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

Kosuke Imai† In Song Kim‡ Erik Wang§

April 28, 2018

Abstract

Matching methods aim to improve the validity of causal inference in observational studies by reducing model dependence and offering intuitive diagnostics. While they have become a part of standard tool kit for empirical researchers across disciplines, matching methods are rarely used when analyzing time-series cross-section (TSCS) data, which consist of a relatively large number of repeated measurements on the same units. We develop a methodological framework that enables the application of matching methods to TSCS data. 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 observation has outcome and covariate histories similar to those of its matched control observations. Assessing the quality of matches is done by examining covariate balance. After the refinement, we estimate both short-term and long-term average treatment effects using the difference-in-differences estimator, accounting for a time trend. We also show that the proposed matching estimator can be written as a weighted linear regression estimator with unit and time fixed effects, providing model-based standard errors. We illustrate the proposed methodology by estimating the causal effects of democracy on economic growth, as well as the impact of inter-state war on inheritance tax. The open-source software is available for implementing the proposed matching methods.

Key Words: covariate balance, difference-in-differences, dynamic treatment, fixed effects, unobserved confounding, weighting

*The methods described in this paper can be implemented via the open-source statistical software, PanelMatch: Matching Methods for Causal Inference with Time-Series Cross-Section Data, available at `https://github.com/insongkim/PanelMatch`. We thank Matt Blackwell, Paul Kellstedt, Anton Strezhnev, and Yiqing Xu for comments and feedback. This paper was previously presented at the Midwest Political Science Association Annual Meeting.

†Professor, Department of Politics and Center for Statistics and Machine Learning, Princeton University, Princeton NJ 08544. Phone: 609–258–6601, Email: kimai@princeton.edu, URL: https://imai.princeton.edu

‡Assistant 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/

§PhD Candidate, Department of Politics, Princeton University, Princeton NJ 08544. Phone: 574-520-9117, Email: haixiaow@princeton.edu, URL: http://erikhw.github.io

1 Introduction

One common 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 (e.g., Rubin, 2006; Stuart, 2010). By making the treatment variable independent of observed confounders, these methods reduce model dependence and further 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). 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, 2016). Moreover, almost all of newly developed 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).

In Section 3, we develop matching methods for TSCS data and estimate the average treatment effect of policy change for the treated (ATT). In the proposed methodological framework, 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 a 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 by adjusting for a possible unobserved time trend. The proposed method can be used to estimate both short-term and long-term ATT and allows for simple diagnostics through the examination of covariate balance. Finally, we establish the equivalence between the proposed matching estimator and the weighted linear regression estimator with unit and time fixed effects. This result enables us to compute model-

based standard errors. The proposed matching methods can be implemented via the open-source statistical software, PanelMatch: Matching Methods for Causal Inference with Time-Series Cross-Section Data, available at `https://github.com/insongkim/PanelMatch`.

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 a 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. Major limitations of this approach are the requirement that only one unit receives the treatment and uncertainty estimates are not available. Xu (2017) overcomes these limitations by generalizing the synthetic control method within 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. In addition, although the possibility of some units receiving the treatment at multiple time periods is noted (see footnote 7), the author assumes that the treatment status never reverses.

Another relevant methodological literature is based on 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). In contrast, our proposed method permits flexible matching and weighting procedures for 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 obser-

vations should be compared with the control observations. This motivates our proposed matching method, which is applied to these empirical studies in Section 4. Finally, Section 5 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. (2017), 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, 2000; Papaioannou and Siourounis, 2008; Gerring, Thacker, and Alfaro, 2012). Acemoglu et al. (2017) 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 a "Free" or "Partially Free" in Freedom House and a positive score in the Polity IV index, which ranges from -10 to 10.[1] 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

[1] 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.

of those who are below16 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.[2]

The authors assume the following standard sequential exogeneity,

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

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. (2017) 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. (2017), 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

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

	Democracy and Growth (Acemoglu et al., 2017)				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	1.745	5.970	1.636
	(0.226)	(0.374)	(0.307)	(0.461)	(2.392)	(0.729)	(2.081)	(0.757)
$\hat{\rho}_1$	1.238	1.204	1.100	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				
	(0.043)	(0.045)	(0.041)	(0.038)				
$\hat{\rho}_3$	−0.026	−0.028	0.005	0.014				
	(0.028)	(0.028)	(0.030)	(0.029)				
$\hat{\rho}_4$	−0.043	−0.036	0.003	−0.018				
	(0.017)	(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	6,161	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. (2017) 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. (2017) 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.

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,

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

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. 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,

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

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

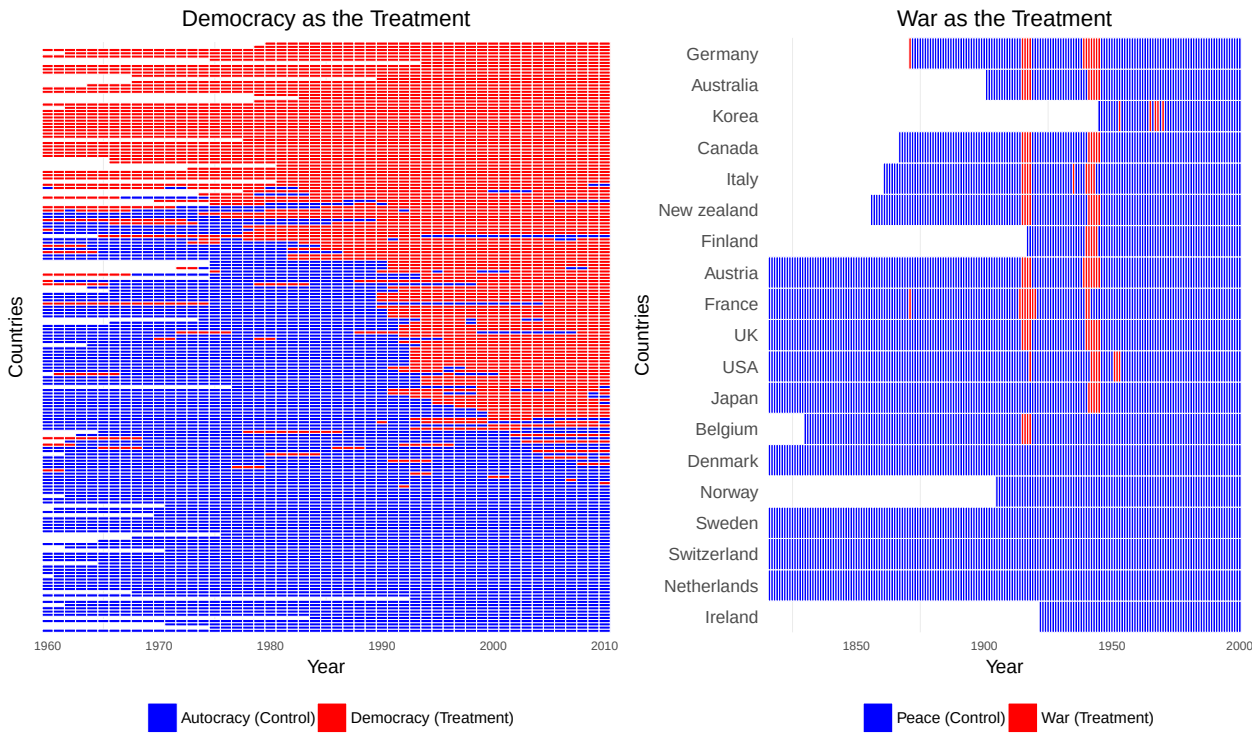

Figure 1: **The Treatment Variation Plots for Visualizing the Distribution of Treatment across Space and Time**. The left panel displays the spatial-temporal distribution of treatment for the study of democracy's effect on economic development (Acemoglu et al., 2017), in which a red (blue) rectangle 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).

(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, 2016). These models offer few diagnostic tools for causal inference. In contrast, matching 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 enables 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. (2017) 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 who 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, 2016). 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.[3]

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 prespecified 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 then estimate its counterfactual outcome under the control condition by using standard matching or weighting techniques, which

[3]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).

further adjust for the differences in confounding variables. Finally, we apply the difference-in-differences estimator in order to account for any underlying time trend. After describing the proposed methodology, we prove that the proposed matching estimator is equivalent to a weighted linear two-way fixed effects regression estimator. Finally, we show how to conduct covariate balance diagnostics and compute standard errors for the estimated ATTs.

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 can be generalized 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 Defining the 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. In addition, researchers must select another non-negative integer L as the number of lags to adjust for. Once these two parameters 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,

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

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.

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 poilcy change.

3.1.2 Assumptions

Under our framework, the choice of F and L implies the assumption that 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., $\{X_{i,t-\ell}\}_{\ell=L+1}^{t-1}$. That is, we assume the absence of spillover effect but allow for some carryover effects (up to L time periods). In many applications, the assumption of no spillover effect may be too restrictive. Although the methodological literature has begun to address this challenge of relaxing 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, 2018), enabling the presence of spillover effects in TSCS data settings is beyond the scope of this paper.

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,

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

where the conditioning set includes the treatment history, the lagged outcomes (except the immediate lag $Y_{i,t-1}$), and the covariate history.

Researchers may choose a relatively large value of L in order to increase the credibility of limited carryover effect and the parallel trend assumptions. In contrast, the choice of F should be substantively motivated.

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 $t-1$ (Imai and Kim, 2016). 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 $X_{i,t-1}=0$. It is possible that the matched set is empty for some treated observations, i.e., $|\mathcal{M}_{it}|=0$. That is, a subset of treated observations may not have any control observations that share the identical treatment history. In such cases, we exclude these treated observations from the subsequent analysis to preserve the internal validity. It is important for researchers to characterize these treated observations that are removed from the analysis so that they understand how the target population of the ATT has changed. Finally, note that this matched set 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 equations (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 $\mathbf{V}_{it'} = (Y_{it'}, \mathbf{Z}_{i,t'+1}^\top)^\top$ and $\boldsymbol{\Sigma}_{it'}$ 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 lagged outcome variable and covariates and average it across time periods.[4]

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,

$$e_{it}(\{\mathbf{V}_{it'}\}_{t'=t-L}^{t-1}) \;=\; \Pr(X_{it}=1 \mid \mathbf{V}_{i,t-1},\ldots,\mathbf{V}_{i,t-L}) \;=\; \frac{1}{1+\exp(-\sum_{\ell=1}^{L}\boldsymbol{\beta}_\ell^\top \mathbf{V}_{i,t-\ell})}. \tag{13}$$

In practice, researchers may assume a more parsimonious model, in which some elements of $\boldsymbol{\beta}$ are 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 outcome $Y_{i,t-1}$. 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

[4]For notational simplicity, this formulation does not adjust for $\mathbf{Z}_{i,t-L}$. This can be done by adding another Mahalanobis distance term with $\mathbf{V}_{i,t-L-1} = \mathbf{Z}_{i,t-L}$. In addition, this formulation adjust for $Y_{i,t-1}$ even though strictly speaking the assumption in given equation (10) does not require it. Again, this can be achieved by removing $Y_{i,t-1}$ from $\mathbf{V}_{i,t-1}$.

by matching on the estimated propensity score, yielding the following distance measure,

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

for each matched control observation $i' \in \mathcal{M}_{it}$ where $\hat{e}_{i't}(\{\mathbf{V}_{i,t-\ell}\}_{\ell=1}^{L})$ 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 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,

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

such that $\sum_{i' \in \mathcal{M}_{it}} w_{it}^{i'} = 1$. Note that the weighting refinement further generalizes the matching refinement since the latter assigns an equal weight to each unit in the refined matched set $\mathcal{M}_{it}^{*}$,

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

Other weighting methods, including the synthetic control method (Abadie, Diamond, and Hainmueller, 2010), 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,

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

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 short or long term treatment effect (i.e., the number of leads F is greater than zero), 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 before the outcome is measured at time $t + F$. Similarly, some treated units may return to the control conditions before 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, 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,

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

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. To estimate this ATT, we use the same estimator as above except that we 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 $t+F$.

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 $\mathbf{V}_{it'} = (Y_{it'}, \mathbf{Z}_{i,t'+1}^{\top})^{\top}$) 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,

$$B_{it}(j,\ell) \;=\; \frac{V_{i,t-\ell,j} - \sum_{i' \in \mathcal{M}_{it}} w_{it}^{i'}\, V_{i',t-\ell,j}}{\sqrt{\frac{1}{N_1 - 1} \sum_{i'=1}^{N} \sum_{t'=L+1}^{T-F} D_{it'} (V_{i',t'-\ell,j} - \overline{V}_{t'-\ell,j})^2}} \tag{20}$$

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.

$$\overline{B}(j,\ell) \;=\; \frac{1}{N_1} \sum_{i=1}^{N} \sum_{t=L+1}^{T-F} D_{it} B_{it}(j,\ell) \tag{21}$$

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. 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.

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,*

$$\hat{\beta}_{\mathsf{DiD}} \;=\; \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}\{(Y_{it} - \overline{Y}_i^* - \overline{Y}_t^* + \overline{Y}^*) - \beta(X_{it} - \overline{X}_i^* - \overline{X}_t^* + \overline{X}^*)\}^2. \tag{22}$$

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,*

$$W_{it} \;=\; \sum_{i'=1}^{N}\sum_{t'=1}^{T} D_{i't'} \cdot v_{it}^{i't'} \quad \text{and} \quad v_{it}^{i't'} \;=\; \begin{cases} 1 & \text{if } (i,t) = (i', t'+F) \\ 1 & \text{if } (i,t) = (i', t'-1) \\ w_{i't'}^{i} & \text{if } i \in \mathcal{M}_{i't'}, t = t'+F \\ -w_{i't'}^{i} & \text{if } i \in \mathcal{M}_{i't'}, t = t'-1 \\ 0 & \text{otherwise.} \end{cases} \tag{23}$$

Proof is in Appendix A. We note that the regression weight W_{it} can take a negative value in some cases. This means that technically the weighted linear two-way fixed effects regression estimator should be considered as the method of moments estimator, where the moment condition is given by the population version of the first order condition of the optimization problem in Theorem 1,

$$\mathbb{E}\{W_{it}\widetilde{X}_{it}(\widetilde{Y}_{it} - \beta\widetilde{X}_{it})\} \;=\; 0 \tag{24}$$

where $\widetilde{Y}_{it} = Y_{it} - \overline{Y}_i^* - \overline{Y}_t^* + \overline{Y}^*$ and $\widetilde{X}_{it} = X_{it} - \overline{X}_i^* - \overline{X}_t^* + \overline{X}^*$.

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, Otsu and Rai (2017) demonstrates that a valid bootstrap inference can be made for matching estimators by treating the implied observation-specific weight, which represents the number of times the observation is used for matching, as a covariate for each observation. For the proposed estimator, this observation-specific weight can be computed as follows,

$$W_{it}^{*} \;=\; \sum_{i'=1}^{N}\sum_{t'=1}^{T} D_{i't'}\cdot v_{it}^{i't'} \quad \text{and} \quad v_{it}^{i't'} \;=\; \begin{cases} 1 & \text{if } (i,t)=(i',t'+F) \\ -1 & \text{if } (i,t)=(i',t'-1) \\ -w_{i't'}^{i} & \text{if } i\in\mathcal{M}_{i't'}, t=t'+F \\ w_{i't'}^{i} & \text{if } i\in\mathcal{M}_{i't'}, t=t'-1 \\ 0 & \text{otherwise.} \end{cases} \tag{25}$$

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.

Alternatively, we can compute the standard error by exploiting the equivalence result given in Theorem 1. That is, we can compute the following cluster-robust variance based on the method of moment estimation,

$$\left(\sum_{i=1}^{N} \widetilde{\mathbf{X}}_i^{\top}\mathbf{W}_i(\mathbf{Y}_i-\hat{\beta}_{\mathsf{DiD}}\widetilde{\mathbf{X}}_i)(\mathbf{Y}_i-\hat{\beta}_{\mathsf{DiD}}\widetilde{\mathbf{X}}_i)^{\top}\mathbf{W}_i\widetilde{\mathbf{X}}_i\right) \Big/ \left(\sum_{i=1}^{N} \widetilde{\mathbf{X}}_i^{\top}\mathbf{W}_i\widetilde{\mathbf{X}}_i\right)^{2} \tag{26}$$

where $\widetilde{\mathbf{X}}_i = (\widetilde{X}_{i1}, \widetilde{X}_{i2}, \ldots, \widetilde{X}_{iT})$, $\widetilde{\mathbf{Y}}_i = (\widetilde{Y}_{i1}, \widetilde{Y}_{i2}, \ldots, \widetilde{Y}_{iT})$, and $\mathbf{W}_i = \mathrm{diag}(W_{i1}, W_{i2}, \ldots, W_{iT})$ with W_{it} defined in Theorem 1. Then, the standard asymptotic properties of the method of moments estimator apply directly to this case (Hansen, 1982). Although this approach is computationally more efficient than bootstrap, it is only applicable to the estimator given in equation (19) when the causal quantity of interest involves a counterfactual scenario about stable policy where no policy change is assumed to occur at least for F time periods after the treatment is administered to the units in the treatment group.

4 Empirical Analyses

We revisit the two motivating studies described in Section 2, one about the effect of democracy on development (Acemoglu et al., 2017), and the other concerning the impact of war on inheri-

tance 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.

4.1 Application of Matching Methods

We demonstrate the use of the proposed methodology. For the Acemoglu et al. (2017) 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,

$$\mathbb{E}\left[Y_{i,t+F}\left(X_{it}=0, X_{i,t-1}=1, \{X_{i,t-\ell}\}_{\ell=2}^{L}\right) - Y_{i,t+F}\left(X_{it}=1, X_{i,t-1}=1, \{X_{i,t-\ell}\}_{\ell=2}^{L}\right) \mid X_{it}=0, X_{i,t-1}=1\right] \tag{27}$$

In addition, we estimate the ATT for stable policy change relative to no policy change, as defined in equation (19), so that we can capture the effect of short or medium term democratic (authoritarian) transition.

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 estimate the ATT of stable policy change, in which countries are involved in war for the pre-specified number of years (equation (19)), as well as the ATT of poilcy change, in which countries may end war during the period of interest (equation (8)).

We use the original studies to guide the specification of matching methods. In their regression models, Acemoglu et al. (2017) 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

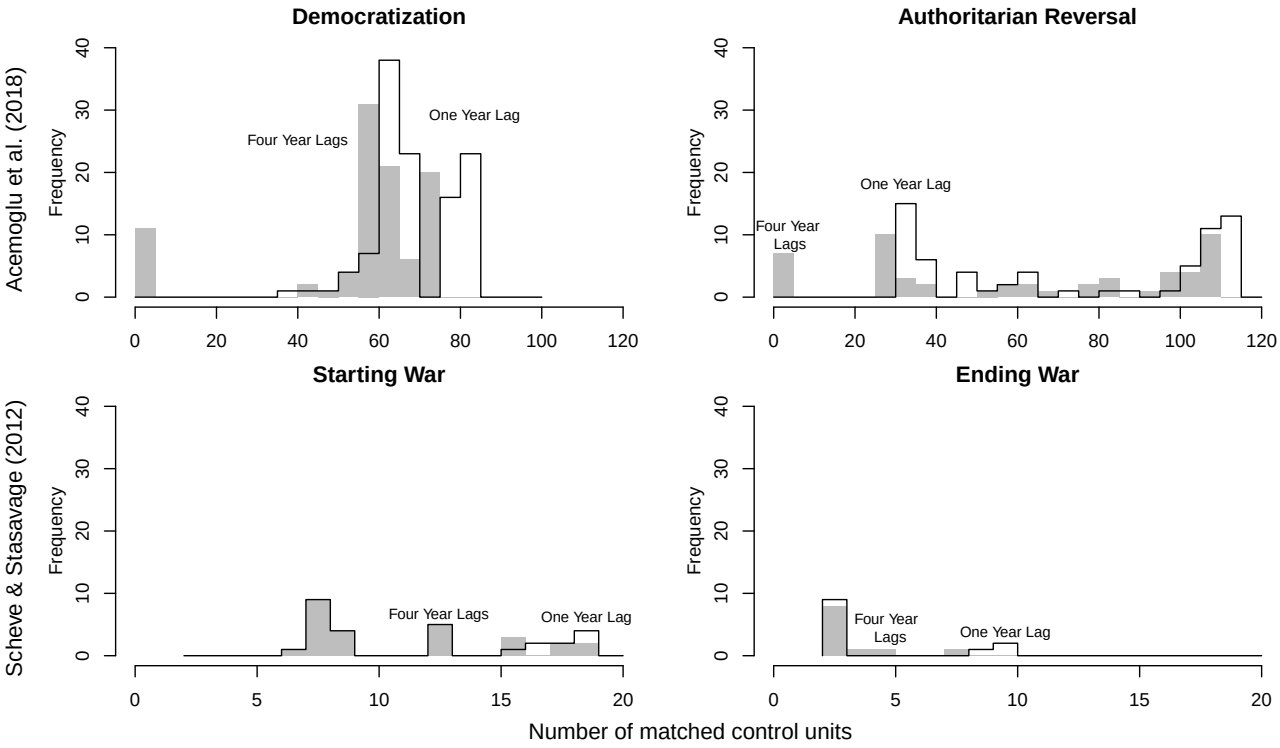

Figure 2: **Frequency Distribution of the Number of Matched Control Units**. The transparent (gray) bar represents the number of matched control units that share the same treatment history as a treated observation for one year (four years) prior to the treatment year. The frequency distribution is presented for each of the two two treatments in the Acemoglu et al. (2017) study (top panel) and the Scheve and Stasavage (2012) study (bottom panel).

(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 2 presents the frequency distribution for the number of matched control units given a treated observation in the case of one and four year lag as grey and transparent 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. (2017) study in the upper panel, there are 18 (13) treated observations for democratization (authoritarian reversal) that have no control unit with the same treatment history when the number of lags is four [5], whereas no such treated observation exists for the case

[5]Such observations for democratization are: Armenia in 1991, Azerbaijan in 1992, Bangladesh in 2009, Belarus in

of one year lag. As noted earlier, for the Acemoglu et al. (2017) study, we have enough matched control units for both democratization and authoritarian reversal: most treated observations have more than 30 matched control units. However, for the Scheve and Stasavage (2012) study in the bottom panel, 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. The number of matched control units is greater for the estimation of effects of starting war, but even in this case, the matched sets are relatively small.

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. (2017) study to examine the sensitivity of empirical findings to the number of matches. For the Scheve and Stasavage (2012) study, we use one-to-one match and up-to-three matching because the matched sets are small. 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 $\mathbf{V}_{it'} = (Y_{i,t'-1}, \ldots, Y_{i,t'-L}, \mathbf{Z}_{i,t'-1}^\top, \ldots, \mathbf{Z}_{i,t'-L}^\top)^\top$. For the Acemoglu et al. (2017) study, the time-varying covariates $\mathbf{Z}_{it}$ includes 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 3 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 (21) before (horizontal axis) and after (vertical axis) the refinement of matched sets. A dot below the 45 degree line implies that the standardized mean balance is improved after the refinement for a particular covariate in $\mathbf{V}_{it'}$. The plots suggest that across almost all variables

1991, Czech Republic in 1993, Guinea-Bissau in 1999, Haiti in 1994, Lesotho in 1999, Lithuania in 1993, Macedonia, FYR in 1991, Niger in 1999, Peru in 1963, Suriname in 1991, Slovenia in 1992, Thailand in 1992, Turkey in 1961 and 1973. Such observations for authoritarian reversal are: Azerbaijan in 1993, Burkina Faso in 1980, Bangladesh in 1974, Republic of Congo in 1963, Ecuador in 1961, Ghana in 1972, Grenada in 1979, Cambodia in 1995, Korea in 1961, Mauritania in 2008, Peru in 1962, Thailand in 1976, and Uganda in 1985.

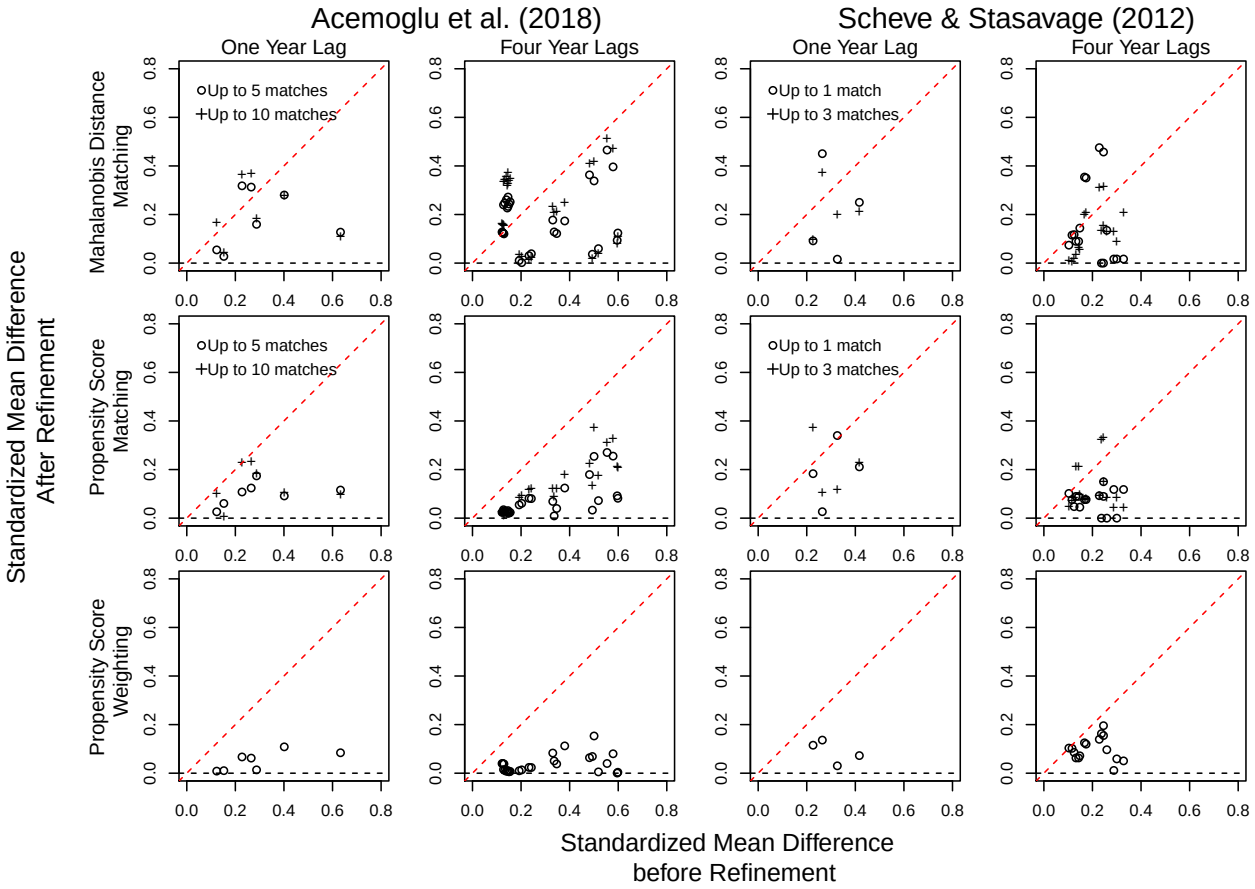

Figure 3: **Improved Covariate Balance due to the Refinement of Matched Sets**. Each scatter plot compares the absolute value of standardized mean difference for each covariate j and lag year ℓ defined in equation (21) before (horizontal axis) and after (vertical axis) the refinement of matched sets. Rows represets the results based on different matching and weighting methods while the columns represent the results using the adjustments for different lag lengths.

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 4 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 lagged outcomes and 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. (2017) 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 history alone) do not dramatically improve the covariate balance. In contrast, the improvement

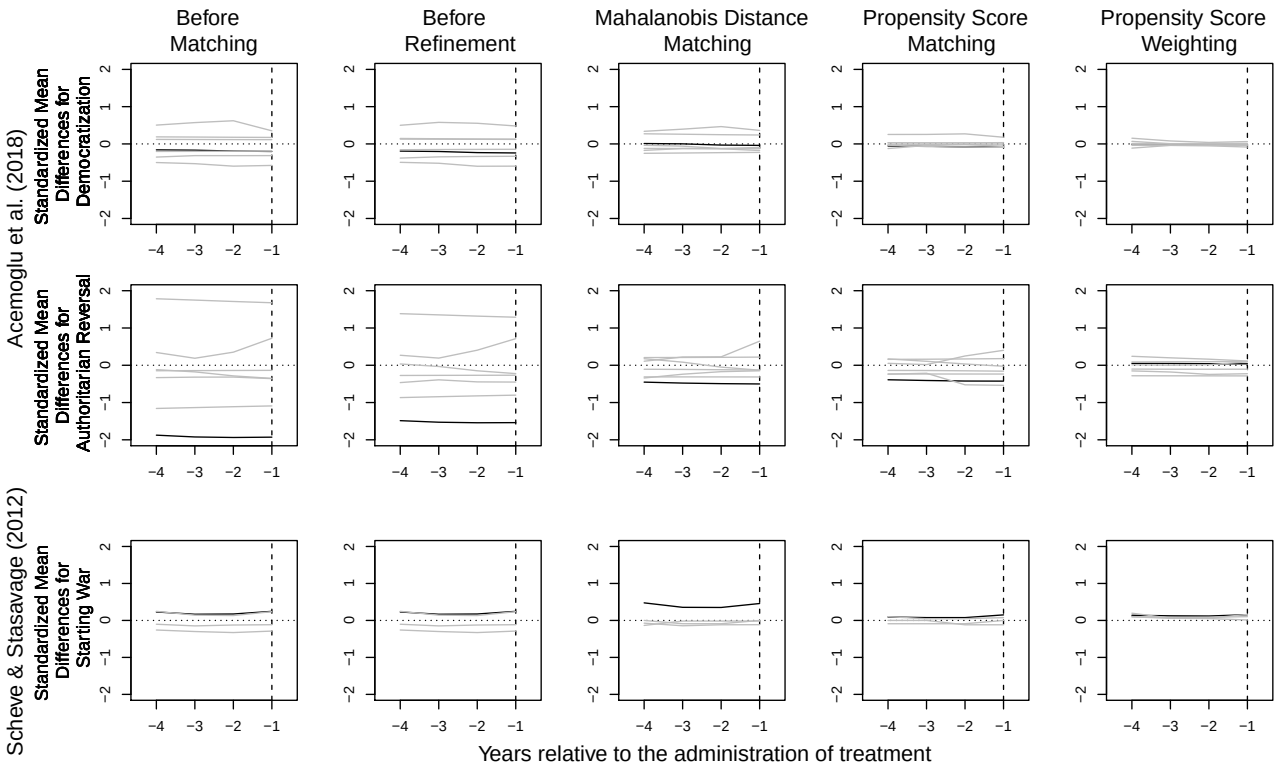

Figure 4: **Improved Covariate Balance due to Matching over the Pre-Treatment Time Period**. Each plot plots the standardized mean difference defined in equation (21) (vertical axis) over the pre-treatment time period of four years (horizontal axis). The left column shows the balance before matching, 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.

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.

4.2 Empirical Findings

We now present the estimated ATTs based on the matching methods. Figure 5 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., $F = 0, 1, \ldots, 4$. Solid circles represent the point estimates of policy change, in which some countries may revert back to the control condition (equation (8)). In contrast, open triangles represent the estimates

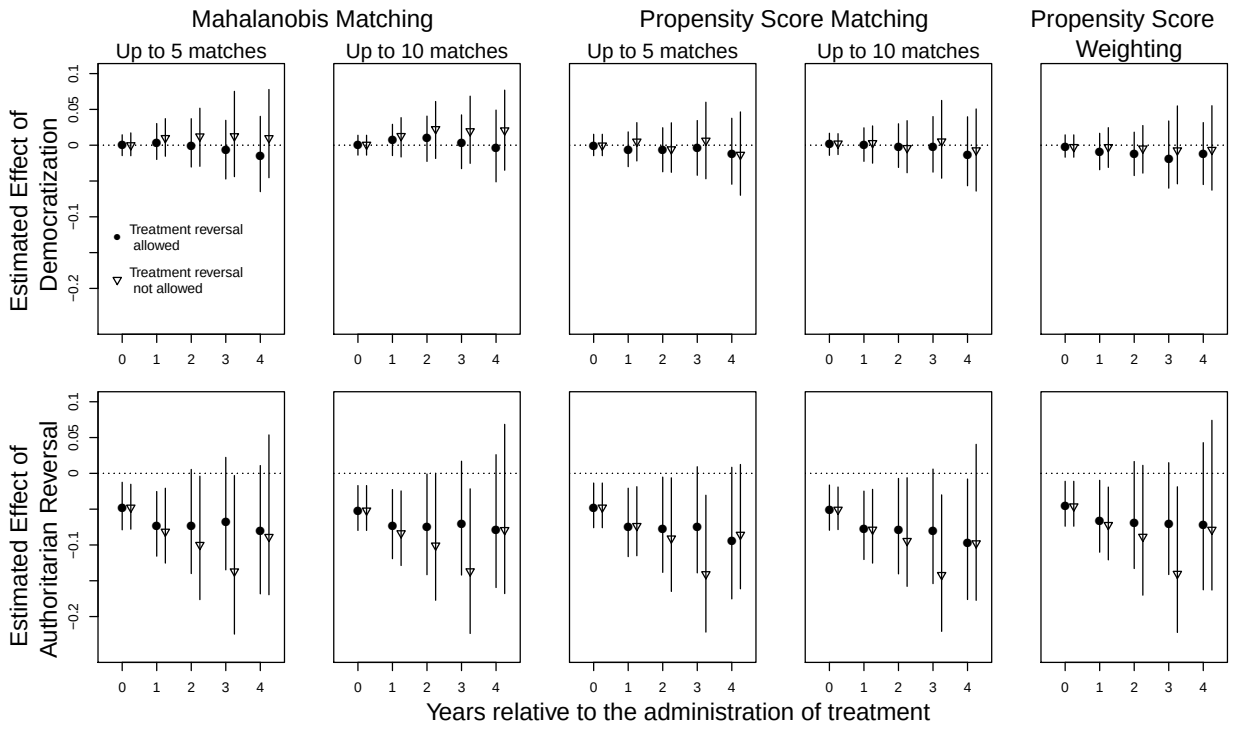

Figure 5: **Estimated Average Effects of Democracy on Logged GDP per Capita**. The estimates are based on the matching method that adjusts for the treatment, outcome and covariate histories during the four year period prior to the treatment, i.e., $L = 4$. The estimates for the average effects of democratization (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. Solid circles (open triangles) represent the estimates when the treatment reversal is (not) allowed.

for stable policy change where no such treatment reversal is allowed (equation (19)). The vertical bars represent 95% confidence intervals based on 500 block bootstrap replicates.

Across all five methods (columns), the substantive results are similar regardless of whether we allow for the treatment reversal (solid circles (open triangles) represent the cases where the treatment reversal is (not) allowed). We find that the point estimates of the effects for democratization are 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 two to three years immediately after the transition when the treatment reversal is allowed. For the scenario where we do not allow treatment reversal, we observe a pattern that is similar in statistical significance and slightly stronger in effect magnitudes. The estimated effects across the two treatment scenarios are substantively large, indicating an approximately 5

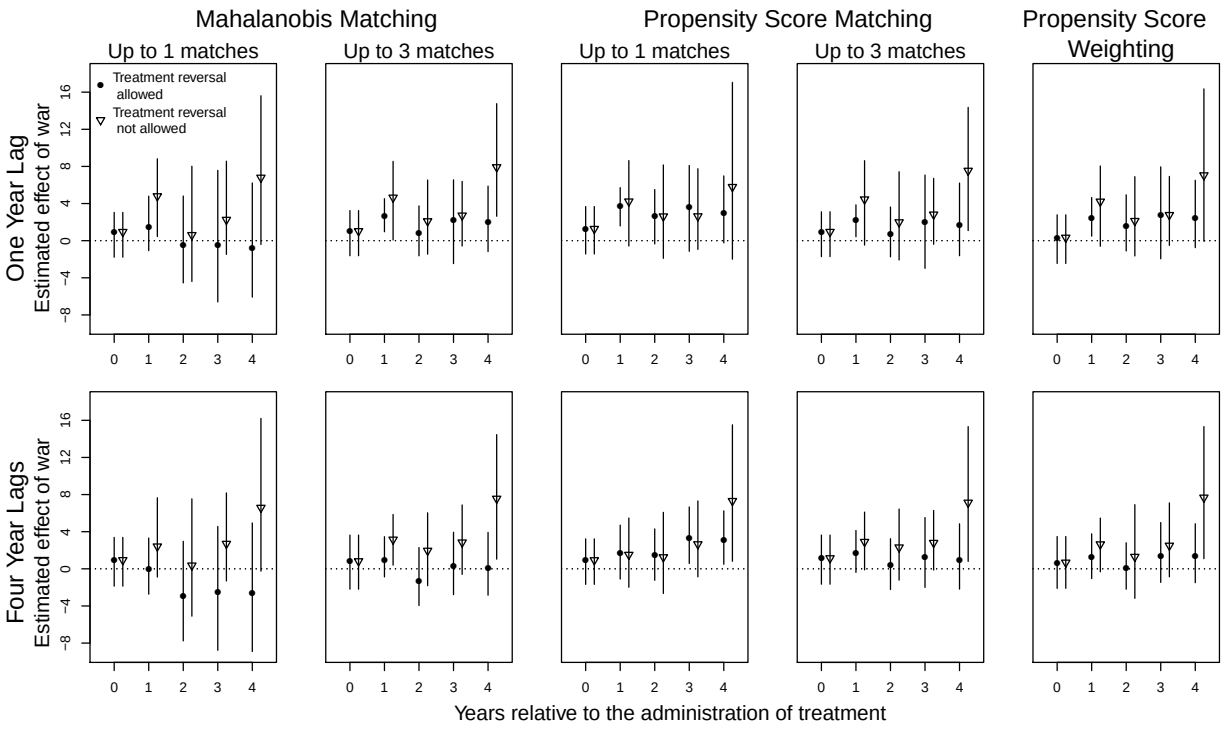

Figure 6: **Estimated Average Effects of Interstate War on Inheritance Tax Rate**. The matching method adjusts for the treatment, outcome and covariate histories during the one (upper panel) or four (lower panel) year period prior to the treatment. The estimated effects are shown for the period of five 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. Solid circles (open triangles) represent the estimates when the treatment reversal is (not) allowed.

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 9 of Appendix B, 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 least in the short and medium term.

Next, Figure 6 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 outcome, treatment and covariate histories of one year

period prior to the treatment as done in the original study. 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, with the point estimates for two different causal quantities of interest represented with different symbols.

For the ATT of policy change where treatment reversal is allowed (solid circles), we find that if we refine the matched set using the one year pre-treatment period, most of the estimated effects are positive and statistically significant at one year after the beginning of war although the effects lose statistical significance for up-to-one Mahalanobis matching and propensity score weighting. This finding is consistent with the results of Scheve and Stasavage (2012), and the effect size of the estimates based on matching methods is somewhere between those of dynamic and static regression models (see Table 1). However, almost all of the estimated causal effects are no longer statistically significant if we refine the matched sets by adjusting for the four year pre-treatment period, suggesting that the findings regarding the short term effects are sensitive to covariate adjustment.

Interestingly, we find somewhat different results for the ATT of stable policy change, in which treatment reversal is not allowed. Our analysis suggests that the continued involvement in war may lead to a large increase in the top marginal rate of inheritance tax. Indeed, we find that a longer involvement in war yields a greater effect on inheritance tax when treatment reversal is not allowed (that is, a country is at war for several consecutive years). This supports the theoretical prediction of Scheve and Stasavage (2012) that the participation of wider population in war puts an electoral pressure on politicians to increase the taxation of the rich.

5 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.

One important limitation of the proposed methodology is the assumption of no interference across units. That is, 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). However, in many social science applications, the assumption of no spillover effects is unrealistic because people affect each other within and across societies. One possible way to address this limitation is to match 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 Californias tobacco control program." *Journal of the American statistical Association* 105 (490): 493–505.

Abadie, Alberto, and Guido W. Imbens. 2008. "On the Failure of the Bootstrap for Matching Estimators." *Econometrica* 76 (6): 1537-1557.

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

Acemoglu, Daron, Suresh Naidu, Pascual Restrepo, and James A. Robinson. 2017. "Democracy Does Cause Growth." NBER Working Paper No. 20004.

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 (2): 277–297.

Aronow, Peter, and Cyrus Samii. 2017. "Estimating Average Causal Effects Under General Interference." *Annals of Applied Statistics* 11 (4): 1912–1947.

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 (September): 634–647.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. 2004. "How Much Should We Trust Differences-in-Differences Estimates?" *Quarterly Journal of Economics* 119 (1): 249–275.

Blackwell, Matthew, and Adam Glynn. 2018. How to Make Causal Inferences with Time-Series Cross-Sectional Data under Selection on Observables. Technical report Harvard University.

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 (3): 932–945.

Gerring, John, Strom C Thacker, and Rodrigo Alfaro. 2012. "Democracy and human development."

The Journal of Politics 74 (1): 1–17.

Hansen, Ben B. 2004. "Full Matching in an Observational Study of Coaching for the SAT." *Journal of the American Statistical Association* 99 (467): 609–618.

Hansen, Lars Peter. 1982. "Large Sample Properties of Generalized Method of Moments Estimators." *Econometrica* 50 (July): 1029–1054.

Hirano, Keisuke, Guido Imbens, and Geert Ridder. 2003. "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score." *Econometrica* 71 (July): 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 (Summer): 199–236.

Hudgens, Michael G., and Elizabeth Halloran. 2008. "Toward Causal Inference with Interference." *Journal of the American Statistical Association* 103 (June): 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 (493): 345–361.

Imai, Kosuke, and In Song Kim. 2016. "When Should We Use Linear Fixed Effects Regression Models for Causal Inference with Longitudinal Data?" Working paper available at `https://imai.princeton.edu/research/FEmatch.html`.

Imai, Kosuke, and Marc Ratkovic. 2014. "Covariate Balancing Propensity Score." *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* 76 (January): 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 (September): 1013–1023.

Imai, Kosuke, Zhichao Jiang, and Anup Malai. 2018. Causal Inference with Interference and Noncompliance in Two-Stage Randomized Experiments. Technical report Princeton University.

Li, Yunfei Paul, Kathleen J. Propert, and Paul R. Rosenbaum. 2001. "Balanced Risk Set Matching." *Journal of the American Statistical Association* 96 (455): 870–882.

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

Otsu, Taisuke, and Yoshiyasu Rai. 2017. "Bootstrap Inference of Matching Estimators for Average Treatment Effects." *Journal of the American Statistical Association* 112 (520): 1720–1732.

Papaioannou, Elias, and Gregorios Siourounis. 2008. "Democratisation and growth." *The Economic Journal* 118 (532): 1520–1551.

Przeworski, Adam. 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 (8): 2379–2412.

Robins, James M., Miguel Ángel Hernán, and Babette Brumback. 2000. "Marginal Structural Models and Causal Inference in Epidemiology." *Epidemiology* 11 (September): 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 (March): 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 (1): 81–102.

Stuart, Elizabeth A. 2010. "Matching Methods for Causal Inference: A Review and a Look Forward." *Statistical Science* 25 (1): 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 (1): 55–75.

Xu, Yiqing. 2017. "Generalized Synthetic Control Method: Causal Inference with Interactive Fixed Effects Models." *Political Analysis* 25 (1): 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 (December): 1360–1371.

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 (23).

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,

$$\begin{aligned}
&\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it} A_{it} \alpha_i^* \\
&= \sum_{i'=1}^{N}\sum_{t'=1}^{T} D_{i't'} \left(\sum_{i=1}^{N}\sum_{t=1}^{T} v_{it}^{i't'} A_{it} \alpha_i^* \right) \\
&= \sum_{i'=1}^{N}\sum_{t'=1}^{T} D_{i't'} \left(1 - 1 - \sum_{i \in \mathcal{M}_{i't'}} \sum_{t=t'+F} A_{it}^2 \cdot w_{i't'}^{i} - \sum_{i \in \mathcal{M}_{i't'}} \sum_{t=t'-1} A_{it} \cdot w_{i't'}^{i} \right) \alpha_i^* \\
&= \sum_{i'=1}^{N}\sum_{t'=1}^{T} D_{i't'} \left(1 - 1 - \sum_{i \in \mathcal{M}_{i't'}} \sum_{t=t'+F} w_{i't'}^{i} + \sum_{i \in \mathcal{M}_{i't'}} \sum_{t=t'-1} w_{i't'}^{i} \right) \alpha_i^* \\
&= \sum_{i'=1}^{N}\sum_{t'=1}^{T} D_{i't'} \left(1 - 1 - 1 + 1 \right) \alpha_i^* \;=\; 0
\end{aligned} \tag{28}$$

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

$$A_{it} - \overline{A}_i^* - \overline{A}_t^* + \overline{A}^* = A_{it} \tag{29}$$

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,

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

$$
\begin{aligned}
&= \sum_{i'=1}^{N}\sum_{t'=1}^{T} D_{i't'}\left(1+1+\sum_{i\in\mathcal{M}_{i't'}}\sum_{t=t'+F} w_{i't'}^{i} - \sum_{i\in\mathcal{M}_{i't'}}\sum_{t=t'-1} w_{i't'}^{i}\right) \\
&= 2\sum_{i'=1}^{N}\sum_{t'=1}^{T} D_{i't'} = 2\sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}. \qquad (30)
\end{aligned}
$$

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 (29), the third equality follows from equation (30), and the fourth equality is implied by equation (28). 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., $A_{it}=2X_{it}-1$. □

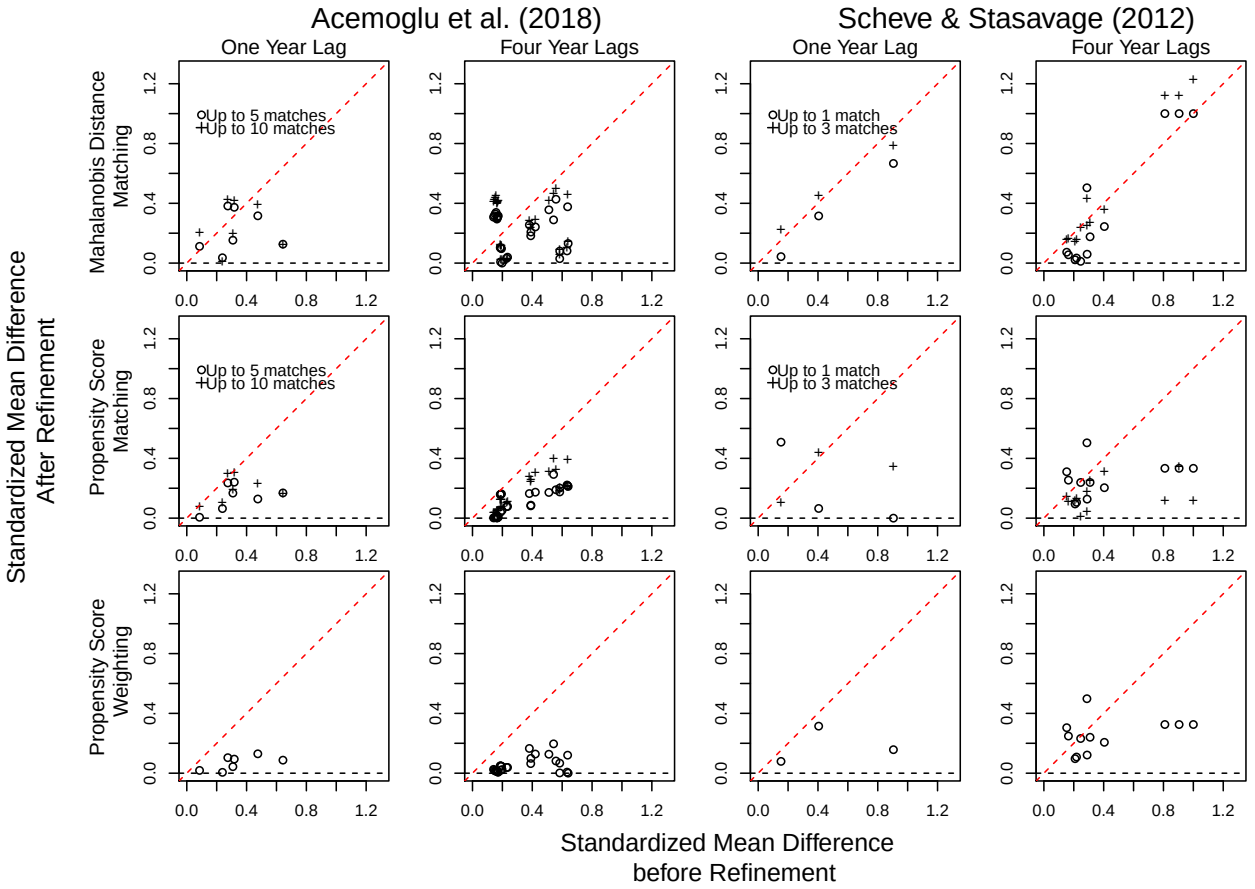

Figure 7: **Improved Covariate Balance due to the Refinement of Matched Sets when Estimating the Average Effects of Stable Policy Change**. Each scatter plot compares the absolute value of standardized mean difference for each covariate j and lag year ℓ defined in equation (21) before (horizontal axis) and after (vertical axis) the refinement of matched sets. Rows represets the results based on different matching and weighting methods while the columns represent the results using the adjustments for different lag lengths.

B Additional Empirical Results

In this appendix, we present additional empirical results. First, Figure 7 shows how the refinement of matched sets improves the covariate balance for the two studies when estimating the ATT of stable policy change (defined in equation (19)) rather than the ATT of policy change, in which treatment reversal is allowed (equation (8)). The analogous results for the latter are given in Figure 3. The two sets of the results are similar. The plots suggest that across almost all variables the refinement results in the improved mean covariate balance for the study in Acemoglu et al. (2017). 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. However, the improvement is much smaller for the study in Scheve and Stasavage (2012), as shown in the right panel of this figure. In addition, we drop the universal male suffrage variable because the standard deviation for this variable across all treated units is 0 for every pre-treatment period. The mean differences in the universal male suffrage variable are large (0.86 for eight out of the nine treated observations, and 1 for the other treated observation), but they are constant over all pre-treatment periods, suggesting that the DiD estimator may be able to eliminate the bias due to the imbalance of this variable.

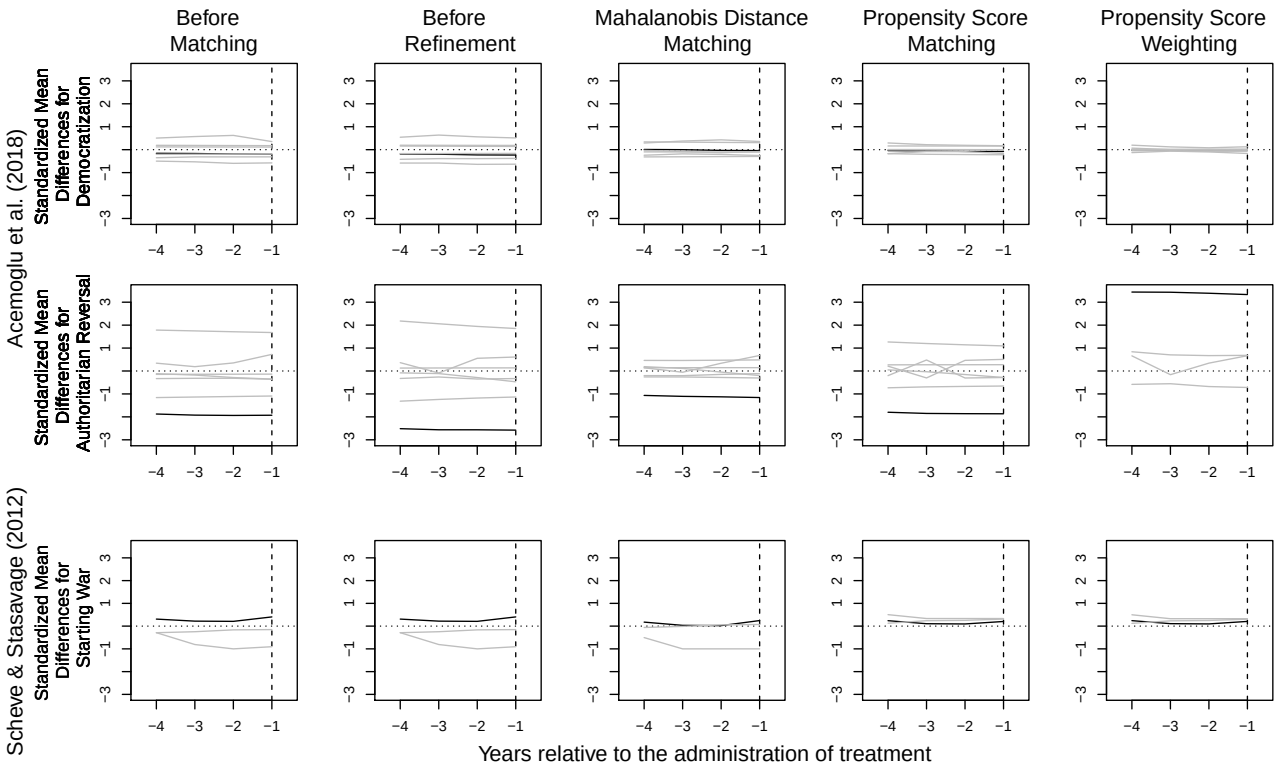

Figure 8: **Improved Covariate Balance due to Matching over the Pre-Treatment Time Period when Estimating the Average Effects of Stable Policy Change**. Each plot plots the standardized mean difference defined in equation (21) (vertical axis) over the pre-treatment time period of four years (horizontal 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.

Figure 8 further illustrates the improvement of covariate balance due to matching over the pre-treatment time period when estimating the ATT of stable policy change, in which the treatment reversal is not permitted. The analogous results for estimating the ATT of policy change with possible treatment reversal appear in Figure 4. The matching and weighting methods have a difficult time balancing covariates when estimating the ATT of authoritarian reversal. This is because for this treatment variable there exists only a small number of matched control units under the scenario of no treatment reversal. Indeed, there exist 9 treated observations, all of which have less than 10 control units in their matched sets. Nevertheless, the imbalance appears to be relatively constant, suggesting that the difference-in-differences estimator may be able to eliminate the bias. For the other two treatments, the results are similar to those in Figure 4. Propensity score matching and weighting are effective in reducing the covariate imbalance.

Finally, Figure 9 presents the estimated ATTs for the Acemoglu et al. (2017) study when adjusting for the treatment, outcome, and covariates of one year pre-treatment period. The analogous results when adjusting for four year pre-treatment period appear in Figure 5. The results are substantively quite similar, suggesting that the effect of democracy is largely due to the negative effect of authoritarian reversal.

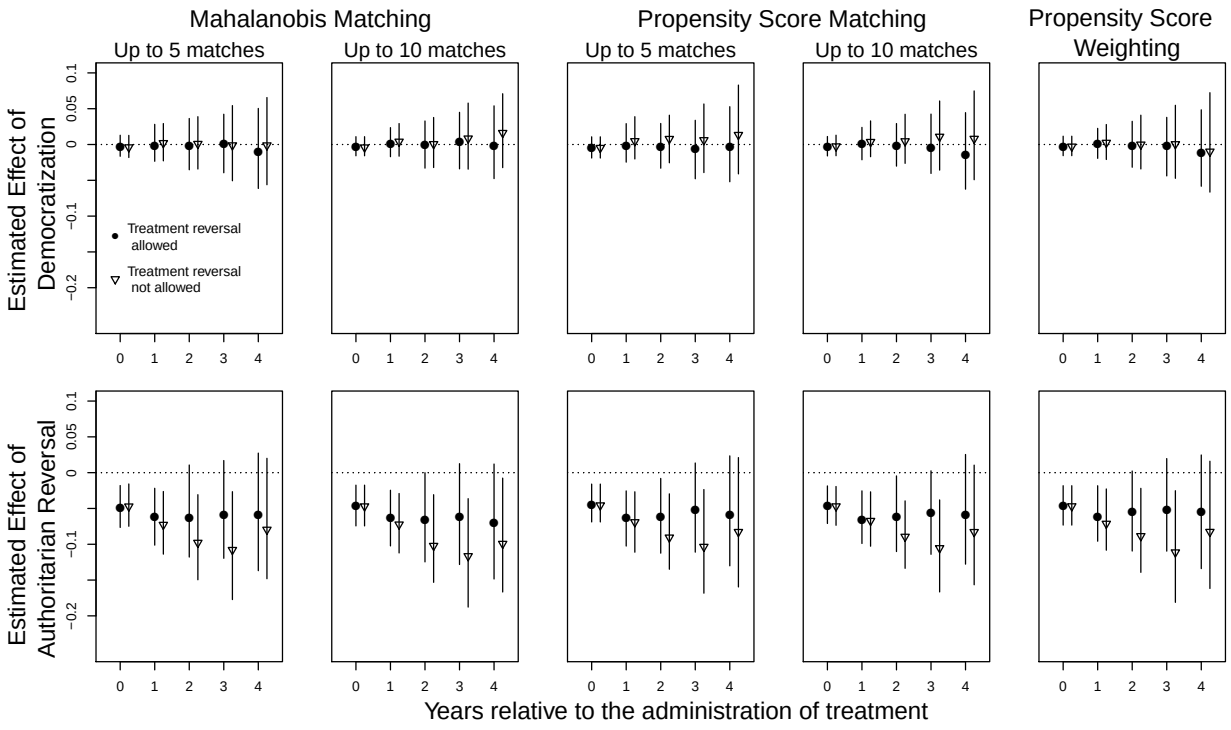

Figure 9: **Estimated Average Effects of Democracy on Logged GDP per Capita when Adjusting for One Year Pre-treatment Period.** The matching method adjusts for the treatment, outcome and covariate histories during the one year period prior to the treatment, i.e., $L = 1$. See the caption of Figure 5.