Handling Limited Overlap in Observational Studies with Cardinality Matching*

Giancarlo Visconti† José R. Zubizarreta‡

Abstract

A common problem encountered in observational studies is limited overlap in covariate distributions across treatment groups. To address this problem, and avoid strong modeling assumptions, it has become common practice to restrict analyses to the portions of the treatment groups that overlap or, ultimately, that are balanced in their covariate distributions. Often, this is done by matching on the estimated propensity score or on coarsened versions of the observed covariates. A recent alternative methodology that, in a sense, encompasses these two approaches is cardinality matching. Cardinality matching is a flexible matching method that uses integer programming to find the largest matched sample that is balanced according to criteria specified before matching by the investigator. In this paper, we apply and illustrate the method of cardinality matching and show how to use it to directly balance several features of the covariates, including their trajectories in time and their distributions, without requiring exact matching. We demonstrate how cardinality matching addresses the problem of limited overlap using the original covariates, as opposed to a summarized or coarsened version of them. We discuss how this method can be extended to build matched samples that are not only balanced but also representative by design. We also show how this method facilitates sensitivity analyses for hidden biases. We explain these advancements through an observational study of the electoral impact of the 2010 earthquake in Chile.

Keywords: Causal inference; Matched sampling; Natural disasters; Natural experiments; Observational studies

*We thank Jake Bowers, Zach Branson, Shigeo Hirano, Gary King, Christopher Lucas, and Luke Miratrix for useful comments and suggestions. All remaining errors are our responsibility.

†Ph.D. Candidate, Department of Political Science, Columbia University, 420 West 118th Street, International Affairs Building, 7th Floor, New York, NY 10027 (email, giancarlo.visconti@columbia.edu).

‡Assistant Professor, Department of Health Care Policy, and Department of Statistics, Harvard University, 180 Longwood Avenue, Office 307-D, Boston, MA 02115 (email, zubizarreta@hcp.med.harvard.edu).

1 Introduction

A guiding principle in the design of observational studies is to approximate the ideal study that would be conducted if it were possible to do it by controlled experimentation (Dorn 1953; Cochran and Rubin 1973; Rosenbaum 2010). In an experiment, randomization tends to produce treatments groups that are comparable in terms of both observed and unobserved covariates, and provides a basis for valid statistical inference that does not rely on distributional assumptions (see, e.g., Armitage 1982). In contrast, in observational studies randomization is absent, the treatment assignment mechanism is unknown, and there may be systematic covariate differences between the treatment groups that can bias effect estimates. In observational studies, matching methods are often used to adjust for differences in observed covariates, emulating the structure of the ideal randomized study, and facilitating sensitivity analyses of the outcome results to biases due to unobserved covariates (see, e.g., Chapter 4 of Rosenbaum 2002).

In practice, a common problem encountered in observational studies is the lack of common support or the limited overlap of the covariate distributions across treatment groups. As discussed by Ho et al. (2007) and Crump et al. (2009), limited overlap can lead to covariate imbalances and produce estimates that are overly sensitive to model misspecification. To address this problem, it is common practice with matching methods to "trim" the treatment and control samples, and restrict the analyses to the portions of the samples that overlap or, ultimately, that are balanced. Often, this is done using a summary measure of the covariates such as the propensity score or the Mahalanobis distance (e.g., Dehejia and Wahba 1999), although alternative matching methods that select a subset of the observations using the original observed covariates instead of a summary of them have also been proposed. Perhaps the most popular of these matching methods is Coarsened Exact Matching, or CEM (Iacus et al. 2011). In short, this method first coarsens the observed covariates, then forms multidimensional strata using the coarsened covariate values, and finally matches the treated and control units in the same stratum, discarding those units with no treated or control counterpart for each stratum.

While matching on the estimated propensity score or coarsened versions of the observed covariates are useful and popular methods, they can fail to balance the original covariates (Diamond and Sekhon 2013) or use very few observations (King et al. 2016). A recent, alternative method that overcomes these limitations is cardinality matching (Zubizarreta et al. 2014).

Taking advantage of recent advancements in mathematical programming (particularly, in integer programming), cardinality matching maximizes the cardinality or size of the matched sample that satisfies the investigators' requirements for covariate balance. As discussed in Zubizarreta (2012), these requirements are flexible and can enforce balance for different features of the empirical distributions of the observed covariates, such as moments, marginals and, in principle, their entire joints. These requirements can also enforce exact matching for coarsened covariates, as in CEM. As we argue in this paper, this flexibility in balancing covariates can be very important for avoiding the curse of dimensionality in exact matching with several covariates and facilitate the use of more observations in the analysis. In this paper, we describe and apply the method of cardinality matching in a study about natural disasters and electoral outcomes.

What are the benefits of cardinality matching? Though cardinality matching is by no means the silver bullet for observational studies, it has several appealing features. First, it handles limited overlap in covariate distributions by using the original covariates as opposed to summaries or coarsened versions of them, directly balancing the covariates, and finding the maximum number of observations for any given covariate balance criteria. Importantly, with cardinality matching these criteria are flexible: one can require different forms of balance covariates from mean balance to exact matching. Thus, in a sense, subject to a matching structure (e.g., a one-to-one matching structure) CEM is a particular case of cardinality matching, where all the balancing constraints enforce exact matching for coarsened covariates. In a study like ours, which examines the impact of an earthquake on electoral outcomes at the county level, where there are not many county-level observations to begin with, the flexibility of cardinality matching allows us to use more observations from the data than other matching methods.

The rest of this paper is organized as follows. In Section 2, we review cardinality matching, explain its mathematical programming formulation, and discuss different forms of covariate balance that can be enforced with it. In Section 3, we describe elections in Chile, the 2010 Chilean earthquake, our longitudinal county data, and the study design. In Section 4, we evaluate the matched sample and report effect estimates. In Section 5, we assess the sensitivity and generalizability of our findings. In Section 6, we compare cardinality matching to other matching techniques. Finally, in Section 7 we close with a summary of the study.

2 Maximizing the cardinality of a balanced matched sample

2.1 Review of cardinality matching

The goal of matching in observational studies is to remove the bias in effect estimates from imbalances in the observed covariates (Rosenbaum 2015b). Though this goal would ideally be accomplished using all of the available treated and control observations, this is rarely possible in practice because of limited overlap in their covariate distributions. In view of this limitation, investigators often subset or trim the treated and control samples and confine their analyses to subsamples for which covariates are balanced.

Perhaps the most common approach to this problem relies on the propensity score, whereby treated (control) units with an estimated propensity score outside the range of the control (treated) units are discarded (Stuart 2010). Afterwards, treated and control units are matched (for example, using nearest neighbor matching (Rubin 1973; Abadie et al. 2004) or optimal matching (Rosenbaum 1989; Hansen 2007)) and then covariate balance is checked. These steps are summarized in Algorithm 1 below. As discussed by Hainmueller (2012) and Diamond and Sekhon (2013), this procedure typically requires many iterations in order to achieve covariate balance.

This procedure can be improved by using the covariate balancing propensity score (Imai and Ratkovic 2014), a propensity score model that penalizes fits for which covariates are imbalanced. Other methods for selecting subsamples of treated and control observations are proposed by Crump et al. (2009) and Rosenbaum (2012). These approaches, however, address the problem of limited overlap using a one-dimensional summary of the covariates rather than the original covariates themselves.

An alternative matching method, coarsened exact matching (CEM; Iacus et al. 2011), selects a subset of the observations but using the original covariates. This method first coarsens the individual covariates (for example, by transforming income into quintiles or deciles), then exactly matches the treated and control units according to the coarsened covariates, and discards units that cannot be exactly matched. While exact matching is the ideal form of covariate balance, it suffers from dimensionality, as in actual practice the number of exact matching categories tends to explode combinatorially. In consequence, exact matching in practice tends to be relegated to a few covariates of overriding importance (see, for example, Chapter 9 of Rosenbaum 2010).

Another recent matching method that selects a subset of the treated and control units using the original covariates is cardinality matching (Zubizarreta et al. 2014). This method solves a linear integer programming problem to maximize the cardinality or size of a matched sample subject to flexible constraints on covariate balance. In their most stringent form, these constraints can require exact matching on the observed covariates, but also other weaker forms of covariate balance: aggregate balance of low-dimensional joint distributions via strength-k balancing (Hsu et al. 2015), balance of marginal distributions by balancing a coarsened version of the Kolmogorov-Smirnov statistic (Zubizarreta 2012) or by fine balancing (Rosenbaum et al. 2007), and balance of moments by mean balancing suitable transformations of the covariates (see Zubizarreta 2012). In this manner, cardinality matching subsets or trims the treatment and control samples and balances multiple covariates in one step, and thus finds the largest matched sample that is balanced.

As described in Section 2.2, cardinality matching finds the largest matched sample that is balanced by solving a constrained linear integer programming problem where the constraints enforce a given matching structure (in our case study, one-to-one or pair matching) and pre-specified forms of covariate balance (ranging from mean balance to exact matching and including different forms of aggregate distributional balance) and the objective function maximizes the total number of matched pairs. Cardinality matching guarantees the optimal solution (or, in some cases, a near-optimal solution within a provable optimality gap) by exploiting recent advancements in algorithms and computation, including advanced heuristics, presolve methods, linear programming, and faster compilers (see Gurobi 2017 for details). In short, cardinality matching directly targets bias removal via flexible covariate balance constraints and variance reduction via an objective function on sample size by solving an optimization problem as opposed to greedy or heuristic methods.

Algorithms 1 and 2 show the basic steps involved in (i) standard matching methods based on the estimated propensity score and (ii) cardinality matching. With standard matching methods, steps 2-5 typically require many repetitions in order to satisfy the covariate balance requirements, whereas with cardinality matching the covariate balance requirements are satisfied in one step, without needing to estimate the propensity score.

In cardinality matching, finding the largest matched sample that satisfies the covariate balance requirements is often followed by rematching the balanced matched sample in order to minimize the covariate distances between matched units. Since the rematched sample is composed of the same treated and control units as the balanced matched sample, the aggregate balance measures are preserved after rematching. However, if the covariates used in calculating the distances are strong predictors of the outcome, then this rematching will reduce the heterogeneity in the matched pair differences in the outcomes. As discussed in Rosenbaum (2005) and Zubizarreta et al. (2014), this translates into greater efficiency and lower sensitivity to unobserved covariates. In Section 5, we illustrate these gains in the context of our case study of the effect of the Chilean earthquake and subsequent reconstruction process on electoral outcomes.

Algorithm 1 Matching with standard matching methods.

---

0. Specify the covariate balance requirements (e.g., mean balance).

Repeat:

1. Estimate the propensity score or another summary of the covariates.
2. Trim extreme observations according to the summary measure.
3. Match on the summary measure (e.g., using nearest neighbor matching).
4. Assess covariate balance.

Until:

The matched sample satisfies the covariate balance requirements.

Algorithm 2 Matching with cardinality matching.

---

0. Specify the covariate balance requirements (e.g., mean balance).
1. Find the largest matched sample that satisfies the covariate balance requirements.
2. Rematch the matched sample to minimize the covariate distances between matched units.

2.2 Mathematical programming formulation

Let $\mathcal{T} = \{1, \ldots, T\}$ and $\mathcal{C} = \{1, \ldots, C\}$ be the sets of treated and control units available before matching, with $T \leq C$. Define $t \in \mathcal{T}$ and $c \in \mathcal{C}$ as the corresponding indexes of these two sets. Write p to index the observed covariates in $\mathcal{P} = \{1, \ldots, P\}$, so that $\boldsymbol{x}_t = \{x_{t1}, \ldots, x_{tp}\}$ and $\boldsymbol{x}_c = \{x_{c1} \ldots, x_{cp}\}$ are the vectors of observed covariates for treated unit t and control unit c, respectively.

In its simplest form, the goal of cardinality matching is to find the largest pair-matched sample that is balanced. Using the binary decision variable $m_{tc} = 1$ if treated unit t is matched to control unit c, and $m_{tc} = 0$ otherwise, we can express this goal as the following objective function (1) subject to the following matching and balancing constraints (2) and (3), respectively. We wish to maximize the size of the matched sample,

$$\underset{\boldsymbol{m}}{\text{maximize}} \quad \sum_{t \in \mathcal{T}} \sum_{c \in \mathcal{C}} m_{tc}, \tag{1}$$

subject to matching constraints that enforce forming pairs,

$$\begin{aligned} \sum_{c \in \mathcal{C}} m_{tc} \leq 1,\ t \in \mathcal{T}, \\ \sum_{t \in \mathcal{T}} m_{tc} \leq 1,\ c \in \mathcal{C}, \end{aligned} \tag{2}$$

and balancing constraints that enforce the pair-matched samples to have similar covariate distributions

$$\left| \sum_{t \in \mathcal{T}} \sum_{c \in \mathcal{C}} m_{tc}(f(x_{tp}) - f(x_{cp})) \right| \leq \varepsilon_p \sum_{t \in \mathcal{T}} \sum_{c \in \mathcal{C}} m_{tc}. \tag{3}$$

In (3), f is a suitable function that transforms the observed covariates, and $\varepsilon_p \geq 0$ is a tolerance, both of which are defined by the investigator. In Section 2.3 below, we discuss different forms of covariate balance, but before a few remarks are in order.

First, in cardinality matching, bias reduction takes precedence over variance reduction. (1)-

(3) above maximize the size of the matched study, subject to removing biases due to imbalances in the observed covariates. This follows the recommendations of Cochran (1965, Section 2.2) and Rosenbaum (2010, Section 8.7), who state that the main objective of matching in observational studies is to balance observed covariates, and that increasing efficiency is typically a second-order concern, as biases that do not decrease as the sample size increases tend to dominate the mean squared error in large samples (see also Haviland et al. 2007 and King et al. 2016).

Second, from a more formal standpoint, it is easy to prove that under an homoskedastic constant additive treatment effect outcome model, cardinality matching minimizes the variance of a difference-in-means effect estimator that is approximately unbiased according to the covariate balancing constraints (see Kilcioglu and Zubizarreta 2016).

Third, as discussed by Imbens (2015, page 382), when confronted with limited overlap in covariate distributions, "with multiple covariates, it is difficult to see what trimming would need to be done." Often, the treated and control samples are trimmed using the propensity score, to then match the observations and check covariate balance on the original covariates (as in Algorithm 1). Cardinality matching essentially performs these three tasks in one step, "trimming" the treated and control samples using the original covariates.

Fourth, from a computational standpoint, cardinality matching solves an integer programming problem, so in the abstract cardinality matching is an NP-hard optimization problem. In pratice, however, relatively large instances of this problem can be solved quickly by using modern solvers, and the running time can be further shortened by using approximation algorithms (Zubizarreta et al. 2017).

Finally, in his influential article on observational studies, Cochran (1965) gives two basic pieces of advice for designing such studies (Rubin 2006, page 15): (i) "when selecting samples for study, make sure that they are large enough and have complete enough data to allow effects of practical importance to be estimated, and avoid treatment and control groups with large initial differences on confounding variables;" (ii) "use both the statistician and the subject-matter expert in the planning stages." Cardinality matching achieves the first goal because it finds the largest matched sample that is balanced, and facilitates the second goal because the knowledge of the subject matter expert can be incorporated into the matching problem through the balancing constraints (3). Again, these constraints are flexible and can require exactly matching all of the covariates or just balancing their means. We describe these constraints in what follows.

2.3 Forms of covariate balance

In our case study about the impact of the earthquake, we enforce different forms of covariate balance via the balancing constraints (3). For some covariates we enforce mean balance. To do this, we let f be equal to the identity so that (3) bounds the differences in means between the matched treated and control groups

$$\left| \frac{\sum_{t\in\mathcal{T}} \sum_{c\in\mathcal{C}} m_{tc} x_{tp}}{\sum_{t\in\mathcal{T}} \sum_{c\in\mathcal{C}} m_{tc}} - \frac{\sum_{t\in\mathcal{T}} \sum_{c\in\mathcal{C}} m_{tc} x_{cp}}{\sum_{t\in\mathcal{T}} \sum_{c\in\mathcal{C}} m_{tc}} \right| \leq \varepsilon_p \tag{4}$$

and choose ε_p so that none of the standardized differences in means are greater than 0.1 (Rosenbaum and Rubin 1985).

For other covariates of greater prognostic importance, we also enforce distributional balance by modifying (3) in two ways. First, to approximately balance the marginal distribution of a continuous covariate, we define indicators for the quantiles of the covariate in the treatment group and then mean balance these indicators via (4). Effectively, this balances a coarsened version of the Kolmogorov-Smirnov statistic (Zubizarreta 2012). Second, to perfectly balance the marginal distribution of a nominal covariate — a form of balance called fine balance (Rosenbaum et al. 2007) — we let $\varepsilon_p = 0$ and $f(x_{\cdot p}) = \mathbb{1}_{\{x_{\cdot p}=b\}}$, where $b \in \mathcal{B}$ indexes the categories of the nominal covariate, so that (3) becomes

$$\sum_{t\in\mathcal{T}}\sum_{c\in\mathcal{C}} m_{tc}\mathbb{1}_{\{x_{tp}=b\}} = \sum_{t\in\mathcal{T}}\sum_{c\in\mathcal{C}} m_{tc}\mathbb{1}_{\{x_{cp}=b\}},\ b\in\mathcal{B}. \tag{5}$$

This constraint makes the counts of treated and control units in the matched sample be the same for each category of the nominal covariate, but without constraining units to be matched within the same category (as in exact matching).

In the context of (3), it is straightforward to enforce exact matching as follows

$$\sum_{t\in\mathcal{T}}\sum_{c\in\mathcal{C}} m_{tc}\mathbb{1}_{\{x_{tp}=b\}\cap\{x_{cp}\neq b\}} = 0,\ b\in\mathcal{B}. \tag{6}$$

Of course, this is the ideal form of covariate balance, but also the most stringent one. See Zubizarreta (2012) and Zubizarreta et al. (2014) for a related discussion on forms of covariate balance.

3 Context, data, and study design

3.1 Elections and Earthquakes in Chile

After the end of the dictatorship in 1990, the center-left coalition "Concertación" won four presidential elections in a row. In January 2010, Sebastián Piñera was elected president, representing the center-right coalition "Coalición por el Cambio." In February 2010, a few days before Piñera's inauguration, the country was shattered by an 8.8 magnitude earthquake. Piñera had to deal with the aftermath of the disaster for the duration of his administration, and the earthquake and its consequences were a central political issue for the next four years (Arana 2016).

The central-southern regions of Chile were severely affected by this 8.8 earthquake that, according to NASA, shortened the length of the day by about 1.26 microseconds (Buis 2010). The earthquake, and the tsunami that followed, killed over 500 people and injured about 12,000 more (Choi 2012). The disaster affected six of the fifteen regions of Chile.[1] The areas that experienced the highest-intensity shaking suffered significant damage to roads, ports, bridges, airports, utilities, and communication networks (Hinrichs et al. 2011). About 220,000 homes, 3,700 schools, 17 hospitals, and 212 bridges were destroyed. The economic losses to the country were estimated at $30 billion, or 18% of the country's gross domestic product (McClean 2012).

People living in the most affected regions directly observed the reconstruction process, such as repairs to schools, hospitals, and roads, and the distribution of financial relief, for a long period of time. In November 2013, three years after the earthquake and mere weeks before the next presidential election, the government was still building new homes for residents in the affected areas,[2] and debate about the success of the reconstruction efforts were an important part of the campaign.[3]

Chile provides a unique opportunity to test the effects of earthquakes on electoral outcomes for three main reasons: (i) the fifteen regions of the country had suffered from earthquakes throughout the 20th century–events that cannot be anticipated; (ii) the country has a centralized political structure; and (iii) the country has low electoral volatility. The first and second points increase the comparability between counties in the exposed and control groups because all are eligible to treatment and have the same political configuration. Regarding the third point, the low electoral volatility in Chile means that previous electoral results (pretreatment measures of the outcomes) have strong predictive power over the subsequent electoral results (outcome of interest). The first post-earthquake presidential election provides the necessary conditions for understanding the effects of this particular natural disaster[4]

[1] These are regions of Valparaiso, O'Higgins, Maule, Biobío and Araucanía, and the Metropolitan Region.

[2] "Inauguran viviendas para damnificados del terremoto en Tomé.", *La Nación*, November 13, 2013.

[3] "Presidente defendió reconstrucción tras crítica de Bachelet.", *La Nación*, August 30, 2013.

[4] It is important to note that the treatment is not just the earthquake, but the subsequent reconstruction process as well. The government's provision of public and private goods and the evaluation of its performance

on voters' electoral choices.[5]

What are the electoral consequences of natural disasters? The literature can be divided according to two main arguments. The first posits that voters will always punish the government for negative shocks, including those whose origin is beyond the current administration's responsibility and scope of action, such as droughts and floods (Achen and Bartels 2004, 2016). In particular, "incumbents will pay at the polls for bad times, even in situations where objective observers can find little rational basis to suppose that those incumbents have had any part in producing the voters' pain" (Achen and Bartels 2004, page 7). Affected voters will simply blame the incumbent party any time their own well-being falls below 'normal' levels, regardless of whether the incumbents have performed well or badly" (Achen and Bartels 2016, page 138). This explanation invokes citizens' emotional distress and tendency to channel their frustration and misfortune into voting for the challenger. The second argument posits that voters may punish or reward the incumbent based on their response to the catastrophe (Healy and Malhotra 2010; Gasper and Reeves 2011). Specifically, "observing that incumbents are adversely affected by natural disasters, does not necessarily mean that voters are irrational. Even though government cannot be blamed for the adverse natural events themselves, they can be held responsible for mitigation, response, and recovery" (Healy and Malhotra 2010, page 195).

In this paper, we focus on effect estimates and not on the causal mechanisms explaining the consequences of the disaster and the reconstruction process. The bulk of the literature on natural catastrophes and electoral outcomes has mainly focused on floods (Bechtel and Hainmueller 2011), hurricanes (Sinclair et al. 2011; Chen 2013), tornadoes (Healy and Malhotra 2010), and severe weather events (Gasper and Reeves 2011). However, there is little evidence of the effects of earthquakes on electoral outcomes.

are crucial variables for understanding the direction of the effects of the earthquake.

[5] Carlin et al. (2014) study the effects of this earthquake on democratic legitimacy, but to our knowledge there have been no studies on its impact on electoral results.

3.2 Longitudinal county data

We include 18 covariates in the analysis. All of these covariates provide county characteristics before the treatment (the 2010 earthquake) and have been extensively studied as predictors of electoral results in Chile.

We use counties as the unit of analysis. Because Chile has low levels of electoral volatility, the most important prognostic covariates are previous electoral results (Roberts 2013). We include the electoral results for the center-right, center-left, and alternative candidates in the last three presidential elections (1999, 2005, and 2009) using official data from the National Electoral Service. This allows us to have longitudinal county data of the pretreatment measures of the outcome variable.

Previous studies have shown that socioeconomic characteristics contribute to Chilean voters' political preferences (Altman 2004; López 2004; Navia et al. 2008; Luna 2010; Calvo and Murillo 2012). We include the Human Development Index (HDI) calculated by the United Nations Development Programme at the county level in 2003[6] and the poverty levels calculated the year before the earthquake by the Ministry of Social Development. Altman (2004) argues that the relationship between voting for the center-right coalition and the HDI follows a U-shape. Therefore, we also balance the quintiles of HDI to constrain the distribution of this covariate and not just its mean.

We also include a number of demographic characteristics at the county level from the 2002 national census that are important predictors of the vote.[7] Urban voters, especially from larger cities, are more likely to vote for the center-left coalition, while voters from smaller cities or rural localities are more likely to vote for the center-right coalition (González 1999). For this reason, we also include the proportion of rural citizens and the total population at the county level.

[6] This index represents the average of three indicators: health, education, and income. The last time the UNDP computed these indices at the county level before the earthquake was in 2003.

[7] This was the last census conducted before the earthquake.

3.3 Matched design

We use two criteria to define the exposed and control groups. The first criterion relies on the official reconstruction plan, in which the government identifies six of fifteen national regions as critically affected by the earthquake. Within these six affected regions, however, there was considerable variation in the intensity of the earthquake: the counties located farther from the epicenter or closer to the Andes were less affected. Furthermore, the official reconstruction plan shows that 87% of emergency houses were built in three of these six regions (Government of Chile 2010). Therefore, we need a second criterion to identify the most affected counties within the six regions.

To do this, we use the peak ground acceleration (PGA) as a measure of the intensity of the earthquake. Following Zubizarreta et al. (2013), in the exposed group we include the counties that had a PGA greater than 0.275 g. In the control group, we include the counties located in the nine unaffected regions, which were not part of the reconstruction plan. We also exclude the counties with a PGA lower than 0.275 g located in the affected regions, because extreme treatment conditions rather than marginal exposure tend to produce results less sensitive to hidden biases (Rosenbaum 2004). This use of extreme exposures is consistent with Zubizarreta et al. (2013), where they were used to study the effects of the 2010 earthquake on posttraumatic stress. As a way to check the suitability of these definitions, the counties placed in the control group had an average PGA of 0.06 g, whereas the counties placed in the exposed group had an average PGA of 0.29 g.

These criteria yield 77 counties in the exposed group and 95 in the control group. Figure 1 depicts the six regions affected by the earthquake in dark gray. The regions in light gray contain all the control counties, where the earthquake was barely felt. In the supplementary appendix, we include a map with the peak ground acceleration at the county level.

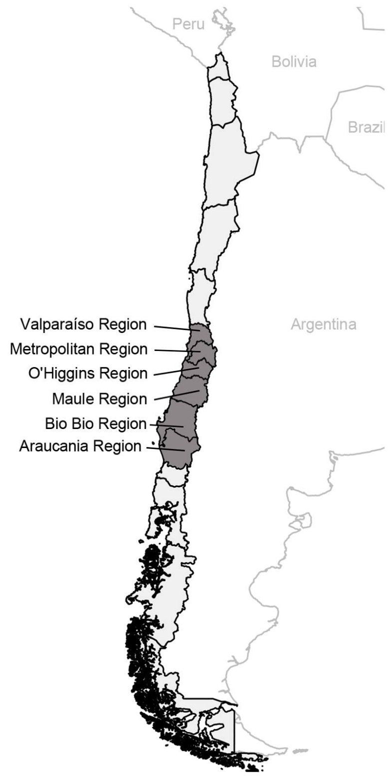


Figure 1: Map of the affected regions.

4 Evaluating and analyzing the matched sample

4.1 Covariate balance

Before matching, there were 77 exposed counties and 95 control ones. After using cardinality matching to find the largest pair-matched sample of counties that was balanced, we obtained 59 pairs of exposed and control counties. In this section, we describe and evaluate covariate balance in the matched sample. In Section 5.2 below, we discuss the extent to which the results from this sample can be generalized.

Figure 2 shows the absolute standardized differences in means before and after matching for the 18 mean balanced covariates. We observe that all the differences are smaller than 0.1 pooled standard deviations, as expected according to the constraints.

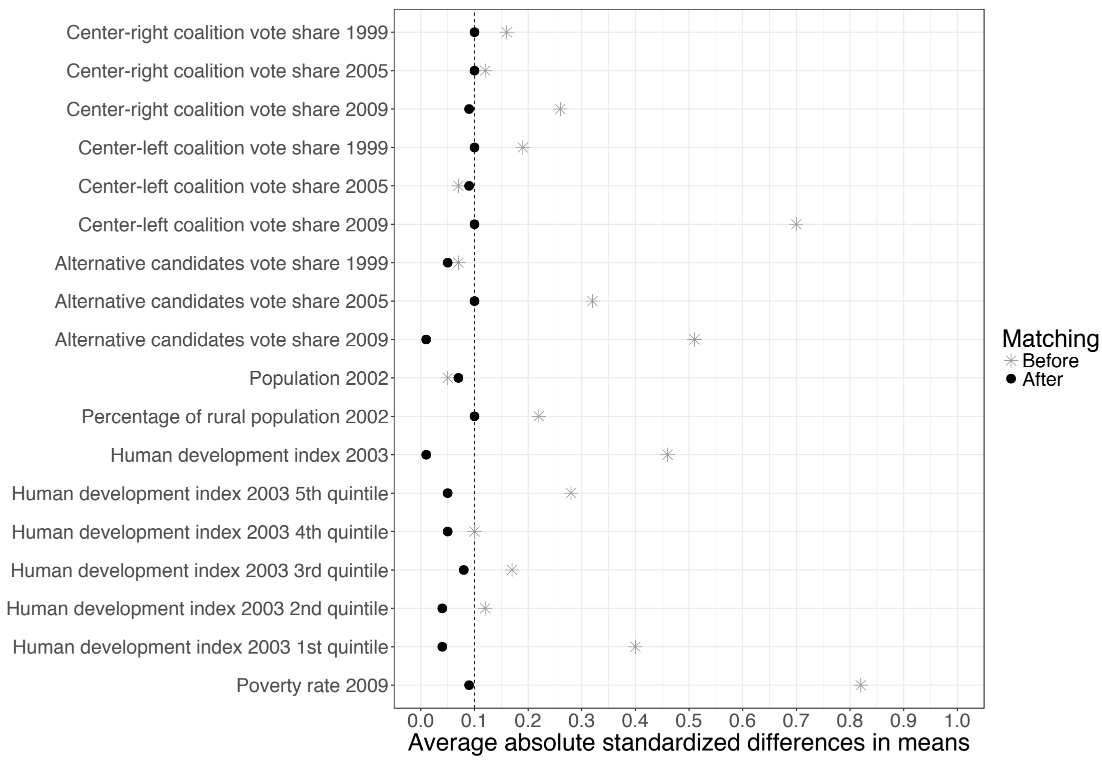
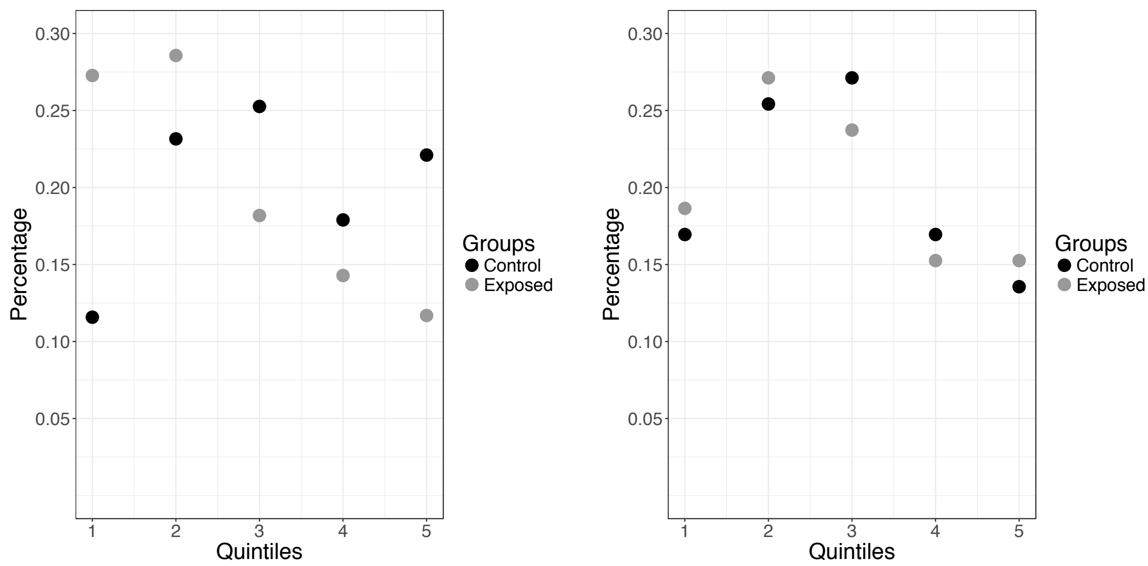


Figure 2: Standardized differences before and after matching.

As discussed in Section 2.3, we include the quintiles of the Human Development Index to balance the distributions. Figure 3 below shows the distribution of the HDI quintiles before and after matching. We constrained the mean and the distribution of this covariate in view of previous evidence of a non-linear (U-shaped) relationship with electoral preferences. In panel (a) of the figure, we observe that before matching there are substantial imbalances in the means by quintiles. In panel (b), we observe that after matching these imbalances were substantially reduced.

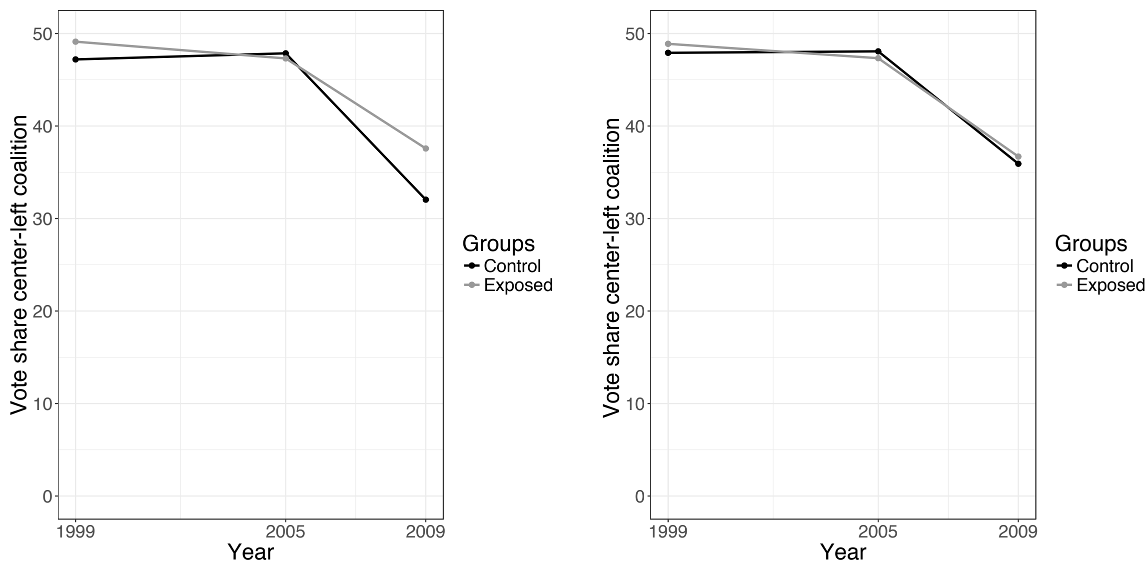


(a) Before matching.

(b) After matching.

Figure 3: Human Development Index quintiles before and after matching.

Figure 4 shows the vote share trajectories for the center-left coalition in the exposed and control groups before and after matching. Before matching, we observe substantial differences in the vote shares, particularly in the elections before the earthquake. After matching, the trajectories are almost indistinguishable between the exposed and control groups. Balancing trajectories allows us to find matched groups that are similar from a dynamic perspective, because they are formed of counties that had similar voting patterns before the earthquake (Haviland et al. 2008).



(a) Before matching. (b) After matching.

Figure 4: Vote share trajectories for the center-left coalition before and after matching.

Ideally, we would have matched exactly for the 18 covariates in our study. However, with 18 covariates, even if these covariates are discretized to be binary and have only two categories, there are 2^{18} or 262,144 possible combinations of types of units, while in our study there are 172 counties (77 exposed and 95 control ones). Furthermore, if the covariates are discretized to have 5 categories, then there are 5^{18} or 3,814,697,265,625 combinations of types of units, and by matching exactly we obtain no pairs. In principle, if we were matching exactly for the 18 covariates, we would be granting the same importance to balancing the 18^{th} interaction of the indicators of the tail category of each of the discretized covariates as to balancing the single indicator of the central category of a particular covariate (aiming to balance the first moment of the covariate, which is the first term that appears in a Taylor expansion of the outcome model). While exact matching is the ideal form of covariate balance, there is a curse of dimensionality in exact matching, and randomization does not produce exact matches but rather balance in aggregate. As discussed previously, flexible forms of covariate balance ranging from exact matching to mean balance can be incorporated within the framework of (1)–(3) to find the largest matched sample that is balanced by design.

4.2 Effect estimates

We estimate the effect of the earthquake on the incumbent vote share using the inferential methods described in Chapter 2 of Rosenbaum (2002). In particular, we use Wilcoxon's signed-rank test statistic to test the sharp null hypothesis of no treatment effect and derive a point estimate by solving Hodges and Lehmann's estimating equation. We obtain a 95% confidence interval by inverting the test. Through this method, we find that the 2010 earthquake increased the incumbent coalition vote share in the 2013 elections by 1.7 percentage points, with an associated p-value of 0.002 and a 95% confidence interval of [0.006, 0.026]. As discussed in Section 3.1, several mechanisms could have led to this effect, such as the strong presence of the government in the affected areas, but those mechanisms remain to be tested. Next, in Section 5, we assess the sensitivity of these findings to biases due to unobserved covariates as well as study their generalizability.

Table 1: Effect of the 2010 earthquake on the center-right coalition vote share in 2013.

Hodges-Lehman point estimate	0.017
95% confidence interval	[0.006, 0.026]
p-value	0.002

5 Assessing the sensitivity and generalizability of the findings

5.1 Rosenbaum bounds

In Section 2.1, we explained that in step 1 of cardinality matching we match for global covariate balance, and thus find the largest matched sample that satisfies the covariate balance requirements specified by the investigator. In step 2 of cardinality matching, we rematch the balanced matched sample from step 1 to reduce heterogeneity in the matched pairs. The rematched sample is composed of the same treated and control units, and the global covariate balance requirements will be preserved; however, the covariate distances between rematched units will be lower. As is argued in Rosenbaum (2005) and Zubizarreta et al. (2014), this translates into less sensitivity to hidden bias if the covariates used to compute the distances are strong predictors of the outcome. In this section, we illustrate this procedure and assess the sensitivity of the findings from Section 4.2.

For this, we use the sensitivity analysis model described in Chapter 4 of Rosenbaum (2002). In this model, each matched pair is composed of two units, i and i', which are identical on their observed covariates, so $\mathbf{x}_i = \mathbf{x}_{i'}$, but perhaps different in terms of an unobserved covariate, so $u_i \neq u_{i'}$. The probabilities of units i and i' of receiving treatment are denoted by π_i and $\pi_{i'}$, respectively. The model posits that the odds of receiving treatment for the two units may differ at most by a factor of $\Gamma \geq 1$ because of the unobserved covariate. More precisely, the model posits that

$$\frac{1}{\Gamma} \leq \frac{\pi_i/(1-\pi_i)}{\pi_{i'}/(1-\pi_{i'})} \leq \Gamma \tag{7}$$

for all matched pairs (i, i'). Clearly, if $\Gamma = 1$, then $\pi_i = \pi_{i'}$, and there is no hidden bias in the study. However, if $\Gamma > 1$, then $\pi_i \neq \pi_{i'}$, and for each $\Gamma > 1$, there is a range of possible inferences summarized by two extreme-case p-values. Using this model, we find the largest value of Γ such that the extreme-case p-values reject the null hypothesis of no treatment effect. Naturally, the larger is Γ, the less sensitive are the effect estimates to the influence of an unobserved covariate. In this sense, Γ measures the degree of departure from a study that is free of biases due to unobserved covariates.

As previous research has documented, Chile enjoys low levels of electoral volatility and stable voter choices over time. In other words, within any county the previous election's vote shares tend to be strong predictors of the vote shares in the next election. For example, the correlation coefficient of the vote share of the incumbent coalition between 2005 and 2013, as well as between 2009 and 2013, is 0.85. In step 2 of cardinality matching, we rematch the balanced matched sample from step 1 to minimize the Mahalanobis distance between matched pairs of the incumbent coalition's vote shares in years 2005 and 2009. In this way, after using cardinality matching for balance, the total sum of covariate distances between matched pairs is 62.7.[8] After rematching the balanced matched sample from cardinality matching, this distance decreases to 11.7. The distribution of covariate distances between matched pairs after matching for balance and rematching for homogeneity is shown in Figure 5. We can observe that the covariate distances between matched pairs after rematching for homogeneity are considerably less dispersed.

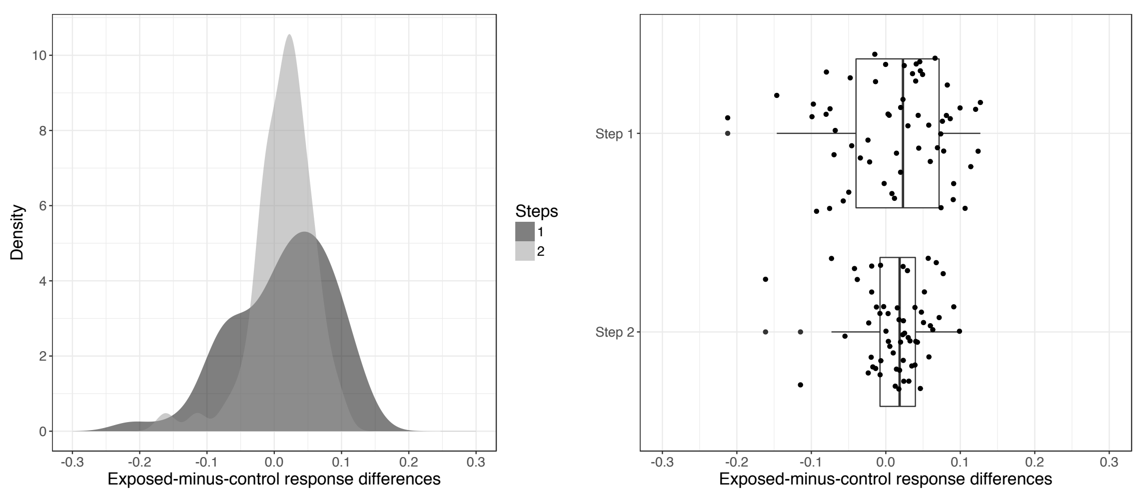


Figure 5: Distribution of exposed-minus-control response differences between matched pairs after matching for balance (step 1) and rematching for homogeneity (step 2).

Table 2 shows the highest value of the extreme-case p-values for each value of Γ in the matched sample after matching for balance (step 1) and rematching the matched sample for homogeneity (step 2). After matching for balance, the results become sensitive for $\Gamma = 1.002$, whereas after rematching for homogeneity they become sensitive for $\Gamma = 1.54$. In other words, for an unobserved covariate to explain away our effect estimate, it would need to increase two

[8] Here, we used the rank-based Mahalanobis distance described in Chapter 8 of Rosenbaum 2010).

identical counties' odds of exposure to the earthquake by a factor of 1.54. This illustrates the value of having pretreatment measures of the outcome in matching, because they tend to be strongly related to the outcome.

Table 2: Upper bounds on the one-sided p-value testing the null hypothesis of no treatment effect using Wilcoxon's signed-rank statistic after steps 1 and 2 of cardinality matching.

Γ	Step 1	Step 2
1.00	0.05	0.001
1.10	0.09	0.003
1.20	0.15	0.01
1.30	0.22	0.01
1.40	0.29	0.03
1.50	0.37	0.04
1.60	0.45	0.06

As discussed in Rosenbaum and Silber (2012), the parameter Γ assumes that the unobserved covariate is a nearly perfect predictor of the outcome, but Γ can be amplified into two parameters Λ and Δ that measure the strength of the relationship between the unobserved covariate and the treatment, and between the unobserved covariate and the outcome, respectively. The results show that $\Gamma = 1.54$ is equivalent to an unobserved covariate that simultaneously increases the odds of exposure to the earthquake and of a greater vote share by $\Lambda = \Delta = 2.71$ times. Similarly, $\Gamma = 1.54$ is equivalent to an unobserved covariate that increases the odds of exposure to the earthquake by $\Lambda = 6$ (or 1.85) times and the odds of a greater vote share by $\Delta = 1.85$ (or 6) times. In other words, for an unobserved covariate to explain away our estimated effect of 1.7%, it would need to have a moderately large effect on the odds of receiving treatment and of exhibiting a positive response.

5.2 Profile of the matched sample

In order to assess the extent to which the outcome results can be generalized, we present a profile of the matched sample and contrast it to a target sample of policy interest: a sample of all the counties in Chile.[9] As we can see from Table 3 below, in relation to all the counties in Chile, the matched counties are slightly less populated and more leftist. Thus, some differences between the groups exist, but only in a few covariates, and none of these differences are particularly large.

Table 3: Description of the samples of counties.

Covariate	All	Exposed matched	Control matched
Center-right coalition vote share 1999	0.49	0.47	0.48
Center-right coalition vote share 2005	0.49	0.49	0.48
Center-right coalition vote share 2009	0.44	0.42	0.43
Center-left coalition vote share 1999	0.47	0.49	0.48
Center-left coalition vote share 2005	0.46	0.47	0.48
Center-left coalition vote share 2009	0.33	0.37	0.36
Alternative candidates vote share 1999	0.04	0.04	0.04
Alternative candidates vote share 2005	0.04	0.04	0.04
Alternative candidates vote share 2009	0.23	0.21	0.21
Population 2002	46,540.49	37,991.29	34,411.47
Percentage of rural population 2002	0.38	0.42	0.45
Human development index 2003	0.69	0.68	0.67
Human development index 2003 5th quintile	0.21	0.15	0.14
Human development index 2003 4th quintile	0.19	0.15	0.17
Human development index 2003 3rd quintile	0.20	0.24	0.27
Human development index 2003 2nd quintile	0.20	0.27	0.25
Human development index 2003 1st quintile	0.20	0.19	0.17
Poverty rate 2009	0.17	0.17	0.16

[9] Chile has 346 counties, but we have excluded a few municipalities that had incomplete data either because they are too small or because they were divided or created after 2004. As a result, we focus on 326 counties.

6 Comparison to other matching methods

We now compare our results with those obtained through propensity score matching (PSM) and coarsened exact matching (CEM). For PSM, we estimated the propensity score using logistic regression, including the covariates as linear terms in the logistic regression model. With these estimated scores, we conducted two analyses: one using all the available exposed counties, and another dropping those exposed (control) counties without a counterpart in the control (exposed) group. In both cases we used the package `optmatch` in `R`.

In the first analysis, we obtained 77 matched pairs of counties, but 13 out of the 18 covariates presented substantial imbalances in means (see Figure 6). In the second analysis, we obtained 59 matched pairs of counties, but still 8 out the 18 covariates exhibited substantial imbalances (Figure 7). Arguably, these imbalances could be improved, but this would require a considerable amount of expertise and iteration as summarized in Algorithm 1.

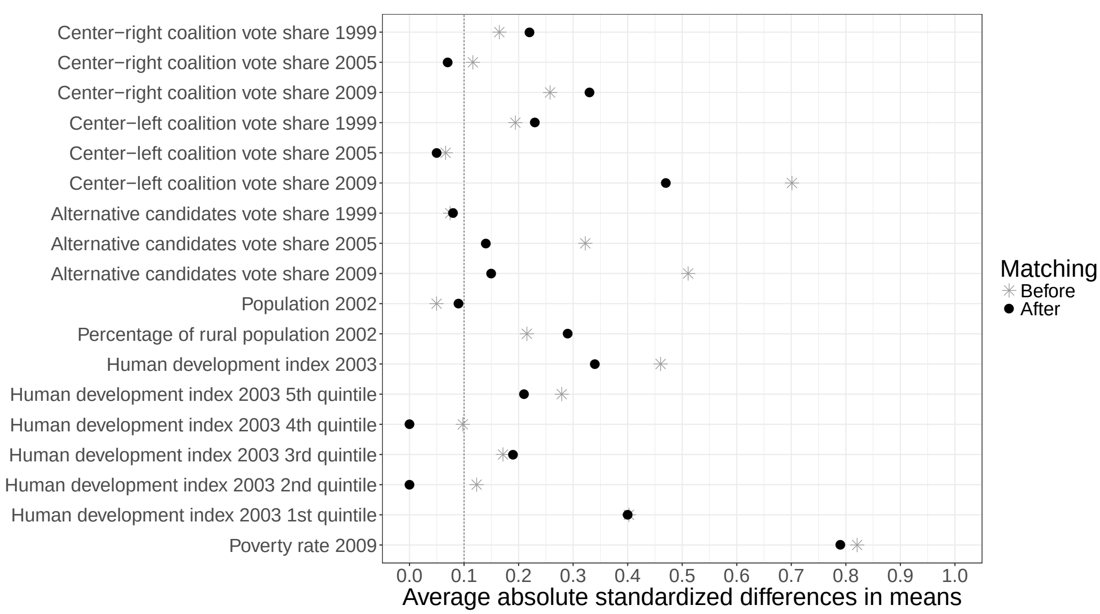


Figure 6: Standardized differences before and after matching on the estimated propensity score without dropping any counties.

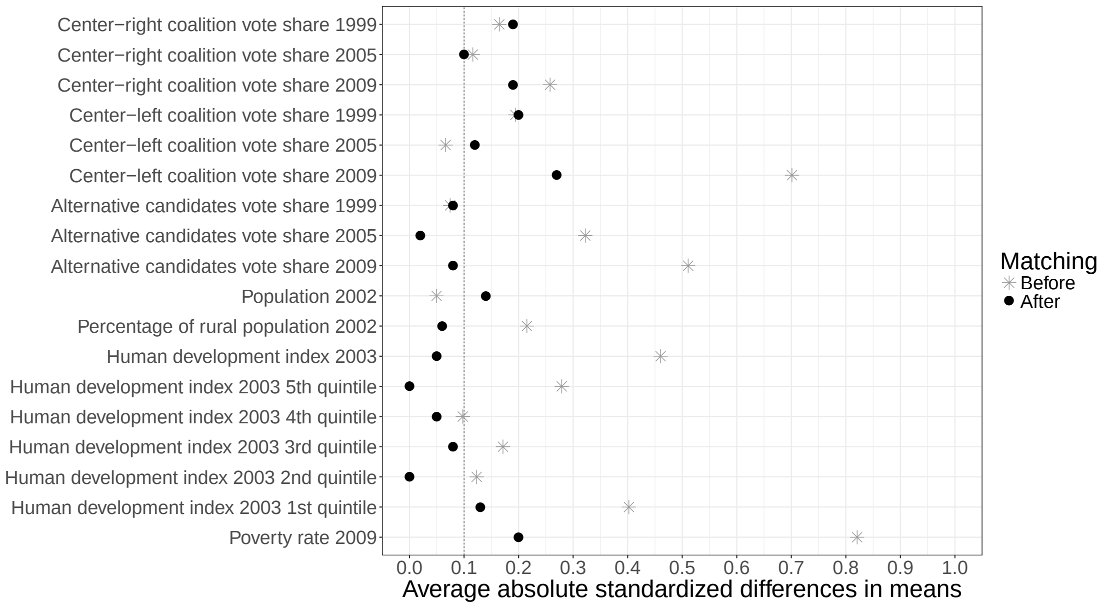


Figure 7: Standardized differences before and after matching on the estimated propensity score after dropping those exposed (control) counties without a counterpart in the control (exposed) group.

For CEM we used the package `cem` in `R`. We first used the default coarsening, but obtained no matched pairs. We then relaxed the coarsening to 9, 8, 7, 6, and 5 bins per covariate, but we still obtained no matched pairs. With 4 bins, we obtained 6 matched pairs, and with 3 and 2 bins we obtained 10 and 30 pairs, respectively. In Figure 8 we show balance for the 30-pair match. Though the results are better than those using PSM, 5 out of 18 covariates still have substantial imbalances.

By contrast, in our matched sample we obtained 59 matched pairs of counties and all the covariates were balanced not only for their means but in some cases also for their distributions. Ideally we would have matched exactly for every covariate, but this is either an infeasible or very difficult constraint in our case study. In other studies with fewer covariates, exact matching may be possible. This illustrates the flexibility of the covariate balancing constraints in cardinality matching, ranging from exact matching to mean balance and including various forms of aggregate distributional balance.

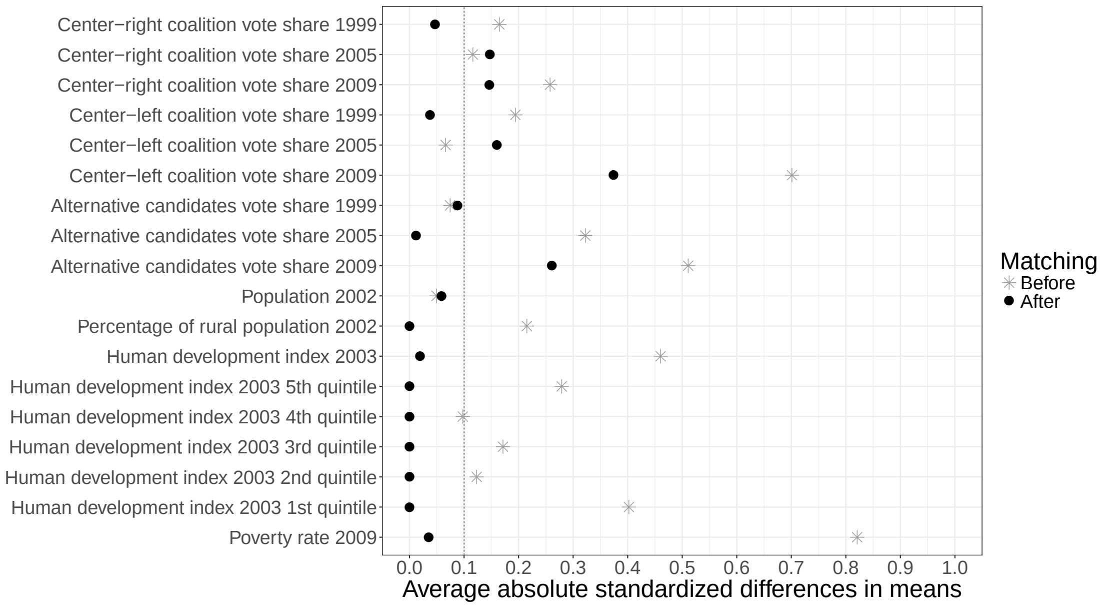


Figure 8: Standardized differences before and after CEM.

7 Summary and remarks

In any observational study, even when observed covariates are precisely balanced, the association between the treatment and the outcome can be ambiguous due to unobserved covariates (Sekhon and Titiunik 2012). However, there are study design strategies that can make the results of an observational study less sensitive to such unobserved covariates (Rosenbaum 2011). In our study, we exploit four strategies. First, we use pretreatment measures of the outcome — in particular, the electoral results for the previous 1999, 2005, and 2009 presidential elections — to compare counties with similar voting patterns before the 2010 earthquake (Rosenbaum 2015a). Second, we reduce pairwise unit heterogeneity (and, therefore, sensitivity to hidden biases; Rosenbaum 2005) by leveraging Chile's low levels of electoral volatility and rematching on previous measures of parties' vote shares. Third, we use very different exposures to the treatment and exclude counties that were partially affected by the earthquake in order to improve the design sensitivity (Rosenbaum 2004). Fourth, we use natural variation in the treatment to study its association with the outcome (Freedman 2006). In our study, all the regions in Chile had been affected by earthquakes in the past, and since these catastrophes cannot be predicted or avoided, the treatment has a haphazard nature.

In addition, we conducted a sensitivity analysis to quantify the association that an unobserved covariate would need to have to explain away the main conclusions of our study. The results of this analysis show that our findings are insensitive to small and medium biases due to an unobserved covariate. Furthermore, we conducted a falsification test to rule out biases due to unobserved covariates that remain constant in time (Keele 2015). To do so, we exploited the fact that the same center-left candidate ran for president both before and after the earthquake. As expected, we found no effect in the pre-earthquake elections which strengthens the evidence of our findings (see Appendix A for details).

In our study, we use cardinality matching to find the largest matched sample of exposed and unexposed counties that is balanced by design. Cardinality matching addresses the problem of limited overlap in covariate distributions using the original covariates, instead of a summary of them, and balances covariates both directly and flexibly. From a statistical standpoint, under a homoskedastic constant additive treatment effect outcome model, cardinality matching minimizes the variance of a difference-in-means effect estimator that is approximately unbiased according to the covariate balancing constraints of the matching problem. From a computational standpoint, cardinality matching takes advantage of recent developments in mathematical programming to solve relatively large matching problems either exactly or approximately. Finally, as illustrated in Section 5, after finding the largest matched sample that is balanced by design, the matched sample can be rematched to reduce heterogeneity in matched groups and thus reduce sensitivity to unobserved covariates. Cardinality matching, as well as other matching methods for instrumental variables, discontinuity designs, and design of experiments, is implemented in the new statistical package `designmatch` for `R` (Zubizarreta et al. 2017).

We apply this method to estimate the effects of the 2010 earthquake in Chile in electoral outcomes. We provide evidence that voters rewarded the incumbent candidate after the 2010 earthquake in Chile. Contrary to the expectations developed by Achen and Bartels (2016), disaster victims do not necessarily blame the incumbent for events that diminish their living conditions. Our findings are in line with studies that show that citizens can either punish or reward incumbents based on their performance handling the consequences of a natural catastrophe (Healy and Malhotra 2010; Gasper and Reeves 2011).

References

Abadie, A., Drukker, D., Herr, J. L., Imbens, G. W., et al. (2004), "Implementing Matching Estimators for Average Treatment Effects in Stata," *Stata journal*, 4, 290–311.

Achen, C. H. and Bartels, L. (2004), "Blind Retrospection. Electoral Responses to Drought, Flu, and Shark Attacks," Working Paper Center for Advanced Study in the Social Sciences.

Achen, C. H. and Bartels, L. M. (2016), *Democracy for Realists: Why Elections Do Not Produce Responsive Government*, Princeton University Press.

Altman, D. (2004), "Redrawing the Chilean Electoral Map: The Presence of Socioeconomic and Gender Factors in the Ballot Box," *Revista de Ciencia Politica*, 24, 49.

Arana, I. (2016), "Aftershocks of Pinochet's Constitution The Chilean Postearthquake Reconstruction," *Latin American Perspectives*, 0094582X16637145.

Armitage, P. (1982), "The Role of Randomization in Clinical Trials," *Statistics in medicine*, 1, 345–352.

Bechtel, M. M. and Hainmueller, J. (2011), "How Lasting Is Voter Gratitude? An Analysis of the Short-and Long-Term Electoral Returns to Beneficial Policy," *American Journal of Political Science*, 55, 852–868.

Buis, A. (2010), "Chilean Quake May Have Shortened Earth Days," NASA.

Calvo, E. and Murillo, M. V. (2012), "When Parties Meet Voters Assessing Political Linkages Through Partisan Networks and Distributive Expectations in Argentina and Chile," *Comparative Political Studies*, 46(7), 851–882.

Carlin, R. E., Love, G. J., and Zechmeister, E. J. (2014), "Natural Disaster and Democratic Legitimacy The Public Opinion Consequences of Chile's 2010 Earthquake and Tsunami," *Political Research Quarterly*, 67, 3–15.

Chen, J. (2013), "Voter Partisanship and the Effect of Distributive Spending on Political Participation," *American Journal of Political Science*, 57, 200–217.

Choi, C. (2012), "Chile Quake and Tsunami Dramatically Altered Ecosystems," Live Science.

Cochran, W. G. (1965), "The planning of observational studies of human populations," *Journal of the Royal Statistical Society. Series A (General)*, 128, 234–266.

Cochran, W. G. and Rubin, D. B. (1973), "Controlling bias in observational studies: A review," *Sankhyā: The Indian Journal of Statistics, Series A*, 417–446.

Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. (2009), "Dealing with Limited Overlap in Estimation of Average Treatment Effects," *Biometrika*, 96, 187–199.

Dehejia, R. H. and Wahba, S. (1999), "Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs," *Journal of the American statistical Association*, 94, 1053–1062.

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.

Dorn, H. F. (1953), "Philosophy of Inferences from Retrospective Studies," *American Journal of Public Health and the Nations Health*, 43, 677–683.

Freedman, D. A. (2006), "Statistical Models for Causation What Inferential Leverage Do They Provide?" *Evaluation review*, 30, 691–713.

Gasper, J. T. and Reeves, A. (2011), "Make it Rain? Retrospection and the Attentive Electorate in the Context of Natural Disasters," *American Journal of Political Science*, 55, 340–355.

González, J. I. (1999), "Geografía Electoral de Chile: Comportamiento del Electorado Chileno entre 1932-1992," *Estudios Geográficos*, 60, 121–138.

Government of Chile (2010), "Plan de Reconstruccion: Terremoto y Maremoto del 27 de Febrero 2010," Resumen Ejecutivo Ministerio de Planificacion.

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

Hansen, B. B. (2007), "Flexible, optimal matching for observational studies," *R News*, 7, 18–24.

Haviland, A., Nagin, D. S., and Rosenbaum, P. R. (2007), "Combining propensity score matching and group-based trajectory analysis in an observational study." *Psychological methods*, 12, 247.

Haviland, A., Nagin, D. S., Rosenbaum, P. R., and Tremblay, R. E. (2008), "Combining Group-Based Trajectory Modeling and Propensity Score Matching for Causal Inferences in Nonexperimental Longitudinal Data." *Developmental psychology*, 44, 422.

Healy, A. and Malhotra, N. (2010), "Random Events, Economic Losses, and Retrospective Voting: Implications for Democratic Competence," *International Quarterly Journal of Political Science*, 5, 193–208.

Hinrichs, R., Jones, L., and Stanley, E. (2011), “Report on the 2010 Chilean Earthquake and Tsunami Response,” U.S. Geological Survey.

Ho, D. E., Imai, K., King, G., and Stuart, E. A. (2007), “Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference,” *Political Analysis*, 15, 199–236.

Hsu, J. Y., Zubizarreta, J. R., Small, D. S., and Rosenbaum, P. R. (2015), “Strong control of the familywise error rate in observational studies that discover effect modification by exploratory methods,” *Biometrika*, 102, 767–782.

Iacus, S. M., King, G., and Porro, G. (2011), “Causal Inference without Balance Checking: Coarsened Exact Matching,” *Political analysis*, mpr013.

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

Imbens, G. W. (2015), “Matching methods in practice: Three examples,” *Journal of Human Resources*, 50, 373–419.

Keele, L. (2015), “The Statistics of Causal Inference: A View from Political Methodology,” *Political Analysis*.

Kilcioglu, C. and Zubizarreta, J. R. (2016), “Maximizing the Information Content of a Balanced Matched Sample in a Study of the Economic Performance of Green Buildings,” *Annals of Applied Statistics*.

King, G., Lucas, C., and Nielsen, R. (2016), “The Balance-Sample Size Frontier in Matching Methods for Causal Inference,” *American Journal of Political Science*.

López, M. Á. (2004), “Conducta Electoral y Estratos Económicos: el Voto de los Sectores Populares en Chile,” *Política*, 43, 285–298.

Luna, J. P. (2010), "Segmented Party–Voter Linkages in Latin America: The case of the UDI," *Journal of Latin American Studies*, 42, 325–356.

McClean, D. (2012), "Chile Still Living with Quake Effects," The United Nations Office for Disaster Risk Reduction.

Navia, P., Izquierdo, J. M., and Morales, M. (2008), "Voto Cruzado en Chile:? Por qué Bachelet Obtuvo Menos Votos que la Concertación en 2005?" *Política y Gobierno*, 15, 35–73.

Roberts, K. M. (2013), "Market Reform, Programmatic (De)Alignment, and Party System Stability in Latin America," *Comparative Political Studies*, 46, 1422–1452.

Rosenbaum, P. and Silber, J. (2012), "Amplification of Sensitivity Analysis in Observational Studies," *Journal of the American Statistical Association*.

Rosenbaum, P. R. (1989), "Optimal matching for observational studies," *Journal of the American Statistical Association*, 84, 1024–1032.

— (2002), *Observational Studies*, Springer.

— (2004), "Design Sensitivity in Observational Studies," *Biometrika*, 91, 153–164.

— (2005), "Heterogeneity and Causality: Unit Heterogeneity and Design Sensitivity in Observational Studies," *The American Statistician*, 59, 147–152.

— (2010), *Design of Observational Studies*, Springer.

— (2011), "What Aspects of the Design of an Observational Study Affect its Sensitivity to Bias from Covariates that Were Not Observed?" in *Looking Back*, Springer, pp. 87–114.

— (2012), "Optimal Matching of an Optimally Chosen Subset in Observational Studies," *Journal of Computational and Graphical Statistics*, 21, 57–71.

— (2015a), "How to See More in Observational Studies: Some New Quasi-Experimental Devices," *Annual Review of Statistics and Its Application*, 2, 21–48.

— (2015b), *Multivariate Matching Methods*, John Wiley and Sons.

Rosenbaum, P. R., Ross, R. N., and Silber, J. H. (2007), "Minimum Distance Matched Sampling with Fine Balance in an Observational Study of Treatment for Ovarian Cancer," *Journal of the American Statistical Association*, 102, 75–83.

Rosenbaum, P. R. and Rubin, D. B. (1985), "Constructing a control group using multivariate matched sampling methods that incorporate the propensity score," *The American Statistician*, 39, 33–38.

Rubin, D. B. (1973), "Matching to remove bias in observational studies," *Biometrics*, 159–183.

— (2006), *Matched Sampling for Causal Effects*, Cambridge University Press.

Sekhon, J. S. and Titiunik, R. (2012), "When Natural Experiments Are Neither Natural Nor Experiments," *American Political Science Review*, 106, 35–57.

Sinclair, B., Hall, T. E., and Alvarez, R. M. (2011), "Flooding the vote: Hurricane Katrina and voter participation in New Orleans," *American politics research*, 39, 921–957.

Stuart, E. A. (2010), "Matching methods for causal inference: A review and a look forward," *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25, 1.

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.

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 (Cambridge, Mass.)*, 24, 79–87.

Zubizarreta, J. R., Kilcioglu, C., and Vielma, J. P. (2017), "designmatch: Matched Samples that are Balanced and Representative by Design," *R package version 0.3*, 1.

Zubizarreta, J. R., Paredes, R. D., and Rosenbaum, P. R. (2014), "Matching for Balance, Pairing for Heterogeneity in an Observational Study of the Effectiveness of For-profit and Not-for-profit High Schools in Chile," *The Annals of Applied Statistics*, 8, 204–231.