Italian Political Science Review/Rivista Italiana di Scienza Politica page 1 of 16
doi:10.1017/ipo.2021.1

IPSR
RISP

RESEARCH ARTICLE

Looking for twins: how to build better counterfactuals with matching

Stefano Costalli[1] and Fedra Negri[2]*

[1]Department of Political and Social Sciences, Università degli Studi di Firenze, Firenze, Italy and [2]Department of Social and Political Sciences, Università degli Studi di Milano, Milano, Italy
*Corresponding author. Email: fedra.negri@unimi.it

(Received 5 June 2020; revised 4 January 2021; accepted 7 January 2021)

Abstract

A primary challenge for researchers that make use of observational data is selection bias (i.e. the units of analysis exhibit systematic differences and dis-homogeneities due to non-random selection into treatment). This article encourages researchers in acknowledging this problem and discusses how and – more importantly – under which assumptions they may resort to statistical matching techniques to reduce the imbalance in the empirical distribution of pre-treatment observable variables between the treatment and control groups. With the aim of providing a practical guidance, the article engages with the evaluation of the effectiveness of peacekeeping missions in the case of the Bosnian civil war, a research topic in which selection bias is a structural feature of the observational data researchers have to use, and shows how to apply the Coarsened Exact Matching (CEM), the most widely used matching algorithm in the fields of Political Science and International Relations.

Key words: causation; coarsened exact matching; peacekeeping; selection bias; statistical matching

Introduction

Observational data often create challenges for social and political scientists willing to detect causation due to the problem of selection bias. Indeed, before interpreting the coefficient of a multivariate regression model as the causal effect of the variable of interest on the outcome, researchers need to check whether their units of analysis exhibit systematic differences and dis-homogeneities able to affect both the variable of interest and, controlling for it, the outcome.

Even recognizing that detecting causation will be always hazardous, we maintain that researchers that make use of observational data cannot give up working on mindful methods to establish causal links. Thus, this article encourages researchers in acknowledging the problem of selection bias and endorses the practice of preprocessing the raw data through statistical matching techniques as a partial solution. Indeed, such techniques are helpful to assess whether systematic differences on observables dimensions between groups exist in the raw data and, if any, to eliminate, or at least reduce, them by generating a well-balanced sample on which to use the same familiar method of estimation they would have used anyway on the raw data without preprocessing.

However, we add the adjective 'partial' as matching techniques, exactly as regression, are grounded on the strong assumption of selection on observables (see the subsection 'The inferential logic behind matching' and footnote 1), whose credibility should be discussed on a case-by-case basis. Once acknowledged this baseline assumption, resorting to matching techniques eliminates, or at least reduces, the selection bias due to the set of variables chosen by the researcher, which in turn makes the causal effect estimate based on the subsequent parametric

analysis more credible and far less sensitive to modeling choices and specifications. Quoting Ho *et al.* (2007: 233): 'Analysts using preprocessing have two chances to get their analyses right, in that if either the matching procedure or the subsequent parametric analysis is specified correctly (and even if one of the two is incorrectly specified), causal estimates will still be consistent'.

The performance of matching estimators in detecting causal effects is a topic with a long history, especially in the field of policy evaluation. The seminal study is LaLonde (1986), which firstly assessed the performance of several non-experimental estimators (among which, matching estimators) by using as benchmark the experimental result of the National Supported Work Demonstration (NSWD), a subsidized work experience program that took place in 1975–1976 in the United States. According to the experimental result, the program was successful as it was estimated to increase post-intervention earnings by $1794. LaLonde compared this experimental result to those obtained from several non-experimental estimators applied to the NSWD individuals that received training and a set of control individuals identified *ex post* from two standard population survey datasets. He concluded that non-experimental estimators were unable to replicate the experimental result. Then, Dehejia and Wahba (1999) used a partition of the LaLonde's dataset to compare the performance of matching estimators to that of a fully saturated in X OLS regression. They concluded that matching estimators dominated a fully saturated in X OLS regression (i.e. the OLS estimate was substantially lower than the experimental benchmark and not statistically significant). Later, Smith and Todd (2005) added that a combination of matching and a subsequent parametric estimator performed better than both a fully saturated in X OLS regression and matching estimators alone. Last, Iacus *et al.* (2019) showed that the result obtained by preprocessing the original LaLonde's dataset with the Coarsened Exact Matching (CEM) and then running an OLS regression on the preprocessed data was closer to the experimental benchmark than a fully saturated in X OLS regression (and outperformed other propensity score-matching algorithms).

This long debate among methodologists in the field of policy evaluation to motivate the usefulness of matching techniques – and to compare their relative performance – exploited an extremely rare case in which the unbiased causal effect of the training on post-intervention earnings was known from a randomized experiment. This is not what usually happens in practice. Given the topics of interest of Political Science and International Relations, far more common are the cases where the variable of interest, be it political parties' location on the left-right ideological spectrum, individuals' values and ideas, or the deployment of troops in a given area, cannot be randomized across units. Moreover, although experiments are sometimes depicted as the 'gold standard' for providing internally valid estimates of causal effects (Duflo *et al.*, 2008), they are not without shortcomings concerning statistical validity (i.e. the ability to produce precise estimates of very small effects) (Young, 2019) and, crucially, external validity (i.e. the explanatory power of a particular causal estimate for times, places, and people beyond those represented in the study that produced it) (Bates and Glennerster, 2017).

Matching techniques can be very useful for social and political researchers that make use of observational data and that have to face several decisions during the implementation of their analyses. For example, Ho *et al.* (2007) replicated Carpenter's (2002) analysis on the causal effect of a Democratic majority in the US Senate on the approval time for a new drug by the Food and Drug Administration (FDA). By using different log-normal survival models, Carpenter (2002) found that a Democratic Senate majority tended to decrease the average approval time of new drugs. However, he was unable to draw solid conclusions as the coefficient was unstable: it switched sign and even lost statistical significance across specifications. Ho *et al.* (2007) replicated this analysis by preprocessing the raw dataset through several matching estimators to obtain a well-balanced sample on key dimensions referring to the diseases drugs are devoted to. Then, they applied the same log-normal survival analysis on the preprocessed dataset and found that a Democratic Senate majority significantly decreases the average approval time of new drugs across model specifications.

These examples taken from the field of policy evaluation and that of partisan determinants of regulatory policy show how the use of matching techniques can be fruitfully combined with more familiar estimators. Indeed, the most desirable feature of matching techniques is that they help researchers to think about selection bias and to evaluate whether they are meeting the necessary conditions for generating valid and reliable results in their analyses or how far they go.

The article is structured as follows. The section that follows puts the reader in front of an empirical research scenario in which randomization is clearly unfeasible but, at the same time, obtaining valid and reliable results is extremely important. In detail, this section engages with the evaluation of the effectiveness of peacekeeping missions by focusing on the case of the Bosnian civil war (Costalli, 2014), a research topic in which selection bias is a structural feature of the observational data researchers have to use. Next, matching techniques are introduced as smart statistical tools able to partially overcome this problem. Then, the article zooms on the CEM algorithm (Iacus *et al.*, 2009, 2019), the latest innovation among matching techniques and the most widely used in the fields of Political Science and International Relations, by showing its implementation steps to evaluate the effectiveness of peacekeeping missions in the Bosnian civil war. The last section hosts concluding remarks.

Assessing the effectiveness of peacekeeping missions: policy relevance and empirical problems

Civil wars have been the prevalent form of armed conflict for decades (Gleditsch *et al.*, 2002), causing millions of deaths (Petterson *et al.*, 2019) and proving to be extremely difficult to end, particularly through negotiations (Walter, 1997). Especially after the end of the Cold War, the international community has placed high expectations on peacekeeping missions as one of the few tools that could effectively stop ongoing civil wars and stabilize conflict-ridden countries after the end of armed combat. Considering these expectations, as well as the political, financial, and organizational costs of peacekeeping missions, it is essential to carefully evaluate their effectiveness with rigorous social scientific methods. Receiving precise and reliable feedback is in fact fundamental for international organizations such as the United Nations (UN) to adjust their peacekeeping policies in terms of mandate, resources, and composition of the troops.

A first wave of empirical studies assessed the effectiveness of peacekeeping missions using cross-national research designs (e.g. Doyle and Sambanis, 2006). However, important contributions demonstrated that civil wars are best understood by looking at local-level issues and variation (e.g. Cederman and Gleditsch, 2009). Indeed, the causes of such conflicts are linked to discriminations and inequalities between specific groups, as well as to local features of the physical and political geography (e.g. Cederman *et al.*, 2013). Moreover, the use of violence can change in a few kilometers' range, depending on the degree of control exercised by the warring factions on specific areas, the organizational features of groups, or the allegiances of the local population (e.g. Kalyvas, 2006; Fjelde and Hultman, 2014; Costalli *et al.*, 2020). Thus, since the situation on the ground can vary at a short distance and quickly, a disaggregated approach is equally essential to investigate the effectiveness of peacekeeping missions in civil wars.

One of the first studies to use spatially disaggregated data to evaluate peacekeeping effectiveness is Costalli (2014), which deals with the long-debated issue of the UN mission in the Bosnian civil war (1992–1995). Indeed, the mission in Bosnia and Herzegovina during the ethnic conflict that plagued the country in the 1990s represents one of the first peacekeeping efforts promoted by the UN after the end of the Cold War and its effectiveness has been at the center of long academic and policy debates.

A crucial problem for all studies willing to assess the effectiveness of peacekeeping missions is that neither peacekeeping missions in a cross-country setting, nor peacekeeping contingents on the ground in spatially disaggregated analyses are randomly allocated, thus causing serious problems of selection bias. In fact, previous research found that UN intervention is more likely in

difficult cases, where the chances of securing peace are low and the risk of renewed conflict higher (Fortna, 2008). The conditions of the ongoing conflict, such as its overall severity or episodes of violence against civilians, as well as organizational and political issues, are likely to have a huge influence on the process of troops' deployment in different areas of a country. Thus, any study on the effectiveness of peacekeeping has to acknowledge the problem of selection bias and address it.

In order to evaluate the effectiveness of the United Nations Protection Force (UNPROFOR) mission in reducing the severity of violence during the Bosnian civil war, Costalli (2014) focuses on the municipal level and operationalizes the severity of violence (the outcome variable) as the logged number of deaths recorded in each municipality. The intervention of peacekeeping forces on the ground (the variable of interest) is operationalized in a twofold way. First, it is operationalized through a dummy variable taking the value of 1 when peacekeeping troops are present in a given municipality and 0 otherwise. Second, it is operationalized through a dummy variable that distinguishes cases in which UN troops are actively involved in the conflict (1) from cases in which they perform pure monitoring activities (0).

To acquire full knowledge of the likely selection bias, and thus to verify whether municipalities with peacekeeping troops exhibit systematic differences from those without peacekeeping troops before the peacekeeping action, Costalli (2014) investigates the determinants of the location of peacekeepers on the ground using a set of logistic regressions. In line with previous cross-national studies (Fortna, 2008), he finds that UNPROFOR units are not randomly deployed across municipalities. Instead, they are more likely to be deployed in the most violent ones, thus empirically confirming that selection bias is a serious risk in his analysis.

Crudely put, given that troops tend to be deployed in the municipalities in which the civil war shows the worst of its brutality, a naïve comparison between the severity of violence recorded in the municipalities with peacekeeping troops and in those without could understate the effectiveness of the peacekeeping action, failing to recognize that the possible scarce effectiveness of UNPROFOR is actually a consequence of the fact that the troops operate in areas that are *ex ante* more violent than the average. Thus, the ideal empirical strategy to properly assess UNPROFOR's effectiveness would be to compare pairs of municipalities with similar observable features (above all, similar severity of violence) before the arrival of peacekeeping troops, where one received the troops and the other did not (see footnote 1 for differences between matching and regression with properly identified control variables). This is the basic inferential logic behind statistical matching techniques. As will be described step by step in the section 'Matching for political scientists', Costalli (2014) employs CEM (Iacus *et al.*, 2009, 2019) to match Bosnian municipalities on two sets of pre-peacekeeping observable variables, thus reducing selection bias by creating a more balanced and comparable subsample of units where to study peacekeeping effectiveness.

What is matching? And why should we use it?

Before moving to the CEM application, we introduce the large family of matching techniques and motivate their usefulness as tools able to downplay the serious problem of selection bias by discarding heterogeneous units from the raw data so that inference is limited to a carefully select subsample. This section is structured as follows. First, we briefly discuss selection bias in terms of potential outcomes notation (Rubin, 1974; Holland, 1986). Then, we describe the inferential logic behind the exact matching technique, which is the ancestor of matching tools, and then we present how such logic can be generalized to approximate matching techniques, thus paving the way to the section 'Matching for political scientists', which is devoted to CEM and its application.

Selection bias as an error in the causal reasoning

The previous section highlighted that the naïve comparison between the severity of violence observed in the municipalities with peacekeeping troops and in those without is likely to be a

biased estimate of the peacekeeping effectiveness due to selection bias. The following equation conveys this message in terms of potential outcomes notation (Rubin, 1974; Holland, 1986):

$$E(Y_i|D_i = 1) - E(Y_i|D_i = 0) = E(Y_i(1) - Y_i(0)|D_i = 1)$$

$$+ E(Y_i(0)|D_i = 1) - E(Y_i(0)|D_i = 0) \quad (1)$$

For the sake of simplicity, let us pretend D_i to be the first dummy variable used by Costalli (2014) to operationalize the intervention of peacekeeping forces in a given municipality. The left-hand side of the equation displays the naïve difference in the average factual outcomes by treatment group: $E(Y_i|D_i = 1)$ is the average severity of violence observed in the municipalities with peacekeeping troops; $E(Y_i|D_i = 0)$ is the average severity of violence observed in those without. Equation 1 clarifies that such a naïve difference is equal to the sum of two quantities, displayed on the right-hand side of the same equation. The first quantity $E(Y_i(1) - Y_i(0)|D_i = 1)$ refers only to the municipalities with peacekeeping troops (as it is conditioned on $D_i = 1$) and corresponds to the Average Treatment Effect for the Treated (ATT). The ATT measures the difference between the average severity of violence observed in the municipalities with peacekeeping troops $E(Y_i(1)|D_i = 1)$ and the average severity of violence that would have been observed in the same municipalities if the peacekeeping operation had not taken place $E(Y_i(0)|D_i = 1)$. Unfortunately, $E(Y_i(0)|D_i = 1)$ is a counterfactual quantity and, as such, cannot be observed. The second quantity on the right-hand side of Equation (1) corresponds to the selection bias and it is equal to the difference between the average severity of violence that would have been observed in the municipalities with peacekeeping troops if the peacekeeping operation had not taken place $E(Y_i(0)|D_i = 1)$, the counterfactual quantity, and the average severity of violence observed in the municipalities without peacekeeping troops $E(Y_i(0)|D_i = 0)$.

This algebraic representation, which dates back to Rubin (1974) and Holland (1986), is 'a fundamental building block in modern research on causal effects' (Angrist and Pischke, 2009: 11) as it makes clear that selection bias is an error in the causal reasoning. Selection bias is different from 0 when, in the absence of peacekeeping troops on the ground, the municipalities that actually received them (with $D_i = 1$) would have recorded systematically different levels of severity of violence from the municipalities without peacekeeping troops (with $D_i = 0$). In Costalli's example (2014), such municipalities would have recorded higher levels of severity of violence than the municipalities without peacekeeping troops in any case.

The message conveyed by Equation (1) is that the naïve comparison of the average factual outcomes by treatment group (i.e. the left-hand side of Equation 1) has no causal interpretation, it is a biased estimate of the ATT (i.e. the first quantity on the right-hand side of Equation 1), unless the selection bias (i.e. the second quantity on the right-hand side of Equation 1) is 0. Thus, Equation (1) is helpful as it reframes the quest for causal effects as a discussion on the existence of selection bias. Intuitively, it suggests that researchers that make use of observational data need tools to make such a biased comparison as close as possible to the ATT by carefully assessing whether systematic compositional differences between units exist and, if this is the case, to adjust for them. This is when matching techniques come into play.

The inferential logic behind matching and its implementation steps

The inferential logic behind the exact matching technique, which is the ancestor of matching tools, mimics the one of randomized experiments with one crucial difference. In randomized experiments, units are assigned to treatment and control groups *ex ante* (i.e. before the intervention) through randomization. Random assignment ensures that treatment and control units are the same in every respect, thus making selection bias equal to 0 by construction. Instead, in the case of the exact matching, the control group is created *ex post* (i.e. after the intervention) by

preprocessing the raw data so that each treated unit is matched with all the available control units having exactly the same values on a set of pre-treatment observable variables X_i carefully selected by the researcher (Arceneaux *et al.*, 2006). In Costalli's case (2014), for example, the use of the exact matching would have guaranteed the equivalence between the two groups composed by municipalities with and without peacekeeping troops only with regard to the two sets of pre-peacekeeping variables chosen by the researcher. We do not know whether the municipalities in the two groups would have differed in terms of other important features (Corbetta, 2003: 102).

It follows that the selection of such a set of variables X_i is a critical step and has to be accomplished through a thought exercise, according to the researcher's previous knowledge. First, to avoid the omitted variable bias, X_i should include all variables that affect both the treatment assignment D_i and, controlling for it, the dependent variable Y_i. Second, to avoid the post-treatment bias (King and Zeng, 2006), variables that may be even remotely consequences of the treatment variable should never be included in X_i (Cox, 1958, section 4.2; Rosenbaum, 1984; Rosenbaum, 2002: 73–74). Back to our example, X_i should include all the variables affecting both the location of peacekeepers on the ground and, controlling for it, the observed severity of violence. Moreover, such variables should not be themselves even remotely caused by the peacekeeping mission (see also Angrist and Pischke, 2009: 50–51).

Once the set of variables X_i has been selected, exactly as when we interpret the coefficient of a multivariate regression model as a causal effect (see Angrist and Pischke, 2009: 51–57 for a discussion of the equivalencies between regression and matching), the inferential logic behind exact matching (and behind any matching tool) builds on the selection on observable assumption (Barnow *et al.*, 1980; Heckman and Robb, 1985 print publication - 2013 online publication). This assumption has a number of different names: alternatives are 'no omitted variable bias', 'conditional ignorability', 'absence of unmeasured confounding', 'unconfoundedness' or 'conditional independence assumption' (Goldberger, 1991). Whatever its name, this assumption means that, in order to attach a causal meaning to the obtained estimates, it should be theoretically plausible that selection into treatment is completely determined by the set of variables X_i selected by the researcher, such that, conditioning on X_i, assignment to treatment is as good as random. To put it differently: it should be theoretically plausible that there are not additional variables able to affect municipalities' likelihood to receive peacekeeping troops.[1]

The selection on observables assumption is non-refutable because it cannot be verified with data (Manski, 2007). As formalized in Equation (2), under this assumption and conditioning on X_i, the average severity of violence observed in the municipalities without peacekeeping troops $(Y_i(0)|D_i = 0, X_i)$ is equal to the average severity of violence that would have been observed in the municipalities with peacekeeping troops if the peacekeeping operation had not taken place

[1]Given that both matching and regression are based on the selection on observables assumption, the reader may wonder whether matching is really different from a regression with properly identified control variables. This question is the object of a heated debate among methodologists. Some maintained that both regression and matching are control strategies and therefore the differences between the two are unlikely to be of major empirical importance (Angrist and Pischke, 2009, section 3.3.1). Others pointed out shortcomings of regression relative to matching: Dehejia and Wahba (1999), for example, found that propensity score-matching procedures more closely approximate results from a randomized experiment than regression alone. Further on this, some underlined that regression is a parametric approach imposing a global linear relationship between Xs and Y and that it uses all the available observations, thus involving a certain amount of extrapolation, while matching is a non-parametric approach that discards observations for which a reasonably close match cannot be found (Martini and Sisti, 2009: 221–225). Others replied that matching involves several choices in its implementation, which could lead to subjectivity of the results. According to Imbens and Wooldridge, 'the best practice is to combine linear regression with either propensity score or matching methods' (2008, 19–20) as in this way the estimated effect will explicitly rely on local, rather than global, linear approximations to the regression function. Even though adjudicating between these views is beyond the scope of this article, the application here discussed embraces this last suggestion and thus combines the CEM algorithm with OLS regression.

$E(Y_i(0)|D_i = 1, X_i)$, the counterfactual quantity of Equation (1), and can be used to estimate the ATT.

$$E(Y_i(0)|D_i = 0,\ X_i) = E(Y_i(0)|D_i = 1,\ X_i) = E(Y_i(0)|\ X_i) \tag{2}$$

Beware that, if some treated units cannot be matched because there is not at least one control unit having exactly the same values on X_i, the exact matching technique drops these treated units. By dropping some treated units, the exact matching technique alters the *estimand*: it is no more the ATT, but a more local version of it (Crump *et al.*, 2009; Rubin, 2010). This choice is reasonable as long as the researcher is transparent about it and its consequences in terms of the new set of treated units over which the causal effect is defined (Iacus *et al.*, 2012: 5). If a large number of treated units are exactly matched with one or more control units, the method of estimation of the ATT can credibly be a simple (weighted) difference between the average outcomes of matched treated and control units.[2] Instead, if an insufficient number of exact matches are found, and thus many treated units have to be discarded, the researcher has to switch to an approximate matching technique that preprocesses the raw data so that each treated unit is matched with all the available control units having 'approximately' the same values on X_i.

The most widely used approximate matching techniques involve three implementation steps. As for the exact matching procedure, the first step asks researchers to establish on which set of pre-treatment variables X_i the degree of closeness between treated and control units has to be evaluated. As anticipated, selecting these dimensions is not an easy task as researchers might be tempted to include several pre-treatment variables. This problem is known as 'the curse of dimensionality' and it has been tackled by Rosenbaum and Rubin (1983). In detail, building on the usual selection on observables assumption, these authors demonstrate that, if potential outcomes are independent of treatment status conditional on the set of variables X_i (see Equation 2), then potential outcomes are also independent of treatment status conditional on a scalar function of the same variables X_i, labelled 'propensity score'. Intuitively, the propensity score is a mono-dimensional variable that measures, for each unit i, the probability of receiving treatment given the values of its set of variables X_i, that is: $P(D_i = 1|X_i)$. Usually, the propensity score is estimated through a logit or a probit function, that regresses D_i on a constant term and the set of variables X_i, without looking at Y_i.

After having estimated the propensity score for each unit i, approximate matching techniques check whether the so-called 'common support assumption' is fulfilled. Such assumption requires that, for any treated unit with given values on X_i, it is also possible to observe a control unit with approximately the same values. To ensuring such common support, the matching algorithms usually drop control units that have a propensity score lower than the minimum or higher than the maximum of the propensity score of the treated units (Khandker *et al.*, 2010).

The second implementation step asks researchers to match treated and control units according to a given metric. Several metrics are available: they vary as for the strategy they follow to select the matches and as for the weight they associate with each match. Table 1 summarizes the commonest approximate matching techniques based on the propensity score and provides references for further readings (see also: Caliendo and Kopeinig, 2008).

Given this non-exhaustive list, how to choose among such approximate matching techniques? The methodological literature suggests only a rule of thumb. Since the main diagnostics of success are balance as well as the number of observations remaining after preprocessing the raw data, researchers have to run as many approximate matching techniques as possible without consulting Y_i. Then, they have to choose the technique that maximizes balance while keeping n as large as

[2]We add 'weighted' in parentheses because, since each treated unit can be matched with more than one control unit, a weighted difference in means across exactly matched subclasses is suggested to account for the difference in the number of treated and control units.

Table 1. The commonest approximate matching techniques based on the propensity score

Technique	Description	Further readings
Nearest neighbor matching	For each treated unit, the algorithm finds the control unit with the nearest propensity score. This can be done with or without replacement. In the former case, an untreated unit can be used more than once as a match. In the latter case, if the nearest control unit has already been matched to another treated unit, the algorithm does not consider it and searches for a new one.	Smith (1997), Smith and Todd (2005)
Caliper and radius matching	For each treated unit, the caliper matching algorithm finds the closest control unit whose propensity score falls within a radius *r* chosen by the researcher. The radius version matches each treated unit with all the control units within the radius *r*.	Smith and Todd (2005), Dehejia and Wahba (2002)
Stratification matching	The algorithm partitions the sample into a set of intervals (strata) so that, in each stratum, the propensity score of treated and control units has the same mean value.	Imbens (2004)
Kernel matching	The algorithm matches every treated unit with a weighted average of (nearly) all control units with weights that are inversely proportional to the distance between the propensity scores.	Heckman *et al*. (1997), (1998)

possible (Ho *et al*., 2007). This search may be tedious and highly time-consuming, as researchers have to manually iterate between the available algorithms (Iacus *et al*., 2009; Heinmueller, 2012; King and Nielsen, 2019).

One might object that increasing balance by throwing away unmatched observations will reduce statistical efficiency (i.e. the mean squared error of the estimated effect might increase). However, 'efficiency should be a secondary concern' for scholars using observational data (Keele, 2015: 325). Indeed, in a randomized experiment, where selection bias is known to be 0, adding observations simply increases power. Rather, in an observational study, increasing the sample size may shrink the confidence intervals to a point that excludes the 'true' treatment effect point estimate (Cochran and Chambers, 1965). Thus, as a rule of thumb, there are reasons for preprocessing raw data through matching techniques even though this may alter the *estimand* by dropping some treated units.

Once the matching algorithm that maximizes balance while keeping n as large as possible is selected, the third and last implementation step asks researchers to move to the usual parametric analysis to estimate the causal effect of their variable of interest on the outcome. As anticipated, whenever the treatment and control groups are not exactly balanced, that is what usually happens, researchers are better off in using the same parametric model they would have used on the raw data without preprocessing. Preprocessing data with matching improve the reliability of the causal effect estimated through the parametric analysis as the latter becomes far less dependent on modeling choices and specifications (Ho *et al*., 2007; Iacus *et al*., 2019).

Matching for political scientists: CEM and the evaluation of peacekeeping effectiveness in Bosnia and Herzegovina

This section focuses on CEM, a matching technique developed by Iacus *et al.* (2009), and constantly refined until 2019, because it displays key advantages.

First, even though approximate matching techniques based on the propensity score have been and still are widely used, the propensity score solution is the object of a heated debate among methodologists (e.g. Becker and Ichino, 2002; King and Nielsen, 2019) as it has been accused of being a tautology. Intuitively, the propensity score was developed because there were too many pre-treatment variables to be controlled for. However, since researchers do not know its 'true' value, the propensity score has to be estimated through a probability model that adds

the same pre-treatment variables on the right-hand side of the equation. Then, to check whether the estimated propensity score is a consistent estimate of the 'true' propensity score, researchers stratify the sample over small propensity score intervals and, for each variable in each interval, test whether the means of the treated and control units are not statistically different. If it is not the case, researchers go back to the specification of the probit or logit function they used to estimate the propensity score and start again (Ho *et al.*, 2007). Instead, CEM does not make use of the propensity score and thus overcomes this debated tautology.

Second, from a practical point of view, as anticipated above, the search for the matching algorithm that maximizes balance while keeping n as large as possible may be tedious and highly time-consuming. CEM overcomes also this practical problem by asking researchers to establish their desired degree of balance before the preprocessing adjustment, thus increasing balance on one variable cannot decrease balance on another (while this can happen with propensity score matching algorithms).

Furthermore, advantages of CEM relative to propensity score matching include greater computational efficiency, ease of implementation, higher flexibility to researchers' needs, less sensitivity to measurement error, and the intuitive appeal of exact matching (Iacus *et al.*, 2011).

The basic inferential logic behind CEM can be summarized by saying that it coarsens each pre-treatment variable into substantively meaningful categories identified *ex ante* by the researcher according to their previous knowledge and then matches the treated and control units exactly on this coarsened scale. Units that cannot be exactly matched are discarded, thus leading CEM to change the *estimand* from the ATT to a more local version of it.

As anticipated in the section 'Assessing the effectiveness of peacekeeping missions', Costalli (2014) investigates the effectiveness of peacekeeping in reducing the severity of violence in the Bosnian civil war by focusing on the local level and using Bosnian municipalities as units of analysis. Reducing military fighting is a key matter for interventions in ongoing civil conflicts and a disaggregated analysis seems crucial because the logic that drives the use of violence is often strictly linked to local factors. According to the expectations of the UN and the overall international community, peacekeeping troops on the ground should be able to reduce violence. For this reason, Costalli (2014) tests the hypothesis that the presence of peacekeepers in a given Bosnian municipality reduces the level of violence recorded in that municipality the following year. The analysis distinguishes instances of active intervention by peacekeepers from simple observation, and expects the former to be more effective in reducing violence than the latter. The severity of violence (the outcome) is expressed by the logged number of deaths recorded in each municipality, while the intervention of peacekeeping forces on the ground is operationalized by two dummy variables of interest. The first one equals 1 when peacekeeping troops are present in a given municipality; the second one detects cases of active involvement of UN troops in the conflict, distinguishing them from pure monitoring activities.

The analysis controls for the major variables that might affect the processes under scrutiny as identified by the empirical literature in the field. To start with, the ethnic dimension could not be excluded from analyses of violence and intervention in the Bosnian civil war, where armed groups mobilized along ethnic lines and the war was fought with symmetric military technologies. Thus, the variable *Ethnic polarization* takes into account the relative size of ethnic groups. Another relevant issue deserving attention when dealing with civil wars is the presence and actions of bordering countries that can intervene directly or indirectly in the conflict (Cederman and Gleditsch, 2009; Gleditsch, 2007). If contiguous states decide to intervene directly with their armed forces, or indirectly supporting one faction, to the conflict, bordering areas could be marked by high levels of violence. Hence, the analysis controls whether the municipalities were on the borders with Serbia and Croatia. Finally, the analysis controls for a set of variables concerning the geographical dimension of the war and the military dynamics of conflict: whether the municipality under scrutiny was on what then became the border between the Federation of Bosnia-Herzegovina and the Republika Srpska, the share of open terrain (i.e. terrain free from

Table 2. Effects of peacekeeping on local violence

	Model 1	Model 2
Presence of peacekeeping t_{-1}	0.202	
	(0.256)	
Active peacekeeping t_{-1}		0.270
		(0.353)
Violence t_{-1}	−0.136	−0.136
	(0.102)	(0.102)
Ethnic polarization	1.507***	1.494***
	(0.526)	(0.511)
Population	1.102***	1.102***
	(0.135)	(0.135)
Contiguity Serbia	0.252*	0.255*
	(0.150)	(0.146)
Contiguity Croatia	−0.590***	−0.588***
	(0.032)	(0.025)
Internal border	0.237	0.232
	(0.219)	(0.221)
Open terrain	0.332	0.346
	(0.608)	(0.604)
Constant	−7.717***	−7.713***
	(1.062)	(1.071)
N	380	380
R^2	0.512	0.513

Note: Time-series cross-sections with lagged dependent variable and panel-corrected standard errors in parentheses. *$P < 0.10$; **$P < 0.05$; ***$P < 0.01$.

dense vegetation), and the population of the municipality calculated by subtracting the number of fatalities that occurred in the previous year.

Table 2 displays two regression models that try to assess whether the level of violence in t0 is affected by the presence and activities of UNPROFOR in t–1 (the year before). Model 1 investigates the influence of the presence of peacekeepers and Model 2 verifies if active intervention provides different results from simple presence. The specifications are time-series cross-sections with lagged dependent variable and panel-corrected standard errors (Beck and Katz, 1995, 1996).

According to Models 1 and 2, the impact of peacekeeping troops on the severity of violence seems to be irrelevant during the war as neither the variable indicating the presence of UN troops, nor the one tracing their active involvement during the previous year reach statistical significance. On the contrary, the level of ethnic polarization seems to drive most of the violence, while municipalities bordering Croatia tend to be relatively less violent than the average and the ones bordering Serbia slightly more violent. Population confirms to be an important control variable when studying the overall level of violence in conflicts that occur amongst the people. However, we cannot forget that according to cross-national empirical studies on peacekeeping, blue helmets tend to go to most violent places (Fortna, 2008) and this might be the cause of having a very unbalanced sample.

For this reason, and as mentioned above, Costalli (2014) investigates the determinants of the location of peacekeepers on the ground through a series of logistic regressions and finds that UNPROFOR units are more likely to be deployed in the most violent municipalities. Thus, in order to address and reduce the likely selection bias, Costalli (2014) uses CEM to obtain a more balanced sample of municipalities where to study the effectiveness of peacekeeping.

As for all the other matching techniques, the preliminary step researchers have to take when using CEM is identifying the set of pre-intervention variables *Xi* they want to match the units in their dataset on. In other words, it is crucial to recognize the observable dimensions on which treated and control units should be similar, by detecting the variables that can affect both the treatment assignment *Di* and, controlling for the treatment, the dependent variable *Yi*. Such

pre-intervention variables are the causes of the selection bias and researchers need a theory to correctly detect them. Only a careful knowledge of the phenomenon under scrutiny and of the data generating process can lead to a proper identification of such a set of variables. One common but serious mistake that could emerge in this phase is matching treated and control units along dimensions that can be possible consequences of the treatment itself. Such procedure would cause a post-treatment bias and it has to be carefully avoided (King and Zeng, 2006). In our example, Costalli (2014) uses CEM to match Bosnian municipalities on two sets of pre-intervention observable variables. In detail, the variables included in the first matching procedure (Coarsening 1) are ethnic polarization, the level of violence recorded in the previous year, the population, the contiguity with Croatia, and the contiguity with Serbia. The second matching procedure (Coarsening 2) also includes whether the municipality is located on what later became the border between the Federation of Bosnia-Herzegovina and the Republika Srpska, and a measure of the roughness of the terrain.

All these variables have been selected based on theoretical expectations and a specific knowledge of the case. Indeed, previous studies found that a high level of ethnic polarization and the contiguity with external and internal borders were major drivers of intense violence during the Bosnian civil war (Costalli and Moro, 2011, 2012). The level of past violence in a given municipality emerged as the primary determinant of UNPROFOR troops' deployment, while theoretical reasons lead to the inclusion of the two remaining variables: on one hand, peacekeeping troops could tend to go to more populous areas in order to protect civilians; on the other hand, the roughness of the terrain can influence both the conduct of the fighting groups and the movements of the peacekeepers. These variables have also been chosen to avoid the risks of post-treatment bias. Indeed, they either do not change after the intervention of peacekeepers (e.g. the percentage of open terrain) or have been recorded before the intervention.

Once the pre-treatment observable variables on which units have to be similar are identified, researchers have to coarsen such variables into substantively meaningful categories. Coarsening variables means creating theoretically meaningful bins and then assigning every value of each variable to a bin, marked by a discrete cut-point. In other words, CEM coarsens the variables chosen by the researcher by recoding them, so that values that are close to each other from a theoretical and substantial point of view are grouped together. For instance, in an individual-level study using survey data, the variable *Age* could be coarsened into four bins along these lines: Age $\leqslant 30 = 1$; $31 \leqslant$ Age $\leqslant 50 = 2$; $51 \leqslant$ Age $\leqslant 70 = 3$; Age $\geqslant 71 = 4$. Similarly, the variable *Geography* could be coarsened into three bins as follows: Plain = 1; Hills = 2; Mountains = 3. Lastly, in studies on civil wars, the variable *Level of violence* could be coarsened in four bins such as: No violence = 1; Low violence = 2; Medium violence = 3; High violence = 4. In CEM, the coarsening is executed variable-by-variable and the algorithm allows researchers to choose the cut-points based on their substantive knowledge of the issue at stake and of the data generating process, thus coarsening each variable into substantively meaningful categories that reduce variability while at the same time preserving information. Coarsening is therefore case-specific, theory-driven, and reflects 'the knowledge the investigator must have' (Iacus *et al.*, 2019: 54). However, CEM may also implement an automated and theory-neutral coarsening strategy, which can be useful as a benchmark.

Having identified the substantively meaningful categories for each pre-treatment variable, CEM performs an exact matching between these categories. For instance, a treated unit with an Age bin value of 1 can only be matched with a control unit with an Age bin value of 1. If Geography and Level of violence are included in the set of pretreatment variables, then the treated and control units must also exactly match on the values of these two dimensions. Thus, a treated unit with a bin signature of: Age (2), Geography (1), Level of violence (0) would only match a control unit with the exact same 2–1–0 bin signature. Such a bin signature constitutes a 'stratum' (in CEM terms). Then, the units are sorted into strata, each of which has identical values for all the coarsened pre-treatment observable variables. Then, CEM drops any observation whose stratum does not contain at least one treated and one control unit.

Table 3. Results of matching

	Coarsening 1	Coarsening 2
Matched units	74	35
Unmatched units	346	385
Multivariate $L1$ distance	0.268	0.029

Treatment = Presence of peacekeepers.
Multivariate $L1$ distance before matching: 0.936.

Table 4. Results of matching

	Coarsening 1	Coarsening 2
Matched units	66	31
Unmatched units	354	389
Multivariate $L1$ distance	0.325	0.033

Treatment = active peacekeeping.
Multivariate $L1$ distance before matching: 0.943.

Intuitively, fewer strata will result in more heterogeneous observations within the same stratum and, thus, higher imbalance. On the other hand, more strata reduce the heterogeneity among the observations within the same stratum, but increase the likelihood of 'non-matching', thus reducing the number of observations for the subsequent analysis. As discussed in the subsection 'The inferential logic behind matching', maximizing balance while keeping a reasonably large number of observations is the key trade-off in matching. However, compared to the other matching techniques, an additional useful feature of CEM consists in the possibility to easily calculate the imbalance in the raw data before preprocessing and compare it with the remaining imbalance in the preprocessed data after matching the units. In detail, the statistic L_1 is an index that measures the global imbalance between the treatment and control groups, ranging from 0 to 1 (Iacus *et al.*, 2011, 2012). $L_1 = 1$ if the empirical distributions of the pre-treatment variables in the two groups are completely separated; $L_1 = 0$ if the empirical distributions coincide exactly. For instance, $L_1 = 0.6$ implies that 40% of the areas under the two histograms overlap.

Tables 3 and 4 show that preprocessing the raw data with CEM strongly reduces the imbalance between the treated and the control groups of Bosnian municipalities. The bottom lines in the tables show the values of the statistic L_1 in the raw data before matching, demonstrating that there was a serious problem of selection bias. More specifically, less than 10% of the distribution of the variables relative to the treated and control groups overlapped before matching. However, CEM helps researchers in dealing with such a selection bias as it strongly reduces the value of L_1, providing a much more balanced sample. Table 3 shows the results of the two matching procedures when the treatment is represented by the presence of peacekeepers in a given municipality, while Table 4 shows the results of the same procedures when the treatment is operationalized as active peacekeeping. Coarsening 2 – with the inclusion of two additional variables – is more effective in terms of imbalance reduction, but it also halves the number of observations compared to Coarsening 1 as the amount of unmatched units is huge.

Finally, the coarsened values are abandoned and the original values of the matched units are maintained for the analysis of the treatment effect. With exact matching between the treated and control groups (with values of L_1 close to 0), a simple weighted difference in means between the observed outcomes of the two groups would be sufficient to estimate the causal effect. However, in cases researchers have not achieved a perfect balance, as in this case, researchers are better off adjusting for the remaining imbalance via a statistical model and taking advantage of the 'CEM weights' (automatically generated and stored by CEM). Tables 5 and 6 show the causal effect of peacekeeping on local violence, estimated through OLS regressions, taking advantage of CEM

Table 5. Causal effect of the presence of peacekeeping on local violence

	Coarsening 1	Coarsening 2
Presence of peacekeeping t_{-1}	0.180	0.163
	(0.222)	(0.281)
N	74	35
R^2	0.367	0.423

Note: Standard errors in parentheses. *$P < 0.10$; **$P < 0.05$; ***$P < 0.01$.

Table 6. Causal effect of active peacekeeping on local violence

	Coarsening 1	Coarsening 2
Active peacekeeping t_{-1}	0.220	0.280
	(0.235)	(0.292)
N	65	31
R^2	0.381	0.432

Note: Standard errors in parentheses. *$P < 0.10$; **$P < 0.05$; ***$P < 0.01$.

weights and including ethnic polarization, the level of previous violence and the population of the municipalities as control variables.[3]

Tables 5 and 6 show that UNPROFOR troops were not effective in reducing the severity of violence and the fact that this lack of effectiveness is confirmed also by using the CEM preprocessed dataset suggests that it is not an artifact of sample selection biases. No significant impact was found, regardless of the criteria of coarsening and of the model specifications. Neither the simple presence, nor active initiatives of peacekeepers on the ground were able to reduce substantially the severity of violence in the Bosnian municipalities. The willingness to compare the most similar municipalities in order to address selection problems and estimate the effect of peacekeeping contingents on violence carefully led to a significant reduction in the number of observations, especially for the models performed after Coarsening 2. Matching must be used cautiously with small samples. However, here we show that it can also be meaningfully used in samples of about 100 observations and the central result of interest appears robust. As we can see by comparing Tables 2, 5 and 6, the main results are robust across model specifications: the peacekeeping troops did not manage to significantly reduce the severity of violence during the Bosnian civil war. However, given the extreme imbalance of the initial sample highlighted in Tables 3 and 4, any naïve analysis that did not address the selection problem would have been unreliable and would have run the risk of providing wrong feedbacks to policy makers. We chose this research as an example of studies where observational data are structurally plagued by selection bias, but causal inference is nonetheless essential. In fact, thanks to careful assessments of their effectiveness, UN peacekeeping missions have changed and similar studies performed on subsequent missions have shown that the problems highlighted in Bosnia have been at least partially addressed, resulting in much better performances (Hultman *et al.*, 2014; Ruggeri *et al.*, 2017).

Concluding remarks

In observational studies, causal inference is always hazardous due to the strong assumption of selection on observables, which is not easily testable by looking at the raw data. However, causal inference can barely be expelled from social and political sciences. In some occasions, it is crucial to direct research efforts toward a careful assessment of potential causal links, for instance when we need to gauge the effects of expensive or hotly debated public policies. Accurate theoretical

[3]The coefficients of the control variables are not shown to highlight the impact of peacekeeping, but are available upon request.

definition of potential causal mechanisms is essential, but researchers cannot avoid improving their toolbox of empirical methods to assess causation as carefully as they can.

In this article, we endorsed the practice of preprocessing the raw data through matching techniques in order to eliminate – or at least downplay – the problem of selection bias by generating well-balanced samples on a set of observable variables chosen by researchers. At this point, they will be able to apply on the preprocessed data the same familiar methods of estimation they would have used anyway on the original dataset, but with a much higher confidence in the reliability of the results. In fact, even if matching techniques will not overcome the selection on observables assumption, they will reduce model dependence for the subsequent estimation of the effect of the variable of interest via parametric analysis (i.e. slightly different model specifications are less likely to alter the substantial empirical conclusion of the analysis).

Matching is not a *panacea*: researchers face a lot of decisions during its implementation and several things may go wrong. For example, researchers may miss a higher dimensional aspect of imbalance when checking lower dimensional summaries. This may affect the estimates. However, since it may happen also without preprocessing, following the steps here suggested should at least not make things worse. Moreover, when the preprocessing implies the loss of some treated unit, researchers should openly discuss the consequences in terms of external validity. However, the implementation steps we suggested in this article should help researchers in openly assessing whether their study is meeting the necessary conditions for generating a valid causal inference or how far they go.

Among the rich family of matching techniques, we focused on CEM, which guarantees a reduction of the imbalance between groups, overcomes the propensity score tautology, and allows a wide range of flexibility to researchers' needs. We reviewed Costalli (2014) by describing in detail the steps he took to assess the severity of the selection bias affecting municipalities with and without peacekeeping troops during the Bosnian civil war. Moreover, we provided practical guidance for the implementation of the CEM algorithm to create a well-balanced sample of municipalities on which to assess the effectiveness of the UN peacekeeping mission.

Lastly, it is worth mentioning that matching is not the only approach in the toolbox of researchers interested in causal inference. For example, we do not discuss research designs that exploit randomness in the assignment to treatment arising indirectly from exogenous factors or events (i.e. instrumental variables models); those in which the assignment to treatment is given by exogenous eligibility criteria making a subset of units as good as randomly assigned to treatment (i.e. regression discontinuity designs); or the ones that rely on assumptions about pre-treatment outcome trends (i.e. difference-in-differences and synthetic control models). Even focusing only on research designs arising from the assumption of selection on observables, matching is not alone: for example, the Entropy Balancing by Heinmueller (2012) is a close cousin of CEM, while Oster (2019) developed an approach to evaluate the robustness of OLS coefficients to the omitted variable bias, provided that such bias is related to the observable controls. This non-exhaustive list of tools should prompt researchers to weigh methods with a proactive attitude: the most credible approach is often a combination of different identification strategies, always grounded on a deep knowledge of the context under investigation.

Funding. The research received no grants from public, commercial or non-profit funding agency.

Data. The replication dataset is available at http://thedata.harvard.edu/dvn/dv/ipsr-risp.

Acknowledgements. An early version of this article has been presented at the 2nd Workshop on Political Science Methods 'Capturing causation: Research and model design', University of Catania, April 26, 2019. We are grateful to the participants and the MetRiSP standing group for their helpful comments and feedbacks. Moreover, we thank Erich Battistin, Marco Bertoni, and the anonymous reviewers for useful comments on earlier versions of this article. All remaining errors are our own.

References

Angrist JD and Pischke JS (2009) *Mostly Harmless Econometrics*. Princeton: Princeton University Press.

Arceneaux K, Gerber AS and Green DP (2006) Comparing experimental and matching methods using a large-scale voter mobilization study. *Political Analysis* **14**, 37–62.

Barnow BS, Cain GG and Goldberger AS (1980) Issues in the analysis of selectivity bias. In Stromsdorfer E and Farkas G (eds), *Evaluation Studies*, vol. **5**. San Francisco: Sage Publications, pp. 43–59.

Bates MA and Glennerster R (2017) The generalizability puzzle. *Stanford Social Innovation Review,* 50–54. https://ssir.org/articles/entry/the_generalizability_puzzle.

Beck N and Katz J (1995) What to do (and not to do) with time-series cross-section data. *American Political Science Review* **89**, 634–47.

Beck N and Katz J (1996) Nuisance vs. substance: specifying and estimating time-series cross-section models. *Political Analysis* **6**, 1–36.

Becker SO and Ichino A (2002) Estimation of average treatment effects based on propensity scores. *The Stata Journal* **2**, 358–77.

Caliendo M and Kopeinig S (2008) Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys* **22**, 31–72.

Carpenter D (2002) Groups, the media, agency waiting costs, and FDA drug approval. *American Journal of Political Science* **46**, 490–505.

Cederman LE and Gleditsch KS (2009) Introduction to special issue on disaggregating civil war. *Journal of Conflict Resolution* **53**, 487–95.

Cederman LE, Gleditsch KS and Buhaug H (2013) *Inequality, Grievances and Civil War*. Cambridge: Cambridge University Press.

Cochran WG and Chambers SP (1965) The planning of observational studies of human populations. *Journal of Royal Statistical Society, Series A* **128**, 234–65.

Corbetta P (2003) *Social Research. Theory, Methods and Techniques*. London, Thousand Oaks, New Delhi: Sage Publications.

Costalli S (2014) Does peacekeeping work? A disaggregated analysis of deployment and violence reduction in the Bosnian war. *British Journal of Political Science* **44**, 357–380.

Costalli S and Moro FN (2011) La violenza nelle guerre civili: un'analisi quantitativa della violenza in Bosnia-Erzegovina. *Rivista Italiana di Scienza Politica* **41**, 5–26.

Costalli S and Moro FN (2012) Ethnicity and strategy in the Bosnian civil war: explanations for the severity of violence in Bosnian municipalities. *Journal of Peace Research* **49**, 801–15.

Costalli S, Moro FN and Ruggeri A (2020) The logic of vulnerability and civilian victimization. Shifting front lines in Italy (1943–1945). *World Politics* **72**, 679–718.

Cox DR (1958) *Planning of Experiments*. New York: John Wiley.

Crump RK, Hotz VJ, Imbens GW and Mitnik O (2009) Dealing with limited overlap in estimation of average treatment effects. *Biometrika* **96**, 187–199.

Dehejia RH and Wahba S (1999) Causal effects in nonexperimental studies: re-evaluating the evaluation of training programs. *Journal of the American Statistical Association* **94**, 1053–1062.

Dehejia RH and Wahba S (2002) Propensity score matching methods for nonexperimental causal studies. *Review of Economics and Statistics* **84**, 151–161.

Doyle M and Sambanis N (2006) *Making War and Building Peace: United Nations Peace Operations*. Princeton: Princeton University Press.

Duflo E, Glennerster R and Kremer M (2008) Using randomization in development economics research: a toolkit. *Handbook of Development Economics* **4**, 3895–3962.

Fjelde H and Hultman L (2014) Weakening the enemy: a disaggregated study of violence against civilians in Africa. *Journal of Conflict Resolution* **58**, 1230–1257.

Fortna VP (2008) *Does Peacekeeping Work? Shaping Belligerents' Choice after Civil War*. Princeton: Princeton University Press.

Gleditsch KS (2007) Transnational dimensions of civil war. *Journal of Peace Research* **44**, 293–309.

Gleditsch NP, Wallensteen P, Eriksson M, Sollenberg M and Strand H (2002) Armed conflict 1946–2001. A new dataset. *Journal of Peace Research* **39**, 615–637.

Goldberger A (1991) *A Course in Econometrics*. Cambridge: Harvard University Press.

Heckman J and Robb R (1985 print publication - 2013 online publication) Alternative methods for evaluating the impacts of interventions. In Heckman J and Singer B (eds), *Longitudinal Analysis of Labor Market Data*. Cambridge: Cambridge University Press, 156-246.

Heckman J, Ichimura H and Todd P (1997) Matching as an econometric evaluation estimator: evidence from evaluating a job training programme. *Review of Economic Studies* **64**, 605–654.

Heckman J, Ichimura H and Todd P (1998) Matching as an econometric evaluation estimator. *Review of Economic Studies* **65**, 261–294.

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

Heinmueller J (2012) Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies. *Political Analysis* **20**, 25–46

Ho DE, Imai K, King G and Stuart EA (2007) Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis* **15**, 199–236.

Holland PW (1986) Statistics and causal inference. *Journal of the American Statistical Association* **8**, 945–60.

Hultman L, Kathman J and Shannon M (2014) Beyond keeping peace: united nations effectiveness in the midst of fighting. *The American Political Science Review* **108**, 737–753.

Iacus SM, King G and Porro G (2009) cem: Software for Coarsened Exact Matching. *Journal of Statistical Software* **30**(9), 1-27. Available at https://www.jstatsoft.org/issue/view/v030.

Iacus SM, King G and Porro G (2011) Multivariate matching methods that are monotonic imbalance bounding. *Journal of the American Statistical Association* **106**, 345–361.

Iacus SM, King G and Porro G (2012) Causal inference without balance checking: coarsened exact matching. *Political Analysis* **20**, 1–24.

Iacus SM, King G and Porro G (2019) A theory of statistical inference for matching methods in causal research. *Political Analysis* **27**, 46–68.

Imbens GM (2004) Nonparametric estimation of average treatment effects under exogeneity: a review. *Review of Economics and Statistics* **86**, 4–29.

Imbens GM and Wooldridge JM (2008) Recent developments in the econometrics of program evaluation. NBER Working Paper No. 14251. Available at http://www.nber.org/papers/w14251.

Kalyvas S (2006) *The Logic of Violence in Civil War*. Cambridge, New York: Cambridge University Press.

Keele L (2015) The statistics of causal inference: a view from political methodology. *Political Analysis* **23**, 313–335.

Khandker SR, Koolwal GB and Samad HA (2010) Handbook on impact evaluation: quantitative methods and practices. World Bank. © World Bank. Available at https://openknowledge.worldbank.org/handle/10986/2693 License: CC BY 3.0 IGO.

King G and Nielsen R (2019) Why propensity scores should not be used for matching. *Political Analysis* **27**, 435–454.

King G and Zeng L (2006) The dangers of extreme counterfactuals. *Political Analysis* **14**, 131–59. Available at http://gking.harvard.edu/files/abs/counterft-abs.shtml.

LaLonde R (1986) Evaluating the econometric evaluations of training programs. *American Economic Review* **76**, 604–620.

Manski CF (2007) *Identification for Prediction and Decision*. Cambridge, MA: Harvard University Press.

Martini A and Sisti M (2009) *Valutare il Successo Delle Politiche Pubbliche*. Bologna: Il Mulino.

Oster E (2019) Unobservable selection and coefficient stability: theory and evidence. *Journal of Business & Economic Statistics* **37**, 187–204.

Petterson T, Hogbladh S and Oberg M (2019) Organized violence 1989–2018 and peace agreements. *Journal of Peace Research* **56**, 589–603.

Rosenbaum PR (1984) The consequences of adjusting for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society, Series A* **147**, 656–66.

Rosenbaum PR (2002) *Observational Studies*. New York: Springer.

Rosenbaum PR and Rubin DB (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.

Rubin DB (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **6**, 688–701.

Rubin DB (2010) On the limitations of comparative effectiveness research. *Statistics in Medicine* **29**, 1991–1995.

Ruggeri A, Dorussen H and Gizelis T (2017) Winning the peace locally: UN peacekeeping and local conflict. *International Organization* **71**, 163–185.

Smith H (1997) Matching with multiple controls to estimate treatment effects in observational studies. *Sociological Methodology* **27**, 325–353.

Smith JA and Todd PE (2005) Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics* **125**, 305–53.

Walter B (1997) The critical barrier to civil War settlement. *International Organization* **51**, 335–364.

Young A (2019) Channelling Fisher: randomization tests and the statistical insignificance of seemingly significant experimental results. *The Quarterly Journal of Economics* **134**, 557–598.

Cite this article: Costalli S, Negri F (2021). Looking for twins: how to build better counterfactuals with matching. *Italian Political Science Review/Rivista Italiana di Scienza Politica* 1–16. https://doi.org/10.1017/ipo.2021.1