Statistical Matching with Time-Series Cross-Sectional Data: Magic, Malfeasance, or Something in Between?

Richard A. Nielsen[1]

Forthcoming in Robert Franzese and Luigi Curini, eds.,*The SAGE Handbook of Research Methods in Political Science and International Relations*, Thousand Oaks, CA: Sage.

In the beginning, there was history. If scholars of International Relations (henceforth IR) wanted to know why something happened, they consulted history for patterns. Under what conditions did that something happen? What if conditions had been different? What if choices had been different? What if timing had been different?

Then, spurred by a combination of factors, some IR scholars discovered that the field of statistics offered new ways to explore history. By compiling history into databases, wrangling it into numerical matrices, and applying mathematical models, these scholars abstracted away from the fine details to see broad patterns that are difficult for even the most systematic historian to pick out. Historical and statistical modes of inquiry seem entirely different; the data rearranged in unfamiliar ways, the skill sets distinct. But the logic of inference can be surprisingly similar, if the language barrier between traditions can be crossed.

Matching is a family of quantitative procedures for creating and analyzing a sample of cases that differ on one key factor (called a "treatment") and are "matched" to be similar on others. It puts historical counterfactual comparison at the center of statistical analysis, providing one "border crossing" between the historical and statistical traditions in IR.

Matching emerges naturally as the statistical analog to the qualitative tradition of paired case comparison (Tarrow 2010). In a qualitative study of a puzzling phenomenon, our first move might be to identify a positive case: one in which the outcome of interest happens. We might then try to infer *why* the outcome happened from the historical record by considering possible counterfactual histories: if various factors had been different, would the outcome have been different also?

[1] Department of Political Science, MIT. Eliza Riley provided research assistance. Luigi Curini, Elizabeth Dekeyser, Rob Franzese, Kosuke Imai, In Song Kim, Kacie Miura, and Eliza Riley generously gave advice.

Relying on subjective intuition about counterfactual histories for a single case can lead us to merely confirm our prior assumptions. Comparison cases provide more objective information for inferring how the positive case might have turned out if some factor had been different. Following the logic of Mill's methods of difference, we might seek a comparison case that is similar to our positive case, but differs in the factor that we suspect matters most. If our cases are otherwise identical and our theories are deterministic, then this comparison can tell us all we need to know. Any difference between the case outcomes is the causal effect of the differing factor. However, our theories are not usually deterministic and our paired cases rarely match in every respect so the difference in outcomes could be due to something else. To be more certain that the factor explains the outcome, we might do another comparison, and another, to see if they support the same conclusion. Soon, we have too many pairs of cases to easily summarize qualitatively, so we might give a numerical summary: perhaps the average difference in outcomes across pairs of cases. We have arrived at matching. Typically, an analyst creates a matched sample by trimming down a quantitative data set, but a researcher employing the method of paired comparison could build the same matched sample "from the ground up," as I have just described. The power of matching comes from the properties of the sample, not how the analyst obtains it.

After constructing a matched sample, an analyst typically estimates the causal effect of treatment on the outcome using a regression model. Regression is a method for calculating correlations that does not, by itself, necessarily produce credible estimates of causal effects. Matching is a method for selecting cases to use in the regression calculation so that the assumptions necessary to infer causation from correlation are more plausible. Matching is case selection for quantitative studies.[2]

In the mid-2000s, matching burst into IR and Political Science in a series of high-profile publications. To chronicle its dramatic rise, I collected a comprehensive list of articles applying matching to IR data the twelve leading IR journals.[3] Following the first application by Simmons and Hopkins (2005), the number of applications has risen exponentially. In 2017, a record high of 28 articles used matching (see Figure 1), and by mid-2018 there were 20, with the year only halfway over (the latest for which I could collect data). Matching now features in at least ten percent of the approximately 200 quantitative IR articles published each year.

[2] Matching is also useful as a case selection strategy for some qualitative studies. See Nielsen (2016) for explanation and Weeks (2018) for an example. However, my focus in this chapter is on matching for statistical inference.

[3] I consider IR data to include either an international predictor variable, an international outcome variable, or both. I use the TRIP article database (Teaching, Research, and International Policy Project 2017; Maliniak et al 2018) to identify IR articles in twelve journals: APSR, AJPS, BJPS, JOP, IO, IS, ISQ, WP, SS, JCR, JPR, and EJIR. Using several methods for searching, I find 124 articles that use matching. A full list is in the online supplemental information.

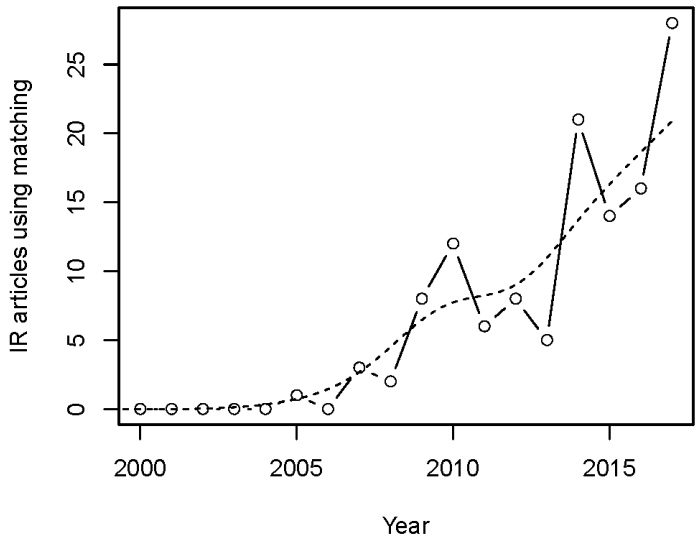

Figure 1: The number of articles per year using statistical matching in the twelve leading IR journals.

The rise of matching has been controversial. Proponents tout the benefits for making causal inferences with non-experimental data. Detractors argue that matching is over-hyped as a magic bullet for causal inference, and that it offers unscrupulous researchers another tool to "p-hack" their way to false discoveries (Arceneaux, Gerber, and Green 2006, Arceneaux, Gerber, and Green 2010, Miller 2019). The time-series cross-sectional (TSCS) data sets common in IR are especially challenging for matching. With little guidance from methodologists, researchers have developed ad hoc approaches to matching with TSCS data, a fact that is equally troubling for the credibility of research findings despite receiving less attention.

The critics are right that matching is not magic. It is helpful for causal inference, but only with strong assumptions. Existing matching methods do not adequately accommodate the complex structure of many IR data sets. And matching does not stop researchers from p-hacking. Like any other method, it can be misused, oversold, and misunderstood. But matching is not just hype either. Matching has proven to be a reliably useful method for causal inference with messy, non-experimental data. It belongs in the toolbox of IR scholars.

In this chapter, I introduce matching for readers with no prior experience and weigh its merits and weaknesses in IR applications using time-series cross-sectional data. There are already several excellent introductions to matching (Ho et al 2007, Sekhon 2009, Stuart 2010), which I aim to complement rather than reproduce. I first introduce the philosophy of causal inference underlying matching. I then explain the mechanics of matching that arise from this philosophy. Matching mechanics are not always well-suited to the time-series cross-sectional data structures

common in IR, so I discuss the challenges and attempts to surmount them. I conclude by returning briefly to the controversies.

The Philosophy of Matching

Matching rests on a counterfactual approach to causation. A variable T (for "treatment") is a cause of variable Y if Y would have been different had T been different. Imagine all possible values of Y for a unit i and call these the potential outcomes of unit i. For simplicity, consider the case where T is a binary indicator for whether the event represented by T happened ($T = 1$) or not ($T = 0$). The causal effect of T on Y for unit i is the potential outcome when T is one minus the potential outcome when T is zero, denoted $Y_i(T = 1) - Y_i(T = 0)$.

The fundamental problem of causal inference is that for any unit, we cannot observe both potential outcomes because we cannot re-run history (Holland 1986). For any causal inference, at least half of the data for our calculation are missing. Thus, we *infer* causal effects rather than *measuring* them; we are always guessing about the counterfactual outcome that we would see if T were different.

Experimental manipulation is the best way to learn about cause and effect (Bowers and Leavitt, this volume). In an experiment, we randomly assign T to learn whether it causes changes in Y. Yet even in an experiment, we cannot observe both potential outcomes for any individual. To impute the missing potential outcomes, most analysts estimate average causal effects over many individuals. With enough units and random assignment of T, we can average Y for the $T = 1$ group and subtract the average of Y for the $T = 0$ group to get an unbiased estimate of the average treatment effect (abbreviated ATE). We can write this difference in means estimator as $Y_{treated} - Y_{control}$.

Despite efforts to bring experiments into IR (Findley, Nielson, and Sharman 2013), there are many situations where randomization is impossible, unethical or both. We are forced to infer cause and effect from an *observational study* using data observed in the natural world. In observational studies, units select their level of treatment or have it assigned to them non-randomly (called "selection effects," "endogeneity," and "reverse causality"). This results in *imbalance*: systematic differences between treated and control groups. If selection into treatment is correlated with potential outcomes of Y, then the average control unit provides a biased estimate of the missing potential outcome for the average treated unit, making the estimator $Y_{treated} - Y_{control}$ biased as well. These factors that are related to both treatment status and the potential outcomes are called *confounders*, and often denoted with the matrix X.

Some observational studies feature a natural experiment in which treatment is assigned "as if" randomly by nature for a subset of the data. Even if treatment assignment in the overall data set

is correlated with possible confounding variables, comparing the subset of units that received as if random treatment assignment can produce credible causal inferences. Techniques such as instrumental variables (Carter and Dunning, this volume) and regression discontinuity designs (Cattaneo, Titiunik, and Vazquez-Bare, this volume) facilitate these comparisons, depending on the type of natural experiment that occurred.

Often, there is no natural experiment for a researcher to exploit. The next-best strategy is to compare outcomes for units with similar values of the confounding variables. This is called a *conditioning strategy* because we condition on the values of X to infer the causal effect of T on Y). Matching is a conditioning strategy X proposed by Rubin and collaborators in a series of papers starting with Rubin (1973).[4] As developed by Rubin, matching identifies treated and control units that have similar values of the confounding variables in X in a data set and then removes the other units, hoping to approximate the data that would have resulted from a randomized experiment. In this matched sample, the variables in X are no longer correlated with treatment assignment so we can obtain unbiased estimates of the ATE using the difference in means: $Y_{treated} - Y_{control}$.

The core insight of matching is that a subset of data may be useful for credible causal inference, even if all available data are not. However, subsetting the data may limit the inferences a researcher can make. In a natural experiment, only some units have treatment randomized by nature, so the estimated quantity of interest is a *local average treatment effect* for only those units (Imbens 2009). Analogously, matching only allows credible causal inference for the types of units retained in the matched sample; all other inferences rely on extrapolation. If an analyst's quantity of interest is the *sample average treatment effect* on the treated units (abbreviated SATT), then she cannot discard any treated units. This quantity of interest answers the question "what would have happened if all treated units in this sample had instead received control?" Answering an alternative question such as "what if all control units received treatment?" requires a different matching scheme that retains all control units but may discard treated units. Sometimes, an analyst may wish to estimate the SATT but there are treated units that do not have good matches and discarding them would greatly reduce imbalance. Analysts may discard these treated observations, but their quantity of interest is now the *feasible sample average treatment effect on the treated* (FSATT): the average treatment effect in the subsample for which causal inference is feasible (King, Lucas, and Nielsen 2017, 475).

Assumptions

The assumptions necessary for causal inference with a conditioning strategy are stringent. First, we require the *stable unit treatment value assumption*, abbreviated SUTVA, which states that the fixed (but generally unknown) potential outcomes for each unit do not depend on the potential

[4] These papers are collected into one volume in Rubin (2006).

outcomes of the other units. Equivalently, this assumption states that there is no interference between units and no hidden versions of treatment.

Second, conditioning strategies require the assumption of *conditional ignorability*: that after conditioning on the X, treatment assignment is independent of the potential outcomes. Alternatively, this assumption is called "selection on observables" or "no omitted variables." If unobserved confounders affect both treatment and outcome, conditioning on X cannot provide unbiased causal estimates. This requires that the analyst know which variables are confounders. Variables that are part of the *assignment mechanism* of T should be included in X. The most compelling matching applications carefully theorize the assignment mechanism of T and then exhaustively measure each variable.

Third, we require *common support*: there must be treated and control units at all levels of X for which we wish to make inference.

Constructing and Analyzing Matched Samples

The core activity of matching is construction a matched sample that plausibly satisfies the assumptions of stable unit treatment values, conditional ignorability, and common support. A *matching algorithm* is the procedure for constructing a matched sample. Matching algorithms typically identify similar treatment and control units and match them to each other, discarding (or "pruning") unmatched units from the data set. This is equivalent to giving discarded units a weight of zero in subsequent statistical calculations, so a matched sample is merely a reweighted version of an unmatched sample.[5] After constructing a matched sample, the analyst can calculate the effect of treatment by calculating the difference of the (weighted) average outcome of the treated units and the weighted average outcome of the control units ($Y_{treated} - Y_{control}$). Or, the analyst may wish to further adjust via regression (Ho et al 2007).

Although matching algorithms are the focus of much of the matching literature, from a certain perspective, the algorithm is not very important. What matters is getting the best possible matched sample by whatever means. What is the best possible matched sample? Matched samples are evaluated on two main criteria: the similarity of treated and control units in the sample (called *balance*) and sample size. A larger sample provides more statistical power for estimating precise results. A balanced sample is more likely to satisfy the assumptions of common support and conditional ignorability. These two criteria are in tension because the way to improve balance is to discard observations. Matching is a constrained optimization problem of

[5] Some matching algorithms also give units non-integer weights, which can be thought of as "partial inclusion" in the matched sample.

maximizing balance subject to a sample size constraint (or maximizing sample size subject to a balance constraint).

Assessing sample size is straightforward, but the notion of balance deserves elaboration. When $X_{treated} = X_{control}$, the data set is unambiguously balanced. However, when $X_{treated} \neq X_{control}$, there is no universal definition of balance. Various balance metrics have been proposed and each may give different answers. Improvement on one balance metric may not correspond to improvements in other balance metrics. Applied scholars often fixate on justifying their choice of matching algorithm, but they should be more concerned with justifying their choice of balance metric. As I show below, these are integrally connected, but the prevalence of nonsensical practices in the literature reveals that analysts may not fully appreciate this connection.

Each balance metric offers different answers to two questions: How should differences between $X_{treated}$ and $X_{control}$ be summarized? And which differences are small enough for a data set to be declared sufficiently balanced? Analysts must inescapably answer these questions by choosing a balance metric, and their choice reflects an implicit set of assumptions about how confounding threatens their inferences.

Analysts who believe that confounding is omnipresent worry that minor differences between $X_{treated}$ and $X_{control}$ could induce substantial confounding. These analysts prefer balance metrics that are sensitive to small differences in many possible dimensions. Other analysts might be less concerned about confounding from small differences between $X_{treated}$ and $X_{control}$ and prefer balance metrics that do not capture these small differences.

One set of balance metrics focus on the discrepancies between pairs (or groups) of treated and control units. The *Average Mahalanobis Imbalance* metric uses the average of the pairwise Mahalanobis distance[6] from each unit to the nearest unit of the opposite treatment status (King, Lucas, Nielsen 2017, 479). The *L1 Imbalance metric* is a discretized version of the same idea. The analyst selects a multivariate histogram binning, calculates the difference in the frequency of treated and control units in each bin, and averages over the bins of the histogram (King, Lucas, Nielsen 2017, 479). For both metrics, larger values indicate more imbalance. These discrepancy-based are sensitive to small differences in many possible dimensions. Both of have the highly desirable ability to detect an exactly matched subset of observations if such a subset exists. However, they offer no definitive threshold for declaring a data set "balanced" short of zero difference ($X_{treated} = X_{control}$).

[6] Mahalanobis distance is a generalization of Euclidean distance that is unit-less, scale-invariant, and accounts for correlated variables.

An alternative style of balance metrics use statistical tests to detect differences between $X_{treated}$ and $X_{control}$. Some analysts calculate the difference in means of the covariates in X and declare the data set balanced if variable-by-variable t-tests fail to reject the null hypotheses that the means are the same. Some analysts prefer Kolmogorov-Smirnov tests for the difference in distributions to t-tests. Hansen and Bowers (2008) propose a global p-value for a test that the combined dimensions of X are detectably different from each other. The metrics rely on widely recognized significance thresholds for declaring a data set sufficiently balanced, even if $X_{treated} \neq X_{control}$. Unfortunately, hypothesis testing depends on the sample size, so "balance" may be achieved once enough observations are removed whether the remaining observations in the matched sample are actually similar or not (Stuart 2010). Also, optimizing these metrics will not reliably uncover an exactly matched subsample, even if one exists.

In addition to these formal approaches, analysts also inspect balance by plotting the difference in means between treated and control groups for each covariate before and after matching (akin to the variable-by-variable t-tests, but without a test).

Matching Algorithms

It does not matter how one constructs a matched sample with sufficient balance and sample size. In theory, analysts could directly optimize their preferred balance metric subject to a sample size constraint. In practice, this is usually not possible because the optimization problem is intractable, though it can be done for some balance metrics (King, Lucas, and Nielsen 2017). One could attempt brute-force optimization by randomly sampling all possible subsets of the data and finding the sufficiently sized subset with the best balance. However, this is infeasible for most data sets because the number of possible combinations is too large. Instead, analysts turn to matching algorithms that dramatically reduce the time necessary to construct candidate matched samples and select the best one.

Ideally, matches should be exact: each treated unit paired to a control unit (or units) with identical values of all confounding pre-treatment covariates. Exact matches can be identified by checking whether each treated unit has a matching control unit with identical values of X (using, for example, the MatchIt library in R).

In most data sets, exact matching leaves too few observations for precise estimation in the subsequent statistical analysis. With continuous covariates, exact matches are unlikely to exist at all. Approximate matching methods offer a way to identify matched subsets of the data that plausibly satisfy the three key assumptions above, even if the treated and control units are not exactly matched.

Each approximate matching algorithm deals with the problem of inexact matches by explicitly or implicitly constructing a distance metric from each treated unit to each control unit and then attempting to identify a matched sample of sufficient size that minimizes that distance metric.

There is no established "best" matching algorithm. Because each algorithm has a different distance metric, each is optimizing a different implied balance metric, and each potentially finds different matched samples. Standard practice is to match with one algorithm and then check balance with one or more metrics *not* implied by that algorithm. If one is lucky, optimizing the balance metric implied by the algorithm will lead to improvements with respect to other balance metrics. In this happy circumstance, the choice of balance metric does not matter much because they all move together.

However, in many cases a matching algorithm does not lead to notable improvements on all balance metrics. Standard practice in this situation is to change the tuning parameters of the algorithm, rematch, and check balance again, repeating until balance on various balance metrics is satisfactory. More generally, analysts often match using a method that minimizes one metric, and then evaluate balance by checking another metric. This is a strange form of indirect optimization. If balance on one metric is really the target, then iteratively optimizing it using hit-or-miss trials from an unrelated matching algorithm seems nonsensical. One justification might be that some balance metrics cannot be directly optimized, but even so, the standard practice of checking balance and manually adjusting the algorithm after each iteration of is inefficient.

Instead, analysts who cannot find exactly matched observations within their data sets must make a philosophical commitment to a particular balance metric and, if possible, chose a matching algorithm that optimizes it. Checking whether matching improves alternative balance metrics may be revealing about the sensitivity of matches to the choice of balance metric, but it is not dispositive about whether the data set is balanced. When two balance metrics diverge meaningfully, there is no way to optimize both at the same time. The choice of one over the other is philosophical.

There is a long and growing list of matching techniques.[7] I'll review the key families of algorithms which are the foundation for many, though not all, of the options available.

Mahalanobis Distance Matching (MDM)

The first matching approach proposed by Rubin (1980) matches treated and control units based on continuous distances from each other in the k-dimensional space defined by the *k* covariates in X. The basic matching machinery works for any continuous distance metric: First, calculate the distance from each treated unit to each control unit. Then, form matches by selecting the

[7] See the online supplemental information for a list with citations.

closest control observation to each treated observation. Finally discard the unmatched control observations. Methodologists prefer Mahalanobis distance (Euclidian distance normalized by the covariance matrix) because it accounts for correlations between the variables and makes them unit-less and scale-invariant. In order to avoid matched pairs that air not adequately similar, analysts often use a *caliper*: a defined maximum distance for a match. If a treated unit does not have a matching control unit within the specified caliper, it is discarded and the quantity of interest shifts from the SATT to the FSATT.

Propensity Score Matching (PSM)

Rubin and Rosenbaum (1983) introduce the *propensity score*, defined as the probability that unit i receives treatment, conditional on X. They show that the propensity score is a sufficient statistic for X; all of the information about X necessary to make the potential outcomes independent of treatment is contained in this score. Assuming conditional ignorability, matching units exactly on their true propensity scores results in unbiased estimates, even if $X_{treated} \neq X_{control}$. The true propensity score is usually unknown, so we must estimate it, typically with predicted values from a logistic regression predicting treatment as a function of X. We then match treated and control units using the similarity of their propensity scores. Again, analysts often define calipers. A common rule of thumb is that observations should be within 0.25 of a standard deviation of the propensity scores to match (Rosenbaum and Rubin 1985, 114, but see King and Nielsen 2019).

The theory underlying PSM is elegant and intuitive, but it requires exact matches on the true propensity score. In practice, we typically must make do with inexact matches on estimated propensity scores, and this can sometimes make causal inferences worse, rather than better (King and Nielsen 2019). Several methods combine information from the propensity score and some other covariate information (Diamon and Sekhon 2013; Imai and Ratkovic 2014), and these methods side-step the problem.

There are a host of complications that can be vexing for both MDM and PSM. Methodologists assume a large pool of control observations from which to match (Rubin and Thomas 1996) but this is not the case in many IR settings. Also, the same control might be the best available match for multiple treated units. Which should get it? Optimal matching (Rosenbaum 1989), optimal full matching (Hansen 2004), and optimal cardinality matching (Zubizaretta, Paredes, and Rosenbaum 2014) are promising solutions that improve global balance as well as local balance.

Coarsened Exact Matching (CEM)

Coarsened exact matching is an extension of exact matching methods to situations where exact matches are not possible (Iacus, King, and Porro 2012), and is the most popular example of a class of matching methods that are monotonically imbalance bounding (Iacus, King, and Porro 2011). Monotonically imbalance bounding methods set a fixed bound on the imbalance that the analyst will permit and achieve that balance, though often at the expense of matched sample size.

CEM "coarsens" each variable into categories that are less restrictive than the measured values. For example, if the democracy level of country is measured on a scale from 21 point scale, as in the Polity IV scores (Marshall, Jaggers, and Gurr 2002), we might coarsen this variable into five categories based on the scores: strongly democratic, leaning democratic, inchoate, leaning autocratic, strongly autocratic. Then, we match units that fall in the same category, rather than requiring that matches have exactly the same score. Observations that do not have matches, both treated and control, are discarded, and the remaining observations are reweighted to account for the fact that matches may not be one-to-one. Then, the analyst "uncoarsens" the data, and estimates causal effects using weighted least squares on the matched data set.

The Matching Frontier

A new class of approaches seeks to improve inferences with each of the prior methods by introducing the "matching frontier" – a set of matched samples that are optimal according to the balance metric they maximize (King, Lucas, and Nielsen 2017). Users of MDM and PSM often use *calipers*, which define the maximum allowable distance between a treated and control match. A stricter caliper will result in a smaller, but more balanced matched sample. Analysts also tend to create one-to-one matched samples out of convenience. However, if more than one control unit is a good match for a treated unit, arbitrarily limiting the number of matches decreases the sample size and decreases the precision of causal estimates without removing bias.

The matching frontier generalizes the idea of a caliper and allows many-to-many matching. The frontier is defined by matched samples of various sizes, from $n = N$ (the full unmatched sample) down to $n = 2$ (the closest matching pair). At each possible sample size n between N and 2, there are $\binom{N}{n}$ possible matched samples. The frontier finds the optimal one for each n, resulting in a series of optimal matched samples of various sizes.

This continuum of matched samples defines the trade-off between bias and variance. Matched samples that discard all but a few extremely close matches will reduce bias the most, but may lack the statistical power to precisely estimate effects. On the other end, those that retain most of the original sample will usually result in more precise estimates, but there may be more bias because the observations are less similar to each other. Any matching solution that is not on this frontier is suboptimal: there is an achievable matched sample that has either more observations or better balance.

Calculating matching frontiers for arbitrary balance metrics is hard because the number of possible matched samples for real-life data sets is extremely large. King, Lucas, and Nielsen (2017) develop fast algorithms that calculate the matching frontier for MDM, PSM, and CEM-style methods.

Categorical and Continuous Treatments

All of these algorithms struggle to accommodate categorical and continuous treatment variables. The problem is philosophical as well as technical. As the number of treatment levels increases, so does the number of counterfactual questions we must answer to describe what would have happened to a given unit if it received each alternative treatment level. One possibility is to construct a matched sample for each possible treatment contrast, but this is difficult if each level of treatment is only assigned to a few observations. With a truly continuous treatment variable, each unit receives a unique level of treatment, so constructing a matched sample for each treatment level is impossible. Instead, methodologists have proposed matching methods that effectively consider similar levels of treatment to be equivalent (Imai and Van Dyk 2004, Hirano and Imbens 2004).

Regression appears to side-step this challenge, effortlessly estimating effects of categorical and continuous treatments. In fact, regression faces the same challenge but limits the analysts ability to diagnose it; regression automatically fits a hyperplane that extrapolates between treatment levels while obscuring the fact that this extrapolation may be based on very little data at any single treatment level. Matching calls attention to the true degree of difficulty in making credible causal inferences when there are many versions of treatment.

Inference

After matching, analysts typically calculate their desired treatment effect by fitting a parametric regression model predicting Y as a function of T, possibly controlling for variables in X if matching was inexact. If the assumptions hold, this results in an unbiased estimate. But analysts are almost always interested in testing whether this estimate is statistically different from some null hypothesis (typically that treatment has no effect). This requires estimates of the uncertainty surrounding an effect estimated via matching and there is substantial debate in the literature about how to do so.

Because it is convenient, standard practice among practitioners is to report the standard errors from regression analysis as measures of uncertainty for the treatment effect. Methodologists worry that these standard errors do not account for the uncertainty of the matching procedure. Abadie and Imbens (2008) show that naive bootstrap standard errors are not asymptotically valid and propose a correction, though recent work proposes an asymptotically valid bootstrap (Otsu and Rai 2017). Others propose algorithm-specific approaches such as a Bayesian estimate incorporating the uncertainty inherent in estimating propensity scores (An 2010). But Iacus, King, and Porro (2019) argue that these concerns are misplaced; if analysts are willing to change their axioms about sampling, then unaltered regression standard errors are correct.

One especially clever approach to justifying measures of effect size uncertainty comes from Imai and Kim (2019). They demonstrate an equivalence between linear models with unit fixed effects

and a class of within-unit matching schemes. From this equivalence, they are able to show that accepted model-based standard errors are valid, without resorting to the more computationally intensive proposed alternatives. While this equivalence approach currently only applies to an uncommon subclass of matching approaches, it might be extensible.

Comparison to Regression

Regression is an alternative approach to conditioning on X that has a much longer history of use in IR. Regression conditions on X by calculating smooth hyperplanes that summarize the central tendency of the data at all levels of the variables in the regression. To make causal inference, a researcher simply compares the average distance between the conditional central tendency for the treated units and the conditional central tendency of the control units. Regression without matching works well for causal inference the assumptions listed above hold, and if the regression hyperplanes accurately reflect the central tendency of X; in other words, if the model fits well. However, matching offers several advantages over traditional regression.

Benefit 1: Matching automatically conditions on complex interactions between covariates. Linear regression can do this in principle, but in practice, most applied researchers specify linear, non-interactive regression models and only check model fit haphazardly. In a balanced matched sample, treatment effect estimates are not dependent on whether the analyst includes interactions and nonlinear terms in a subsequent regression (Ho et al 2007).

Benefit 2: Matching identifies and corrects issues of "common support," where some subset of treated or control units is so dissimilar from any units with other treatments that any inference about the outcomes of these units is determined almost entirely by modeling assumptions rather than data. Matching also draws attention to the related challenge of estimating the effects of categorical and continuous treatments. Linear regression obscures these challenges.

Benefit 3: Analysts have intuitions about cases, but regression makes it difficult for analysists to tell which cases are really being compared. Analysts can examine matched cases to directly assess the quality of their counterfactual comparisons.

Matching with Time-series Cross-sectional Data

IR researchers frequently use time-series cross-sectional (TSCS) data to estimate causal effects. With the exception of a few working papers (Nielsen and Sheffield 2009, Imai, Kim, and Wang 2018), there has been very little attention to TSCS data in the matching literature, leaving applied researchers with few guidelines. My survey of the applied matching literature in IR shows that 65% of IR articles that use matching use it on TSCS data, but only half of these do anything to account for the structure of the data, and few analysts defend their choices.

In this section, I explain the challenges of matching with TSCS data and describe current best practices. This is an area of active research (Imai, Kim, and Wang 2018), so future scholarship may offer amendments and improvements to the approach I endorse here.

To match TSCS data, IR scholars must be attentive to their unit of analysis. Outside of IR, matching is typically applied to cross-sectional data, in which each row in a data matrix records information about a single unit (e.g., a patient in a medical study). Analysts generally assume that these units are independent, so each row of the matrix is exchangeable. If this assumption holds, then every treated unit may be matched to any control unit without fear of violating the stable unit treatment value assumption or the assumption of conditional ignorability. Virtually all matching algorithms have been designed to take the data matrix provided by the analyst and match each treated row to the nearest available control row, without constraint. In this setting, each row of the data set matches the analyst's conceptual unit of analysis.

TSCS data are typically formatted in a "long" format matrix where each column is a variable and each row contains information about a given unit in a given time period, such as a country-year. These rows are no longer plausibly independent observations. Repeated observations of the same unit are probably correlated, as are observations of different units in the same time period. Applying standard matching algorithms to these data sets often results in matches that are likely to be dependent (e.g., if the country of Ghana received a treatment in the year 2005, the row of the data matrix with the most similar X values is likely to be Ghana in 2004). This is not necessarily bad. In fact, Imai and Kim (2019) show that estimates from a within-unit matched sample is equivalent to a popular fixed effects model. But generally, if countries (rather than country-years) are the analyst's unit of analysis, then I advise modifying the TSCS data matrix to correspond. For most TSCS matching applications, the fix is simple: transform the data matrix from "long" to "wide" format and then perform matching.

The cross-sectional data sets used in traditional matching applications can be thought of as TSCS datasets in "wide" format with only two time-periods: pretreatment (covariates) and posttreatment (outcome). Extending matching methods to data sets with a longer pretreatment time-series component merely requires appending this information in the right way.[8] Each prior time period simply adds to the available set of covariates on which to match treated and control units. Rather than representing this additional time-period with a new row in the data matrix X, represent it with a new column. Similarly, lagged values of the outcome variable should be represented as additional columns on the data matrix.

[8] There are many complicated issues involved with making causal inferences about time-varying (dynamic) treatments and treatment regimes (Murphy, van der Laan, and Robins 2001). But most IR scholars using matching have worked with static treatments, so I focus on how to extend the standard matching apparatus for static treatments to TSCS data.

Analysts may be interested in long-range outcomes. If so, long-range measurements of the outcome variable should be included in a single row as well. However analysts should not adjust for post-treatment covariates that might affect the outcome, because they might also be a result of treatment, and would then bias the estimated treatment effect. The difficulty of adjusting for post-treatment confounding means that inferences about immediate outcomes will be much more precise than inferences about long-range outcomes.

Several complications remain. First, should TSCS matching include every observed lag of every covariate? Can units provide multiple observations? And if so, how should treatment and control units be defined? Finally, should analysts remain concerned about interference between units inducing SUTVA violations? I discuss each of these issues in turn.

Should matching include every lag of every covariate?

With a TSCS set in "wide" format, analysts will suddenly feel that they have "lost" most of their data because they have exchanged a data matrix with many rows and few columns for one with few rows and many columns. This "loss" is illusory; the "N" of the data matrix in "long" format gives a highly inflated sense of the effective number of observations. The "wide" data matrix encodes the same information, but reflects dependencies between repeated observations of the same unit.

With TSCS data in "wide" format, the analyst may have more variables than observations. Suppose she observes 180 countries over, say, 74 years since the end of World War II and wishes to "control for" 20 variables (e.g., GDP, Democracy, Trade). Structured as country-years, the X matrix would have 13,320 observations and 20 variables. But reformatted to be "wide," the same data set now has 180 observations and 1,480 variables (e.g., GDP_{1946},…, GDP_{2019}; $Democracy_{1946}$,…, $Democracy_{2019}$; $Trade_{1946}$,…, $Trade_{2019}$). Most matching methods fail when the number of covariates exceeds the number of observations (Roberts et al 2018).

If a country receives treatment in 2015, does the analyst really need to consider each observation of each control variable back to the end of World War II? Probably not. Determining how many lags of a given covariate to use depends on how treatment is assigned. Analysts should choose the smallest set of lags that ensures conditional ignorability. Ideally, theory and prior evidence should guide this choice, but they are rarely precise enough to dictate whether $democracy_{t-4}$ is a confounder after including $democracy_{t-1}$, $democracy_{t-2}$, and $democracy_{t-3}$. Without strong theory to guide the choice about lags, analysts can reasonably turn to heuristics. For example, if the analyst assumes that treatment assignment is largely a function of recent events, she might match placing greater weight on the recent past while still placing a small weight on the distant past (see Nielsen 2016, 588 for an example).

One practical reason to omit unnecessary lags is that some are likely to be missing. Analysts can impute these missing values, but it complicates matching (D'Agostino Jr and Rubin 2000). If treatment assignment is not a function of these missing lags, then avoid the complication.

Can units provide multiple observations?

Considering each row of a "long" TSCS data set as an independent observation dramatically overstates the amount of information the data contain. But transforming it to "wide" format implicitly assumes that repeated observations are completely dependent, which is probably too conservative. There is often independent information in repeated observations of each unit that could be exploited.

This motivates some IR scholars to use a different unit of analysis – the country-block (see for example Simmons and Hopkins 2005, Hollyer and Rosendorf 2012, Nielsen and Simmons 2015). The analysts selects a number of l lags which they consider to be sufficient for conditional ignorability on all variables, and a number of m post-treatment periods for measuring the outcome. Each treatment block is defined by the l time-period observations before treatment and the m time periods after. Control units are also divided into blocks of $l+m$ consecutive observation periods. If the analyst wishes to match exactly on time, then time subscripts of treated and control blocks must match. If exactly matching time is unnecessary, then treated blocks can match control blocks with different time subscripts, greatly increasing the number of available controls. Each unit that never receives treatment can potentially offer multiple control blocks, rather than one. Additional control blocks can come from the pre-treatment life histories of units that eventually get treatment. In some IR data sets, almost every unit eventually receives treatment, so drawing from the pretreatment histories is the only source of control blocks.

Approximately 15% percent of the TSCS matching papers in IR have landed on this strategy, including the pioneering paper of Simmons and Hopkins (2005). Simmons and Hopkins use the following procedure. First, for each treated unit, they drop all lags except for the four years prior to treatment, the year of treatment, and the year after. They divide the control units into blocks of six consecutive observations, using both countries that never receive treatment and countries that receive treatment later. They then average the lagged covariates over the first four years of both treated and control blocks, and do a propensity score matching with these averages, reporting improvements in balance and, presumably, a reduction in bias.

The Potential for Violations of SUTVA

Many applications in international relations and comparative politics are likely to violate the stable unit treatment value assumption (SUTVA). This is not a problem induced by matching; matching merely illuminates a problem that standard TSCS regression techniques often overlook.

One obvious violation of SUTVA arises from the dependence between repeated observations of the same unit that I have just discussed; there may be "interference" between the observations. Causal inference with regression requires the same SUTVA assumptions as causal inference with matching, so it is puzzling that IR scholars sometimes wring their hands about matching repeated observations of the same unit to each other but then blithely throw them all into a pooled regression. If pooled matching makes the analyst uncomfortable, then pooled regression is inappropriate. Matching TSCS data in the "wide" format avoids the problem of dependence within repeated observation of a single unit, but interactions between units are also threaten to violate SUTVA. Unlike a medical study in which subjects can be isolated, it is not generally reasonable to assume that there is no interference between units in the International System. There is currently no widely accepted solution to this problem, though causal inference under networked interference is an active area of research (Aronow and Samii 2017).

Hidden versions of treatment are another other major source of SUTVA violations. Common TSCS practices for estimating the effects of "sticky" treatments can inadvertently create hidden versions of treatment. For example, when estimating the effect of a binary democracy variable on an outcome like trade flows, researchers typically estimate a model that compares each observed year of democracy to each year of non-democracy. However, the first year of democracy after democratization may not be equivalent to a year of mature democracy many years after democratization. Conceptualizing the treatment variable as a *transition to treatment* and using countries, rather than country-years, as the unit of analysis avoids this problem.

SUTVA is a strong assumption that may not be plausible in many TSCS applications. If so, TSCS matching will still mitigate model specification dependence but the results should probably not be interpreted causally. There has been very little research exploring when and how violations of SUTVA are problematic and methods for proceeding without SUTVA are in their infancy (Tchetgen Tchetgen and Vanderweele 2012). The standard practice for applied researchers facing possible SUTVA violations is to simply proceed as if they didn't exist. Rather than ignoring SUTVA violations entirely, first order violations of SUTVA may be avoided by following the advice above.

Estimation after Matching

What should analysts do after create a matched TSCS sample? Following Ho et al (2007), analysts may be able to proceed with the model they would have used on unmatched data. The data may need to be transformed back into "long" format for use with popular TSCS model software. However, analysts may realize in the course of matching that the pooled regression they initially wanted to estimate probably violates the assumptions above.

If so, reasonable approach is to estimate treatment effects using the difference-in-differences estimator (Keele, this volume). This estimator assumes that units observed have parallel over-

time trajectories prior to treatment. If so, differences in the differences of their outcome trajectories after treatment of some panels are the effect of treatment (hence the name). However, it is difficult to find settings where the parallel trends assumption is plausible. Heckman, Ichimura and Todd (1997) have suggested that researchers use matching to select units with similar trends prior to difference-in-differences estimation. The TSCS matching methods outlined above and in Imai, Kim, and Wang (2018) offer a way to perform this matching.

Match or Mismatch?

Will matching help me get better, more reliable answers to my research questions? This is a crucial question for any International Relations scholar considering using statistical matching in their work. With over a decade of matching under our belts, we can step back and evaluate its usefulness for IR. Potential criticism comes in two varieties: matching is generally problematic and matching is especially problematic when applied to the TSCS data sets common in IR.

To review, matching is a method for selecting cases to use in a subsequent statistical analysis. A matched sample is fully characterized by a data matrix of cases and a set of weights (generally derived from a matching algorithm) for each case. Observations with weights of zero are considered "pruned" from the data set, resulting in a data set that is more "balanced" than the original (meaning treated and control observations are more similar on a set of possible confounders represented by X). Once a matched sample is in hand, a researcher can typically use any appropriate regression technique to analyze the data without much modification.

As Geddes (1990) reminds us, the cases we choose affect the results we get. Because matching is case selection based on X, calculations based on the matched sample are conditional on X whether the subsequent estimation is explicitly conditional or not. This means that matching "controls" for X in ways similar to regression, with added flexibility for interactions and ensuring common support. It also means that statistical results from matched data will depend less on which specific regression model an analyst chooses than results from unmatched data (Ho et al 2007).

Critics of matching warn that it can produce misleading results (Arceneaux, Gerber, and Green 2006) and facilitate false discovery through p-hacking (Miller 2019). At seminars and conferences, I have seen critics scold matching advocates for overstating the benefits (and benefitting professionally from those overstatements!). It is fair to say that some proponents of matching (including myself) have at times been overly sanguine about the method, though the same could be said for virtually every statistical technique. Perceptions that researchers benefit professionally from applying matching in their research are also probably correct. Using the data set of IR articles I described above, I estimate that articles using matching have gotten 10 more

citations on average than comparable articles that did not use matching.[9] It is worth asking whether this additional attention is warranted.

The critics have a point. Matching is not a statistical truth serum to extract causal information from even the most recalcitrant data sets. There is no secret "causality" sauce in the inner workings of matching methods. Matching is for researchers who want to make causal inferences, but it won't necessarily allow them to do so. Matching cannot transform your unruly data into a neat and tidy experimental study. Moreover, existing matching methods have been used in an off-the-shelf manner not appropriate for the TSCS structure of many IR data sets. And it is unlikely that matching alleviates the risk of fishing for desired results despite early optimism on this score. Now that a decade has passed, my sense is that the critics of matching have successfully tempered these overly-optimistic claims about what matching can do.

Matching is not magic, but it is not all hype either. The upside of matching is that it is likely to help you understand your data better, and make clear to you and your readers what assumptions underlie your conclusions. Fitting a good regression model for causal inference can be very hard, and bad regression models abound. My experience is that matching makes the assumptions necessary for reliable causal inference more transparent to researchers. The intuition of matching is easier for many people than the intuition of regression, and understanding the intuition can help analysts avoid pitfalls. Matching has proven to be a reliably useful method for causal inference with messy, non-experimental data, precisely the sort that most of us in IR are used to dealing with. Although conditioning with regression can get the same result as conditioning with matching, I generally find that someone using matching is going to have a better sense for whether the counterfactuals implied by their model make any sense.

Yet, even if matching is useful for cross-sectional data analysis, it may be that the dependencies of time-series cross-sectional data are too challenging for existing matching methods to tackle. I acknowledge that the challenges are formidable. It is difficult to confidently assert that the stable unit treatment value assumption and conditional ignorability have been satisfied in any cross-national comparison study. But these assumptions are necessary for making credible causal inferences from regressions analysis without matching. Experiments are currently infeasible for many pressing IR questions and genuine natural experiments are rare. This leaves us with the necessity of extracting the most credible inferences we can from the observational data provided by history. Matching offers a transparent, useful framework for doing so while foregrounding the challenges.

[9] See the online Supplemental Information (Nielsen 2019) at https://doi.org/10.7910/DVN/HEFNHA.

So, where does this leave us? Matching is here to stay and rightfully belongs in the tool kit of IR scholars. It is no silver bullet. Like any other method, it can be misused, oversold, and misunderstood. It shouldn't be, if we care about learning the truth about international relations.

References

Abadie, Alberto, and Guido W. Imbens. "On the failure of the bootstrap for matching estimators." Econometrica 76.6 (2008): 1537-1557.

An, Weihua. "Bayesian Propensity Score Estimators: Incorporating Uncertainties in Propensity Scores into Causal Inference." Sociological Methodology 40.1 (2010): 151-189.

Arceneaux, Kevin; Gerber, Alan S.; Green, Donald P. (2006). "Comparing Experimental and Matching Methods Using a Large-Scale Field Experiment on Voter Mobilization". Political Analysis. 14 (1): 37–62.

Arceneaux, Kevin; Gerber, Alan S.; Green, Donald P. (2010). "A Cautionary Note on the Use of Matching to Estimate Causal Effects: An Empirical Example Comparing Matching Estimates to an Experimental Benchmark". Sociological Methods & Research. 39 (2): 256–282.

Aronow, Peter M., and Cyrus Samii. "Does regression produce representative estimates of causal effects?" American Journal of Political Science 60, no. 1 (2016): 250-267.

Aronow, Peter M., and Cyrus Samii. "Estimating average causal effects under general interference, with application to a social network experiment." The Annals of Applied Statistics 11.4 (2017): 1912-1947.

Bowers, Jake, and Thomas Leavitt. "Causal-Inference & Design-Based Inferential Methods". This volume.

Carter, Christopher L., and Thad Dunning. "Instrumental variables: From structural equation models to design-based causal inference." This volume.

Cattaneo, Matias D., Rocío Titiunik, and Gonzalo Vazquez-Bare. "The Regression Discontinuity Design." This volume.

D'Agostino Jr, Ralph B., and Donald B. Rubin. "Estimating and using propensity scores with partially missing data." Journal of the American Statistical Association 95, no. 451 (2000): 749-759.

Diamond, Alexis, and Jasjeet S. Sekhon. "Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies." Review of Economics and Statistics 95, no. 3 (2013): 932-945.

Findley, Michael G., Daniel L. Nielson, and Jason C. Sharman. "Using field experiments in international relations: A randomized study of anonymous incorporation." International Organization 67, no. 4 (2013): 657-693.

Geddes, Barbara. "How the cases you choose affect the answers you get: Selection bias in comparative politics." Political analysis 2 (1990): 131-150.

Gerring, John. "What is a case study and what is it good for?" American political science review 98, no. 2 (2004): 341-354.

Hansen, Ben B. "Full matching in an observational study of coaching for the SAT." Journal of the American Statistical Association 99.467 (2004): 609-618.

Hansen, Ben B., and Jake Bowers. "Covariate balance in simple, stratified and clustered comparative studies." Statistical Science (2008): 219-236.

Heckman, James, Hidehiko Ichimura and Petra Todd. 1997. "Matching as an Econometric Evaluation Estimator." Review of Economic Studies 65 (2):261–294.

Hirano, Keisuke, and Guido W. Imbens. "The propensity score with continuous treatments." Applied Bayesian modeling and causal inference from incomplete-data perspectives 226164 (2004): 73-84.

Ho, Daniel E., Kosuke Imai, Gary King, and Elizabeth A. Stuart. "Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference." Political Analysis 15, no. 3 (2007): 199-236.

Holland, Paul W. "Statistics and causal inference." Journal of the American Statistical Association 81, no. 396 (1986): 945-960.

Hollyer, James R., and B. Peter Rosendorff. "Leadership survival, regime type, policy uncertainty and PTA accession." International Studies Quarterly 56.4 (2012): 748-764.

Iacus, Stefano M., Gary King, and Giuseppe Porro. "Multivariate matching methods that are monotonic imbalance bounding." Journal of the American Statistical Association 106, no. 493 (2011): 345-361.

Iacus, Stefano M., Gary King, and Giuseppe Porro. "Causal inference without balance checking: Coarsened exact matching." Political analysis 20, no. 1 (2012): 1-24.

Iacus, Stefano M., Gary King, and Giuseppe Porro. "A theory of statistical inference for matching methods in causal research." Political Analysis 27.1 (2019): 46-68.

Imai, Kosuke, and Marc Ratkovic. "Covariate balancing propensity score." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76, no. 1 (2014): 243-263.

Imai, Kosuke, and In Song Kim. "When Should We Use Unit Fixed Effects Regression Models for Causal Inference with Longitudinal Data?" American Journal of Political Science (2019).

Imai, Kosuke, In Song Kim, and Erik Wang. 2018. "Matching Methods for Causal Inference with Time-Series Cross-Section Data."

Imai, Kosuke, and David A. Van Dyk. "Causal inference with general treatment regimes: Generalizing the propensity score." Journal of the American Statistical Association 99.467 (2004): 854-866.

Imbens, Guido W. "Better LATE than nothing: Some comments on Deaton (2009) and Heckman and Urzua (2009)." Journal of Economic literature 48.2 (2010): 399-423.

King, Gary, Christopher Lucas, and Richard A. Nielsen. "The balance-sample size frontier in matching methods for causal inference." American Journal of Political Science 61, no. 2 (2017): 473-489.

Keele, Luke. "Difference-in-Differences: Neither Natural nor an Experiment." This volume.

King, Gary, and Richard Nielsen. Forthcoming. "Why propensity scores should not be used for matching." Political Analysis.

Maliniak, Daniel, Susan Peterson, Ryan Powers, and Michael J. Tierney (2018). "Is International Relations a Global Discipline? Hegemony, Insularity, and Diversity in the Field." Security Studies, 27 (3): 448-484.

Marshall, Monty G., Keith Jaggers, and Ted Robert Gurr. "Polity IV project: Dataset users' manual." College Park: University of Maryland (2002).

Miller, Michael K. “The Uses and Abuses of Matching in Political Science.” Unpublished. https://sites.google.com/site/mkmtwo/Miller-Matching.pdf (accessed 2 July, 2019).

Murphy, Susan A., Mark J. van der Laan, James M. Robins, and Conduct Problems Prevention Research Group. "Marginal mean models for dynamic regimes." Journal of the American Statistical Association 96, no. 456 (2001): 1410-1423.

Nielsen, Richard A. "Case selection via matching." Sociological Methods & Research 45, no. 3 (2016): 569-597.

Nielsen, Richard, 2019, "Supplemental Information for: "Statistical Matching with Time-Series Cross-Sectional Data: Magic, Malfeasance, or Something in Between?"", https://doi.org/10.7910/DVN/HEFNHA, Harvard Dataverse.

Nielsen, Richard A., and Beth A. Simmons. "Rewards for Ratification: Payoffs for Participating in the International Human Rights Regime?." International Studies Quarterly 59.2 (2015): 197-208.

Nielsen, Richard, and John Sheffield. "Matching with time-series cross-sectional data." Polmeth XXVI. Yale University (2009).

Otsu, Taisuke, and Yoshiyasu Rai. "Bootstrap inference of matching estimators for average treatment effects." Journal of the American Statistical Association 112.520 (2017): 1720-1732.

Roberts, Margaret E., Brandon M. Stewart, and Richard Nielsen. Adjusting for Confounding with Text Matching. Working paper, 2018.

Rosenbaum, Paul R. "Optimal matching for observational studies." Journal of the American Statistical Association 84, no. 408 (1989): 1024-1032.

Rosenbaum, Paul R., and Donald B. Rubin. "The central role of the propensity score in observational studies for causal effects." Biometrika 70, no. 1 (1983): 41-55.

Rosenbaum, Paul. R., and Donald. B. Rubin. 1985. “The Bias Due to Incomplete Matching.” Biometrics 41(1):103–116.

Rubin, Donald B. "Matching to remove bias in observational studies." Biometrics (1973): 159-183.

Rubin, Donald B. "Bias reduction using Mahalanobis-metric matching." Biometrics (1980): 293-298.

Rubin, Donald B. Matched sampling for causal effects. Cambridge University Press, 2006.

Rubin, Donald B., and Neal Thomas. "Matching using estimated propensity scores: relating theory to practice." Biometrics (1996): 249-264.

Sekhon, Jasjeet S. "Opiates for the matches: Matching methods for causal inference." Annual Review of Political Science 12 (2009): 487-508.

Simmons, Beth A., and Daniel J. Hopkins. "The constraining power of international treaties: Theory and methods." American Political Science Review 99.4 (2005): 623-631.

Stuart, Elizabeth A. "Matching methods for causal inference: A review and a look forward." Statistical science: a review journal of the Institute of Mathematical Statistics 25.1 (2010): 1.

Tarrow, Sidney. "The strategy of paired comparison: toward a theory of practice." Comparative political studies 43.2 (2010): 230-259.

Tchetgen Tchetgen, Eric J., and Tyler J. VanderWeele. "On causal inference in the presence of interference." Statistical methods in medical research 21.1 (2012): 55-75.

Teaching, Research, and International Policy Project. (2017). TRIP Journal Article Database Release (Version 3.1). Available at https://trip.wm.edu/.

VanderWeele, Tyler J. 2010. "Direct and Indirect Effects for Neighborhood-Based Clustered and Longitudinal Data." Sociological Methods and Research 38 (4):515–544.

Weeks, Ana Catalano. "Why Are Gender Quota Laws Adopted by Men? The Role of Inter-and Intraparty Competition." Comparative Political Studies (2018).

Zubizarreta, José R., Ricardo D. Paredes, and Paul R. Rosenbaum. "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 8.1 (2014): 204-231.