Matching Methods for Causal Inference with Time-Series Cross-Sectional Data*

Kosuke Imai† In Song Kim‡ Erik Wang§

First Draft: April 28, 2018
This Draft: September 7, 2020

Abstract

Matching methods improve the validity of causal inference by reducing model dependence and offering intuitive diagnostics. While they have become a part of the standard tool kit across disciplines, matching methods are rarely used when analyzing time-series cross-sectional data. We fill this methodological gap. In the proposed approach, we first match each treated observation with control observations from other units in the same time period that have an identical treatment history up to the pre-specified number of lags. We use standard matching and weighting methods to further refine this matched set so that the treated and matched control observations have similar covariate values. Assessing the quality of matches is done by examining covariate balance. Finally, we estimate both short-term and long-term average treatment effects using the difference-in-differences estimator, accounting for a time trend. We illustrate the proposed methodology through simulation and empirical studies. An open-source software package is available for implementing the proposed methods.

Key Words: difference-in-differences, fixed effects, observational studies, unobserved confounding, weighting

*The methods described in this paper can be implemented via an open-source statistical software package, PanelMatch: Matching Methods for Causal Inference with Time-Series Cross-Sectional Data, available at `https://CRAN.R-project.org/package=PanelMatch`. We thank Adam Rauh for superb research assistance. Thanks also go to Neal Beck, Matt Blackwell, David Carlson, Robert Franzese, Paul Kellstedt, Anton Strezhnev, Vera Troeger, James Raymond Vreeland, and Yiqing Xu who provided useful comments and feedback.

†Professor, Department of Government and Department of Statistics, Harvard University. 1737 Cambridge Street, Institute for Quantitative Social Science, Cambridge MA, 02138. Phone: 617–384–6778, Email: Imai@Harvard.Edu, URL: https://imai.fas.harvard.edu

‡Associate Professor, Department of Political Science, Massachusetts Institute of Technology, Cambridge MA 02142. Phone: 617–253–3138, Email: insong@mit.edu, URL: http://web.mit.edu/insong/www/

§Research Fellow, Institute for Advanced Study in Toulouse (IAST) & Visiting Fellow, Department of Political and Social Change at Australian National University (ANU) Email: erik.wang@iast.fr, URL: http://erikhw.github.io

1 Introduction

One common and effective strategy to estimating causal effects in observational studies is the comparison of treated and control observations who share similar observed characteristics. Matching methods facilitate such comparison by selecting a set of control observations that resemble each treated observation and offering intuitive diagnostics for assessing the quality of resulting matches (e.g., Rubin, 2006; Stuart, 2010). By making the treatment variable independent of observed confounders, these methods reduce model dependence and improve the validity of causal inference in observational studies (e.g., Ho et al., 2007). For these reasons, matching methods have become part of the standard tool kit for empirical researchers across social sciences.

Despite their popularity, matching methods have been rarely used for the analysis of time-series cross section (TSCS) data, which consist of a relatively large number of repeated measurements on the same units. In such data, each unit may receive the treatment multiple times and the timing of treatment administration may differ across units. Perhaps, due to this complication, we find few applications of matching methods to TSCS data, and an overwhelming number of social scientists use linear regression models with fixed effects (e.g., Angrist and Pischke, 2009). Unfortunately, these regression models heavily rely on parametric assumptions, offer few diagnostic tools, and make it difficult to intuitively understand how counterfactual outcomes are estimated (Imai and Kim, 2019, 2020).[1] Moreover, almost all of the existing matching methods assume a cross-sectional data set (e.g., Hansen, 2004; Rosenbaum, Ross and Silber, 2007; Abadie and Imbens, 2011; Iacus, King and Porro, 2011; Zubizarreta, 2012; Diamond and Sekhon, 2013).[2]

We fill this methodological gap by developing matching methods for TSCS data. In the proposed approach (Section 3), for each treated observation, we first select a set of control observations from other units in the same time period that have an identical treatment history for a pre-specified time span. We further refine this matched set by using standard matching or weighting methods so that matched control observations become similar to the treated observation in terms of outcome and covariate histories. After this refinement step, we apply a difference-in-differences estimator that

[1] There is a growing body of literature that shows the limitations of the standard two-way fixed effects regressions for causal inference with panel data (see e.g., Chaisemartin and D'Haultfœuille, 2018; Goodman-Bacon, 2018). However, these papers focus on the interpretation of fixed effects regression and do not consider general matching methods, which we develop in this paper.

[2] A notable exception we found is a working paper by Nielsen and Sheffield (2009). But, as the authors acknowledge in the paper, their proposed algorithm was still in development at the time of their writing. It is also substantially different from the matching algorithm we propose in this paper.

adjusts for a possible time trend. The proposed method can be used to estimate both short-term and long-term average treatment effect of policy change for the treated (ATT) and allows for simple diagnostics through the examination of covariate balance. Finally, we establish the formal connection between the proposed matching estimator and the linear regression estimator with unit and time fixed effects. All together, the proposed methodology provides a design-based approach to causal inference with TSCS data.[3] The proposed matching methods can be implemented via the open-source statistical software in R language , PanelMatch: Matching Methods for Causal Inference with Time-Series Cross-Sectional Data, available at `https://CRAN.R-project.org/package=PanelMatch`.

In Appendix B, we conduct a simulation study to evaluate the finite sample performance of the proposed matching methodology relative to the standard linear regression estimator with unit and time fixed effects. We show that the proposed matching estimators are more robust to model misspecification than this standard two-way fixed effects regression estimator. The latter is generally more efficient but suffers from a substantial bias unless the model is correctly specified. In contrast, our methodology yields estimates that are stable across simulation scenarios considered here. We also find that our block-bootstrap based confidence interval has a reasonable coverage.

Our work builds upon the growing methodological literature on causal inference with TSCS data. In an influential work, Abadie, Diamond and Hainmueller (2010) focuses on the setting, in which only one unit receives the treatment and the data are available for a long time period prior to the administration of treatment. The authors propose the synthetic control method, which constructs a weighted average of pre-treatment outcomes among control units such that it approximates the observed pre-treatment outcome of the treated unit. A major limitation of this approach is the requirement that only one unit receives the treatment. Even when multiple treated units are allowed, they are assumed to receive the treatment at a single point in time (see also Doudchenko and Imbens, 2017; Ben-Michael, Feller and Rothstein, 2019a). In addition, the synthetic control method and its extensions require a relatively long pre-treatment time period for good empirical performance.

Recently, a number of researchers have extended the synthetic control method. For example, Xu (2017) proposes a generalized synthetic control method based on the framework of linear models with interactive fixed effects. This method, however, still requires a relatively large number of control units that do not receive the treatment at all. Furthermore, although the possibility of

[3] In epidemiology, such an approach is called trial emulation as it attempts to emulate a randomized experiment in an observational study (Hernán and Robins, 2016).

some units receiving the treatment at multiple time periods is noted (see footnote 7), the author assumes that the treatment status never reverses. Indeed, such "staggered adoption" assumption is common even among the recently proposed extensions of the synthetic control method (e.g., Ben-Michael, Feller and Rothstein, 2019b). In contrast, our methods allow multiple units to receive the treatment at any point in time, and units can switch their treatment status multiple times over time. Moreover, the proposed methodology can be used to estimate causal effects using a panel data with a relatively small number of time periods.

Another relevant methodological literature is the model-based approaches such as the structural nested mean models (Robins, 1994) and marginal structural models (Robins, Hernán and Brumback, 2000). These models focus on estimating the causal effect of treatment sequence while avoiding the post-treatment bias due to the fact that future treatments may be caused by past treatments (see Blackwell and Glynn, 2018, for an introduction). These approaches, however, require the modeling of potentially complex conditional expectation functions and propensity score for each time period, which can be challenging for TSCS data that often have a large number of time periods (e.g., Imai and Ratkovic, 2015). Our proposed method can incorporate these model-based approaches within the matching framework, permitting more robust confounding adjustment when estimating short-term and long-term treatment effects.

In the next section, we introduce two motivating empirical applications, one estimating the causal effects of democracy on economic growth and the other examining whether interstate war increases inheritance tax. These two studies represent typical observational studies that analyze TSCS data (spanning over 50 and 180 years, respectively) to estimate causal effects. The original authors use various linear regression models with country and year fixed effects that are extremely popular among social scientists. These models, however, do not make explicit which control units are used to estimate counterfactual outcomes. We introduce the treatment variation plot which visualizes the distribution of treatment so that researchers can understand how the treated observations should be compared with the control observations. This motivates our proposed matching method, which is applied to these empirical studies in Section 5. Finally, Section 6 gives concluding remarks.

2 Motivating Applications

In this section, we introduce two influential studies that motivate our methodology and briefly review the original empirical analyses. The first study is Acemoglu et al. (2019), which examines

the causal effect of democracy on economic development. Our second application is Scheve and Stasavage (2012), which investigates whether war mobilization leads countries to introduce significant taxation of inherited wealth. Both studies use linear regression models with fixed effects to estimate the causal effects of interest. After we briefly describe the original data and analysis for each study, we visualize the variation of treatment across time and space for each data set and motivate the proposed methodology, which exploits this variation.

2.1 Democracy and Economic Growth

The relationship between political institutions and economic well-being is a central question in the field of political economy. In particular, scholars have long debated whether democracy promotes economic development (e.g., Przeworski et al., 2000; Papaioannou and Siourounis, 2008; Gerring, Thacker and Alfaro, 2012). Acemoglu et al. (2019) conducts an up-to-date and comprehensive empirical study to investigate this question. The authors analyze an unbalanced TSCS data set, which consists of a total of 184 countries over a half century from 1960 to 2010.

The main results presented in the original study are based on the following dynamic linear regression model with country and year fixed effects,

$$Y_{it} \;=\; \alpha_i + \gamma_t + \beta X_{it} + \sum_{\ell=1}^{4} \left\{ \rho_\ell Y_{i,t-\ell} + \zeta_\ell^\top \mathbf{Z}_{i,t-\ell} \right\} + \epsilon_{it} \tag{1}$$

for $i = 1, \ldots, N$ and $t = 5, \ldots, T$ (the notation assumes a balanced panel for simplicity) where Y_{it} is logged real GDP per capita, and X_{it} represents the democracy indicator variable that is equal to 1 if country i in year t receives both a "Free" or "Partially Free" in Freedom House and a positive score in the Polity IV index, and 0 otherwise.[4] The model also includes four lagged outcome variables, $Y_{i,t-\ell}$ for $\ell = 1, \ldots, 4$, as well as a set of time-varying covariates $\mathbf{Z}_{it}$ and their lagged values. For the basic model specification, $\mathbf{Z}_{it}$ includes the log population, the log population of those who are below 16 years old, the log population of those who are above 64 years old, net financial flow as a fraction of GDP, trade volume as a fraction of GDP, and a dichotomous measure of social unrest.[5] The choice of four lags is particularly important, specifying how far back in time one needs to consider when adjusting for confounding factors.

The authors assume the following standard sequential exogeneity,

$$\mathbb{E}(\epsilon_{it} \mid Y_{i,t-1}, Y_{i,t-2}, \ldots, Y_{i1}, X_{it}, X_{i,t-1}, \ldots, X_{i1}, \mathbf{Z}_{it}, \mathbf{Z}_{i,t-1}, \ldots, \mathbf{Z}_{i1}, \alpha_i, \gamma_t) \;=\; 0 \tag{2}$$

[4] There exist a small number of observations where data are missing for either Freedom House score or Polity IV score. The original authors hand-code these observations.

[5] In the original study, the authors include one covariate at a time rather than including them all together.

	Democracy and Growth (Acemoglu et al., 2019)				War and Taxation (Scheve and Stasavage, 2012)			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
ATE ($\hat{\beta}$)	0.787	0.875	0.666	0.917	6.775	1.737	5.532	1.539
	(0.226)	(0.374)	(0.306)	(0.461)	(2.392)	(0.729)	(2.091)	(0.753)
$\hat{\rho}_1$	1.238	1.204	1.098	1.046		0.908		0.904
	(0.038)	(0.041)	(0.042)	(0.043)		(0.014)		(0.014)
$\hat{\rho}_2$	−0.207	−0.193	−0.133	−0.121				
	(0.046)	(0.045)	(0.040)	(0.038)				
$\hat{\rho}_3$	−0.026	−0.028	0.005	0.014				
	(0.029)	(0.028)	(0.030)	(0.029)				
$\hat{\rho}_4$	−0.043	−0.036	−0.031	−0.018				
	(0.018)	(0.020)	(0.024)	(0.023)				
country FE	Yes	Yes	Yes	Yes	Yes	No	Yes	No
time FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
time trends	No	No	No	No	Yes	Yes	Yes	Yes
covariates	No	No	Yes	Yes	No	No	Yes	Yes
estimation	OLS	GMM	OLS	GMM	OLS	OLS	OLS	OLS
N	6,336	4,416	4,416	4,245	2,780	2,537	2,779	2,536

Table 1: **Regression Results from the Two Motivating Empirical Applications**. The estimated coefficients for the treatment variable and lagged outcome variables are presented with standard errors in parentheses. For the Acemoglu et al. (2019) study, we show four models based on equation (1) using OLS or GMM estimation and with or without covariates. The estimated coefficients and standard errors are multiplied by 100 for the ease of interpretation. For the Scheve and Stasavage (2012) study, we show two statistic models based on equation (4) and the dynamic models defined in equation (6), with or without covariates. The standard errors are in parentheses. For the Acemoglu et al. (2019) study, we use the heteroskedasticity-robust standard errors. For the Scheve and Stasavage (2012) study, we cluster standard errors by countries for the static models while the panel corrected standard errors are used for the dynamic models.

which implies that the error term is independent of past outcomes, current and past treatments and covariates. It is well known that the ordinary least squares (OLS) estimate of β has an asymptotic bias of order $1/T$ (Nickell, 1981). To address this problem, Acemoglu et al. (2019) also fit the model in equation (1) using the generalized method of moments (GMM) estimation (Arellano and Bond, 1991) with the following moment conditions implied by equation (2),

$$\mathbb{E}\{(\epsilon_{it} - \epsilon_{i,t-1})Y_{is}\} \;=\; \mathbb{E}\{(\epsilon_{it} - \epsilon_{i,t-1})X_{i,s+1}\} \;=\; 0 \tag{3}$$

for all $s \leq t-2$. The error terms are assumed to be serially uncorrelated, and the authors use the heteroskedasticity-robust standard errors.

Table 1 presents the estimates of the coefficients of this model given in equation (1). Following the original paper, the estimated coefficients and standard errors are multiplied by 100 for the ease of interpretation. The results in the first two columns are based on the model without the time-varying covariates $\mathbf{Z}$ whereas the next two columns are those from the model with the covariates. For each model, we use both OLS (columns (1) and (3)) and GMM (columns (2) and (4)) estimation as explained above. As shown in Acemoglu et al. (2019), the effect of democracy on logged GDP per capita is positive and statistically significant across all four models. Based on this finding, the authors conclude that in the year of democratization the GDP per capita increases more than 0.5 percent. This is a substantial effect given that the democratization may have a long term effect on economic growth.

2.2 War and Taxation

As a central element of redistributive policies, inheritance taxation plays an essential role in wealth accumulation and income inequality. The merits and pitfalls of estate tax have been heavily featured in academic and policy debates. Scheve and Stasavage (2012) is among the first to empirically investigate this normatively controversial subject by examining the political conditions that underpin progressive inheritance taxation. The study documents that participation in inter-state war propels countries to increase inheritance taxation. The key proposed mechanism is that war mobilization leads to a widespread willingness to share financial burden of war among the public.

Scheve and Stasavage (2012) analyzes an unbalanced TSCS data set of 19 countries repeated over 185 years, from 1816 to 2000. The treatment variable of interest X_{it} is binary, indicating whether country i experiences an inter-state war in year t, whereas the outcome variable Y_{it} represents top rate of inheritance taxation for country i in year t. The study measures the outcome variable for each country in a given year using the top marginal rate for a direct descendant who inherits an estate. Although the authors of the original study aggregate the data into five-year or decade intervals, we analyze the annual data in order to avoid any aggregation bias.

The authors fit the following static linear regression model with country and time fixed effects as well as country-specific linear time trends,

$$Y_{it} \;=\; \alpha_i + \gamma_t + \beta X_{i,t-1} + \delta^\top \mathbf{Z}_{i,t-1} + \lambda_i t + \epsilon_{it} \tag{4}$$

where $\mathbf{Z}_{it}$ represents a set of the time-varying covariates, including an indicator variable for a leftist executive, a binary variable for the universal male suffrage, and logged real GDP per capita. The authors use the lagged values of the treatment variable and time-varying covariates in order to avoid

the issue of simultaneity. However, unlike the Acemoglu et al. study, they exclude lagged outcome variables and only include one period lag of time-varying confounders. The OLS estimation is used for fitting the model, requiring the following strict exogeneity assumption,

$$\mathbb{E}(\epsilon_{it} \mid \mathbf{X}_i, \mathbf{Z}_i, \alpha_i, \gamma_t, \lambda_i) \;=\; 0 \tag{5}$$

where $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{iT})$ and $\mathbf{Z}_i = (\mathbf{Z}_{i1}^\top, \mathbf{Z}_{i2}^\top, \ldots, \mathbf{Z}_{iT}^\top)^\top$. The authors use the cluster-robust standard error to account for the auto-correlation within each country.

Recognizing the limitation of such static models and yet wishing to avoid the bias of dynamic models with unit fixed effects mentioned above, Scheve and Stasavage (2012) also fit the following model with the lagged outcome variable and country specific time trends but without country fixed effects,

$$Y_{it} \;=\; \gamma_t + \beta X_{i,t-1} + \rho Y_{i,t-1} + \delta^\top \mathbf{Z}_{i,t-1} + \lambda_i t + \epsilon_{it} \tag{6}$$

where the strict exogeneity assumption is now given by,

$$\mathbb{E}(\epsilon_{it} \mid \mathbf{X}_i, \mathbf{Z}_i, Y_{i,t-1}, \gamma_t, \lambda_i) \;=\; 0 \tag{7}$$

The OLS estimation is employed for model fitting while panel-corrected standard errors are used to account for correlation across countries within a time period (Beck and Katz, 1995).

The last four columns of Table 1 present the results. Column (5) and (7) report the results obtained using the static model given in equation (4) without and with the time-varying covariates, respectively. Similarly, columns (6) and (8) are based on the dynamic model specified in equation (6) without and with the time varying covariates, respectively. These results show that war has a positive estimated effect of several percentage points on inheritance taxation although the magnitude for contemporaneous effect in dynamic models is much smaller.

2.3 The Treatment Variation Plot

A variety of linear regression models with fixed effects used by these studies represent the most commonly used methodological approaches to causal inference with TSCS data in the social sciences (e.g., Angrist and Pischke, 2009). However, a major drawback of these approaches is that they completely rely on the framework of linear regression models with fixed effects. In addition to the fact that the linearity assumption may be too stringent, it is also difficult to understand how these models use observed data to estimate relevant counterfactual quantities (Imai and Kim, 2019, 2020). These models offer few diagnostic tools for causal inference. In contrast, matching

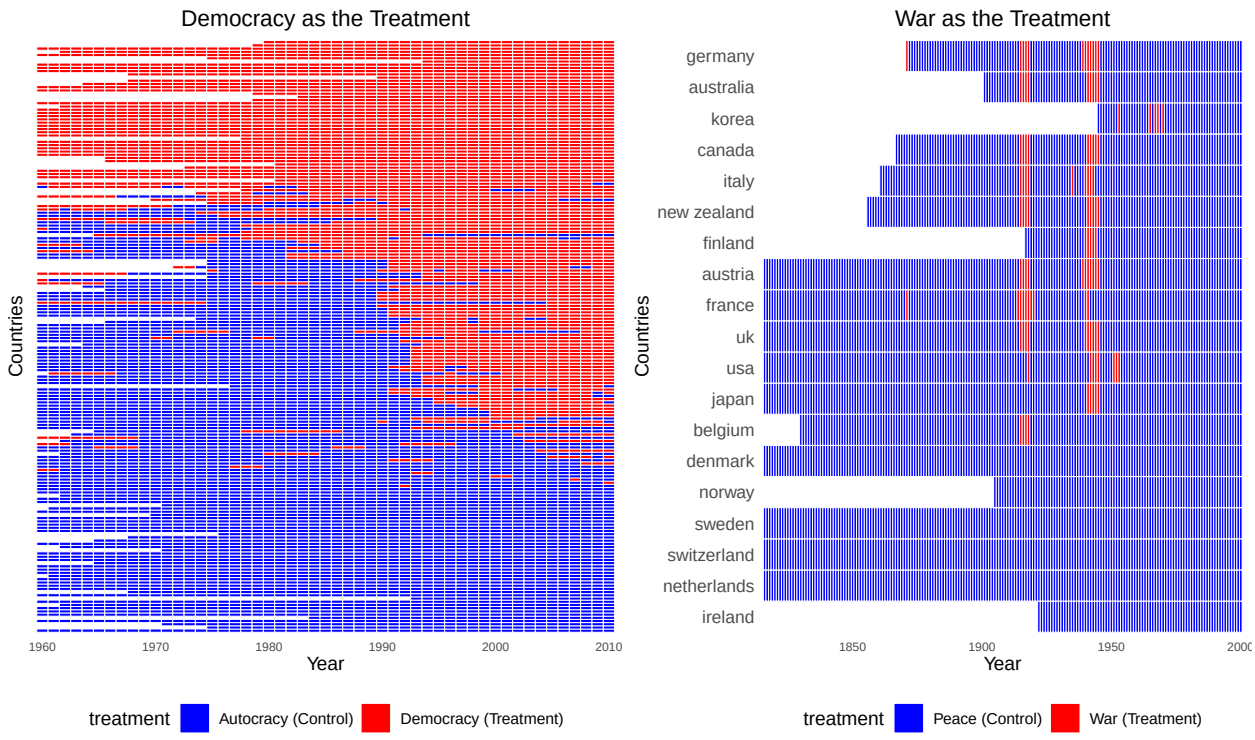


Figure 1: **The Treatment Variation Plots for Visualizing the Distribution of Treatment across Space and Time**. The left panel displays the spatial-temporal distribution of treatment for the study of democracy's effect on economic development (Acemoglu et al., 2019), in which a red (blue) rectangle represents a treatment (control) country-year observation. A white area represents the years when a country did not exist. The right panel displays the treatment variation plot for the study of war's effect on inheritance taxation (Scheve and Stasavage, 2012).

methods, that have been developed in the causal inference literature and are extended to TSCS data in Section 3, clearly specify a set of control observations used to estimate the counterfactual outcomes of treated observations and enable the assessment of credibility of such comparisons.

Before describing our proposed methodology, we introduce the *treatment variation plot*, which visualizes the variation of treatment across space and time, in order to help researchers build an intuition about how comparison of treated and control observation can be made. In the left panel of Figure 1, we present the distribution of the treatment variable for the Acemoglu et al. (2019) study where a red (blue) rectangle represents a treated (control) country-year observation. White areas indicate the years when countries did not exist. We observe that many countries stayed either democratic or autocratic throughout years with no regime change. Among those that experienced a regime change, most have transitioned from autocracy to democracy, but some of them have gone back and forth multiple times. When ascertaining the causal effects of democratization, therefore, we may consider the effect of a transition from democracy to autocracy as well as that of a transition from autocracy to democracy.

The treatment variation plot suggests that researchers can make a variety of comparisons between the treated and control observations. For example, we can compare the treated and control observations within the same country over time, following the idea of regression models with unit fixed effects (Imai and Kim, 2019). With such an identification strategy, it is important not to compare the observations far from each other to keep the comparison credible. We also need to be careful about potential carryover effects where democratization may have a long term effect, introducing post-treatment bias. Alternatively, researchers can conduct comparison within the same year, which would correspond to the identification strategy of year fixed effects models. In this case, we wish to compare similar countries with one another for the same year and yet we may be concerned about unobserved differences among those countries.

The right panel of Figure 1 shows the treatment variation plot for the Scheve and Stasavage (2012) study, in which a treated (control) observation represents the time of interstate war (peace) indicated by a red (blue) rectangle. As in the left plot of the figure, a white area represent the time period when a country did not exist. We observe that most of the treated observations are clustered around the time of two world wars. This implies that although the data set extends from 1816 to 2000, most observations in earlier and recent years would not serve as comparable control observations for the treated country-year observations.[6] As a result, it may be difficult to generalize the estimates obtained from this data set beyond the two world wars.

In sum, the treatment variation plot is a useful graphical tool for visualizing the distribution of treatment across time and units. Researchers should pay special attention to whether the treatment sufficiently varies both over time and across units as in the Acemoglu et al. study or the treatment variation is concentrated in a relatively small subset of the data as in the Scheve and Stasavage study. Since the internal and external validity of causal effect estimation with TSCS data critically rely upon such variation, the treatment variation plot plays an essential role when considering the causal identification strategies.

3 The Proposed Methodology

In this section, we propose a general matching method for causal inference with TSCS data. The proposed methodology can be summarized as follows. For each treated observation, researchers first find a set of control observations that have the identical treatment history up to the pre-specified

[6] The treatment variation plot is also useful for detecting potential anomalies in data. For example, the right panel of Figure 1 shows that Korea is coded to be in war only in 1953 during the course of the Korean War (1950–1953).

number of periods. We call this group of matched control observations a *matched set*. Once a matched set is selected for each treated observation, we further refine it by adjusting for observed confounding via standard matching and weighting techniques so that the treated and matched control observations have similar covariate values. Finally, we apply the difference-in-differences estimator in order to account for an underlying time trend. At the end of this section, we establish the exact relationship between the proposed matching estimator and the standard linear fixed effects regression estimator. We also discuss how to conduct covariate balance diagnostics and compute standard errors.

3.1 Matching Estimators

Consider a TSCS data set with N units (e.g., countries) and T time periods (e.g., years). For the sake of notational simplicity, we assume a balanced TSCS data set where the data are observed for all N units in each of T time periods. However, all the methods described below are applicable to an unbalanced TSCS data set. For each unit $i = 1, 2, \ldots, N$ at time $t = 1, 2, \ldots, T$, we observe the outcome variable Y_{it}, the binary treatment indicator X_{it}, and a vector of K time-varying covariates $\mathbf{Z}_{it}$. We assume that within each time period the causal order is given by $\mathbf{Z}_{it}$, X_{it}, and Y_{it}. That is, these covariates $\mathbf{Z}_{it}$ are realized before the administration of the treatment in the same time period X_{it}, which in turn occurs before the outcome variable Y_{it} is realized.

3.1.1 Causal Quantity of Interest

The first step of the proposed methodology is to define a causal quantity by choosing a non-negative integer F as the number of *leads*, which represents the outcome of interest measured at F time periods after the administration of treatment. For example, $F = 0$ represents the contemporaneous effect while $F = 2$ implies the treatment effect on the outcome two time periods after the treatment is administered. Specifying $F > 0$ allows researchers to examine a cumulative (or long-term) effect. In addition, researchers must select another non-negative integer L as the number of *lags* to adjust for. As in the regression approach, the choice of L is important and faces a bias-variance tradeoff. While a greater value improves the credibility of the unconfoundedness assumption introduced below, it also reduces the efficiency of the resulting estimates by reducing the number of potential matches.

We assume the absence of spillover effect but allow for some carryover effects (up to L time periods). That is, the potential outcome for unit i at time $t + F$ depends neither on the treatment status of other units, e.g., $X_{i't'}$ with $i' \neq i$ and for any t', nor the previous treatment status of

the same unit after L time periods, i.e., $\{X_{i,t-\ell}\}_{\ell=L+1}^{t-1}$. In many applications, the assumption of no spillover effect may be too restrictive. Although the methodological literature has begun to relax the assumption of no spillover effect in experimental settings (e.g., Hudgens and Halloran, 2008; Tchetgen Tchetgen and VanderWeele, 2010; Aronow and Samii, 2017; Imai, Jiang and Malai, 2020). We will leave the challenge of enabling the presence of spillover effects in TSCS data settings to future research.

Once these two parameters, L and F, are selected, we can define a causal quantity of interest. We first consider the average treatment effect of policy change among the treated (ATT), which is defined as,

$$\begin{aligned}\delta(F,L) \;=\; & \mathbb{E}\left\{Y_{i,t+F}\left(X_{it}=1, X_{i,t-1}=0, \{X_{i,t-\ell}\}_{\ell=2}^{L}\right) - \right. \\ & \left. Y_{i,t+F}\left(X_{it}=0, X_{i,t-1}=0, \{X_{i,t-\ell}\}_{\ell=2}^{L}\right) \mid X_{it}=1, X_{i,t-1}=0\right\} \qquad (8)\end{aligned}$$

where the treated observations are those who experience the policy change, i.e., $X_{i,t-1}=0$ and $X_{it}=1$. In this definition, $Y_{i,t+F}\left(X_{it}=1, X_{i,t-1}=0, \{X_{i,t-\ell}\}_{\ell=2}^{L}\right)$ is the potential outcome under a policy change, whereas $Y_{i,t+F}\left(X_{it}=0, X_{i,t-1}=0, \{X_{i,t-\ell}\}_{\ell=2}^{L}\right)$ represents the potential outcome without the policy change, i.e., $X_{i,t-1}=X_{it}=0$. In both cases, the rest of the treatment history, i.e., $\{X_{i,t-\ell}\}_{\ell=2}^{L}=\{X_{i,t-2},\ldots,X_{i,t-L}\}$, is set to the realized history. For example, $\delta(1,5)$ represents the average causal effect of policy change on the outcome one time period after the treatment while assuming that the potential outcome only depends on the treatment history up to five time periods back.[7]

This causal quantity allows for a future treatment reversal in a sense that the treatment status could go back to the control condition before the outcome is measured, i.e., $X_{i,t+\ell}=0$ for some ℓ with $1\leq\ell\leq F$. Later in this section, we discuss an alternative quantity of interest, which does not permit treatment status reversal, and define the ATT of stable policy change. This represents a counterfactual scenario, in which the treatment is in place at least for F time periods after policy change (see Section 3.1.5 for a discussion of this alternative causal quantity).

How should researchers choose the values of L and F? A large value of L improves the credibility of the aforementioned limited carryover effect assumption because it allows a greater number of past treatments (i.e., those up to time $t-L$) to affect the outcome of interest (i.e., $Y_{i,t+F}$). However, this

[7] Alternatively, one may be interested in the average treatment effect of *policy reversal* among the control (ATC) defined as, $\xi(F,L) = \mathbb{E}\left\{Y_{i,t+F}\left(X_{it}=0, X_{i,t-1}=1, \{X_{i,t-\ell}\}_{\ell=2}^{L}\right) - Y_{i,t+F}\left(X_{it}=1, X_{i,t-1}=1, \{X_{i,t-\ell}\}_{\ell=2}^{L}\right) \mid X_{it}=0, X_{i,t-1}=1\right\}$. This quantity corresponds to the effects of authoritarian reversal estimated in Section 5.

may reduce the number of matches and yield less precise estimates. We emphasize that choosing an appropriate number of lags is as important for our methods as for regression models. In practice, we recommend that researchers choose the number of lags based on their substantive knowledge and examine the sensitivity of empirical results to this choice. Similarly, the choice of F should be substantively motivated as it determines whether one is interested in short-term or long-term causal effects. We note that a large value of F may make the interpretation of causal effects difficult if many units switch the treatment status during the F lead time periods.

3.1.2 Identification Assumption

Given the values of F and L and the causal quantity of interest, we need an additional identification assumption. One possibility is to assume that conditional on the treatment, outcome, and covariate history up to time $t - L$, the treatment assignment is unconfounded. This assumption is called sequential ignorability in the literature (e.g., Robins, Hernán and Brumback, 2000),

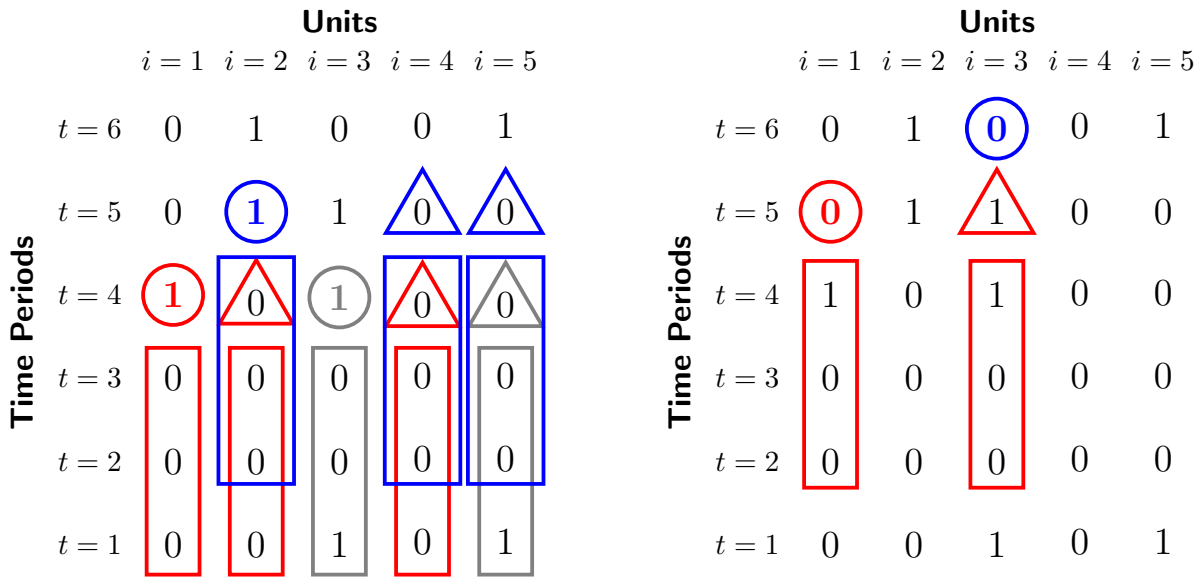
$$\begin{aligned}\left\{Y_{i,t+F}\left(X_{it}=1, X_{i,t-1}=0, \{X_{i,t-\ell}\}_{\ell=2}^{L}\right), Y_{i,t+F}\left(X_{it}=0, X_{i,t-1}=0, \{X_{i,t-\ell}\}_{\ell=2}^{L}\right)\right\} \\ \perp\!\!\!\perp X_{it} \;\mid\; X_{i,t-1}=0, \{X_{i,t-\ell}\}_{\ell=2}^{L}, \{Y_{i,t-\ell}\}_{\ell=1}^{L}, \{\mathbf{Z}_{i,t-\ell}\}_{\ell=0}^{L} \qquad (9)\end{aligned}$$

where $\mathbf{Z}_{it}$ is a vector of observed time-varying confounders for unit i at time period t. The assumption will be violated if there exist unobserved confounders. The violation also occurs if the treatment, outcome, and covariate histories before time $t - L$ confound the causal relationship between X_{it} and $Y_{i,t+F}$.

In many practical applications with TSCS data, however, researchers are concerned about the potential existence of unobserved confounding variables. Therefore, instead of the unconfoundedness assumption given in equation (9), we adopt the difference-in-differences (DiD) design (e.g., Abadie, 2005). Specifically, we make the following parallel trend assumption after conditioning on the treatment, outcome, and covariate histories,

$$\begin{aligned}&\mathbb{E}[Y_{i,t+F}\left(X_{it}=0, X_{i,t-1}=0, \{X_{i,t-\ell}\}_{\ell=2}^{L}\right) - Y_{i,t-1} \mid X_{it}=1, X_{i,t-1}=0, \{X_{i,t-\ell}, Y_{i,t-\ell}\}_{\ell=2}^{L}, \{\mathbf{Z}_{i,t-\ell}\}_{\ell=0}^{L}] \\ =\;& \mathbb{E}[Y_{i,t+F}\left(X_{it}=0, X_{i,t-1}=0, \{X_{i,t-\ell}\}_{\ell=2}^{L}\right) - Y_{i,t-1} \mid X_{it}=0, X_{i,t-1}=0, \{X_{i,t-\ell}, Y_{i,t-\ell}\}_{\ell=2}^{L}, \{\mathbf{Z}_{i,t-\ell}\}_{\ell=0}^{L}] \qquad (10)\end{aligned}$$

where the conditioning set includes the treatment history, the lagged outcomes (except the immediate lag $Y_{i,t-1}$), and the covariate history. It is well known that this parallel trend assumption cannot account for unobserved time-varying confounders. As such, it is important to examine whether the outcome time trends are indeed parallel on average between the treated and matched control units, using the data from the pre-treatment periods.



(a) Matched Sets for ATT

(b) Matched Sets for ATC

Figure 2: **An Example of Matched Sets with Five Units and Six Time Periods**. Panels (a) and (b) illustrate how matched sets are chosen for the ATT (as defined in equation (11)) and the ATC (see footnote 7), respectively, when $L = 3$. For each treated observation (colored circles), we select a set of control observations from other units in the same time period (triangles with the same color) that have an identical treatment history (rectangles with the same color).

3.1.3 Constructing the Matched Sets

The next step of the proposed methodology is to construct for each treated observation (i, t), the *matched set* of control units that share the identical treatment history from time $t - L$ to $t - 1$. We choose to match exactly on the treatment history because this allows us to partially control for carryover effects. We also believe that in many cases the past treatments are among the most important confounders as they are likely to affect both the current treatment and outcome. It is also important to note that the matched sets only include observations from the same time period, implying exact matching on time period. We do this in order to adjust for time-specific unobserved confounders. Partially relaxing these matching restrictions is straightforward. For example, we can match each treated observation with control observations that have a similar treatment history, where the degree of similarity is defined by researchers. The consequences of such relaxation needs to be carefully investigated in future research.

Figure 2 illustrates how the matched sets, with the identical treatment history with the treated observations, are constructed when $L = 3$. For example, in the left panel (the ATT), the control

observations $(i,t) = (2,4)$ and $(4,4)$ (red triangles) are matched to the treated observation $(1,4)$ (red circle) as they share the identical treatment history at $t = 1,2,3$ (red rectangles). The right panel, on the other hand, shows the matched set for the ATC where the treated observation (red triangle) is matched to the control observation (red circle). Another control observation highlighted by a blue circle has an empty matched set because no treated observation shares the same treatment history. We exclude these observations from the subsequent analysis to preserve the internal validity. It is important for researchers to examine the characteristics of these removed observations as this modifies the target population.

Formally, the matched set is defined as,

$$\mathcal{M}_{it} \;=\; \{i' : i' \neq i, X_{i't} = 0, X_{i't'} = X_{it'} \text{ for all } t' = t-1, \ldots, t-L\} \tag{11}$$

for the treated observations with $X_{it} = 1$ and $X_{i,t-1} = 0$. For the ATC (see footnote 7), we define the matched set as $\mathcal{M}_{it} = \{i' : i' \neq i, X_{i't} = 1, X_{i't'} = X_{it'} \text{ for all } t' = t-1, \ldots, t-L\}$. The observations in this set are matched to the control observations with $X_{it} = 0$ and $X_{i,t-1} = 1$.

Finally, we note that unlike the existing methods for staggered adoption, units are allowed to switch their treatment status multiple times over time. This matched set also differs from the risk set of Li, Propert and Rosenbaum (2001). The latter only includes units who have not received the treatment in the previous time periods. Instead, we allow for the possibility of a unit receiving the treatment multiple times, which is common in many TSCS data sets.

3.1.4 Refining the Matched Sets

The matched sets, defined above in equation (11), only adjust for the treatment history. However, the parallel trend assumption, defined in equation (10), demands that we also adjust for other confounders such as past outcomes and (possibly time-varying) covariates. Below, we discuss examples of matching and weighting methods that can be used to make additional adjustments by further refining the matched sets.

We first consider the application of matching methods. Suppose that we wish to match each treated observation with at most J control units from the matched set with replacement, i.e., $|\mathcal{M}_{it}| \leq J$. For example, we can use the Mahalanobis distance measure although other distance measure can also be used (see e.g., Rubin, 2006; Stuart, 2010). Specifically, we compute the average Mahalanobis distance between the treated observation and each control observation over time,

$$S_{it}(i') \;=\; \frac{1}{L}\sum_{\ell=1}^{L} \sqrt{(\mathbf{V}_{i,t-\ell} - \mathbf{V}_{i',t-\ell})^\top \boldsymbol{\Sigma}_{i,t-\ell}^{-1} (\mathbf{V}_{i,t-\ell} - \mathbf{V}_{i',t-\ell})} \tag{12}$$

for a matched control unit $i' \in \mathcal{M}_{it}$ where $\mathbf{V}_{it'}$ represents the time-varying covariates one wishes to adjust for and $\mathbf{\Sigma}_{it'}$ is the sample covariance matrix of $\mathbf{V}_{it'}$. That is, given a control unit in the matched set, we compute the standardized distance using the time-varying covariates and average it across time periods.[8]

Alternatively, we can use the distance measure based on the estimated propensity score. The propensity score is defined as the conditional probability of treatment assignment given pre-treatment covariates (Rosenbaum and Rubin, 1983). To estimate the propensity score, we first create a subset of the data, consisting of all treated observations and their matched control observations from the same year. We then fit a treatment assignment model to this data set. For example, we may use the following logistic regression model,

$$e_{it}(\{\mathbf{U}_{i,t-\ell}\}_{\ell=1}^{L}) \;=\; \Pr(X_{it} = 1 \mid \mathbf{U}_{i,t-1}, \ldots, \mathbf{U}_{i,t-L}) \;=\; \frac{1}{1 + \exp(-\sum_{\ell=1}^{L} \boldsymbol{\beta}_{\ell}^{\top} \mathbf{U}_{i,t-\ell})}. \quad (13)$$

where $\mathbf{U}_{it'} = (X_{it'}, \mathbf{V}_{it'}^{\top})^{\top}$. In practice, researchers may assume a more parsimonious model, in which some elements of $\boldsymbol{\beta}$ are set to zero. For example, setting $\boldsymbol{\beta} = 0$ for $\ell < t-1$ means that the model only includes the contemporaneous covariates $\mathbf{Z}_{it}$ and the previous value of the treatment variable. In addition, alternative robust estimation procedures such as the covariate balancing propensity score (CBPS) of Imai and Ratkovic (2014) can be used.

Given the fitted model, we compute the estimated propensity score for all treated observations and their matched control observations. Then, we adjust for the lagged outcomes and covariates by matching on the estimated propensity score, yielding the following distance measure,

$$S_{it}(i') \;=\; |\text{logit}\{\hat{e}_{it}(\{\mathbf{U}_{i,t-\ell}\}_{\ell=1}^{L})\} - \text{logit}\{\hat{e}_{i't}(\{\mathbf{U}_{i',t-\ell}\}_{\ell=1}^{L})\}| \quad (14)$$

for each matched control observation $i' \in \mathcal{M}_{it}$ where $\hat{e}_{i't}(\{\mathbf{U}_{i,t-\ell}\}_{\ell=1}^{L})$ is the estimated propensity score.

Once the distance measure $S_{it}(i')$ is computed for all control units in the matched set, then we refine the matched set by selecting up to J most similar control units that satisfy a caliper constraint C specified by researchers and giving zero weight to the other matched control units. In this way, we choose a subset of control units within the original matched set that are most similar

[8] For example, we might use all the observed time-varying covariates by setting $\mathbf{V}_{it'} = \mathbf{Z}_{i,t'+1}$. It is also possible to adjust for the lagged outcome variable by setting $\mathbf{V}_{it'} = (Y_{it'}, \mathbf{Z}_{i,t'+1}^{\top})^{\top}$ though typically researchers prefer to adjust for the differences in the lagged outcomes through assuming the parallel trend under the difference-in-differences design.

to the treated unit in terms of the observed confounders. Formally, the refined matched set for the treated observation (i,t) is given by,

$$\mathcal{M}_{it}^{*} \;=\; \{i' : i' \in \mathcal{M}_{it}, S_{it}(i') < C, S_{it}(i') \leq S_{it}^{(J)}\} \tag{15}$$

where $S_{it}^{(J)}$ represents the Jth order statistic of $S_{it}(i')$ among the control units in the original matched set $\mathcal{M}_{it}$.

Instead of matching, we can also use weighting to refine the matched sets. The idea is to construct a weight for each control unit i' within a matched set of a given treated observation (i,t) where a greater weight is assigned to a more similar unit. For example, we can use the inverse propensity score weighting method (Hirano, Imbens and Ridder, 2003), based on the propensity score model given in equation (13). In this case, the weight for a matched control unit i' is defined as,

$$w_{it}^{i'} \;\propto\; \frac{\hat{e}_{i't}(\{\mathbf{U}_{i,t-\ell}\}_{\ell=1}^{L})}{1 - \hat{e}_{i't}(\{\mathbf{U}_{i,t-\ell}\}_{\ell=1}^{L})} \tag{16}$$

such that $\sum_{i' \in \mathcal{M}_{it}} w_{it}^{i'} = 1$ and $w_{it}^{i'} = 0$ for $i' \notin \mathcal{M}_{it}$. Note that the model should be fitted to the entire sample of treated and matched control observations.

The weighting refinement further generalizes the matching refinement since the latter assigns an equal weight to each unit in the refined matched set $\mathcal{M}_{it}^{*}$,

$$w_{it}^{i'} \;=\; \begin{cases} \frac{1}{|\mathcal{M}_{it}^{*}|} & \text{if } i' \in \mathcal{M}_{it}^{*} \\ 0 & \text{otherwise} \end{cases} \tag{17}$$

In addition to propensity score weighting, other weighting methods such as calibration weights can also be used to refine each matched set.

3.1.5 The Difference-in-Differences Estimator

Given the refined matched sets, we estimate the ATT of policy change defined in equation (8). To do this, for each treated observation (i,t), we estimate the counterfactual outcome $Y_{i,t+F}(X_{it} = 0, X_{i,t-1} = 0, X_{i,t-2}, \ldots, X_{i,t-L})$ using the weighted average of the control units in the refined matched set. We then compute the difference-in-differences estimate of the ATT for each treated observation and then average it across all treated observations. Formally, our ATT estimator is given by,

$$\hat{\delta}(F,L) \;=\; \frac{1}{\sum_{i=1}^{N}\sum_{t=L+1}^{T-F} D_{it}} \sum_{i=1}^{N}\sum_{t=L+1}^{T-F} D_{it} \left\{ (Y_{i,t+F} - Y_{i,t-1}) - \sum_{i' \in \mathcal{M}_{it}} w_{it}^{i'} \left(Y_{i',t+F} - Y_{i',t-1}\right) \right\} \tag{18}$$

where $D_{it} = X_{it}(1 - X_{i,t-1}) \cdot \mathbf{1}\{|\mathcal{M}_{it}| > 0\}$, and $w_{it}^{i'}$ represents the non-negative normalized weight such that $w_{it}^{i'} \geq 0$ and $\sum_{i' \in \mathcal{M}_{it}} w_{it}^{i'} = 1$. Note that $D_{it} = 1$ only if observation (i, t) changes the treatment status from the control condition at time $t-1$ to the treatment condition at time t and has at least one matched control unit.

Specifying the future treatment sequence. When researchers are interested in a non-contemporaneous treatment effect (i.e., $F > 0$), the ATT defined in equation (8) does not specify the future treatment sequence. As a result, the matched control units may include those units who receive the treatment after time t but before the outcome is measured at time $t + F$. Similarly, some treated units may return to the control conditions between time t and time $t + F$. However, in certain circumstances, researchers may be interested in the ATT of stable policy change where the counterfactual scenario is that a treated unit does not receive the treatment before the outcome is measured. We can modify the ATT by specifying the future treatment sequence so that the causal quantity is defined with respect to the counterfactual scenario of interest.

For example, in the left panel of Figure 2, unit 1 receives the treatment at time 4 but reverts to the control condition at time 5. In contrast, unit 2 is a potential matched control unit who has the same treatment history from time 1 to 3 as unit 1, but receives the treatment at time 5 and 6. In this case, researchers may prefer to exclude unit 2 from the matched set of unit 1 and instead focus on unit 4 who shares the same treatment history and does not receive the treatment after time 4.

Formally, suppose that after a policy change, for some observations, the treatment will be in place at least for F time periods. We may be interested in estimating the ATT of stable policy change relative to no policy change among these treated observations. In this case, the ATT can be defined as,

$$\begin{aligned}
&\mathbb{E}\left[Y_{i,t+F}\left(\{X_{i,t+\ell}\}_{\ell=1}^{F} = \mathbf{1}_F, X_{it} = 1, X_{i,t-1} = 0, \{X_{i,t-\ell}\}_{\ell=2}^{L}\right) \; - \right. \\
&\quad \left. Y_{i,t+F}\left(\{X_{i,t+\ell}\}_{\ell=1}^{F} = \mathbf{0}_F, X_{it} = 0, X_{i,t-1} = 0, \{X_{i,t-\ell}\}_{\ell=2}^{L}\right) \mid \{X_{i,t+\ell}\}_{\ell=1}^{F} = \mathbf{1}_F, X_{it} = 1, X_{i,t-1} = 0\right]
\end{aligned} \tag{19}$$

where $\mathbf{1}_F$ and $\mathbf{0}_F$ are F dimensional vectors of ones and zeros, respectively.

The difference between equations (8) and (19) is that the latter specifies the future treatment sequence. The treated (matched control) observations are those who remain under the treatment (control) condition throughout F time periods after the administration of the treatment whereas the matched control units receive no treatment at least for F time periods after the treatment is

given. Thus, the matched set changes to,

$$\mathcal{M}_{it} \;=\; \{i' : i' \neq i, X_{i't} = X_{i't+1} = \ldots = X_{i't+F} = 0, X_{i't'} = X_{it'} \text{ for all } t' = t-1, \ldots, t-L\} \quad (20)$$

To estimate this ATT, we apply the idea of marginal structural models (MSMs) in order to make covariate adjustments while avoiding post-treatment bias (Robins, Hernán and Brumback, 2000). Note that the identification assumption is unchanged. We first constrain the matched set for each treated observation (i, t) such that the matched control units do not receive the treatment at least after time $t + F$. We then estimate the propensity score by modeling the treatment assignment, for example, using the logistic regression as follows,

$$e_{it}(\{\mathbf{U}_{i,t-\ell}\}_{\ell=1}^{L}) \;=\; \Pr(X_{it} = 1 \mid \mathbf{U}_{i,t-1}, \ldots, \mathbf{U}_{i,t-L}) \;=\; \frac{1}{1 + \exp(-\sum_{\ell=1}^{L} \boldsymbol{\beta}_\ell^\top \mathbf{U}_{i,t-\ell})}. \quad (21)$$

Unlike the above setting, the model must be fit to all observations including those who are not in the matched sets in order to model the entire treatment sequence. Using the result from MSMs, the weights are then computed as,

$$w_{it}^{i'} \;=\; \prod_{f=0}^{F} \frac{e_{i,t+f}(\{\mathbf{U}_{i,t+f-\ell}\}_{\ell=1}^{L})}{1 - e_{i,t+f}(\{\mathbf{U}_{i,t+f-\ell}\}_{\ell=1}^{L})} \quad (22)$$

for $i' \in \mathcal{M}_{it}$ and $w_{it}^{i'} = 0$ if $i' \notin \mathcal{M}_{it}$. Finally, we apply the DiD estimator in equation (18) to obtain an estimate of the long term ATT under the specified treatment sequence as defined in equation (19).

3.2 Checking Covariate Balance

One advantage of the proposed methodology, over regression methods, is that researchers can examine the resulting covariate balance between treated and matched control observations, enabling the investigation of whether the treated and matched control observations are comparable with respect to observed confounders. Under the proposed methodological framework, the examination of covariate balance is straightforward once the matched sets are determined and refined.

We propose to examine the mean difference of each covariate (e.g. $V_{it'j}$, which represents the jth variable in $\mathbf{V}_{it'}$) between a treated observation and its matched control observations at each pre-treatment time period, i.e. $t' < t$. We further standardize this difference, at any given pre-treatment time period, by the standard deviation of each covariate across all treated observations in the data so that the mean difference is measured in terms of standard deviation units. Formally, for each treated observation (i, t) with $D_{it} = 1$, we define the covariate balance for variable j at

the pre-treatment time period $t - \ell$ as,

$$B_{it}(j, \ell) \;=\; \frac{V_{i,t-\ell,j} - \sum_{i' \in \mathcal{M}_{it}} w_{it}^{i'} \, V_{i',t-\ell,j}}{\sqrt{\frac{1}{N_1 - 1} \sum_{i'=1}^{N} \sum_{t'=L+1}^{T-F} D_{i't'} (V_{i',t'-\ell,j} - \overline{V}_{t'-\ell,j})^2}} \tag{23}$$

where $N_1 = \sum_{i'=1}^{N} \sum_{t'=L+1}^{T-F} D_{i't'}$ is the total number of treated observations. We then further aggregate this covariate balance measure across all treated observations for each covariate and pre-treatment time period.

$$\overline{B}(j, \ell) \;=\; \frac{1}{N_1} \sum_{i=1}^{N} \sum_{t=L+1}^{T-F} D_{it} B_{it}(j, \ell) \tag{24}$$

Finally, we emphasize that one must examine the balance of the lagged outcome variables over multiple pre-treatment periods as well as that of time-varying covariates. This helps us evaluate the appropriateness of the parallel trend assumption used to justify the proposed DiD estimator.

3.3 Relations with Linear Fixed Effects Regression Estimators

It is well known that the standard DiD estimator is numerically equivalent to the linear two-way fixed effects regression estimator if there are two time periods and the treatment is administered to some units only in the second time period. Unfortunately, this equivalence result does not generalize to the multi-period DiD design that we consider in this paper, in which the number of time periods may exceed two and each unit may receive the treatment multiple times (see e.g., Imai and Kim, 2011, 2020; Abraham and Sun, 2018; Athey and Imbens, 2018; Chaisemartin and D'Haultfœuille, 2018; Goodman-Bacon, 2018). Nevertheless, researchers often motivate the use of the two-way fixed effects estimator by referring to the DiD design (e.g., Angrist and Pischke, 2009). Bertrand, Duflo and Mullainathan (2004), for example, call the linear regression model with two-way fixed effects "a common generalization of the most basic DiD setup (with two periods and two groups)" (p. 251).

The following theorem establish the algebraic equivalence between the proposed matching estimator given in equation (18) and *weighted* two-way fixed effects estimator. Our estimand is the ATT of stable policy change relative to no policy change as defined in equation (19), in which the treatment will be in place at least for F time periods. This generalizes the result of Imai and Kim (2011, 2020). Specifically, we allow for estimating both short-term and long-term average treatment effects with nonparametric covariate adjustment.

Theorem 1 (Difference-in-Differences Estimator as a Weighted Two-way Fixed Effects Estimator) *Assume that there is at least one treated and control unit, i.e.,* $0 < \sum_{i=1}^{N} \sum_{t=1}^{T} X_{it} <$

NT, and that there is at least one unit with $D_{it} = 1$, i.e., $0 < \sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}$. The difference-in-differences estimator, $\hat{\delta}(F, L)$ defined in equation (18), is equivalent to $\hat{\beta}_{\mathsf{DiD}}$ where $\hat{\beta}_{\mathsf{DiD}}$ is the following weighted two-way fixed effects regression estimator,

$$\hat{\beta}_{\mathsf{DiD}} \;=\; \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}\{(Y_{it} - \overline{Y}_i^* - \overline{Y}_t^* + \overline{Y}^*) - \beta(X_{it} - \overline{X}_i^* - \overline{X}_t^* + \overline{X}^*)\}^2. \tag{25}$$

The asterisks indicate weighted averages, i.e., $\overline{Y}_i^* = \sum_{t=1}^{T} W_{it}Y_{it}/\sum_{t=1}^{T} W_{it}$, $\overline{Y}_t^* = \sum_{i=1}^{N} W_{it}Y_{it}/\sum_{i=1}^{N} W_{it}$, $\overline{X}_i^* = \sum_{t=1}^{T} W_{it}X_{it}/\sum_{t=1}^{T} W_{it}$, $\overline{X}_t^* = \sum_{i=1}^{N} W_{it}X_{it}/\sum_{i=1}^{N} W_{it}$, $\overline{Y}^* = \sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}Y_{it}/\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}$, $\overline{X}^* = \sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}X_{it}/\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}$, and the regression weights are given by,

$$W_{it} \;=\; \sum_{i'=1}^{N}\sum_{t'=1}^{T} D_{i't'} \cdot v_{it}^{i't'} \quad \text{and} \quad v_{it}^{i't'} \;=\; \begin{cases} 1 & \text{if } (i,t) = (i', t'+F) \\ 1 & \text{if } (i,t) = (i', t'-1) \\ w_{i't'}^{i} & \text{if } i \in \mathcal{M}_{i't'}, t = t'+F \\ -w_{i't'}^{i} & \text{if } i \in \mathcal{M}_{i't'}, t = t'-1 \\ 0 & \text{otherwise.} \end{cases} \tag{26}$$

Proof is in Appendix A.

We note that the regression weight W_{it} can take a negative value in many cases. This implies that the two-way fixed effects regression estimator critically relies upon its parametric assumption. Although many applied researchers motivate the use of two-way fixed effects regression by the DiD design, Theorem 1 shows that such an argument is invalid unless the modeling assumption is correct.

3.4 Standard Error Calculation

To compute the standard errors of the proposed estimator given in equation (18), we use a block-bootstrap procedure specifically designed for matching with TSCS data. Abadie and Imbens (2008) shows that a standard bootstrap procedure yields an invalid inference for matching estimators. However, we can get around this problem by conditioning on the weights implied by the matching procedure (Imbens and Rubin, 2015). Much like the conditional variance in regression models, the resulting standard errors do not account for the uncertainty about a matching procedure, but can be interpreted as the uncertainty measure conditional upon it (Ho et al., 2007). In particular, we treat the implied observation-specific weight, which represents the number of times an observation is used for matching, as an observed variable and do not recompute it for each bootstrapped sample (see also Otsu and Rai, 2017).

For the proposed estimator, this observation-specific weight can be computed as follows,

$$W_{it}^{*} \;=\; \sum_{i'=1}^{N}\sum_{t'=1}^{T} D_{i't'}\cdot v_{it}^{i't'} \quad \text{and} \quad v_{it}^{i't'} \;=\; \begin{cases} 1 & \text{if } (i,t)=(i',t'+F) \\ -1 & \text{if } (i,t)=(i',t'-1) \\ -w_{i't'}^{i} & \text{if } i\in\mathcal{M}_{i't'}, t=t'+F \\ w_{i't'}^{i} & \text{if } i\in\mathcal{M}_{i't'}, t=t'-1 \\ 0 & \text{otherwise.} \end{cases} \tag{27}$$

which differs from the weight defined in Theorem 1. Note that $\hat{\delta}(F,L)$ defined in equation (18) can be attained by applying the weights directly to each observation: $\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}^{*}Y_{it}/\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}$. We treat this weight as a covariate and apply the block bootstrap procedure to account for within-unit time dependence. That is, we sample each unit, which consists of a sequence of T observations, with replacement, and compute $\sum_{i'=1}^{N}\sum_{t=1}^{T} W_{i't}^{*}Y_{i't}/\sum_{i'=1}^{N}\sum_{t=1}^{T} D_{i't}$ for the bootstrap sample units i' in each iteration.

4 A Simulation Study

We conduct a simulation study to examine the finite sample properties of the proposed matching estimator by comparing its empirical performance with the standard linear regression models with fixed effects. Specifically, we assess the robustness of the estimators to various degrees of model misspecification. We do so by gradually omitting the lagged covariates and their interaction terms. This setup is designed to replicate the common difficulty, faced by applied researchers, of determining the number of lags when analyzing TSCS data. Due to the space constraint, all the details and results of the simulation study are given in Appendix B. As expected, we show that the proposed matching estimator is more robust to the omission of relevant lags but is less efficient than the standard fixed effects regression estimator.

5 Empirical Analyses

We revisit the two motivating studies described in Section 2, one about the effect of democracy on development (Acemoglu et al., 2019), and the other concerning the impact of war on inheritance taxation (Scheve and Stasavage, 2012). We reanalyze their data by applying the proposed methodology described in Section 3 and illustrate how it can be used in practice. We find that the (negative) effect of authoritarian reversal on economic growth is more pronounced than the (positive) effect of democratization, and that war appears to increase inheritance tax rate but the effects are not precisely estimated.

5.1 Application of Matching Methods

We demonstrate the use of the proposed methodology. For the Acemoglu et al. (2019) study, we estimate the two effects of democracy on economic growth, the effect of democratization and that of authoritarian reversal. Since the treatment variable X_{it} takes the value of one (zero) if country i is democratic (autocratic) at year t, the average effect of democratization for the treated is defined by equation (8). The average effect of autocratic reversal for the treated, on the other hand, is defined as,

$$\mathbb{E}\left[Y_{i,t+F}\left(X_{it}=0, X_{i,t-1}=1, \{X_{i,t-\ell}\}_{\ell=2}^{L}\right) - Y_{i,t+F}\left(X_{it}=1, X_{i,t-1}=1, \{X_{i,t-\ell}\}_{\ell=2}^{L}\right) \mid X_{it}=0, X_{i,t-1}=1\right] \tag{28}$$

In addition, one may also be interested in the ATT of stable policy (regime) change relative to no policy (regime) change, as defined in equation (19). We present the covariate balance for this alternative quantity of interest in Appendix C.

As shown in the left panel of Figure 1, although most countries transition from autocracy to democracy, we also observe enough cases of authoritarian reversal, suggesting that we may have sufficient data to estimate both effects. In contrast, for the Scheve and Stasavage (2012) study, we focus on the effect of involvement in a war on inheritance tax rather than the effect of ending a war since the latter lacks enough control countries (i.e., countries still in a war when a treated country ends a war). This is because most war observations come from two world wars (see the right panel of Figure 1). Again, we present the covariate balance in the case of an alternative quantity of interest in Appendix C.

We use the original studies to guide the specification of matching methods. In their regression models, Acemoglu et al. (2019) include four years of lag for the outcome and time-varying covariates (see equation (1)). Therefore, when estimating the ATTs of democratization and authoritarian reversal, we also condition on four years of lag, i.e., $L = 4$, and estimate the ATT up to four years after regime change, i.e., $F = 1, 2, 3, 4$. In contrast, the dynamic model of Scheve and Stasavage (2012) adjusts only for one year lag of the outcome variable (see equation (6)). Since one year lag may not be sufficient, we conduct an analysis based on four year lags as well as one year lag when estimating the effect of war on inheritance tax.

To illustrate the proposed methodology, we begin by constructing the matched set for each treated observation based on the treatment history. Figure 3 presents the frequency distribution for the number of matched control units given a treated observation in the case of one and four

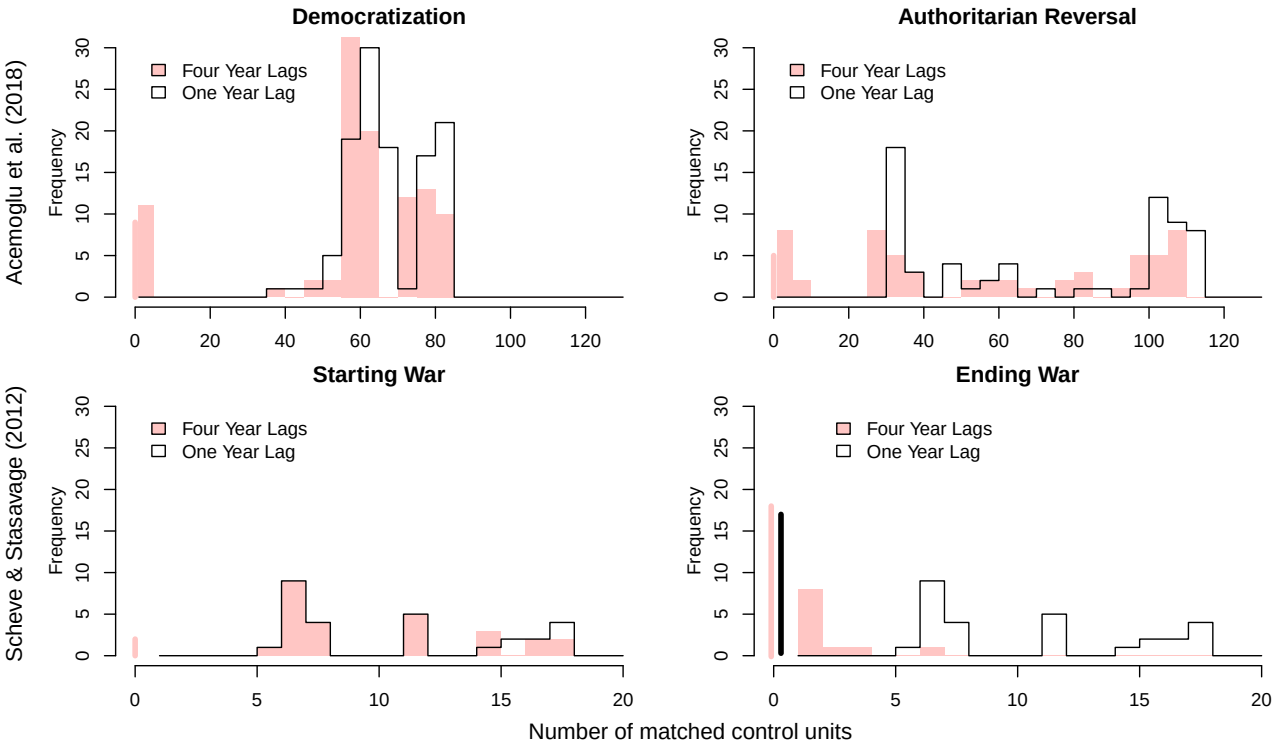


Figure 3: **Frequency Distribution of the Number of Matched Control Units**. The transparent (red) bar represents the number of matched control units that share the same treatment history as a treated observation for one year (four years) prior to the treatment year. The frequency distribution is presented for each of the two treatments in the Acemoglu et al. (2019) study (top panel) and the Scheve and Stasavage (2012) study (bottom panel). Thinner vertical bars at zero represent the number of treated observations that have no matched control units.

year lag as transparent and red bars, respectively. The distribution is presented for the transition from the control to treatment conditions (left column) and that from the treatment to control conditions (right column). As expected, the number of matched control units generally decreases when we adjust for the treatment history of four year period rather than that of one year period.

For the Acemoglu et al. (2019) study in the upper panel, there are 9 (5) treated observations for democratization (authoritarian reversal) that have no control unit with the same treatment history when the number of lags is four (represented by a thin red vertical bar at zero),[9] whereas no such treated observation exists for the case of one year lag. As noted earlier, for the Acemoglu et al. (2019) study, we have enough matched control units for both democratization and authoritarian reversal: most treated observations have more than 30 matched control units.

[9]Such observations for democratization are: Bangladesh in 2009, Guinea-Bissau in 1999, Haiti in 1994, Lesotho in 1999, Niger in 1999, Peru in 1993, Suriname in 1991, Thailand in 1992, and Turkey in 1973. Such observations for authoritarian reversal are: Burkina Faso in 1980, Bangladesh in 1974, Comoros in 1976, Ghana in 1972, and Mauritania in 2008.

However, for the Scheve and Stasavage (2012) study, most treated observations have less than five observations when studying the effect of ending war, suggesting that causal inference is more challenging in this setting. In addition, there are also unmatched treated observations. For starting war as the treatment, there are 2 treated observations without any matched control units if we match on 4 lags, as represented by a thinner red vertical bar at zero.[10] For ending war as the treatment, the use of 4 (1) lags leads to the number of unmatched treated observations to 18 (17), as represented by a thinner red (black) vertical bar at zero.[11] Thus, causal inference is challenging especially when estimating the effects of ending war. Below, we do not estimate the effects of ending war because the validity of such estimates is likely to be low.

To refine the matched sets, we apply Mahalanobis distance matching, propensity score matching, and propensity score weighting so that we can compare the performance of each refinement method. For matching, we apply up-to-five matching and up-to-ten matching for the Acemoglu et al. (2019) study to examine the sensitivity of empirical findings to the maximum number of matches. For the Scheve and Stasavage (2012) study, we use one-to-one match and up-to-three matches because the matched sets are smaller to begin with. Mahalanobis distance is defined in equation (12), while we use the logistic regression model estimated with just identified CBPS for propensity score matching (equation (14)) and weighting (equation (16)).

When specifying the Mahalanobis distance and the propensity score model, we use all time-varying covariates. For the Acemoglu et al. (2019) study, the time-varying covariates include the log population, the log population of age below 16 years, the log population of age above 64 years, net financial flow as a fraction of GDP, trade volume as a fraction of GDP, and a dichotomous measure of social unrest (though the original authors do not include all variables at once in their regression model). Similarly, for the Scheve and Stasavage (2012) study, we use all available time-varying covariates, i.e., an indicator variable for leftist executive, a binary variable for the universal male suffrage, and logged GDP per capita.

Figure 4 shows how the refinement of matched sets improves the covariate balance for the two studies. In each scatter plot, we compare the absolute value of standardized mean difference defined in equation (24) before (horizontal axis) and after (vertical axis) the refinement of matched sets.

[10] They are Korea in 1967 and Korea in 1970.

[11] The treated observations without any matched control units for 4 lags are: USA in 1919, 1946, and 1954; Canada in 1946; UK in 1946; France in 1872 and 1921; Germany in 1872 and 1946; Austria in 1946; Italy in 1946; Korea in 1954, 1966, 1969, and 1971; Japan in 1946; Australia in 1946; New Zealand in 1946. The same list applies to 1 lag except for USA 1919.

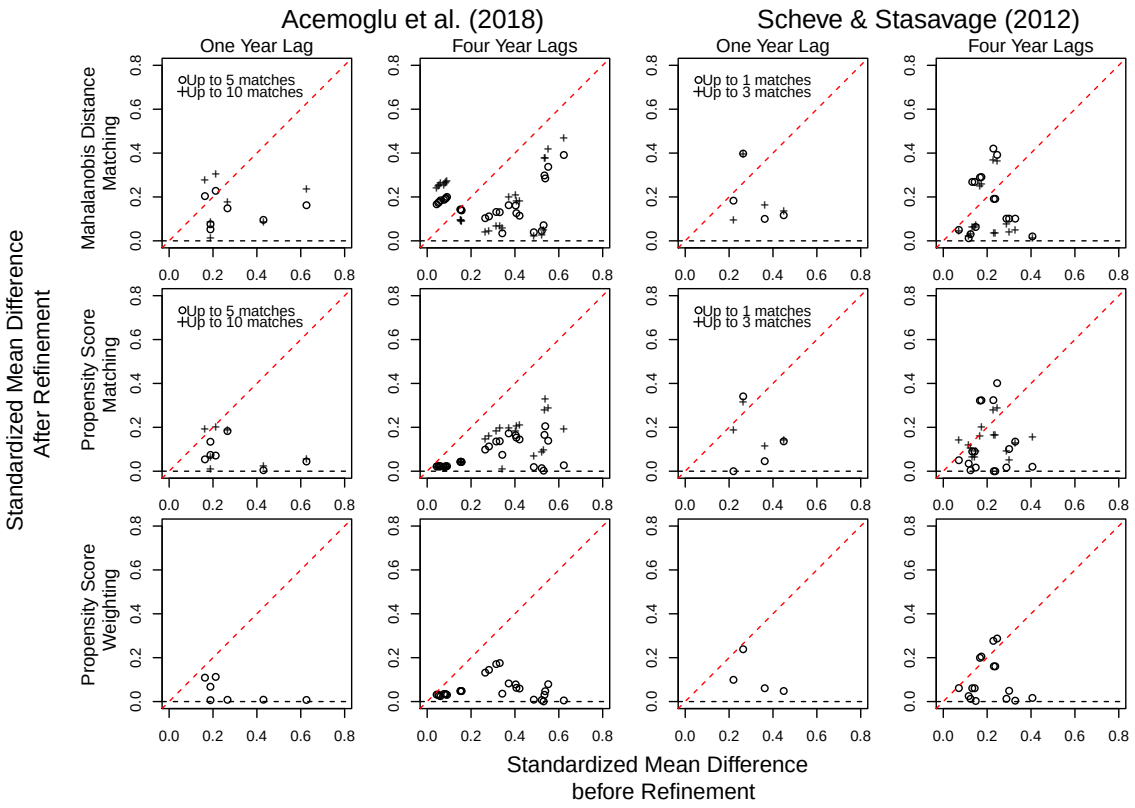


Figure 4: **Improved Covariate Balance due to the Refinement of Matched Sets**. Each scatter plot compares the absolute value of standardized mean difference for each covariate j and lag year ℓ defined in equation (24) before (horizontal axis) and after (vertical axis) the refinement of matched sets. Rows represents the results based on different matching and weighting methods while the columns represent the results using the adjustments for different lag lengths.

A dot below the 45 degree line implies that the standardized mean balance is improved after the refinement for a particular time-varying covariate. The plots suggest that across almost all variables the refinement results in the improved mean covariate balance. The amount of improvement is the greatest for propensity score weighting (bottom row) whereas Mahalanobis matching (top row) achieves only the modest degree of improvement.

Figure 5 further illustrates the improvement of covariate balance due to matching over the pre-treatment time period. We focus on the results for matching methods that adjust for time-varying covariates during the four year period prior to the administration of treatment. The top two rows present the standardized mean covariate balance for the two treatments of the Acemoglu et al. (2019) study whereas the bottom row shows that for the treatment of starting war in the Scheve and Stasavage (2012) study. The solid line represents the balance of the lagged outcome whereas grey lines show the balance of other covariates.

In all three cases, we find that the construction of matched sets (i.e., the adjustment of treatment

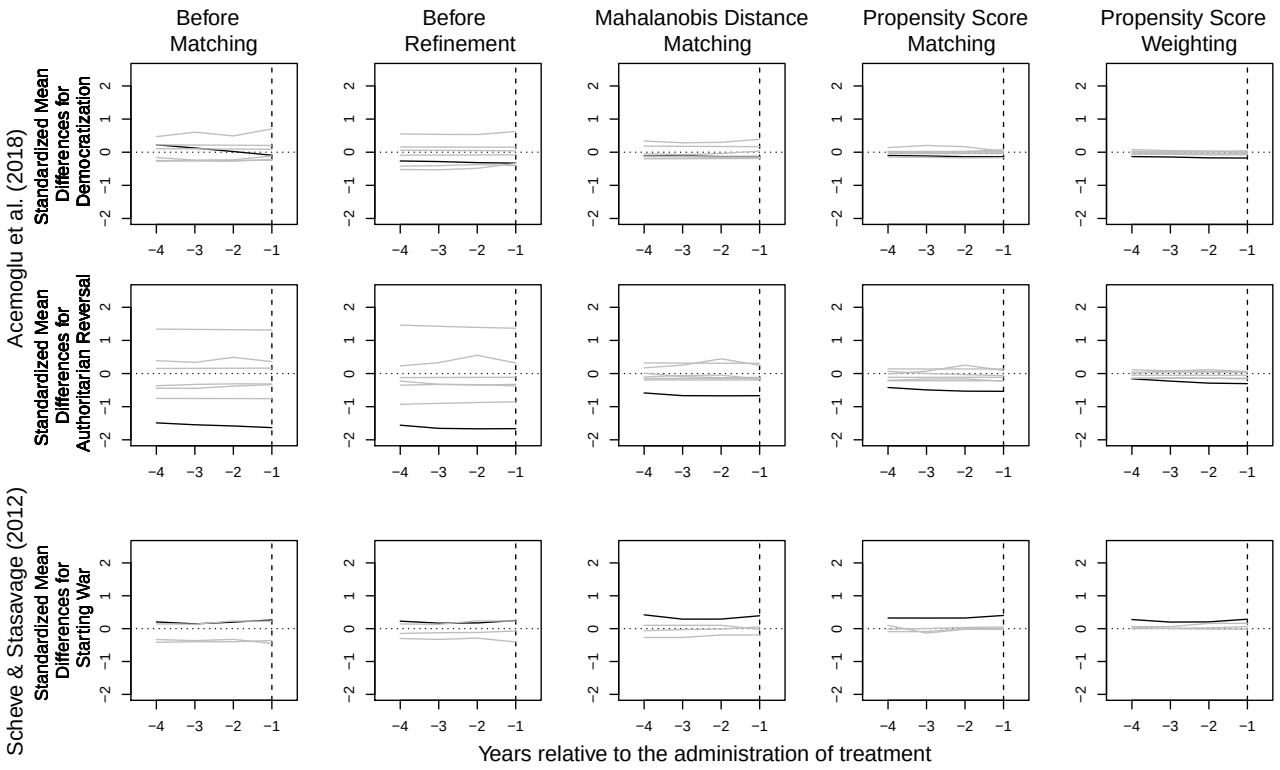


Figure 5: **Improved Covariate Balance due to Matching over the Pre-Treatment Time Period**. Each plot plots the standardized mean difference defined in equation (24) (vertical axis) over the pre-treatment time period of four years (horizontal axis). The left column shows the balance before matching, while the next column shows that before refinement but after the construction of matched sets. The remaining three columns present the covariate balance after applying different refinement methods. The solid line represents the balance of the lagged outcome variable whereas the grey lines represent that of time-varying covariates.

history alone) do not dramatically improve the covariate balance. In contrast, the improvement due to the refinement of matched sets is substantial. In particular, propensity score weighting essentially eliminates almost all imbalance in confounders. Although some degree of imbalance remains for Mahalanobis distance and propensity score matching, the standardized mean difference for the lagged outcome stays relatively constant over the entire pre-treatment period. This suggests that the assumption of parallel trend for the proposed difference-in-difference estimator may be appropriate.

5.2 Empirical Findings

We now present the estimated ATTs based on the matching methods. Figure 6 shows the matching estimates of the effects of democratization (upper panel) and authoritarian reversal (lower panel) on logged GDP per capita for the period of five years after the transition, i.e., $F = 0, 1, \ldots, 4$. Across all five methods (columns), we find that the point estimates of the effects for democratization

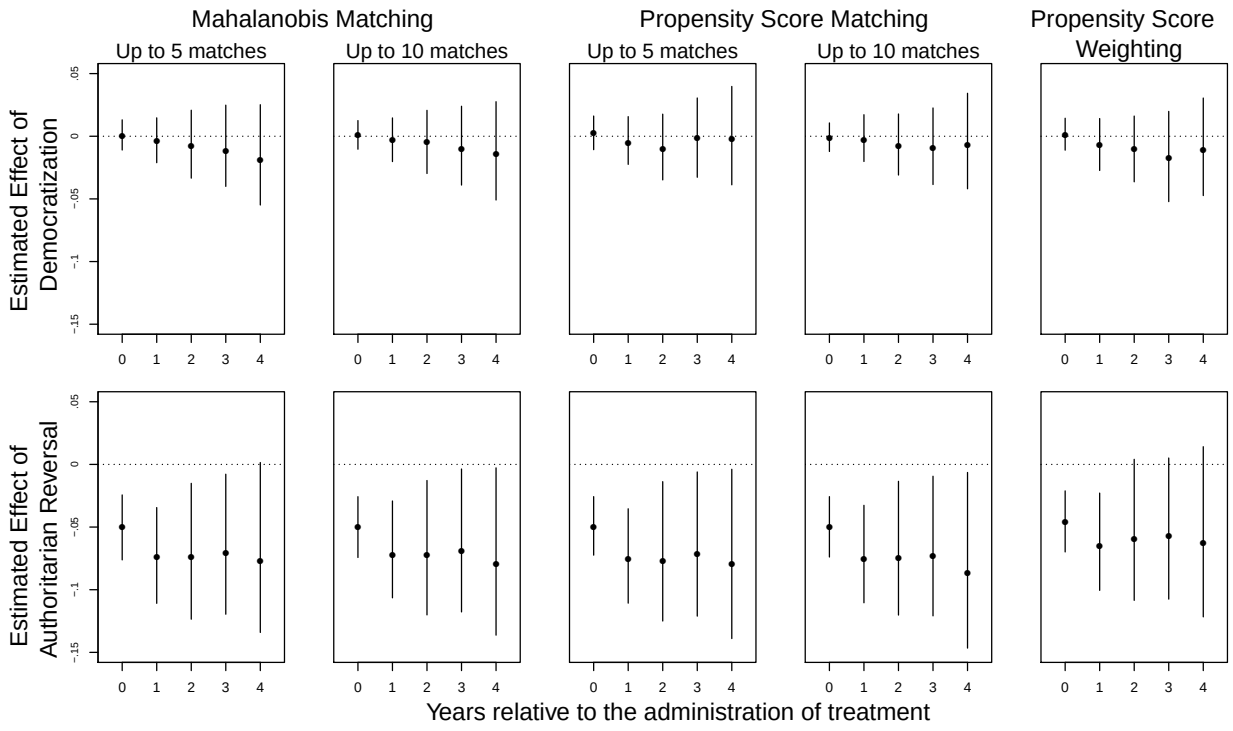


Figure 6: **Estimated Average Effects of Democracy on Logged GDP per Capita**. The estimates are based on the matching method that adjusts for the treatment and covariate histories during the four year period prior to the treatment, i.e., $L = 4$. The estimates for the average effects of democratization (upper panel) and authoritarian reversal (lower panel) are shown for the period of five years after the transition, i.e., $F = 0, 1, \ldots, 4$, with 95% bootstrap confidence intervals as vertical bars. Five different refinement methods are considered and their results are presented in different columns.

are mostly close to zero over the five year time period. On the other hand, the estimated effects of authoritarian reversal are negative and statistically significant across most refinement methods during the year of transition and the one to three years immediately after the transition when the treatment reversal is allowed. The estimated effects are substantively large, indicating an approximately 5 to 8 percent reduction of GDP per capita. This effect size is greater than the estimated effect of one percent found in the original analysis (see Table 1). In Figure A15 of Appendix D, as a robustness check, we show that the same analysis with the refinement based on one year period yields essentially the same results.

In sum, our analysis implies that the positive effect of democracy is driven by the negative effect of authoritarian reversal. In other words, we find that the transition into democracy from autocracy does not necessarily lead to a higher level of development. Rather, the treatment of backsliding into autocracy from democracy has a pronounced negative effect on development at

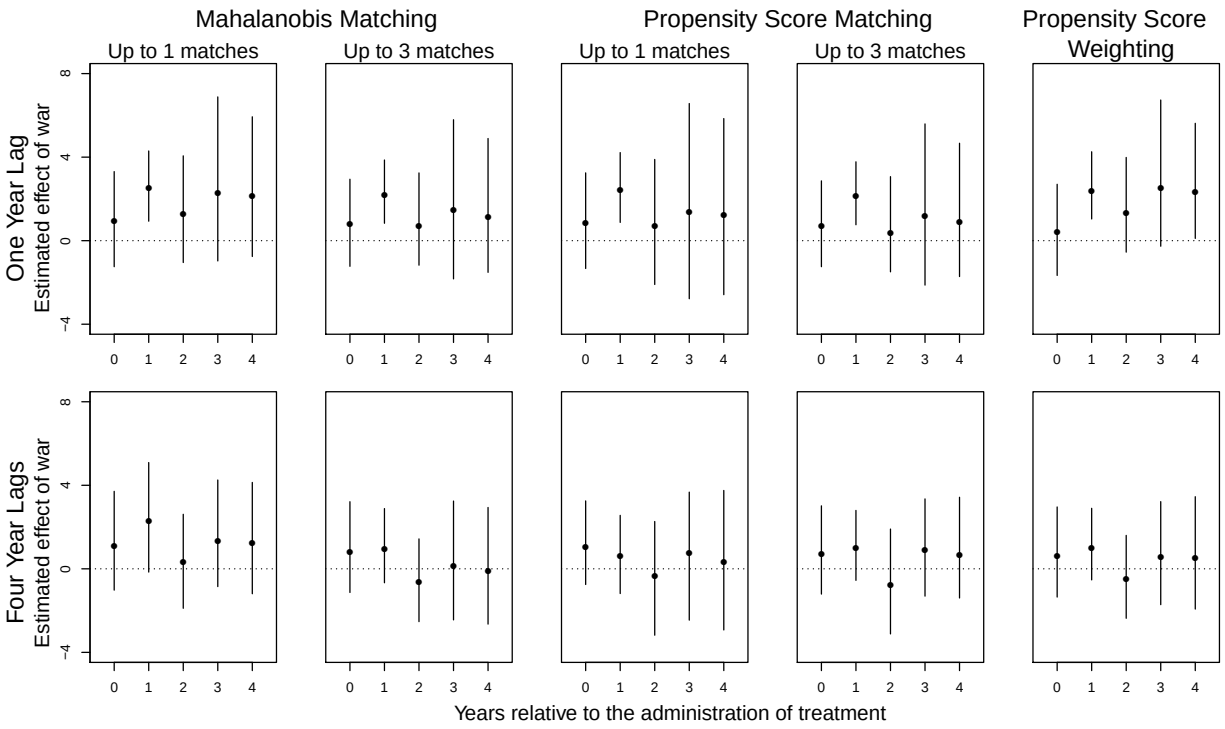


Figure 7: **Estimated Average Effects of Interstate War on Inheritance Tax Rate**. The matching method adjusts for the treatment and covariate histories during the one (upper panel) or four (lower panel) year period prior to the treatment. The estimated effects are shown for the period of five years after the war, i.e., $F = 0, 1, \ldots, 4$, with 95% bootstrap confidence intervals as vertical bars. Five different matching/weighting methods are considered and their results are presented in different columns.

least in the short and medium term.[12]

Next, Figure 7 shows the results based on matching methods for estimating the ATT of interstate war on inheritance tax. The upper panel shows the estimates based on the refinement of matched sets while adjusting for the treatment and covariates from one year period prior to the treatment. In contrast, the lower panel presents the estimates based on the adjustment for the four year pre-treatment period. As in the previous figure, each column represents the results based on a different matching/weighting method, and the vertical bars indicate the 95% confidence intervals based on block bootstrap.

We find that if we refine the matched set using the one year pre-treatment period, most of the estimated effects are not statistically significant for Mahalanobis and propensity score matching methods. In contrast, the results for propensity score weighting show larger point estimates.

[12]The original authors also seek to separately estimate the effects of democratic transition and authoritarian reversal, using the linear regression models. In Appendix E discusses this approach in detail. As shown in the appendix, the empirical results obtained from this approach substantively differ from those presented here.

However, all of the estimated causal effects are not statistically significant if we refine the matched sets by adjusting for the four year pre-treatment period. This sensitivity may come from the fact that as shown in the right panel of Figure 1 there is little variation in the treatment variable of this study. Given that the results based on the four year adjustment are likely to be more credible, our analysis suggests that it is difficult to conclusively establish the positive effects of war on inheritance tax rate.

6 Concluding Remarks

Due to its simplicity and transparency, matching methods have become part of tool kit for empirical researchers across different disciplines who wish to estimate causal effects in observational studies. Yet, most matching methods have been developed for causal inference with cross-sectional data. And even a small number of existing matching and weighting methods focus on simple settings in which each unit receives the treatment at most once and there exists no treatment reversal and are often based on linear models.

In the current paper, we fill this gap in the methodological literature by developing a methodological framework that enables the application of matching methods to causal inference with time-series cross section (TSCS) data. A main advantage of the proposed methodology over popular linear regression models with fixed effects is that it clarifies the source of information used to estimate counterfactual outcomes. In addition to transparency, our methods also offer simple diagnostics through balance checking.

The proposed methodology can be extended in a number of ways. First, while we focus on the binary treatment variable in this paper, the method can be extended to deal with a non-binary (e.g., continuous) treatment variable by possibly combining it with a model-based approach. Second, it is of interest to relax the assumption of no interference across units. While we allow for some degree of carryover effects (i.e., the possibility that past treatments affect future outcomes), the proposed methodology assumes the absence of spillover effects (i.e., one unit's treatment does not affect the outcomes of other units). Within the proposed matching framework, we can address this limitation by, for example, matching on the treatment history of one's neighbors as well as its own treatment history. We plan to explore such extensions of the proposed methods in our future research.

References

Abadie, Alberto. 2005. "Semiparametric Difference-in-Differences Estimators." *Review of Economic Studies* 72:1–19.

Abadie, Alberto, Alexis Diamond and Jens Hainmueller. 2010. "Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program." *Journal of the American statistical Association* 105:493–505.

Abadie, Alberto and Guido W. Imbens. 2008. "On the Failure of the Bootstrap for Matching Estimators." *Econometrica* 76:1537–1557.

Abadie, Alberto and Guido W. Imbens. 2011. "Bias-Corrected Matching Estimators for Average Treatment Effects." *Journal of Business and Economic Statistics* 29:1–11.

Abraham, Sarah and Liyang Sun. 2018. Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects. Technical Report. Department of Economics, Massachusetts Institute of Technology.

Acemoglu, Daron, Suresh Naidu, Pascual Restrepo and James A. Robinson. 2019. "Democracy Does Cause Growth." *Journal of Political Economy* 127:47–100.

Angrist, Joshua D. and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press.

Arellano, Manuel and Stephen Bond. 1991. "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations." *Review of Economic Studies* 58:277–297.

Aronow, Peter and Cyrus Samii. 2017. "Estimating Average Causal Effects Under General Interference." *Annals of Applied Statistics* 11:1912–1947.

Athey, Susan and Guido Imbens. 2018. Design-based Analysis in Difference-in-Differences Settings with Staggered Adoption. Technical Report. Stanford Graduate School of Business `https://arxiv.org/abs/1808.05293`: .

Beck, Natheniel and Jonathan N. Katz. 1995. "What to do (and not to do) with time-series cross-section data." *American Political Science Review* 89:634–647.

Ben-Michael, Eli, Avi Feller and Jesse Rothstein. 2019a. The Augmented Synthetic Control Method. Technical Report. arXiv:1811.04170.

Ben-Michael, Eli, Avi Feller and Jesse Rothstein. 2019b. Synthetic Controls and Weighted Event Studies with Staggered Adoption. Technical Report. arXiv:1912.03290.

Bertrand, Marianne, Esther Duflo and Sendhil Mullainathan. 2004. "How Much Should We Trust Differences-in-Differences Estimates?" *Quarterly Journal of Economics* 119:249–275.

Blackwell, Matthew and Adam Glynn. 2018. "How to Make Causal Inferences with Time-Series Cross-Sectional Data under Selection on Observables." *American Political Science Review* Forthcoming.

Chaisemartin, Clément de and Xavier D'Haultfœuille. 2018. Two-way Fixed Effects Estimators with Heterogeneous Treatment Effects. Technical Report. Department of Economics, University of California, Santa Barbara `https://arxiv.org/abs/1803.08807`: .

Diamond, Alexis and Jasjeet Sekhon. 2013. "Genetic Matching for Estimating Causal Effects: A New Method of Achieving Balance in Observational Studies." *Review of Economics and Statistics* 95:932–945.

Doudchenko, Nikolay and Guido Imbens. 2017. Balancing, Regression, Difference-In-Differences and Synthetic Control Methods: A Synthesis. Technical Report. arXiv:1610.07748.

Gerring, John, Strom C Thacker and Rodrigo Alfaro. 2012. "Democracy and human development." *The Journal of Politics* 74:1–17.

Goodman-Bacon, Andrew. 2018. Difference-in-Differences with Variation in Treatment Timing. Working Paper No. 25018. National Bureau of Economic Research.

Hansen, Ben B. 2004. "Full Matching in an Observational Study of Coaching for the SAT." *Journal of the American Statistical Association* 99:609–618.

Hernán, Miguel A. and James M. Robins. 2016. "Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available." *American Journal of Epidemiology* 183:758–764.

Hirano, Keisuke, Guido Imbens and Geert Ridder. 2003. "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score." *Econometrica* 71:1307–1338.

Ho, Daniel E., Kosuke Imai, Gary King and Elizabeth A. Stuart. 2007. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." *Political Analysis* 15:199–236.

Hudgens, Michael G. and Elizabeth Halloran. 2008. "Toward Causal Inference with Interference."

Journal of the American Statistical Association 103:832–842.

Iacus, Stefano, Gary King and Giuseppe Porro. 2011. "Multivariate Matching Methods That Are Monotonic Imbalance Bounding." *Journal of the American Statistical Association* 106:345–361.

Imai, Kosuke and In Song Kim. 2011. On the Use of Linear Fixed Effects Regression Models for Causal Inference. Technical Report. Princeton University.

Imai, Kosuke and In Song Kim. 2019. "When Should We Use Linear Unit Fixed Effects Regression Models for Causal Inference with Longitudinal Data?" *American Journal of Political Science* 63:467–490.

Imai, Kosuke and In Song Kim. 2020. "On the Use of Two-way Fixed Effects Regression Models for Causal Inference with Panel Data." *Political Analysis*.

Imai, Kosuke and Marc Ratkovic. 2014. "Covariate Balancing Propensity Score." *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* 76:243–263.

Imai, Kosuke and Marc Ratkovic. 2015. "Robust Estimation of Inverse Probability Weights for Marginal Structural Models." *Journal of the American Statistical Association* 110:1013–1023.

Imai, Kosuke, Zhichao Jiang and Anup Malai. 2020. "Causal Inference with Interference and Noncompliance in Two-Stage Randomized Experiments." *Journal of the American Statistical Association*.

Imbens, Guido W. and Donald B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.

Li, Yunfei Paul, Kathleen J. Propert and Paul R. Rosenbaum. 2001. "Balanced Risk Set Matching." *Journal of the American Statistical Association* 96:870–882.

Nickell, Stephen. 1981. "Biases in Dynamic Models with Fixed Effects." *Econometrica* 49:1417–1426.

Nielsen, Rich and John Sheffield. 2009. Matching with Time-Series Cross-Sectional Data. Technical Report. Harvard University. `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.510.7097&rep=rep1&type=pdf`.

Otsu, Taisuke and Yoshiyasu Rai. 2017. "Bootstrap Inference of Matching Estimators for Average Treatment Effects." *Journal of the American Statistical Association* 112:1720–1732.

Papaioannou, Elias and Gregorios Siourounis. 2008. "Democratisation and growth." *The Economic*

Journal 118:1520–1551.

Przeworski, Adam, R Michael Alvarez, Michael E Alvarez, Jose Antonio Cheibub, Fernando Limongi et al. 2000. *Democracy and Development: Political Institutions and Well-being in the World, 1950-1990*. Vol. 3 Cambridge University Press.

Robins, James M. 1994. "Correcting for non-compliance in randomized trials using structural nested mean models." *Communications in Statistics – Theory and Methods* 23:2379–2412.

Robins, James M., Miguel Ángel Hernán and Babette Brumback. 2000. "Marginal Structural Models and Causal Inference in Epidemiology." *Epidemiology* 11:550–560.

Rosenbaum, P. R. and D. B. Rubin. 1983. "Assessing Sensitivity to An Unobserved Binary Covariate in An Observational Study With Binary Outcome." *Journal of the Royal Statistical Society, Series B, Methodological* 45:212–218.

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

Rubin, Donald B. 2006. *Matched Sampling for Causal Effects*. Cambridge: Cambridge University Press.

Scheve, Kenneth and David Stasavage. 2012. "Democracy, war, and wealth: lessons from two centuries of inheritance taxation." *American Political Science Review* 106:81–102.

Stuart, Elizabeth A. 2010. "Matching Methods for Causal Inference: A Review and a Look Forward." *Statistical Science* 25:1–21.

Tchetgen Tchetgen, Eric J. and Tyler J. VanderWeele. 2010. "On causal inference in the presence of interference." *Statistical Methods in Medical Research* 21:55–75.

Xu, Yiqing. 2017. "Generalized Synthetic Control Method: Causal Inference with Interactive Fixed Effects Models." *Political Analysis* 25:57–76.

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

Supplementary Appendix

A Proof of Theorem 1

Let $A_{it} = 2X_{it} - 1$. We consider the following a general definition of the weights,

$$W_{it} \;=\; \sum_{i'=1}^{N}\sum_{t'=1}^{T} D_{i't'} \cdot v_{it}^{i't'} \quad \text{and} \quad v_{it}^{i't'} \;=\; \begin{cases} A_{it} & \text{if } (i,t) = (i', t'+F) \\ 1 & \text{if } (i,t) = (i', t'-1) \\ -A_{it} \cdot w_{i't'}^{i} & \text{if } i \in \mathcal{M}_{i't'}, t = t'+F \\ -w_{i't'}^{i} & \text{if. } i \in \mathcal{M}_{i't'}, t = t'-1 \\ 0 & \text{otherwise.} \end{cases}$$

Note that the quantity of interest given in equation (19) implies that $A_{it} = 1$ if $(i,t) = (i', t'+F)$, and $A_{it} = -1$ if $(i,t) \in \mathcal{M}_{i't'}, t = t'+F$ as the treatment status does not change for at least F time periods once treatment is administered at time t. This gives the weights in equation (26).

We begin this proof by establishing the following algebraic equality. Specifically, we prove that for any unit-specific constant α_i^*, the following equality holds,

$$\begin{aligned}
&\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it} A_{it} \alpha_i^* \\
&= \sum_{i'=1}^{N}\sum_{t'=1}^{T} D_{i't'} \left(\sum_{i=1}^{N}\sum_{t=1}^{T} v_{it}^{i't'} A_{it} \alpha_i^* \right) \\
&= \sum_{i'=1}^{N}\sum_{t'=1}^{T} D_{i't'} \left(1 - 1 - \sum_{i \in \mathcal{M}_{i't'}} \sum_{t=t'+F} A_{it}^2 \cdot w_{i't'}^{i} - \sum_{i \in \mathcal{M}_{i't'}} \sum_{t=t'-1} A_{it} \cdot w_{i't'}^{i} \right) \alpha_i^* \\
&= \sum_{i'=1}^{N}\sum_{t'=1}^{T} D_{i't'} \left(1 - 1 - \sum_{i \in \mathcal{M}_{i't'}} \sum_{t=t'+F} w_{i't'}^{i} + \sum_{i \in \mathcal{M}_{i't'}} \sum_{t=t'-1} w_{i't'}^{i} \right) \alpha_i^* \\
&= \sum_{i'=1}^{N}\sum_{t'=1}^{T} D_{i't'} \left(1 - 1 - 1 + 1\right) \alpha_i^* \;=\; 0
\end{aligned} \tag{A1}$$

where the second equality follows from the fact that $A_{it} = 1$ if $(i,t) = (i', t'+F)$, $A_{it} = -1$ if $(i,t) = (i', t'-1)$, and $A_{it} = -1$ if $(i,t) \in \mathcal{M}_{i't'}, t = t'-1$ as given by equation (19). The last equality if from $\sum_{i \in \mathcal{M}_{i't'}} w_{i't'}^{i} = 1$.

Following the same logic, it is straightforward to show that $\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it} A_{it} \gamma_t^* = 0$ for any time-specific constant γ_t^* and $\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it} A_{it} K^* = 0$ for any constant K^*. This implies that

$$A_{it} - \overline{A}_i^* - \overline{A}_t^* + \overline{A}^* = A_{it} \tag{A2}$$

where $\overline{A}_i^* = \sum_{t=1}^{T} W_{it} A_{it} / \sum_{t=1}^{T} W_{it}$, $\overline{A}_t^* = \sum_{i=1}^{N} W_{it} A_{it} / \sum_{i=1}^{N} W_{it}$, $\overline{A}^* = \sum_{i=1}^{N}\sum_{t=1}^{T} W_{it} A_{it} / \sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}$.

Second, we show the following algebraic equality,

$$\begin{aligned}
&\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it} \\
&= \sum_{i=1}^{N}\sum_{t=1}^{T} \left(\sum_{i'=1}^{N}\sum_{t'=1}^{T} D_{i't'} \cdot v_{it}^{i't'} \right)
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i'=1}^{N}\sum_{t'=1}^{T} D_{i't'}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} v_{it}^{i't'}\right) \\
&= \sum_{i'=1}^{N}\sum_{t'=1}^{T} D_{i't'}\left(1+1+\sum_{i\in\mathcal{M}_{i't'}}\sum_{t=t'+F} w_{i't'}^{i} - \sum_{i\in\mathcal{M}_{i't'}}\sum_{t=t'-1} w_{i't'}^{i}\right) \\
&= 2\sum_{i'=1}^{N}\sum_{t'=1}^{T} D_{i't'} = 2\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}. \qquad \text{(A3)}
\end{aligned}
$$

Finally, we can derive the desired result,

$$
\begin{aligned}
\hat{\beta}_{\mathsf{DiD}} &= \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}(A_{it}-\overline{A}_i^*-\overline{A}_t^*+\overline{A}^*)(Y_{it}-\overline{Y}_i^*-\overline{Y}_t^*+\overline{Y}^*)}{\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}(A_{it}-\overline{T}_i^*-\overline{T}_t^*+\overline{T}^*)^2} \\
&= \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}A_{it}(Y_{it}-\overline{Y}_i^*-\overline{Y}_t^*+\overline{Y}^*)}{\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}} \\
&= \frac{1}{2\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}}\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}A_{it}(Y_{it}-\overline{Y}_i^*-\overline{Y}_t^*+\overline{Y}^*) \\
&= \frac{1}{2\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}}\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}A_{it}Y_{it} \\
&= \frac{1}{2\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{i'=1}^{N}\sum_{t'=1}^{T} D_{i't'}\cdot v_{it}^{i't'}A_{it}Y_{it} \\
&= \frac{1}{2\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}}\sum_{i'=1}^{N}\sum_{t'=1}^{T} D_{i't'}\sum_{i=1}^{N}\sum_{t=1}^{T} v_{it}^{i't'}A_{it}Y_{it} \\
&= \frac{1}{2\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}}\sum_{i'=1}^{N}\sum_{t'=1}^{T} D_{i't'}\left(Y_{i',t'+F}-Y_{i',t'-1}-\sum_{i\in\mathcal{M}_{i't'}}\sum_{t=t'+F} w_{i't'}^{i}Y_{it}+\sum_{i\in\mathcal{M}_{i't'}}\sum_{t=t'-1} w_{i't'}^{i}Y_{it}\right) \\
&= \frac{1}{2\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}}\sum_{i'=1}^{N}\sum_{t'=1}^{T} D_{i't'}\left(Y_{i',t'+F}-Y_{i',t'-1}-\sum_{i\in\mathcal{M}_{i't'}} w_{i't'}^{i}Y_{i,t'+F}+\sum_{i\in\mathcal{M}_{i't'}} w_{i't'}^{i}Y_{i,t'-1}\right) \\
&= \frac{1}{2\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}}\sum_{i=1}^{N}\sum_{t=L+1}^{T-F} D_{it}\left\{(Y_{i,t+F}-Y_{i,t-1})-\sum_{i'\in\mathcal{M}_{it}} w_{it}^{i'}\left(Y_{i',t+F}-Y_{i',t-1}\right)\right\} \\
&= \hat{\delta}(F,L)/2
\end{aligned}
$$

where the second equality follows from equation (A2), the third equality follows from equation (A3), and the fourth equality is implied by equation (A1). The second from the last equality follows from the fact that $D_{it}=0$ for $t<L+1$ and $t>T-F$ for any unit i, because there will be no matched for such units by construction. This concludes the proof because $2\hat{\beta}_{\mathsf{DiD}}=\hat{\delta}(F,L)$ (see Theorem 1). Note that the multiplication by 2 is required due to the change of the variable of the original treatment variable, i.e., $A_{it}=2X_{it}-1$. □

B A Simulation Study

B.1 The Setup

To make our simulation studies realistic, we use the original data from Acemoglu et al. (2019). For simplicity, we begin by creating a balanced TSCS data set with $N = 162$ units and $T = 51$ time periods although our method can handle missing and/or unbalanced data. Since the original data set is unbalanced, we impute missing values for continuous (binary) variables based on linear (logistic) regression models.[13] We emphasize that the proposed methodology does not require the data to be balanced. Next, we generate the binary treatment variable X_{it} and the outcome variable Y_{it}, with true data generating process given by

$$X_{it} \sim \text{Benoulli}\left(\text{logit}^{-1}\left\{\tilde{\alpha}_i + \tilde{\gamma}_t + \sum_{\ell=1}^{L} \tilde{\beta}_\ell^\top X_{i,t-\ell} + \sum_{\ell=0}^{L}\left(\tilde{\zeta}_\ell^\top \mathbf{Z}_{i,t-\ell} + \tilde{\phi}_\ell^\top [\mathbf{Z}^{(1)}_{i,t-\ell} : \mathbf{Z}^{(3)}_{i,t-\ell}]\right)\right\}\right) \quad \text{(A4)}$$

$$Y_{it} = \alpha_i + \gamma_t + \sum_{\ell=0}^{L} \beta_\ell^\top X_{i,t-\ell} + \sum_{\ell=0}^{L}\left(\zeta_\ell^\top \mathbf{Z}_{i,t-\ell} + \phi_\ell^\top [\mathbf{Z}^{(1)}_{i,t-\ell} : \mathbf{Z}^{(3)}_{i,t-\ell}]\right) + \epsilon_{it}, \quad \text{(A5)}$$

where $\epsilon \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$, α_i and γ_t are unit and time fixed effects, and $\mathbf{Z}_{i,t-\ell} = (\mathbf{Z}^{\top(1)}_{i,t-\ell}, \mathbf{Z}^{\top(2)}_{i,t-\ell}, \mathbf{Z}^{\top(3)}_{i,t-\ell})^\top$ is the lagged covariates consisting of continuous variables $\mathbf{Z}^{(1)}_{i,t-\ell}$, binary variables $\mathbf{Z}^{(2)}_{i,t-\ell}$, and a set of other continuous variables $\mathbf{Z}^{(3)}_{i,t-\ell}$ that have interactive effects with $\mathbf{Z}^{(1)}_{i,t-\ell}$ with the interaction being represented by a colon.[14] We set the values of all the parameters to the actual estimates obtained from fitting the above treatment and outcome models to the imputed data set, including the true contemporaneous treatment effect β_0 to -7.5. Finally, we set σ to the sample variance of the outcome variable.

We use $L = 3$ such that the true dynamic data generating process includes the lagged treatment, covariates, and the interaction between the covariates across three time periods. For each of 1,000 independent Monte Carlo replications, we generate the treatment and outcome variables according to the models above while keeping the covariates $\mathbf{Z}_{it}$ fixed. To evaluate the robustness of the proposed methodology to model misspecification, we consider three scenarios: (1) severe misspecification with only one period of lagged variables (i.e., $L = 1$), (2) moderate misspecification with only two periods of lagged variables (i.e., $L = 2$), and (3) correct specification (i.e., $L = 3$). Under each scenario, we compare the performance of the ordinary least squares (OLS) estimator with that of three diffrent refinement methods: (1) Mahalanobis distance matching with at most 10 maches (i.e., $J = 10$), (2) propensity score matching with $J = 10$, and (3) propensity score weighting. To make a fair comparison across all four methods, we use the same covariate information by using the identical set of lagged variables for both computing the OLS estimator and refining matched sets. This means that under the two scenarios of model misspecification, the propensity score model, which is estimated via CBPS, is also misspecified. Finally, we also compute the 95% confidence interval for each method in order to evaluate the its coverage rate.

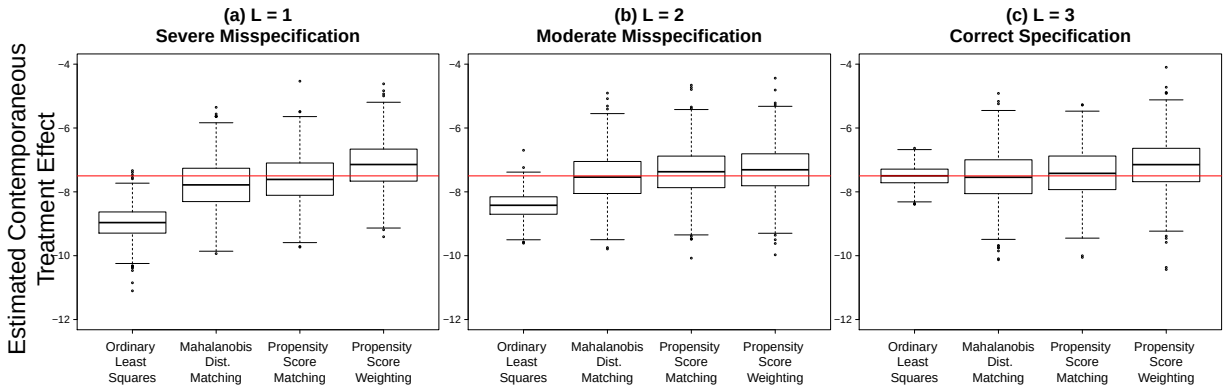


Figure A1: **Robustness of the Proposed Methodology to Model Misspecification with** $N = 162$. This figure summarizes the simulation studies across three levels of misspecification. Panels (a) and (b) show that the ordinary least squares (OLS) estimator with unit and time fixed effects yields a significant bias as the severity of misspecification increases. In contrast, the proposed methodology, based on three different refinement methods (Mahalanobis matching, Propensity score matching, and Propensity score weighting), returns similar estimates of the quantity of interest under the two different model misspecifications. Panel (c) shows that when the model is correctly specified, the OLS estimator is unbiased and most efficient.

B.2 Results

Figure A1 presents the boxplot of each estimator across 1,000 Monte Carlo simulations under each scenario. Panels (a) and (b) show that the proposed methodology significantly less biased than the OLS estimator when the model is misspecified. That is, the OLS estimates tend to substantially underestimate the coefficient of interest (the horizontal red line at -7.5) whereas our matching estimators yield roughly unbiased estimates regardless of the degree of model misspecification. Although the OLS is generally more efficient than the matching estimators, its variance increases as the degree of model misspecification increases. As expected, when the model is correctly specified, the OLS is unbiased and most efficient. In contrast, the matching estimators have similar variances across all scenarios. Thus, the matching estimators are less sensitive to model misspecification in terms of both bias and variance.

Table A1 further investigates the bias-variance tradeoff across the estimators. We find that under the two model misspecification scenarios considered here, the root mean squared error (RMSE) of the OLS estimator exceeds those of the matching estimators. This suggests that although the OLS estimator is more efficient than the matching estimators, its bias increases quickly once the model becomes misspecified. In contrast, the RMSE of the matching estimators stays relatively stable across model (mis)specifications considered here. Finally, the 90% confidence intervals of the proposed methodology maintain a reasonable coverage rate (Cov.) whereas the corresponding confidence intervals of the OLS estimator have a poor coverage unless the model is correctly specified. Overall, we find that the matching estimators outperform the OLS estimator unless the model is correctly specified.

[13] We use a linear time trend for variables when 35% or less of the data is missing, while including a quadratic time trend when the missingness is more severe. We added a small error term from a normal distribution by setting the standard error to the standard deviation of the difference in the variable between two consecutive time periods.

[14] For the simulation analysis, $\mathbf{Z}_{it}$ includes the log population, log population of age below 16 years, the log population of age above 64 years, net financial flow as a fraction of GDP, trade volume as a fraction of GDP, and a dichotomous measure of social unrest. We use the log population as the single variable $\mathbf{Z}_{it}^{(3)}$ for the interaction term.

Method	L = 1 Severe Misspecification				L =2 Moderate Misspecification				L = 3 Correct Specification			
	Bias	SD	RMSE	Cov.	Bias	SD	RMSE	Cov.	Bias	SD	RMSE	Cov.
Ordinary Least Squares	−1.46	0.51	1.55	0.141	−0.93	0.42	1.02	0.39	0.00	0.31	0.31	0.96
Mahalanobis Dist. Matching	−0.27	0.75	0.80	0.90	−0.04	0.75	0.75	0.92	−0.03	0.79	0.79	0.93
Propensity Score Matching	−0.11	0.74	0.75	0.94	0.12	0.75	0.76	0.93	0.07	0.77	0.78	0.94
Propensity Score Weighting	0.35	0.75	0.83	0.93	0.2	0.77	0.79	0.91	0.33	0.84	0.90	0.93

Table A1: **Results of Simulation Studies with** $N = 162$**.** When the model is misspecified, the proposed methodology exhibits a smaller bias and Root Mean Square Error (RMSE), compared to the ordinary least squares OLS estimator. The OLS estimator is generally more biased but less variable than the matching and weighting methods, as shown by the smaller standard deviation (SD). The 90% confidence intervals of the proposed methodology produces reasonable coverage rates (Cov.) under all simulation scenarios whereas the OLS estimator results in substantial under-coverage unless the model is correctly specified.

Why does the matching estimator perform well even under the model misspecification? To examine this question, we investigate the covariate balance achieved by our matching methods. Figure A2 shows covariate balance for each refinement method (Mahalanobis distance matching in the top row, propensity score matching in the middle row, and propensity score weighting in the bottom row) under three scenarios (columns). In each plot, we present the distribution of the average absolute standardized mean difference across covariates for contemporaneous ("L0") and each of the three lags ("L1" through "L3"). We find that across three methods the covariate balance between correct specification and moderate misspecification is similar. The covariate balance under the severe misspecification is noticeably worse than the other two scenarios. This is consistent with the performance of the matching estimators presented in Figure A1 and Table A1.

The bias reduction of our matching methods does not come free. Indeed, as can be seen from Figure A1 and Table A1, the variance of the proposed matching estimators is typically greater than that of the least squares estimator. We illustrate this point by computing the statistical power of each estimator with the 90% confidence level as shown in Figure A3 while varying the true value across various values from −1 to 1. We find that although the proposed matching estimators are less powerful than the least squares estimator across three scenarios, the statistical power of the latter depends on the true value in an asymmetrical fashion when the model is misspecified.

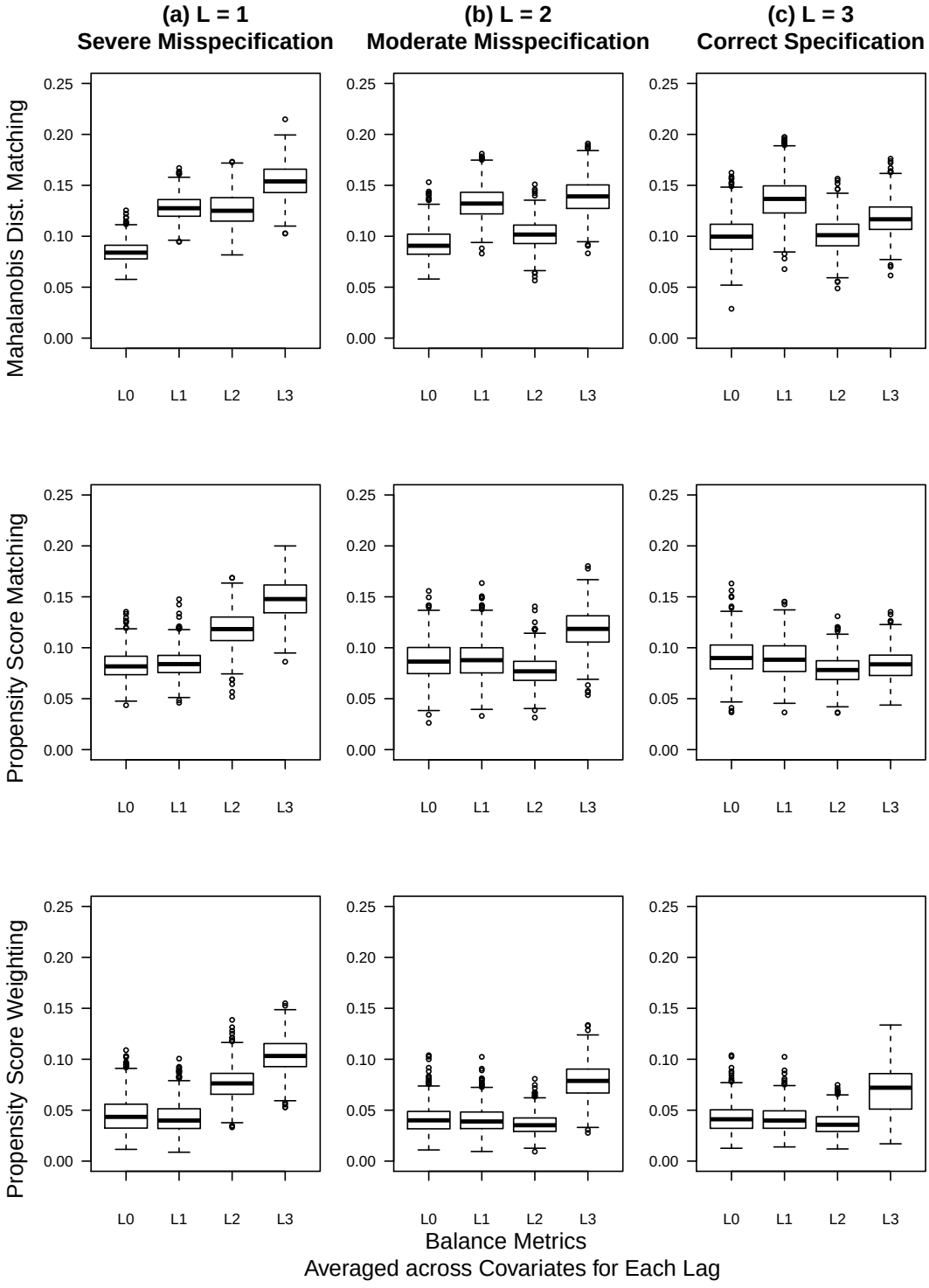


Figure A2: **Covariate Balance Achieved by the Proposed Methodology under Model Misspecification with $N = 162$.** This figure summarizes the simulation studies of Mahalanobis distance matching, propensity score matching, and propensity score weighting in first through third row across three levels of misspecification in columns (a) through (c). For each method under each scenario, a plot shows the distribution of the average absolute standardized mean difference across all covariates for the contemporaneous period ("L0") and three lag periods ("L1" through "L3").

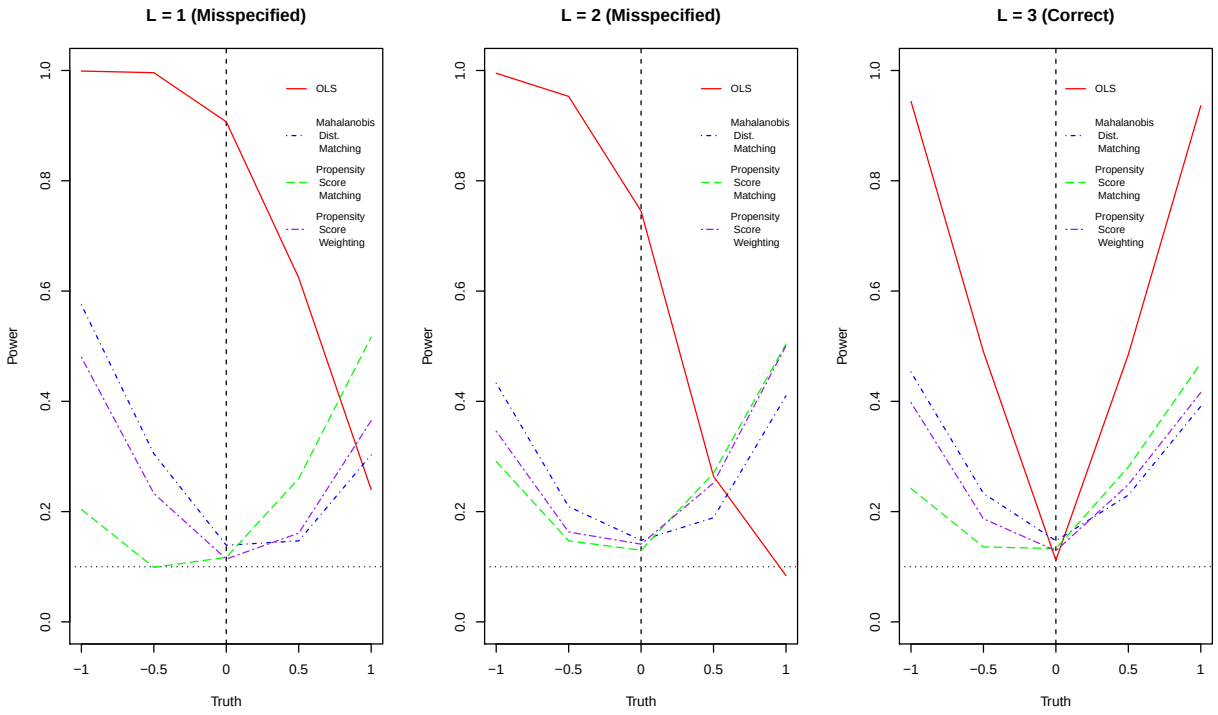


Figure A3: **Statistical Power of the Proposed Methodology to Model Misspecification with $N = 162$.** This figure summarizes the simulation studies across three levels of misspecification. It shows the statistical power of each estimator with the 90% confidence level across five different truths, $-1, -0.5, 0, 0.5, 1$. The results indicate that the proposed methodology is more conservative than OLS.

We find similar results when the sample size is smaller, $N = 50$. As shown in Figure A4 and Table A2, the least squares are much more sensitive to the model misspecification than the proposed matching estimators. Figure A5 shows that the covariate balance is reasonable so long as the model is not severely misspecified. As before, a better balance leads to a better performance of matching estimator. Finally, when the sample size is small, the statistical power of the proposed estimators further deteriorates (see Figure A6) and the coverage of the confidence interval diverges from the nominal coverage even under correct model specification (see also Table A2). This suggests that a large sample size is necessary for obtaining better uncertainty estimates.

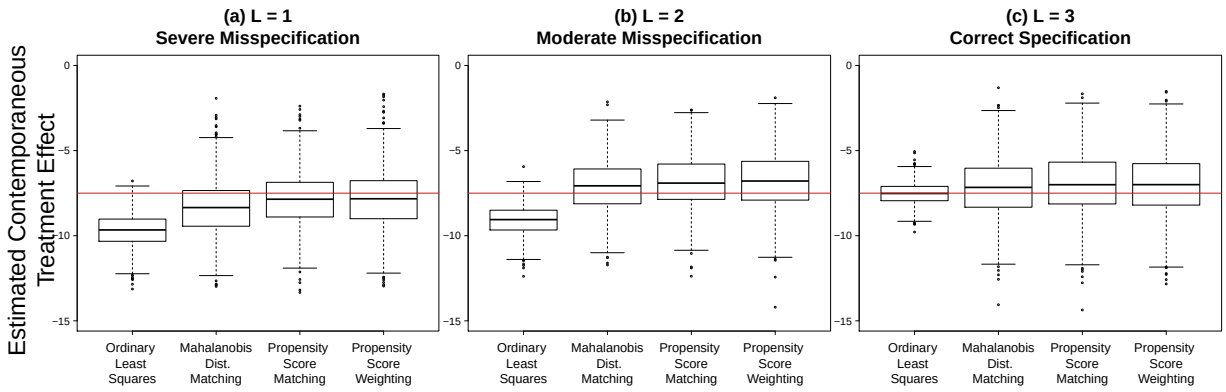


Figure A4: **Robustness of the Proposed Methodology to Model Misspecification with $N = 50$.** This figure summarizes the simulation studies across three levels of misspecification. Panels (a) and (b) show that the ordinary least squares (OLS) estimator with unit and time fixed effects yields a significant bias as the severity of misspecification increases. In contrast, the proposed methodology, based on three different refinement methods (Mahalanobis matching, Propensity score matching, and Propensity score weighting), returns similar estimates of the quantity of interest under the two different model misspecifications. Panel (c) shows that when the model is correctly specified, the OLS estimator is unbiased and most efficient.

Method	**L = 1 Severe Misspecification**				**L =2 Moderate Misspecification**				**L = 3 Correct Specification**			
	Bias	SD	RMSE	Cov.	Bias	SD	RMSE	Cov.	Bias	SD	RMSE	Cov.
Ordinary Least Squares	−2.18	0.50	2.40	0.35	−1.59	0.87	1.81	0.52	0.00	0.62	0.62	0.88
Mahalanobis Dist. Matching	−0.83	1.58	1.79	0.90	0.42	1.51	1.57	0.83	0.35	1.71	1.75	0.80
Propensity Score Matching	−0.34	1.53	1.56	0.91	0.66	1.56	1.69	0.80	0.58	1.77	1.86	0.81
Propensity Score Weighting	−0.36	1.70	1.74	0.92	0.70	1.66	1.81	0.80	0.52	1.84	1.91	0.78

Table A2: **Results of Simulation Studies.** When the model is misspecified, the proposed methodology exhibits a smaller bias and Root Mean Square Error (RMSE), compared to the ordinary least squares OLS estimator. The OLS estimator is generally more biased but less variable than the matching and weighting methods, as shown by the smaller standard deviation (SD). The 90% confidence intervals of the proposed methodology produces reasonable coverage rates (Cov.) under all simulation scenarios whereas the OLS estimator results in substantial under-coverage unless the model is correctly specified.

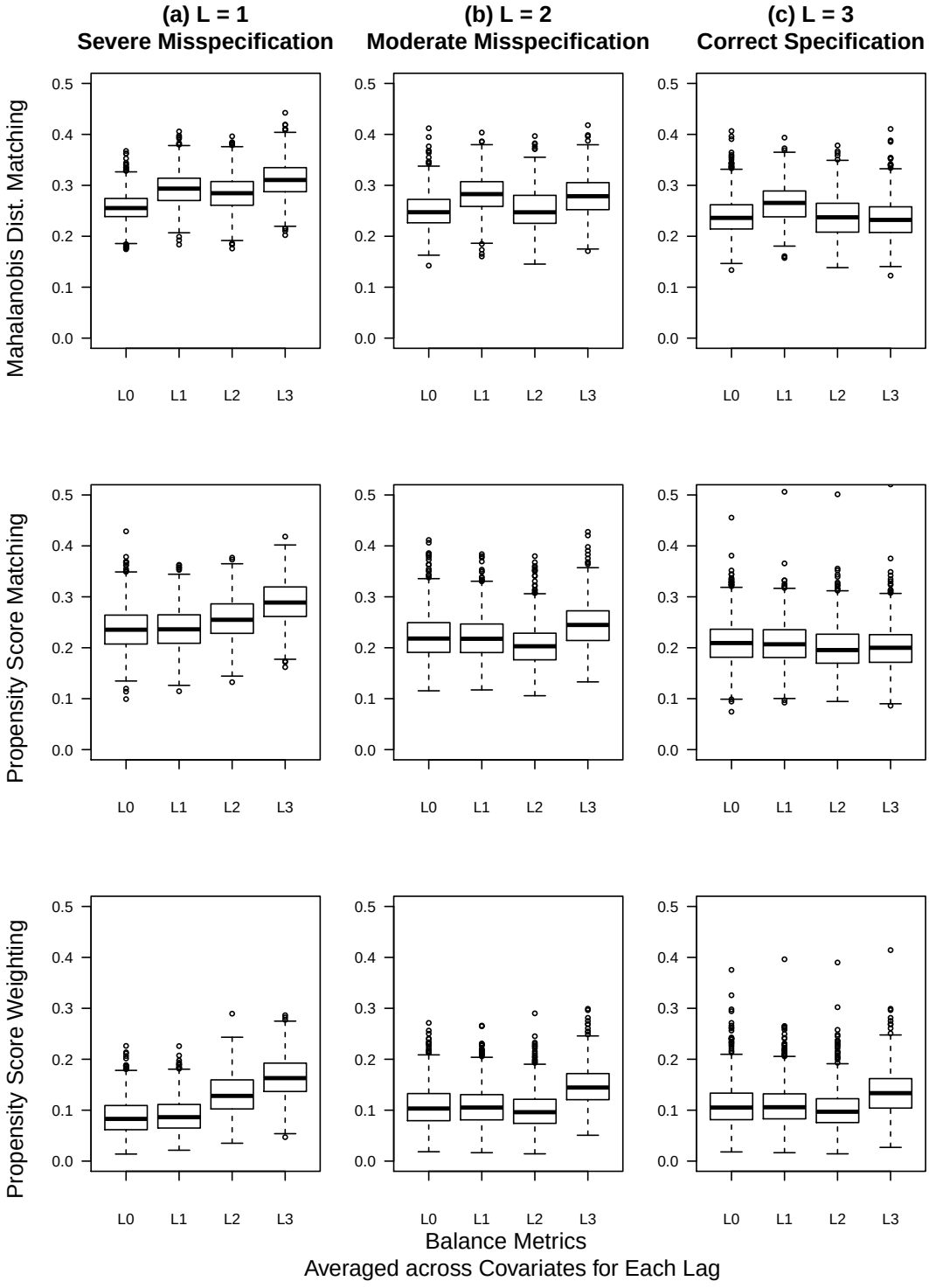


Figure A5: **Covariate Balance Achieved by the Proposed Methodology under Model Misspecification, TRUTH = -7.5 and N = 50.** This figure summarizes the simulation studies of Mahalanobis distance matching, propensity score matching, and propensity score weighting in first through third row across three levels of misspecification in panels (a) through (c), respectively. For each method under each misspecification scenario, there is a graph showing the absolute standardized mean difference across all covariates at each l, summarized across simulations in a boxplot.

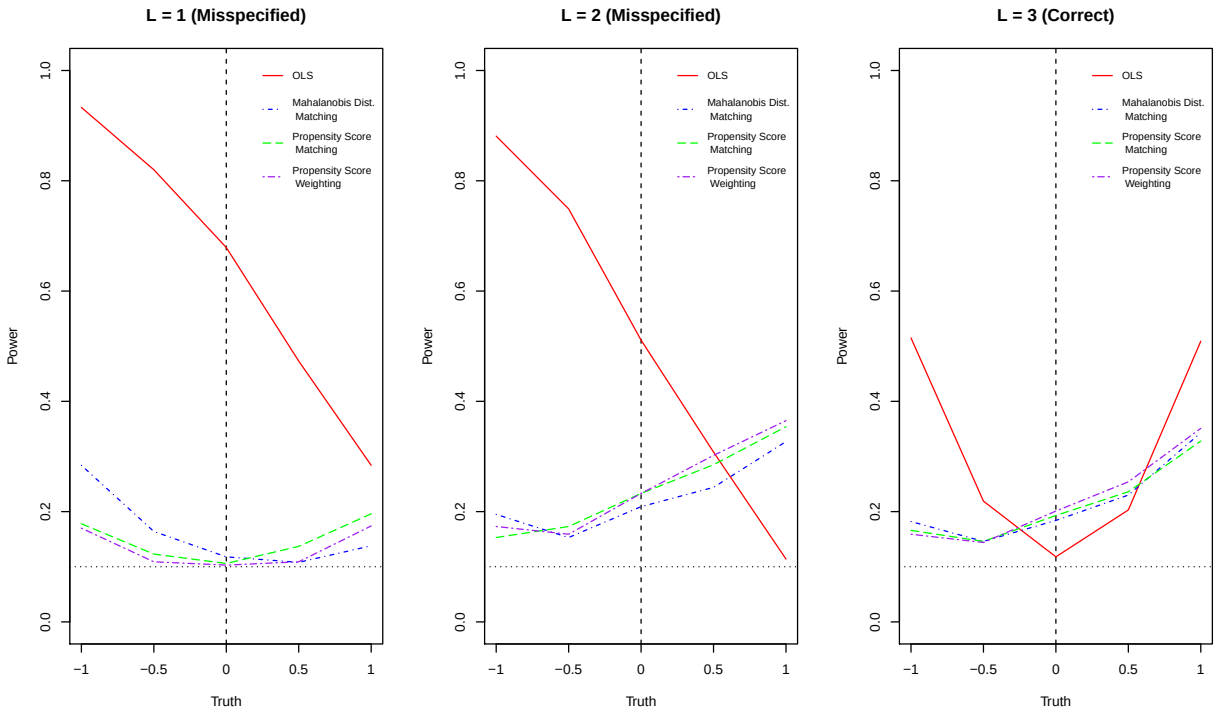


Figure A6: **Statistical Power of the Proposed Methodology to Model Misspecification, N = 50.** This figure summarizes the simulation studies across three levels of misspecification. It shows the rate of Type II error with 90% confidence interval across five different truths, -1, -0.5, 0, 0.5, 1. The results indicate that the proposed methodology is more conservative than OLS.

C Covariate Balance when the Treatment Reversal is Not Allowed

This appendix presents the covariate balance for the two empirical applications in the case where treatment reversal is not allowed (as opposed to the cases in the main text, which allow for the treatment reversal). That is, we present the covariate balance for "stable policy change" as described in equation (19). Below, we find that the covariate balance for stable policy change is far from satisfactory. As such, the resulting causal estimates are likely to be less credible than those presented in the main text where the treatment reversal is allowed.

First, we present scatter plots for $F = 1, 2, 3, 4$ in Figures A7— A10.[15] Notice that the covariate balance for stable policy change in the case of $F = 1$ (shown in Figures A7) already slightly deteriorates relative to its counterpart that allows for the treatment reversal in Figure 4. Notably, for the Scheve and Stasavage (2012) study, Figure A7 shows that the off-diagonal post-refinement covariate balance (those above the 45-degree line) with "Four Year Lags" tends to be further away from the 45-degree line compared to its counterpart in Figure 4, while the balance in general gets worse with propensity score matching (second row) with "One Year Lag."

Moverover, the covariate balance for the same study in the case of $F = 4$ (see Figure A10) exhibits further deterioration regardless of the matching methods and the choice of lags. For instance, several covariates have balances that are outside the range of the graph for all three methods with

[15] Note that when $F = 0$, we are estimating the contemporaneous effects and hence the treatment reversal does not matter. In this case, therefore, the covariate balance figures (both scatter and line plots) are identical to Figures 4 and 5.

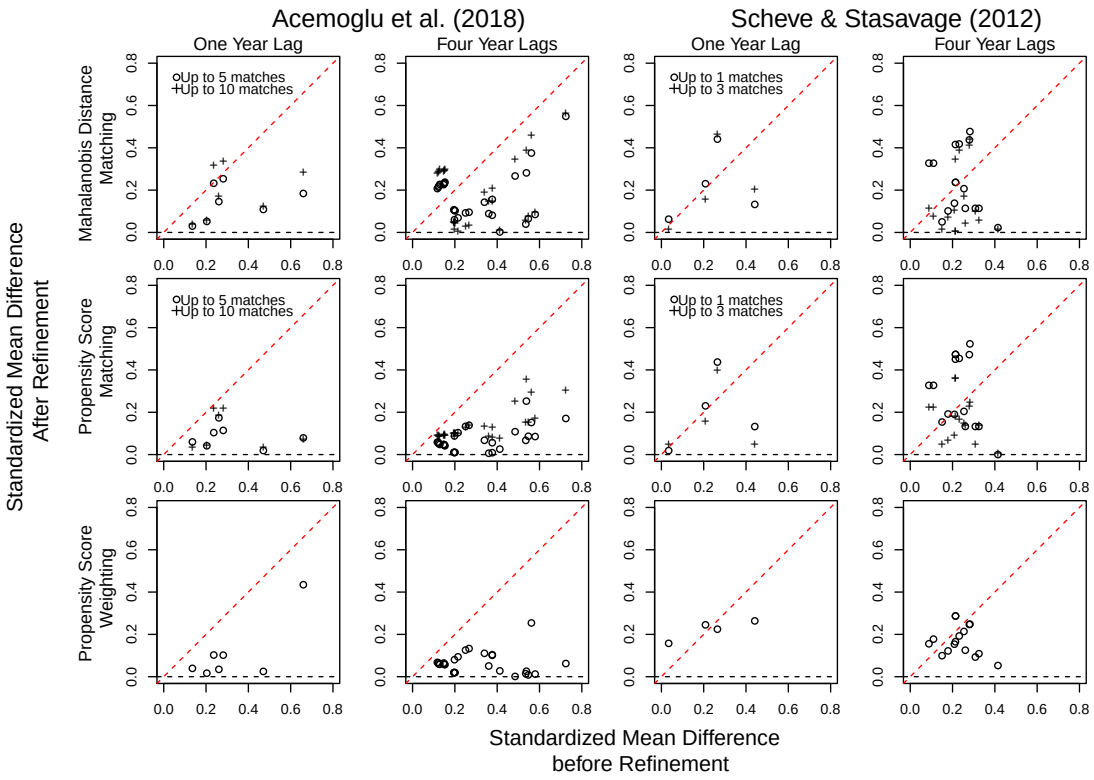


Figure A7: **Covariate Balance due to the Refinement of Matched Sets when Estimating the Average Effects of Stable Policy Change, with** $F = 1$. See the caption of Figure 4.

"One Year Lag." As for "Four Year Lags," the deterioration is also clear across methods for the Scheve and Stasavage (2012) study. Note that in contrast, balances for the Acemoglu et al. (2019) study show deterioration that is less severe across methods and the choices of lag.

Next, we present the line plots in Figures A11— A14. To begin, a comparison between Figures A11 and 5 demonstrates that the covariate balance based on Mahalanobis matching for the authoritarian reversal treatment in the Acemoglu et al. (2019) study is substantially worse. Notice that covariate imbalance exacerbates further as F increases to 2, 3, and 4. For example, when $F = 4$ (see Figure A14), the covariate balance lines for Mahalanobis distance matching, propensity score matching, and propensity score weighting are much further away from zero when compared to their counterparts in Figure 5 for the case of authoritarian reversal (second row).

Similarly, we observe a clear deterioration in covariate balance for the Scheve and Stasavage (2012) study (third row) when $F = 4$ for Mahalanobis distance matching and propensity score weighting. In addition, the number of unmatched treated observations increases when we do not allow for treatment reversal because a unit with treatment reversal no longer qualifies as a control unit. In the authoritarian reversal scenario, for example, the number of unmatched treated observations is 11, 15, 19, 22 for F = 1, 2, 3, 4, respectively. In the case of starting war as the treatment, the number of unmatched treated observations is 7, 8, 9, 19 for F = 1, 2, 3, 4 using four year lags.[16]

[16] Using one year lag, the numbers are 6, 8, 9, 19 for the four post-treatment periods, respectively.

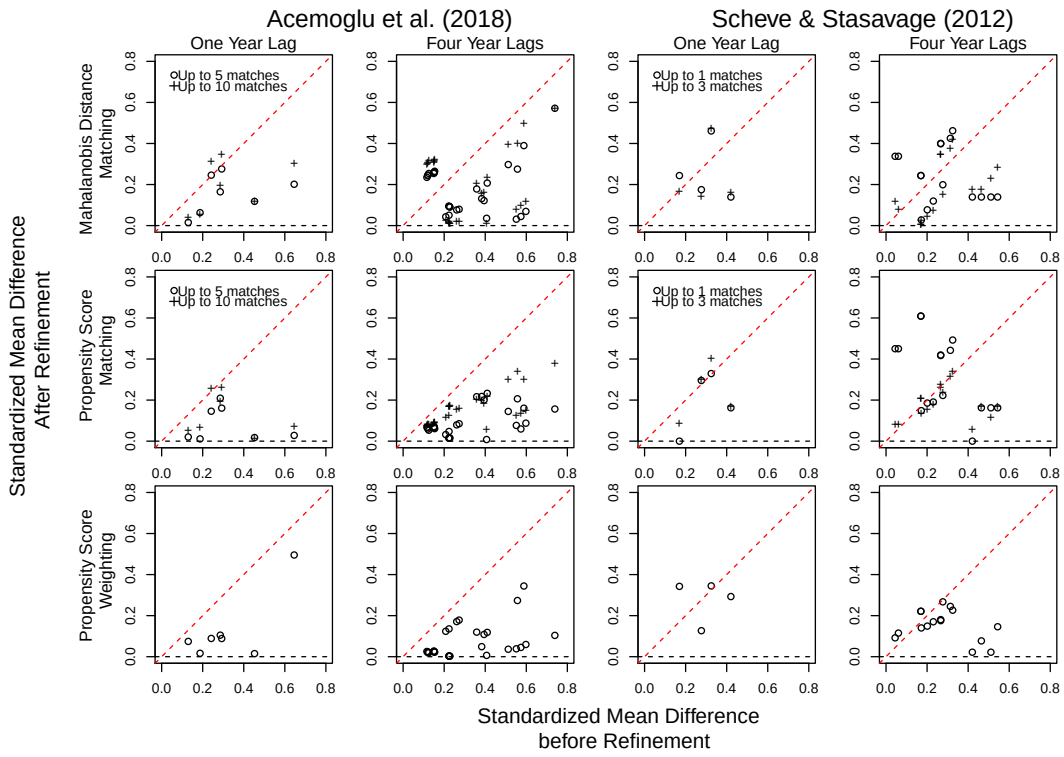


Figure A8: **Covariate Balance due to the Refinement of Matched Sets when Estimating the Average Effects of Stable Policy Change, with $F = 2$.** See the caption of Figure 4.

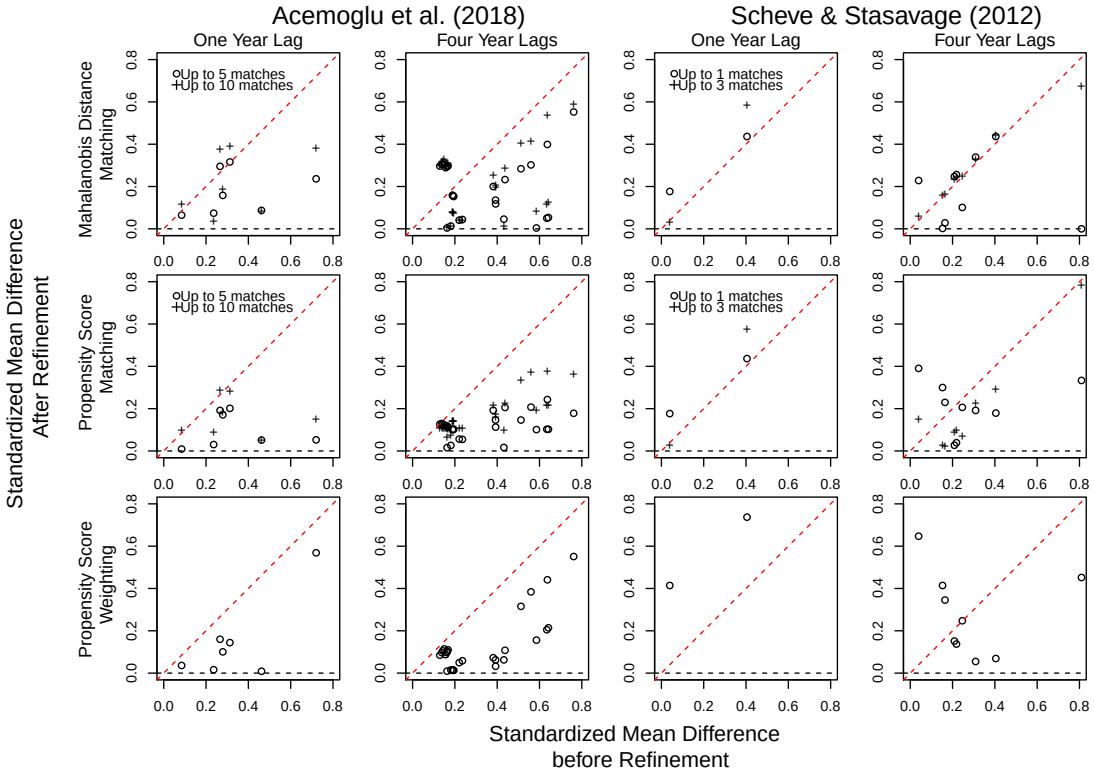


Figure A10: **Covariate Balance due to the Refinement of Matched Sets when Estimating the Average Effects of Stable Policy Change, with $F = 4$.** See the caption of Figure 4.

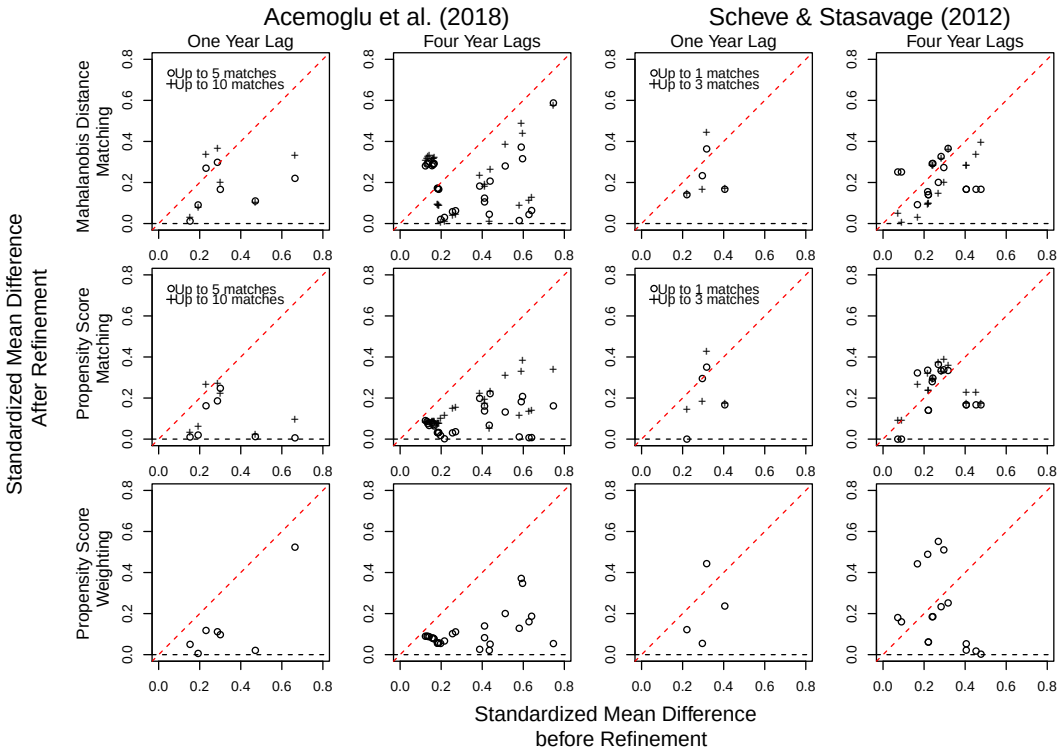


Figure A9: **Covariate Balance due to the Refinement of Matched Sets when Estimating the Average Effects of Stable Policy Change, with $F = 3$.** See the caption of Figure 4.

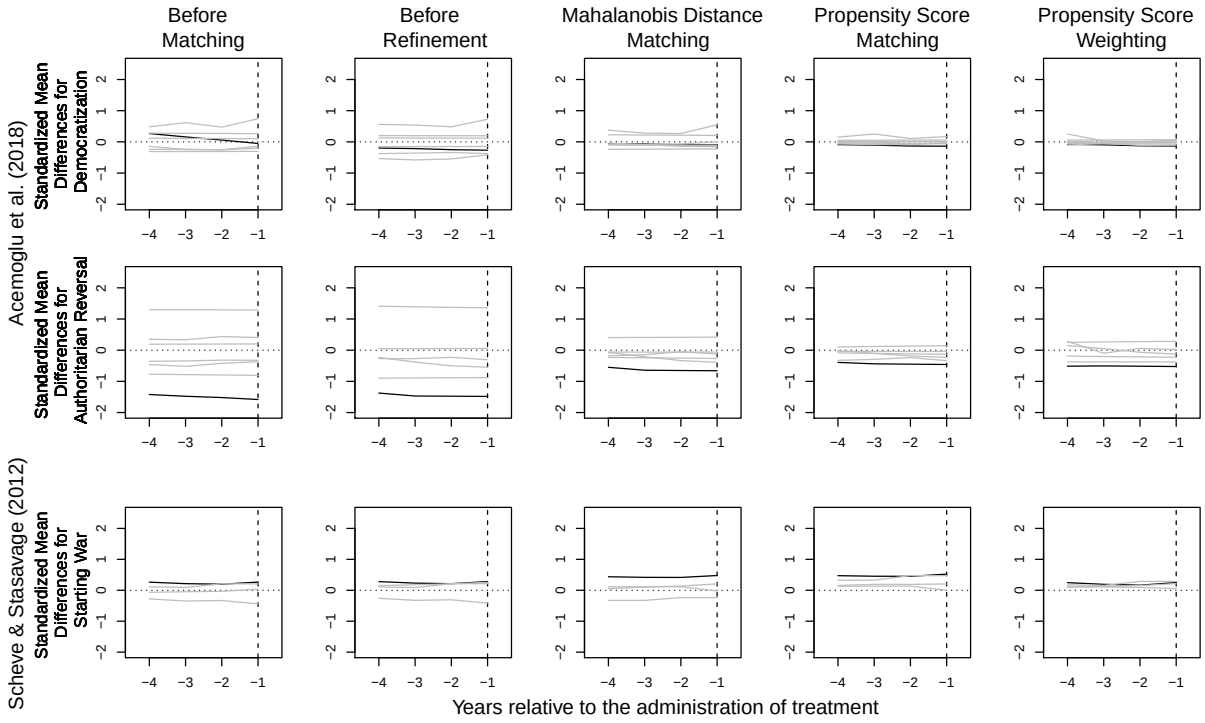


Figure A11: **Improved Covariate Balance due to Matching over the Pre-Treatment Time Period when Estimating the Average Effects of Stable Policy Change, $F = 1$.** See the caption of Figure 5.

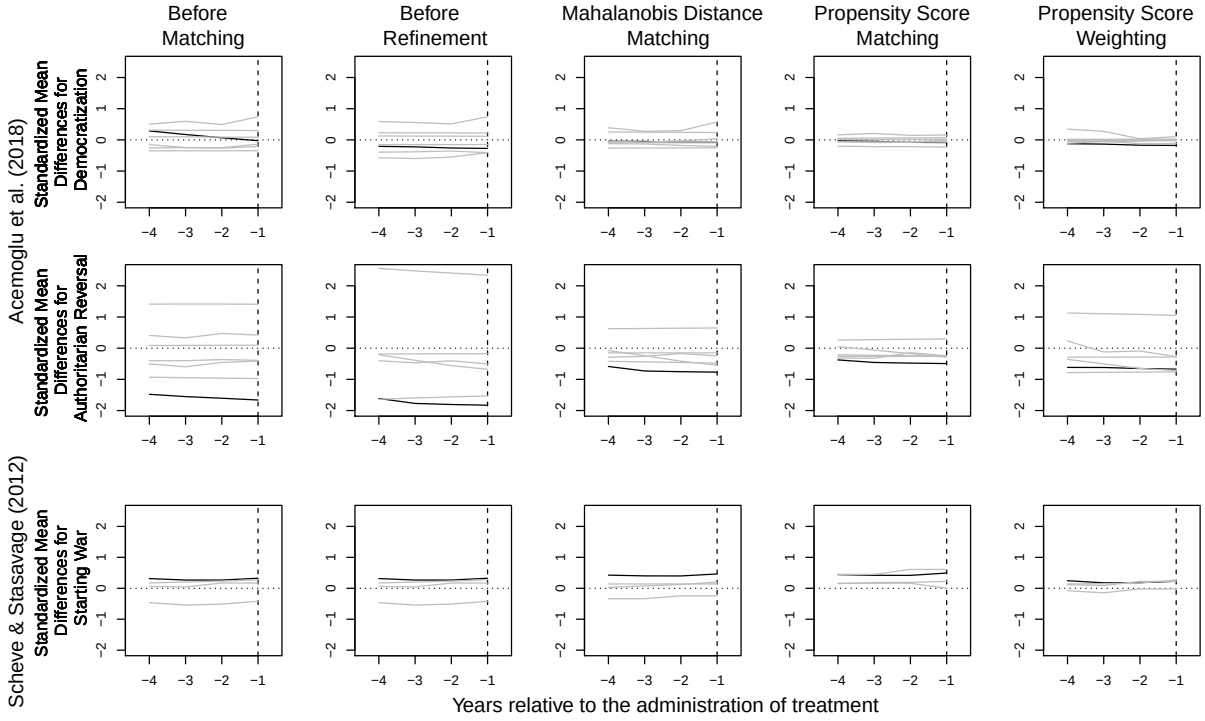


Figure A12: **Improved Covariate Balance due to Matching over the Pre-Treatment Time Period when Estimating the Average Effects of Stable Policy Change,** $F = 2$. See the caption of Figure 5.

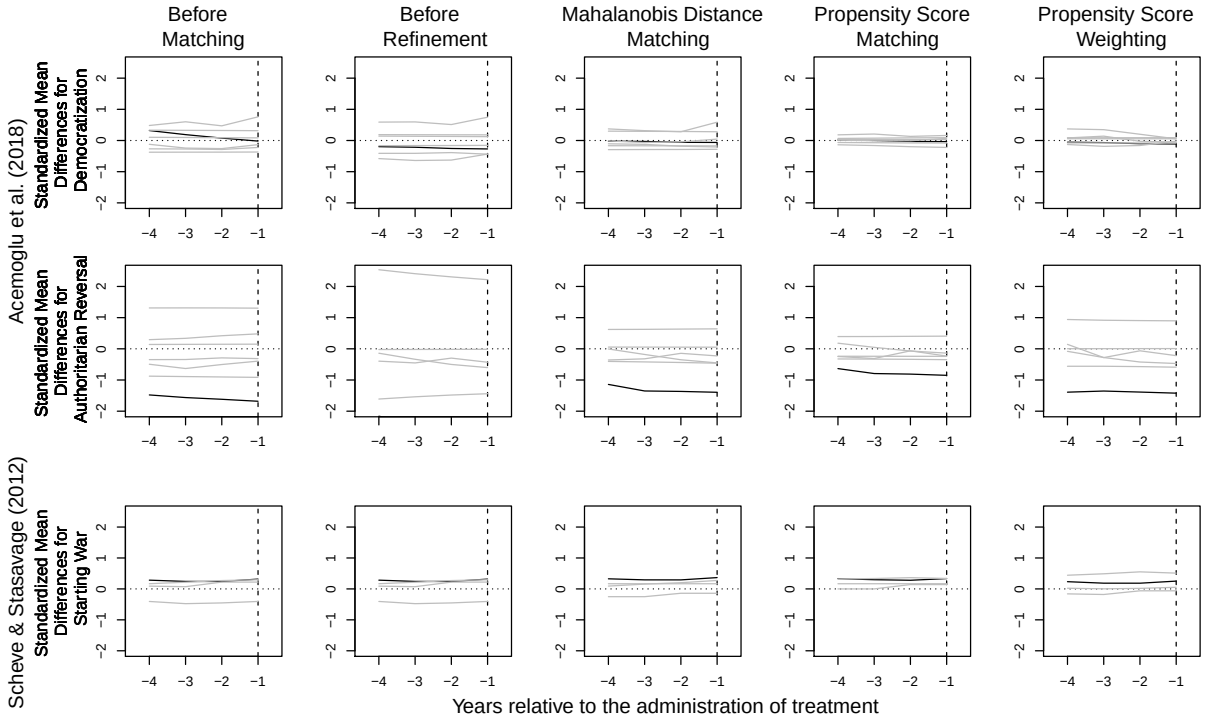


Figure A13: **Improved Covariate Balance due to Matching over the Pre-Treatment Time Period when Estimating the Average Effects of Stable Policy Change,** $F = 3$. See the caption of Figure 5.

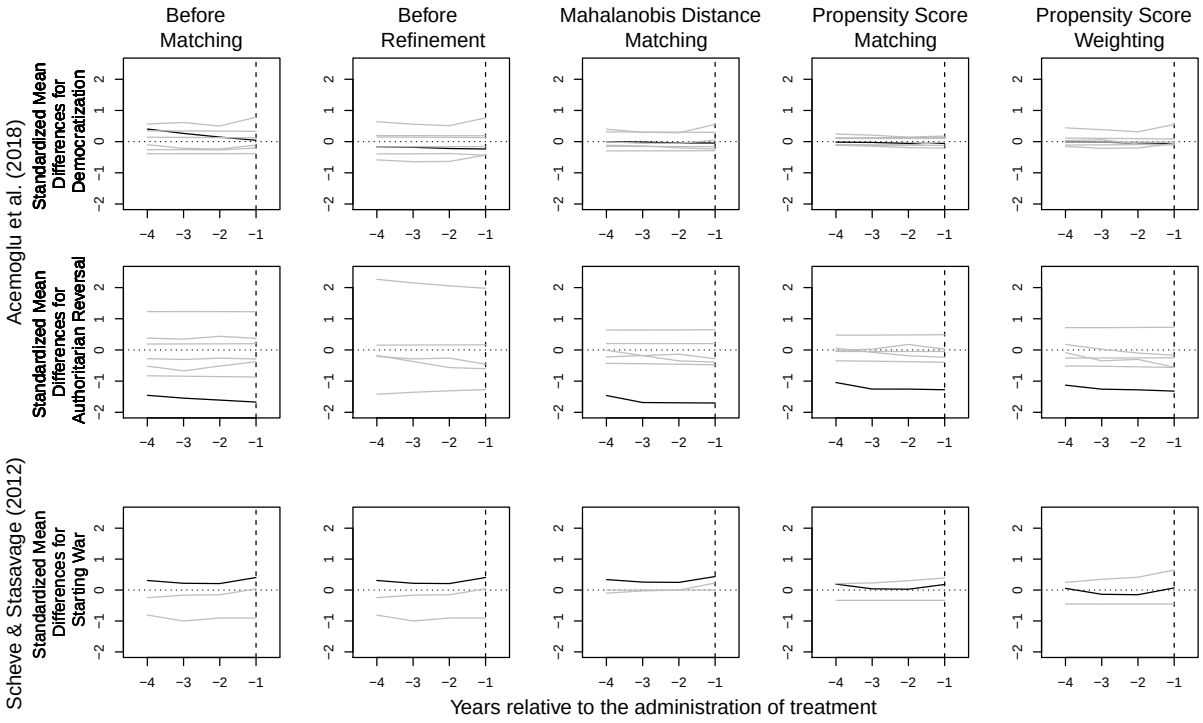


Figure A14: **Improved Covariate Balance due to Matching over the Pre-Treatment Time Period when Estimating the Average Effects of Stable Policy Change,** $F = 4$. See the caption of Figure 5.

D The Results based on One Year Lag

This section presents the estimated effects of democracy on development using one year lag instead of four year lags as shown in Figure 6). Our findings remain substantively unchanged. We find that democracy has a positive effect on economic development not because transitioning to democracy improves a country's economic prospects, but because backsliding into autocracy worsens a country's economic development.

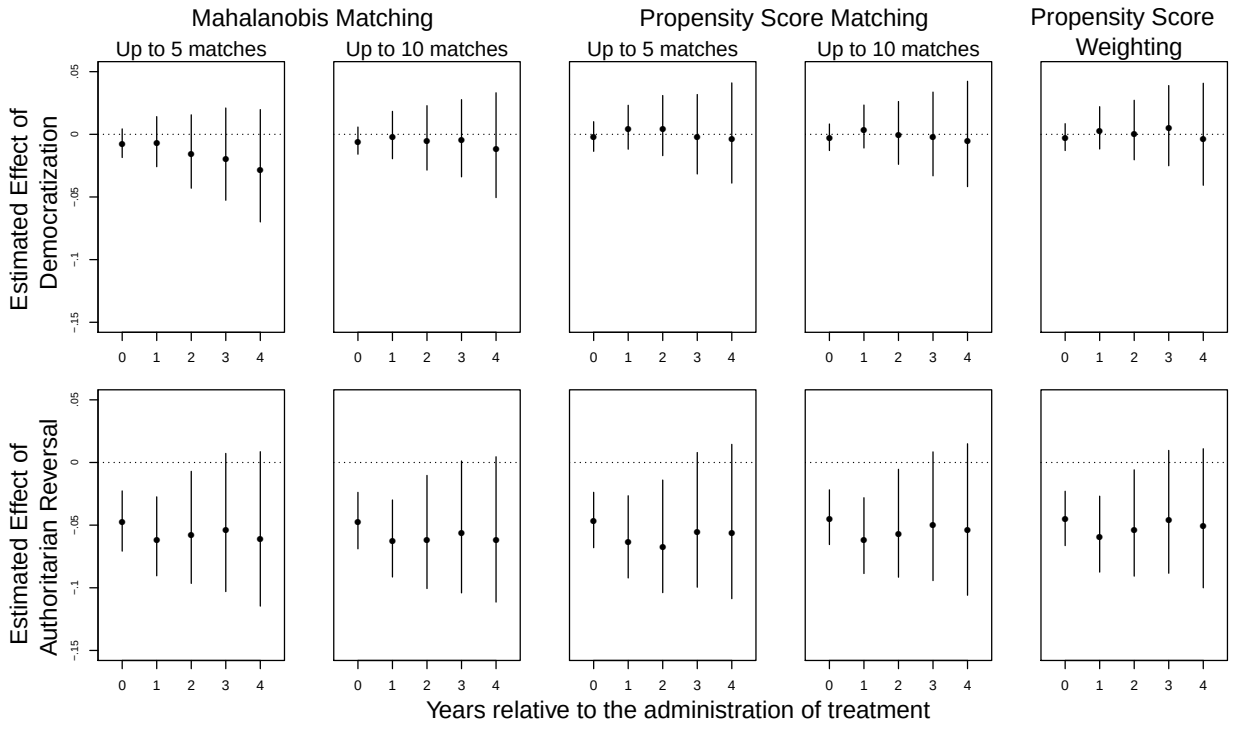


Figure A15: **Estimated Average Effects of Democracy on Logged GDP per Capita when the Treatment Reversal is Allowed and Adjusting for One Year Pre-treatment Period.** The matching method adjusts for the treatment and covariate histories during the one year period prior to the treatment, i.e., $L = 1$. See the caption of Figure 6.

E The Estimated Effects of Democratization and Authoritarian Reversal based on the Linear Regression Models

This Appendix presents the estimated effects of democratic transition and authoritarian reversal, using the linear regression approach of the original analysis (Acemoglu et al., 2019). We begin by replicating the results reported in Section A6.3 of the appendix of the original study. In that analysis, democratization and authoritarian reversal are coded as follows. For each country, both variables take the value of zero in the beginning period. Throughout the subsequent periods, whenever X_{it} changes from 0 to 1, the value of the democratization variable will increase by one and stays at that value until X_{it} changes from 0 to 1 again. Similarly, the authoritarian reversal variable starts with the value of zero and increases by one whenever X_{it} changes from 0 to 1.

In the original study, the authors then fit the two-way fixed effects linear regression models using the least squares and GMM estimation. They include the lagged outcome variables but no other

covariate. The estimated coefficients for the democratization and authoritarian reversal variables are then interpreted as their respective average causal effects. One issue with this approach is the assumption that the effects of democratization as well as those of authoritarian reversal are additive, regardless of the past history of regime changes. This contrasts with our approach where we nonparametrically adjust for the past treatment history.

The first and second columns of Table A3 reproduces the estimates reported in the original study whereas the third and fourth columns report the estimates based on the models that include additional covariates. The results suggest that the effects of democratization are similar, in their magnitude, to those of authoritarian reversal (their signs are opposite as expected). However, the former is more precisely estimated than the latter. These results are qualitatively different from those based on our approach. We find that the economic effects of democracy are largely driven by the negative effects of authoritarian reversal rather than the positive effects of democratization. In contrast, the original analysis shows that the positive effects of democratization plays a more significant role.

	(1)	(2)	(3)	(4)
ATT	0.8033^{***}	1.4697^{***}	0.6706^{**}	0.9548^{**}
	(0.2381)	(0.5425)	(0.3139)	(0.4723)
ATC	-0.7054^{**}	-1.3125	-0.6459	-0.8247
	(0.3398)	(0.9570)	(0.4487)	(0.5712)
$\hat{\rho}_1$	1.2381^{***}	1.2038^{***}	1.0981^{***}	1.0463^{***}
	(0.0381)	(0.0463)	(0.0416)	(0.0426)
$\hat{\rho}_2$	-0.2065^{***}	-0.1916^{***}	-0.1331^{***}	-0.1206^{***}
	(0.0464)	(0.0276)	(0.0405)	(0.0379)
$\hat{\rho}_3$	-0.0261	-0.0276	0.0053	0.0139
	(0.0286)	(0.0276)	(0.0296)	(0.0286)
$\hat{\rho}_4$	-0.0424^{**}	-0.0382	-0.0311	-0.0175
	(0.0176)	(0.0210)	(0.0239)	(0.0232)
country FE	Yes	Yes	Yes	Yes
time FE	Yes	Yes	Yes	Yes
covariates	No	No	Yes	Yes
estimation	OLS	GMM	OLS	GMM
N	6,336	6,336	4,416	4,416

Notes: ***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.
robust standard errors clustered by prefecture in parentheses

Table A3: **The Effects of Democracy on Growth with Lagged Treatments**: This table presents the estimated effects of transition to democracy from authoritarian regime (labeled as "ATT") and vice versa (labeled as "ATC"). The control variables in Columns (2) and (4) are those in column (3) of Table 1