0.18	0.00	0.06
SBW	0.01	0.15	0.00	0.05

Table 8: Coverage (%) and average length of 95% confidence intervals using the robust sandwich variance estimator in the second setting, first scenario.

Method	β_0		β_1	
	Coverage	Length	Coverage	Length
Unwt	95.00	0.52	9.80	0.15
GLM	95.20	0.68	93.70	0.25
GBM	95.00	0.63	82.10	0.19
CBPS	94.70	0.69	95.30	0.24
SBW	95.50	0.61	95.30	0.21

Finally, in the second scenario of this second setting (where the treatment is continuous and the models are misspecified), none of the methods produce the desired results; see Figure 5, and tables 9 and 11. In terms of bias and RMSE for the parameter of interest, β_1, GLM and CBPS appear to be more sensitive to model misspecification, whereas SBW performs slightly better and GBM produces the best results. Still, under misspecification, the coverage of the confidence intervals of all the methods is extremely low (Table 11). Their performance is improved when including improved transformations of the covariates, with CBPS exhibiting lowest RMSE and SBW the highest coverage; see tables 10 and 12. It is worth noting that both GBM and SBW are based on a ten-level categorical version of the treatment as opposed to its original, continuous version, as in the other two methods. With SBW, we also categorized the treatment into five and twenty levels (for its quantiles and ventiles) but the results were not sensitive to this degree of categorization.

Figure 5: Boxplots of $\hat{\beta}_1$ in the second setting, second scenario.



(a) All the methods include the original covariates as linear terms.

(b) All the methods include transformations of the covariates.

Table 9: Bias and RMSE of the estimated parameters of the MSM in the second setting, second scenario.

Method	$\hat{\beta}_0$		$\hat{\beta}_1$	
	Bias	RMSE	Bias	RMSE
Unwt	1.79	1.80	0.57	0.57
GLM	1.99	2.01	0.59	0.70
GBM	1.77	1.78	0.21	0.22
CBPS	1.92	1.94	0.41	0.43
SBW	1.93	1.94	0.32	0.33

Table 10: Bias and RMSE of the estimated parameters of the MSM in the second setting, second scenario. Here, all the methods include more accurate transformations of the covariates, improving the results in Table 9.

Method	$\hat{\beta}_0$		$\hat{\beta}_1$	
	Bias	RMSE	Bias	RMSE
Unwt	1.80	1.81	0.56	0.57
GLM	2.41	2.47	0.10	0.25
GBM	1.79	1.80	0.21	0.22
CBPS	2.05	2.19	0.08	0.11
SBW	2.14	2.17	0.11	0.15

Table 11: Coverage (%) and average length of 95% confidence intervals using the robust sandwich variance estimator in the second setting, second scenario.

Method	β_0		β_1	
	Coverage	Length	Coverage	Length
Unwt	0.00	0.60	0.00	0.20
GLM	0.30	0.82	3.40	0.53
GBM	0.00	0.72	3.30	0.22
CBPS	0.60	0.76	1.30	0.30
SBW	0.00	0.67	0.00	0.23

Table 12: Coverage (%) and average length of 95% confidence intervals using the robust sandwich variance estimator in the second setting, second scenario (model misspecification); here again, all the methods include more accurate (but not correct) transformations of the observed covariates, improving the results in Table 11.

Method	β_0		β_1	
	Coverage	Length	Coverage	Length
Unwt	0.00	0.60	0.00	0.20
GLM	0.10	1.35	69.40	0.41
GBM	0.00	0.71	2.30	0.22
CBPS	0.20	1.17	77.10	0.31
SBW	1.10	1.54	85.80	0.46

5 Case study

5.1 The EPT data

The earthquake struck Chile on February 27, 2010. Before the earthquake, between November and December of 2009, Chile's Ministry of Social Development collected a new version of its main household survey, the National Socioeconomic Characterization Survey or CASEN (MDS 2011b). Between May and June of 2010, in order to evaluate the impact of the earthquake, the Ministry re-interviewed a representative subsample of 22,456 of the original 71,460 households in CASEN. The resulting longitudinal survey is called Post Earthquake Survey or EPT (MDS 2011a). Perhaps for the first time, the EPT provides detailed measurement of the same individuals before and after a great earthquake, that is, of magnitude 8 or above. Using this longitudinal structure, we may compare, without recall bias, individuals with dif-

ferent levels of exposures to the earthquake but who appeared similar in terms of covariates prior to the earthquake. See Table 13 for a list of these covariates.

Table 13: Covariates in the EPT.

Demographic covariates	Housing structure
Age (years)	Acceptable
Women	Reparable
Indigenous ethnic group	Irreparable
Household size	Overcrowding
Marital status	No
Married or cohabitating	Medium
Divorced or widow	Critical
Singe	Health before the earthquake
Socioeconomic covariates	Health problem (last month)
Education (years)	Hospitalized (last year)
Employment status	Has a psychiatric problem
Employed	Self-rated health
Unemployed	Poor
Inactive	Fair
Individual work income (1000 pesos)	Good
Household per capita income (1000 pesos)	Missing
Household total income (1000 pesos)	Health insurance
Poor	Public (FONASA)
Housing before the earthquake	Private (ISAPRE)
Housing status	Other
Own housing or paying to own it	No
Rented housing	Unknown
Ceded housing	Disability
Irregular use of housing	Self-sufficient or low
Housing rent per year (pesos)	Moderate or severe
0-25,000	No
25,001-50,000	Other
50,001-75,000	Rural zone
>75000	

5.2 Intensity of the earthquake

Following Zubizarreta et al. (2013), we use peak ground acceleration (PGA; USGS 2011b, 2014) to measure the strength of the earthquake. Unlike other earthquake scales, such as the Modified Mercalli or the Richter scales, the PGA is an entirely physical measure of how

strongly the earth shakes in different geographic areas. While Zubizarreta et al. (2013) used matching to compare similar individuals that experienced very high and very low PGA (this is, a binary exposure), in this study we evaluate the effect of the earthquake on PTSD as a function of the earthquake intensity, both as a categorical and continuous exposures. In particular, the categorical exposure has three values: low [0, 0.08), medium [0.08, 0.25), and high [0.25,) PGA; whereas the continuous exposures corresponds to the actual PGA.

5.3 Outcomes

To assess the impact of the earthquake on PTSD we use the total score from the self-rated Davidson Trauma Scale (Davidson et al. 1997) questions in the post-earthquake survey. These include two items rated on a five-point scale for each one of the 17 symptoms of PTSD from the Diagnostic and Statistical Manual of Mental Disorders. For each pair of items, one referred to frequency (1 = "not at all" to 5 = "every day") and the other referred to severity (1 = "not at all distressing" to 5 = "extremely distressing"). Adding the responses of every item, the total score ranged from 34 to 170.

5.4 Study design

In this study, for the continuous covariates (age, household size, education, individual work income, household per capita income, and household total income), we included distributional balance constraints for the deciles of the covariates, in addition to mean balance constraints. We excluded from the study subjects for which no treatment or outcome variables were available, leaving a sample size of $n = 23322$ subjects.

We focused on estimating $\tau_{12} = \mathrm{E}[Y_i(2) - Y_i(1)]$ and $\tau_{23} = \mathrm{E}[Y_i(3) - Y_i(2)]$ for the categorical version of the treatment, and β_0 and β_1 from the MSM $\mathrm{E}[Y_i(z_i)] = \beta_0 + \beta_1 z_i$ for the continuous version, where Y_i represents the PTSD score of subject i. Again, for the continuous version

of the treatment we constructed the weights by categorizing the treatment into deciles, and then used these weights to estimate the parameters of $\mathrm{E}[Y_i(z)] = \beta_0 + \beta_1 z$ by weighted least squares. We obtained 95% confidence intervals using the robust sandwich variance estimator.

5.5 Results

Figure 6 summarizes covariate balance before and after weighting for the categorical version of the exposure. Specifically, the figure shows the distribution of absolute standardized mean differences in the covariates, defined as $|\bar{x}_{w,z} - \bar{x}_{\mathrm{tar}}|/s_{\mathrm{tar}}$, where $\bar{x}_{w,z}$ is the weighted mean of covariate x in treatment category z, and $\bar{x}_{\mathrm{tar}}$ and s_{tar} are the mean and standard deviation of the target population of inference. (Here, $\bar{x}_{\mathrm{tar}}$ and s_{tar} are computed from the overall (random) sample in order to estimate the effect of each treatment category on the overall population.) The figure shows the distribution of absolute standardized mean differences in the covariates for all treatment categories z relative to the target. Before weighting, most of the covariates are not drastically imbalanced; however, there are several covariates for which the absolute standardized differences in means are beyond the commonly accepted level of 0.1 absolute standardized mean differences. After weighting, we have close to perfect balance for all the means of the covariates.

Having adjusted for these imbalances, we compute the effect estimates and their corresponding confidence intervals. Table 14 shows these results. We see that on average the second level of exposure to the earthquake increased the PTSD score by 7.70 points relative to the first (lowest) level. This effect is somewhat higher when experiencing the third (highest) exposure level relative to the second one, with a value of 8.37 points on average. In other words, the average cumulative effect of the third exposure level relative to the first one is 16.07 points. All these effect estimates are statistically significant at the 5% level.

Figure 7 presents a summary of the mean covariate balance before and after weighting when

Figure 6: Boxplots of absolute standardized differences in means before and after weighting for the categorical version of the exposure with three categories.

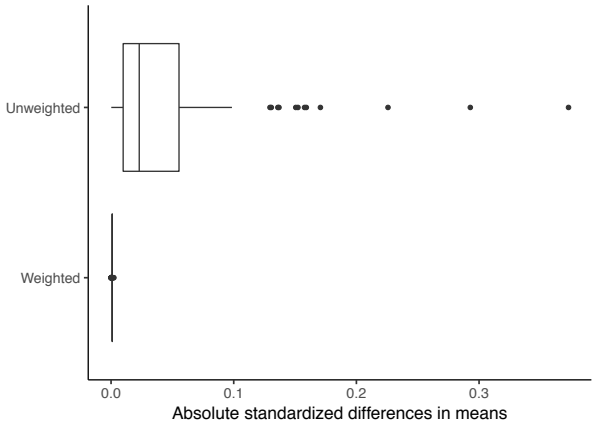


Table 14: Effect estimates for the categorical version of the exposure.

	τ_{12}	τ_{23}
Estimate	7.70	8.37
95% C.I.	(6.95, 8.45)	(7.49, 9.26)

the continuous exposure is transformed into a categorical variable with ten levels. Before weighting, there are more imbalances across exposure levels than in the case with three levels. However, after weighting all these imbalances are practically removed with the SBW.

Figure 7: Boxplots of absolute standardized differences in means before and after weighting for the exposure with ten categories.

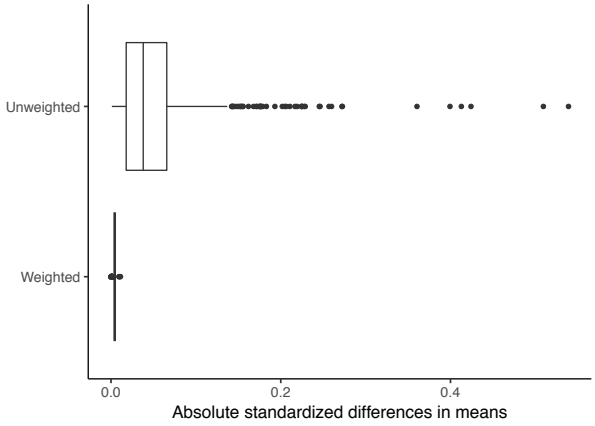


Table 15 shows the estimated average effect for the continuous exposure and its corresponding 95% confidence interval. Each additional unit of PGA has an average effect of 60.88 points on the PTSD score. The PGA of the 2010 Chilean earthquake ranged between 0 to 0.32 [g], with a mean value of 0.2 [g] and a standard deviation of 0.1 [g]. This estimate represents a maximum average increase of 20 points on the PTSD score for the earthquake. These results are consistent with the results of Zubizarreta et al. (2013), who used multivariate matching methods to estimate the impact of two extreme levels exposures to the earthquake.

For illustration purposes, in this application we have considered a linear specification for the MSM for the continuous treatment. A more general specification is $\mathrm{E}[Y_i(z_i)] = \sum_{r=0}^{R} \beta_r z_i^r$ for $R = 2$ or 3, in order to capture a non-linear effect of the continuous treatment.

Table 15: Effect estimate for the continuous version of the exposure.

	β_0	β_1
Estimate	3.37	60.88
95% C.I.	(2.86, 3.88)	(58.04, 63.72)

6 Summary and concluding remarks

In this paper, we extended the stable balancing weights of Zubizarreta (2015) to observational studies with multi-valued treatments and suggested how to use them with continuous treatments. We determined the conditions that the weights need to satisfy in order to provide close to unbiased treatment effect estimates with reduced variability, and posited a convex optimization problem that can be solved to obtain them.

We conducted a simulation study to compare the proposed method with the standard approach to estimate the generalized propensity score (GLM) and other recent methods that target covariate balance in the estimation process in different ways (CBPS and GBM). The

simulation included two different settings, one with a categorical treatment with three categories and one with a continuous treatment. We analyzed both settings in a scenario where the probability and/or outcome model were correctly specified and a scenario where they were both misspecified. In general, the exact specification of CBPS and the multi-valued SBW produced the best results.

Finally, we used the proposed method to estimate the impact of levels of exposure to the 2010 Chilean earthquake on PTSD. The earthquake intensity was considered both as a multi-valued categorical variable and a continuous variable. In both cases, we concluded that there was a statistically significant positive effect of the intensity level of the earthquake on posttraumatic stress disorder.

While our approach is designed for multi-valued treatments, it appears to perform well with continuous treatments too. With a continuous treatment, the approach requires categorizing the continuous treatment variable and finding the weights of minimum variance that approximately balance the covariates for each treatment category. Then, these weights are used to estimate the parameters of an MSM for the original continuous treatment. Ideally, the weights will be found for the continuous treatment variable itself, but with many covariates, this will often entail solving a complex multivariate density estimation problem for which the resulting weights may fail to balance the covariates and yield unstable effect estimates. In practice, with our approach we recommend categorizing the continuous treatment into quantiles (e.g., deciles, as in our empirical studies, and further breaking the two extreme deciles into two ventiles, if the data permits). While this approach appears to perform well in practice, a formal extension of our approach to continuous treatments is yet to be developed. Other directions of future research include variance estimators that better take advantage of our proposed weights (while the resulting confidence intervals are well-calibrated and often narrower than with alternative methods, they tend to be conservative; here, bootstrap approaches are to be explored). Further directions of future research include the development

weighting approaches for factorial treatments and effect modification.

References

Athey, S., Imbens, G. W., and Wager, S. (2018), "Approximate residual balancing: debiased inference of average treatment effects in high dimensions," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Davidson, J. R., Book, S. W., Colket, J. T., Tupler, L. A., Roth, S., David, D., Hertzberg, M., Mellman, T., Beckham, J. C., Smith, R. D., Davison, R. M., Katz, R., and Feldman, M. E. (1997), "Assessment of a new self-rating scale for post-traumatic stress disorder," *Psychological Medicine*, 27, 153–160.

Diamond, A. and Sekhon, J. S. (2013), "Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies," *Review of Economics and Statistics*, 95, 932–945.

Fong, C., Hazlett, C., Imai, K., et al. (2018), "Covariate balancing propensity score for a continuous treatment: application to the efficacy of political advertisements," *The Annals of Applied Statistics*, 12, 156–177.

Goenjian, A. K., Steinberg, A. M., Najarian, L. M., Fairbanks, L. A., Tashjian, M., and Pynoos, R. S. (2000), "Prospective study of posttraumatic stress, anxiety, and depressive reactions after earthquake and political violence," *The American Journal of Psychiatry*, 6, 911–895.

Hainmueller, J. (2012), "Entropy balancing for causal effects: a multivariate reweighting method to produce balanced samples in observational studies," *Political Analysis*, 20, 25–46.

Hirshberg, D. A. and Wager, S. (2018), "Augmented minimax linear estimation," *arXiv preprint arXiv:1712.00038*.

Holland, P. W. (1986), "Statistics and causal inference," *Journal of the American statistical Association*, 81, 945–960.

Imai, K. and Ratkovic, M. (2014), "Covariate balancing propensity score," *Journal of the Royal Statistical Society: Series B*, 76, 243–263.

Imbens, G. W. (2000), "The role of the propensity score in estimating dose-response functions," *Biometrika*, 87, 706–710.

Imbens, G. W. and Rubin, D. B. (2015), *Causal inference in statistics, social, and biomedical sciences*, Cambridge University Press.

Kallus, N. (2020), "Generalized optimal matching methods for causal inference," *Journal of Machine Learning Research*, *in press*.

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 (with discussion)," *Statistical Science*, 22, 523–539.

Li, F. and Li, F. (2019), "Propensity score weighting for causal inference with multiple treatments," *Annals of Applied Statistics*, 13, 2389–2415.

Lopez, M. J., Gutman, R., et al. (2017), "Estimation of causal effects with multiple treatments: a review and new ideas," *Statistical Science*, 32, 432–454.

Mattei, A. and Mealli, F. (2015), "Discussion of "on bayesian estimation of marginal structural models"," *Biometrics*, 71, 293–296.

McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R., and Burgette, L. F. (2013), "A tutorial on propensity score estimation for multiple treatments using generalized boosted models," *Statistics in medicine*, 32, 3388–3414.

McCaffrey, D. F., Ridgeway, G., and Morral, A. R. (2004), "Propensity score estimation with boosted regression for evaluating causal effects in observational studies," *Psychological methods*, 9, 403–425.

MDS (2011a), "Ficha técnica encuesta post terromoto," *http://www.ministeriodesarrollosocial.gob.cl/encuesta-post-terremoto/index.html*.

— (2011b), "The national socioeconomic characterization survey," .

Neria, Y., Nandi, A., and Galea, S. (2008), "Post-traumatic stress disorder following disasters: a systematic review," *Psychological Medicine*, 38, 467–480.

Neyman, J. (1923, 1990), "On the application of probability theory to agricultural experiments," *Statistical Science*, 5, 463–480.

Pimentel, S. D., Kelz, R. R., Silber, J. H., and Rosenbaum, P. R. (2015), "Large, sparse optimal matching with refined covariate balance in an observational study of the health outcomes produced by new surgeons," *Journal of the American Statistical Association*, 110, 515–527.

Pollitz, F. F., Brooks, B., Tong, X., Bevis, M. G., Foster, J. H., Bürgmann, R., R. Smalley, J., Vigny, C., Socquet, A., Ruegg, J. C., Campos, J., Barrientos, S., Parra, H., Soto, J. C. B., Cimbaro, S., and Blanco, M. (2011), "Coseismic slip distribution of the February 27, 2010 Mw 8.8 Maule, Chile earthquake," *Geophysical Research Letters*, 38.

Resa, M. A. and Zubizarreta, J. R. (2016), "Evaluation of subset matching methods and forms of covariate balance," *Statistics in Medicine*, 35, 4961–4979.

Ridgeway, G., McCaffrey, D., Morral, A., Burgette, L., and Griffin, B. A. (2006), "Toolkit for weighting and analysis of nonequivalent groups: A tutorial for the twang package," *Santa Monica, CA: RAND Corporation*.

Robins, J. (1986), "A new approach to causal inference in mortality studies with a sus-

tained exposure period — application to control of the healthy worker survivor effect," *Mathematical Modelling*, 7, 1393–1512.

Robins, J. M., Hernán, M. A., and Brumback, B. (2000), "Marginal structural models and causal inference in epidemiology," *Epidemiology*, 550–560.

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.

Roussos, A., K., A., Goenjian, Steinberg, A. M., Sotiropoulou, C., Kakaki, M., Kabakos, C., Karagianni, S., and Manouras, V. (2005), "Posttraumatic stress and depressive reactions among children and adolescents after the 1999 earthquake in Ano Liosia, Greece," *The American Journal of Psychiatry*, 162, 530–537.

Rubin, D. B. (1974), "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of Educational Psychology*, 66, 688.

— (1980), "Randomization analysis of experimental data: the Fisher randomization test comment," *Journal of the American Statistical Association*, 75, 591–593.

— (1986), "Comment: which ifs have causal answers," *Journal of the American Statistical Association*, 81, 961–962.

Sharan, P., Chaudhary, G., Kavathekar, S. A., and Saxena, S. (1996), "Preliminary report of psychiatric disorders in survivors of a severe earthquake," *The American Journal of Psychiatry*, 153, 556–558.

USGS (2011a), "Magnitude 8.8 - Offshore Bio-Bio, Chile," `http://earthquake.usgs.gov/earthquakes/recenteqsww/Quakes/us2010tfan.php`.

— (2011b), "Shakemap 2011," `http://earthquake.usgs.gov/earthquakes/shakemap/global/shake/2010tfan/download`.

— (2014), "Largest earthquakes in the world since 1900," http://earthquake.usgs.gov/earthquakes/world/10_largest_world.php.

Wang, X., Gao, L., Shinfuku, N., Zhang, H., Zhao, C., and Shen, Y. (2000), "Longitudinal study of earthquake-related PTSD in a randomly selected community sample in north China," *The American Journal of Psychiatry*, 157, 1260–1266.

Wang, Y. and Zubizarreta, J. R. (2019), "Minimal dispersion approximately balancing weights: Asymptotic properties and practical considerations," *Biometrika*, in press.

Wong, R. K. and Chan, K. C. G. (2018), "Kernel-based covariate functional balancing for observational studies," *Biometrika*, 105, 199–213.

Yang, S., Imbens, G. W., Cui, Z., Faries, D. E., and Kadziola, Z. (2016), "Propensity score matching and subclassification in observational studies with multi-level treatments," *Biometrics*, 72, 1055–1065.

Yehuda, R., Bierer, L. M., Lopez-Ibor, J., Christodoulou, G., Maj, M., Sartorius, N., and Okasha, A. (2005), "Re-evaluating the link between disasters and psychopathology," *Disasters and mental health*, 65–80.

Zubizarreta, J. R. (2012), "Using mixed integer programming for matching in an observational study of kidney failure after surgery," *Journal of the American Statistical Association*, 107, 1360–1371.

— (2015), "Stable weights that balance covariates for estimation with incomplete outcome data," *Journal of the American Statistical Association*, 110, 910–922.

Zubizarreta, J. R., Cerdá, M., and Rosenbaum, P. R. (2013), "Effect of the 2010 Chilean earthquake on posttraumatic stress: Reducing sensitivity to unmeasured bias through study design," *Epidemiology*, 24, 79–87.

Zubizarreta, J. R., Reinke, C. E., Kelz, R. R., Silber, J. H., and Rosenbaum, P. R. (2011),

"Matching for several sparse nominal variables in a case-control study of readmission following surgery," *The American Statistician*, 65, 229–238.