Constructing Generalizable Geographic Natural Experiments

Journal Title
XX(X):1–8

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Owura Kuffuor, Giancarlo Visconti, and Kayla Young[1]

Abstract

A natural experiment is a real-world situation that generates as-if random or haphazard assignment to treatment. Geographic or administrative boundaries can be exploited as natural experiments to construct treated and control groups. Previous research has demonstrated that matching can help enhance these designs by reducing imbalances on observed covariates. An important limitation of this empirical approach, however, is that the results are inherently local. While the treated and control groups may be quite similar to each other, they could be substantially different from the target population of interest (e.g., a country). We propose a simple design inspired by the idea of template matching to construct generalizable geographic natural experiments. By matching our treated and control groups to a template (i.e., the target population), we obtain groups that are similar to the target population of interest and to each other, which can increase both the internal and external validity of the study.

Keywords

Generalizability, Geographic Natural Experiments, Matching

Introduction

Natural experiments offer a unique opportunity to explore some of the most elusive questions about the political world, characterized by circumstances that allow researchers to assume as-if random or haphazard treatment conditions even though treatment allocation is not defined by a random device (Rosenbaum 2010; Dunning 2012; Keele 2015). One of the most popular types of natural experiments uses geographic or administrative boundaries to construct treated and control groups by exploiting certain geographical features that generate as-if random variation in the treatment assignment (Keele and Titiunik 2016). Such geographic natural experiments (GNEs) have been used to study topics such as ethnic relations in Zambia and Malawi (Posner 2004), political polarization in the United States (Nall 2018), and support for authoritarian regimes in East Germany (Kern and Hainmueller 2009) that might otherwise escape attempts to establish causal inference.

However, as with any methodological approach, there are important limitations to (geographic) natural experiments. For instance, because randomization is not guaranteed, researchers must provide a compelling justification for the as-if random assumption (Dunning 2012; Sekhon and Titiunik 2012). Even with empirical and theoretical justification, though, "the strong possibility that unobserved differences across groups may account for difference in average outcomes is always omnipresent in observational studies" (Dunning 2008, 289). This concern may obscure important relationships and even undermine the validity of causal claims.

The local geographic ignorability design (LGID) therefore emerges as an attractive empirical approach to limit potential unobservable factors from biasing results. Under the assumption that the treatment was as-if randomly assigned to units that are especially close to a given geographic or administrative boundary, there can be greater confidence in the assumed independence of potential outcomes (Keele and Titiunik 2016).[1] For example, when studying the effects of a policy intervention, a LGID would examine differences between residents living within a small buffer area from the border. Presumably, these residents would be more similar to each other than to those living far away from

[1]Department of Political Science, Purdue University, West Lafayette, IN, USA

Corresponding author:
Giancarlo Visconti Department of Political Science, Purdue University, 100 N. University Street, West Lafayette, IN, 47907, USA.
Email: gviscont@purdue.edu

the administrative boundary. The LGID approach, however, might still require adjustment for pretreatment covariates. One solution to this problem is to enhance geographic designs by using matching as a flexible form of statistical adjustment (Keele et al. 2015).[2]

While the latter is an undoubtedly powerful research design, it is also accompanied by an important limitation- its inherent locality. Although matched treated and control groups may, in fact, be quite similar to each other, they could also be markedly distinct from a larger population of interest (e.g., a city or state). Given an unrepresentative sample, "the estimate of a causal effect may fail to characterize how effects operate in the population of interest" (Aronow and Samii 2016, 250). Such external validity concerns are often of particular interest for political scientists (McDermott 2002), as it may be difficult to determine whether any causal effects identified must be restricted to only areas within the narrowly defined boundary or if they can be generalized across cases to answer fundamental questions about broader political phenomena.

We present an approach that addresses this problem, inspired by the idea of template matching (Silber et al. 2014) as well as by recent advances in optimal matching and the construction of representative matched samples (Visconti and Zubizarreta 2018; Bennett et al. 2019). Using a target population as a template to implement the matching, such as a city, state, or country, matched treated and control groups will not only be similar to each other but also similar to the population of interest. This can increase the generalizability of causal evidence from GNEs, providing a kind of external validity check. By implementing this method, researchers would not have to only rely on collecting multiple studies conducted in diverse contexts to learn about the generalizability of an effect since template matching reveals the hidden studies that resemble other populations within the original study. We see this strategy as a second step to be implemented after the main analysis to explore whether results are consistent across samples that look like the populations of interest. In the following sections we describe the assumptions and the methodology for this approach and provide an empirical illustration.

Notation and Assumptions

When using a sample to draw causal inference, the evidence can be generalized to a target population only when that sample was randomly selected from the target population of interest. In the case of geographic natural experiments, the sample (e.g., the buffer from either side of the administrative boundary) is not constructed by randomly selecting people from the target population (e.g., the city). As a consequence, generalizability efforts must rely on an observational data analysis assumption (Stuart et al. 2018).

In randomized experiments, the most common quantity of interest is the average treatment effect (ATE). Let $Y_i(1)$ denote the potential outcome if subject i were treated and $Y_i(0)$ if subject i were not treated. The average treatment effect or ATE $= E(Y_i(1)) - E(Y_i(0))$. In observational studies, the estimand of interest is usually the average treatment effect on the treated (ATT), which can be expressed as: ATT $= E(Y_i(1)|T_i = 1) - E(Y_i(0)|T_i = 1)$. The counterfactual control units are not observed. As a result, it is necessary to construct a control group by using two assumptions: conditional independence and common support (Hidalgo and Sekhon 2011).

In this paper, we instead focus on a different estimand: the target average treatment effect on the treated (TATT), which will inform us about how the treatment effects operate on the target population of interest. In a sample of n units, TATT = $= \frac{1}{n}\Sigma_{i=1}^{n}(E(Y_i(1)|T_i = 1) - E(Y_i(0)|T_i = 1))$. In this case, the sample of n units needs to resemble the target population of interest.

We propose a design based on template matching to extend beyond the local effects estimated when using local geographic ignorability designs and to recover the target average treatment effect on the treated (TATT). Template matching was developed by Silber et al. (2014) to make standardized comparisons based on observed characteristics. Their study randomly selected 300 patients (i.e., the template) and used them to match 300 patients at 217 hospitals, constructing a sample that resembled the template used to implement the multivariate matching.

Two assumptions are needed to claim that the matched sample resembles the population of interest and to provide causal evidence after adjusting on observables. The first, the *ignorability of sample selection*, states that after adjusting for the relevant observed covariates, treatment effects are the same in the matched sample and the target population (Visconti and Zubizarreta 2018; Stuart et al. 2018). Specifically, the target average treatment effect on the treated (TATT) and the population (of interest) average treatment effect on the treated (PATT) need to be equivalent. In that case, we expect that $\frac{1}{n}\Sigma_{i=1}^{n}(E(Y_i(1)|T_i = 1) - E(Y_i(0)|T_i = 1)) = E(Y_i(1)|T_i = 1) - E(Y_i(0)|T_i = 1)$. Second, the *conditional geographic ignorability in local neighborhood assumption*, holds that within a neighborhood the potential outcomes are independent of treatment assignment conditional on observed covariates (Keele et al. 2015). In this case,

every unit i has a score defined $S_j = (S_{j1}, S_{j2})$ that refers to the geographic location of the subject, which will be used to compute the distance to any point (b_1, b_2) located on the boundary. A collection of points within a small geographic neighborhood is defined as $N(b_1, b_2)$. The set of covariates used to obtain covariate balance is defined as X_i. Therefore, for each point (b_1, b_2) located on the boundary, we can find a neighborhood $N(b_1, b_2)$ where $(Y_i(1), Y_i(0)) \perp\!\!\!\perp T_i | X_i$ for all subjects i with score (S_{j1}, S_{j2}) in $N(b_1, b_2)$.

A key question is how to define what is the appropriate template or target population. Recent research has advocated for a stronger connection between theory and causal identification. Scholars point to the advantages of theory-driven endeavors, which can help to better recognize undefined potential outcomes (Slough 2022), to improve covariate balance (Resa and Zubizarreta 2016), and to generalize a causal effect to other contexts (Gailmard et al. 2021). While the nature of causal identification strategies may require a narrow focus, the theories researchers wish to test may be far more extensive. When constructing a generalizable geographic natural experiment, we argue that researchers should ask not only what identification strategy is best to recover causal effects, but also what template or population they wish to mimic that would best test their broader theory.

For example, Posner (2004) takes advantage of the border between Zambia and Malawi to study the political salience of a cultural cleavage. Chewa and Tumbuka people live on both sides of the border. While their cultural differences are identical on both sides of the border, their political differences are more salient in Malawi than Zambia. The rationale behind exploiting this distinction is that Chewas and Tumbukas are large groups relative to the country as a whole in Malawi and, therefore, can be used as a base for coalition-building. Meanwhile, in Zambia, Chewas and Tumbukas are small relative to the country as a whole, creating little incentive to rely on them for coalition-building.

As a result, Posner (2004)'s theory directly connects with a population of interest (i.e., the entire Chewa and Tumbuka people in Malawi and Zambia) rather than just four villages along the border used in the study. If people from these villages have different distributions of observed characteristics than in the entire country,[3] using a traditional geographical experiment might generate estimates that do not speak to the theory. Thus, we would advocate implementing a generalizable geographic natural experiment to improve the connection between theory and causal identification.

The utility of using template matching is also evidenced in more recent implementations of geographic natural experiments. Keele and Titiunik (2018) aim to uncover the effects of all-mail voting on turnout. To do so, they rely on data from two counties, one that used only in-person voting and one that used all-mail voting. While the resulting estimates can tell us about turnout effects at a local scale, they may not be able to extend to the true populations of interest, Colorado, and even the United States as a whole. Employing template matching in this case would provide a kind of external validity check on how well the theory underlying the paper connects with the analysis and results of the causal identification strategy.

It is important to note that we do not equate external validity and representativeness. Our goal is to show that a treatment effect can be generalized across different populations of interest (i.e., external validity). We use template matching to construct representative matched samples that are similar to the population of interest (i.e., representativeness). Using template matching to build representative matched samples can improve the limited external validity of studies that have an especially local nature, often a result of researchers' efforts to reduce heterogeneity and decrease sensitivity to hidden biases (Rosenbaum 2005). In observational studies, reducing heterogeneity often means decreasing the sample size to improve comparability between units (Keele 2015). Therefore, we could end up with a treated and control group that allows us to make credible inferences but that might be substantially different from the target population.

Method

To implement template matching, we use mean balance constraints, with the goal of reducing the standardized differences or difference-in-means in standard deviation units between the treated and control groups. Though stricter balance constraints, such as fine balance, can also be used.[4] In this case, we use matching to restrict the standardized differences (i) between the treated group and our target population and (ii) between the control group and our target population to be no larger than 0.05 pooled standard deviations. This ensures that the standard deviations between the matched treated and control groups cannot be larger than 0.1: a traditional threshold used in the literature to demonstrate covariate balance (see Zubizarreta (2012) and Pimentel et al. (2015) for examples).

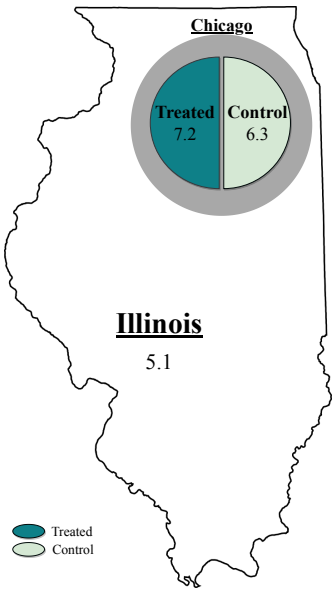
To generate covariate balance, we use cardinality matching, which allows for different types of balance, such

as aggregate balance of low-dimensional joint distributions, marginal distributions, and moments such as the means, among other forms (Visconti and Zubizarreta 2018). Even though we recommend cardinality matching because it maximizes the size of the matched sample based on flexible constraints on covariate balance, we acknowledge that template matching can also be implemented using other matching techniques such as genetic matching (Diamond and Sekhon 2013) or matching frontier (King et al. 2017), or by using weighting approaches such as entropy balance (Hainmueller 2012) or minimal weights (Wang and Zubizarreta 2020).

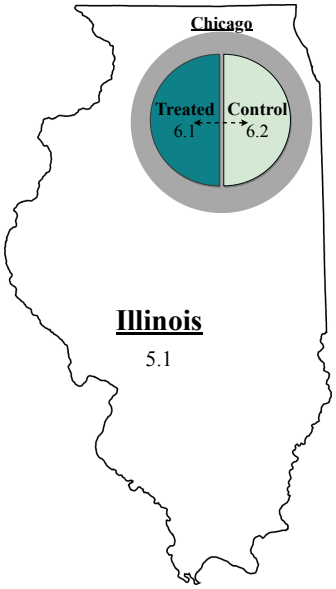
As an illustration of the structure of the design, the first panel of Figure 1 depicts a state that contains a group of hypothetical people living to the east and another group to the west of a geographic boundary within the city of Chicago. If we are interested in covariate x, the average for the treatment group is 7.2, which is quite different from an average of 6.3 in the control group. Following Keele et al. (2015)'s approach, some adjustment for the covariate x may then be necessary.

The second panel uses a regular matching approach to decrease imbalances in the observed covariate x. Even though the standardized differences between the matched treatment and control groups are balanced in terms of covariate x, they vary considerably from the state average of 5.1 for that same observed characteristic. Therefore, while we may be able to make a credible claim about a relationship between the matched treated and control groups at the boundary, we nonetheless cannot make more generalizable inferences beyond the border and, certainly, not about the state in its entirety. This reduces the external validity of our hypothetical study.

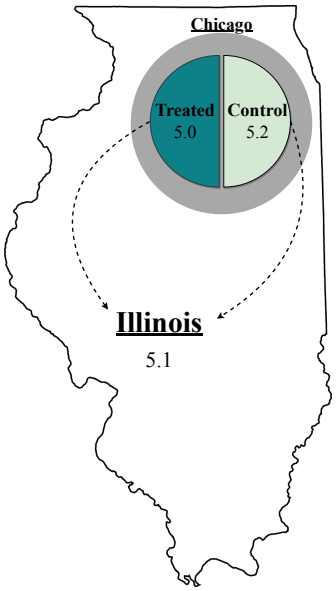
We further illustrate our approach, using Illinois as a template, in the third panel. We choose how the standardized differences between the treated group and the state should be restricted, making them no larger than 0.05 pooled standard deviation unitsthe same for the difference between the control and the state. Therefore, by construction, the treatment and control groups cannot have imbalances greater than 0.1 pooled standard deviation units.



(a) No Matching



(b) Regular Matching



(c) Template Matching

Figure 1. Different types of matching

Table 1. Before matching

Covariates	Treated	Control
Housing prices	$157,823.10	$151,354.50
Turnout 2004	0.81	0.80
Turnout 2006	0.59	0.57
Male	0.45	0.48
Age	40.67	40.41
Observations	73052	15965

As we have seen, template matching provides a flexible approach for making multiple estimations based on the target population of interest. The standard matching approach, in contrast, achieves balance between the matched treated and control groups, but cannot necessarily be used to make inferences outside of a given administrative or geographic boundary.

Example

We provide a more concrete illustration of our approach by extending Keele et al. (2015)'s study on the role of ballot initiatives in voter turnout. The authors draw on a natural experiment in the city of Milwaukee, Wisconsin, wherein a ballot initiative was established in 2008 but not implemented in the seventeen surrounding areas. This draws on a local geographic ignorability design, where units "within a narrow band around the border are assumed to be good counterfactuals for each other" (Keele and Titiunik 2016, 3), and offers a unique opportunity to understand whether such initiatives do, in fact, foster greater political participation. Keele et al. (2015) do not find enough evidence to claim that the ballot initiative has increased turnout.

Table 1 reports the averages for the treated and control groups within a 1000 meter buffer of either side of the boundary. To compare the treatment group (i.e., residents with the ballot initiative) to the control group in their original study, Keele et al. (2015) balance on observable characteristics: age, gender, voting history, and housing values (see appendix A for details).[5] While the unmatched treated and control groups are very similar to each other for four out of five covariates, the difference for housing prices is greater than 0.1 standard deviations. We use cardinality matching, which finds the largest matched sample that achieves the covariate balance requirements imposed by the researchers (Zubizarreta et al. 2014), to reduce imbalances. In this case, standardized differences cannot be larger than 1/10th standard deviations. Table 2 summarizes the results after matching and also compares the new means with the means for the city, state, and country (see appendix B for all of the standardized differences before and after matching).

As expected, matching generates balance for all of the covariates. This provides a compelling way to improve causal inferences by combining a natural experiment based on geography and matching to improve covariate balance. However, the results might be highly local and not necessarily reflect the effects of initiatives on voter turnout for an average American citizen or even a typical resident of Milwaukee, Wisconsin. In fact, considering the averages for the city, state, and country shown in Table 2, the matched treated and control groups do not look, on average, very similar to these potential populations of interest.

To address this concern, in this paper we combine a geographic natural experiment with template matching. Template matching is critical for achieving covariate balance not only between the treated and control groups but also with a given a target population. We use as a template the city (Milwaukee), the state (Wisconsin), and the country (United States) to match with the treated and control groups (see appendix C for more information on the city-, state-, and country-level characteristics we use).

We report results using the state as the template for the matching in table 3 (see appendix D for the results using Milwaukee and the United States as templates). This demonstrates that the matched treated and control groups are not just similar to each other, but also similar to Wisconsin. Finally, in appendix E we provide an illustration about how to implement a generalizable geographic natural experiment using cardinality matching in *R*.

We use a permutational t-test in matched pairs with an embedded sensitivity analysis (Rosenbaum 2015).[6] The parameter Γ represents the odds of differential assignment to the treatment due to an unobserved factor u. $\Gamma = 1$ means that two subjects in a matched pair have the same probability of getting the treatment (i.e., there are no hidden biases). $\Gamma = 1.1$ means that in a matched pair, one of the subjects is 1.1 more likely than the other to get the treatment because of an unmeasured covariate u. We provide the point estimates (when $\Gamma = 1$) and the p-values for different values of Γ.

Table 4 shows that results do not stand to a $\Gamma = 1.2$ for any of the studies. These findings illustrate that there is not enough evidence to claim that ballot initiatives can increase participation since the conclusions are sensitive to small biases and the point estimate changed direction in one of the studies. Since we understand external validity as being able to hold the conclusions of the study in other contexts or populations, the evidence when using template matching improves the external validity of the findings reported using regular matching. There is not strong evidence in any of the four analyses to claim that the ballot initiative has increased

Table 2. After regular matching

Covariates	Treated	Control	City	State	Country
Housing prices	$153,737.90	$151,354.50	$152,996	$162,407	$214,546
Turnout 2004	0.78	0.80	0.70	0.73	0.64
Turnout 2006	0.55	0.57	0.57	0.51	0.48
Male	0.46	0.48	0.48	0.49	0.49
Age	39.80	40.41	30	36	35.30
Observations	15965	15965			

Table 3. After template matching

Covariates	Treated	Control	State
Housing prices	$160,108.40	$160,010.60	$162,407
Turnout 2004	0.75	0.75	0.73
Turnout 2006	0.53	0.53	0.51
Male	0.47	0.48	0.49
Age	36.87	36.86	36
Observations	9924	9924	

turnout, which provides extra support to Keele et al. (2015)'s main findings.

Conclusion

Natural experiments can be powerful designs for exploring causal relationships that might otherwise be confined to correlations. Local geographic ignorability designs, for example, allow researchers to focus on differences between treatment and control groups that are in close proximity to one another (Keele and Titiunik 2016), and recent methodological advances have blended designs based on geographic distance and covariate balance (Keele et al. 2015). Although this approach has increased internal validity because the treatment has a justifiably as-if random or haphazard nature, it is highly local by design and may raise concerns about external validity. We therefore suggest a generalizable natural experiment approach, using template matching as a solution to ensure that the matched sample is similar to the population of interest. This combines both the strong internal validity of LGIDs with an external validity check to provide insight into how generalizable the results may be to other contexts. Additionally, we recommend that theory or previous knowledge is used to define the appropriate template or target population.

While we focus on LGID designs here, it is important to note that this approach could be extended to other designs such as more standard natural experiments that do not rely on buffer zones or can be implemented using other adjustment techniques such as weighting method to obtain covariate balance. We believe that the main implication of this design is to allow for credible inferences while characterizing how the results can operate in other populations of interest.

Acknowledgements

Authors are listed in alphabetical order. We thank Tom Leavitt, Jay McCann, Tara Slough, Logan Strother, José Zubizarreta, and anonymous reviewers for valuable comments and suggestions. All errors are our own.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Supplemental material

Supplemental material for this article is available online.

Notes

1. An alternative design is a geographic regression discontinuity design (GRDD), which relies on a continuity assumption (Keele and Titiunik 2015).
2. In appendix F, we provide examples of studies published using different types of natural experiments based on geography.
3. This is presented solely as an illustrative example and may not be the case in this circumstance.
4. Researchers may use stricter balance requirements when they expect a non-linear relationship between the covariate and the outcome (e.g., U-shaped), since constraining the mean in a case like that is not a meaningful decision (Resa and Zubizarreta 2016; Visconti and Zubizarreta 2018).
5. The main source of data is the Wisconsin voter file.
6. To address possible biases from unmeasured covariates, we recommend the use of a sensitivity analysis to assess how sensitive our findings are to the incorporation of hidden confounders that change the odds of treatment assignment.

References

Aronow PM and Samii C (2016) Does regression produce representative estimates of causal effects? *American Journal of Political Science* 60(1): 250–267.

Table 4. Permutational t-test and sensitivity analysis

Type of matching	Point estimate	P-value Γ=1	P-value Γ=1.1	P-value Γ=1.2
Regular matching	0.02	0.00	0.23	0.99
Template matching: city	0.03	0.00	0.02	0.75
Template matching: state	0.02	0.00	0.54	1.00
Template matching: country	-0.03	0.98	1.00	1.00

Bennett M, Vielma JP and Zubizarreta JR (2019) Building representative matched samples with multi-valued treatments in large observational studies. *Working paper, Harvard University* .

Diamond A and Sekhon JS (2013) Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies. *Review of Economics and Statistics* 95(3): 932–945.

Dunning T (2008) Improving causal inference: Strengths and limitations of natural experiments. *Political Research Quarterly* 61(2): 282–293.

Dunning T (2012) *Natural Experiments in the Social Sciences: A Design-Based Approach*. Cambridge: Cambridge University Press.

Gailmard S et al. (2021) Theory, history, and political economy. *Journal of Historical Political Economy* 1(1): 69–104.

Hainmueller J (2012) Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political analysis* 20(1): 25–46.

Hidalgo FD and Sekhon JS (2011) Causality. In: Badie, Bertrand, Berg-Schlosser, Dirk and Morlino, Leonardo (ed.) *International Encyclopedia of Political Science*. Thousand Oaks: SAGE Publications, Inc.

Keele L (2015) The statistics of causal inference: A view from political methodology. *Political Analysis* 23(3): 313–335.

Keele L and Titiunik R (2016) Natural experiments based on geography. *Political Science Research and Methods* 4(01): 65–95.

Keele L and Titiunik R (2018) Geographic natural experiments with interference: The effect of all-mail voting on turnout in colorado. *CESifo Economic Studies* 64(2): 127–149.

Keele L, Titiunik R and Zubizarreta JR (2015) Enhancing a geographic regression discontinuity design through matching to estimate the effect of ballot initiatives on voter turnout. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 178(1): 223–239.

Keele LJ and Titiunik R (2015) Geographic boundaries as regression discontinuities. *Political Analysis* 23(1): 127–155.

Kern HL and Hainmueller J (2009) Opium for the masses: How foreign media can stabilize authoritarian regimes. *Political Analysis* 17(4): 377–399.

King G, Lucas C and Nielsen RA (2017) The balance-sample size frontier in matching methods for causal inference. *American Journal of Political Science* 61(2): 473–489.

McDermott R (2002) Experimental methods in political science. *Annual Review of Political Science* 5(1): 31–61.

Nall C (2018) *The Road to Inequality: How the Federal Highway Program Polarized America and Undermined Cities*. Cambridge: Cambridge University Press.

Pimentel SD, Kelz RR, Silber JH and Rosenbaum PR (2015) Large, sparse optimal matching with refined covariate balance in an observational study of the health outcomes produced by new surgeons. *Journal of the American Statistical Association* 110(510): 515–527.

Posner DN (2004) The political salience of cultural difference: Why chewas and tumbukas are allies in zambia and adversaries in malawi. *American Political Science Review* 98(4): 529–545.

Resa M and Zubizarreta JR (2016) Evaluation of subset matching methods and forms of covariate balance. *Statistics in medicine* 35(27): 4961–4979.

Rosenbaum PR (2005) Heterogeneity and causality: Unit heterogeneity and design sensitivity in observational studies. *The American Statistician* 59(2): 147–152.

Rosenbaum PR (2010) *Design of Observational Studies*. Cham: Springer.

Rosenbaum PR (2015) Two r packages for sensitivity analysis in observational studies. *Observational Studies* 1: 1–17.

Sekhon JS and Titiunik R (2012) When natural experiments are neither natural nor experiments. *American Political Science Review* 106(01): 35–57.

Silber JH, Rosenbaum PR, Ross RN, Ludwig JM, Wang W, Niknam BA, Mukherjee N, Saynisch PA, Even-Shoshan O, Kelz RR et al. (2014) Template matching for auditing hospital cost and quality. *Health services research* 49(5): 1446–1474.

Slough T (2022) Phantom counterfactuals. *American Journal of Political Science Forthcoming* .

Stuart EA, Ackerman B and Westreich D (2018) Generalizability of randomized trial results to target populations: design and analysis possibilities. *Research on social work practice* 28(5): 532–537.

Visconti G and Zubizarreta JR (2018) Handling limited overlap in observational studies with cardinality matching. *Observational*

Studies 4: 217–249.

Wang Y and Zubizarreta JR (2020) Minimal dispersion approximately balancing weights: asymptotic properties and practical considerations. *Biometrika* 107(1): 93–105.

Zubizarreta JR (2012) Using mixed integer programming for matching in an observational study of kidney failure after surgery. *Journal of the American Statistical Association* 107(500): 1360–1371.

Zubizarreta JR, Paredes RD and Rosenbaum PR (2014) Matching for balance, pairing for heterogeneity in an observational study of the effectiveness of for-profit and not-for-profit high schools in chile. *The Annals of Applied Statistics* 8(1): 204–231.