AJPS AMERICAN JOURNAL of POLITICAL SCIENCE

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

Kosuke Imai Harvard University
In Song Kim Massachusetts Institute of Technology
Erik H. Wang Australian National University

Abstract: *Matching methods improve the validity of causal inference by reducing model dependence and offering intuitive diagnostics. Although they have become a part of the standard tool kit across disciplines, matching methods are rarely used when analysing 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 prespecified 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.*

Verification Materials: The data and materials required to verify the computational reproducibility of the results, procedures and analyses in this article are available on the *American Journal of Political Science* Dataverse within the Harvard Dataverse Network, at: https://doi.org/10.7910/DVN/ZTDHVE.

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).

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, 2021). Moreover, almost all of the existing matching methods

Kosuke Imai, Professor, Department of Government and Department of Statistics, Harvard University. 1737 Cambridge Street, Institute for Quantitative Social Science, Cambridge, MA 02138 (Imai@Harvard.Edu). In Song Kim, Associate Professor, Department of Political Science, Massachusetts Institute of Technology. 77 Massachusetts Avenue E53-407, Cambridge, MA 02142 (insong@mit.edu). Erik H. Wang, Lecturer, Department of Political and Social Change at Australian National University (ANU). Hedley Bull Building, 130 Garran Road, Acton ACT 2600 Australia (erik.wang@anu.edu.au).

The methods described in this article 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. Imai thanks the Sloan Foundation for partial support (Economics Program; 2020–13946). Wang acknowledges funding from the French National Research Agency (ANR) under the Investments for the Future program (Investissements d'Avenir, grant ANR-17-EURE-0010).

American Journal of Political Science, Vol. 0, No. 0, xxxx 2021, Pp. 1–19

DOI: 10.1111/ajps.12685

assume a cross-sectional data set (e.g. Abadie and Imbens 2011; Diamond and Sekhon 2013; Hansen 2004; Iacus et al. 2011; Rosenbaum et al. 2007; Zubizarreta 2012).[1]

We fill this methodological gap by developing matching methods for TSCS data. In the proposed approach, 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 prespecified 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 covariate histories. After this refinement step, we apply a difference-in-differences (DiD) estimator that 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.[2] 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.

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 asymptotic 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 et al. (2010) propose the synthetic control method, which constructs a weighted average of pretreatment outcomes among control units such that it approximates the observed pretreatment 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 Ben-Michael et al. 2019a; Doudchenko and Imbens 2017). In addition, the synthetic control method and its extensions require a long pretreatment 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 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 et al. 2019b). In contrast, our methods allow multiple units to be treated 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 et al. 2000). These models focus on estimating the causal effect of treatment sequence while avoiding posttreatment bias (as future treatments may be caused by past treatments) (see Blackwell and Glynn 2018, for an introduction). These approaches, however, require the modelling of potentially complex conditional expectation functions and propensity score for each time period, which can be challenging for TSCS data that 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.

Motivating Applications

This section introduces two influential studies that motivate our methodology. The first study is Acemoglu et al.

[1] An exception is an unpublished paper by Nielsen and Sheffield (2009). Their matching method is substantially different from our methodology.

[2] 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).

(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 briefly describing the original 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.

Democracy and Economic Growth

Scholars have long debated whether democracy promotes economic development. Acemoglu et al. (2019) conduct an up-to-date and comprehensive empirical study to investigate this question. The authors analyse 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,

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

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 equals 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. 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 below 16 years old, the log population above 64 years old, net financial flow as a fraction of GDP, trade volume as a fraction of GDP and a binary measure of social unrest.[3] 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,

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

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. 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),

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

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 the original study, 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%, a substantial effect given that democratization may have a long-term effect on economic growth.

War and Taxation

As a central element of redistributive policies, inheritance taxation plays an essential role in wealth accumulation and income inequality. 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 interstate war propels countries to increase inheritance taxation.

Scheve and Stasavage analyse 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 interstate 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

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

TABLE 1 Regression Results from the Two Motivating Empirical Applications

	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
	(0.230)	(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		0.909		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	6,161	4,416	4,245	2,780	2,779	2,537	2,536

Note: The estimated coefficients for the treatment variable and lagged outcome variables are presented with standard errors in parentheses. For the Acemoglu et al. 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 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. study, we use the heteroskedasticity-robust standard errors. For the Scheve and Stasavage study, we cluster standard errors by countries for the static models whereas the panel corrected standard errors are used for the dynamic models.

5-year or decade intervals, we analyse the annual data 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,

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

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,

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

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 mentioned above, Scheve and Stasavage also fit the following model with the lagged outcome variable and

country-specific time trends but without country fixed effects,

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

where the strict exogeneity assumption is now given by,

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

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

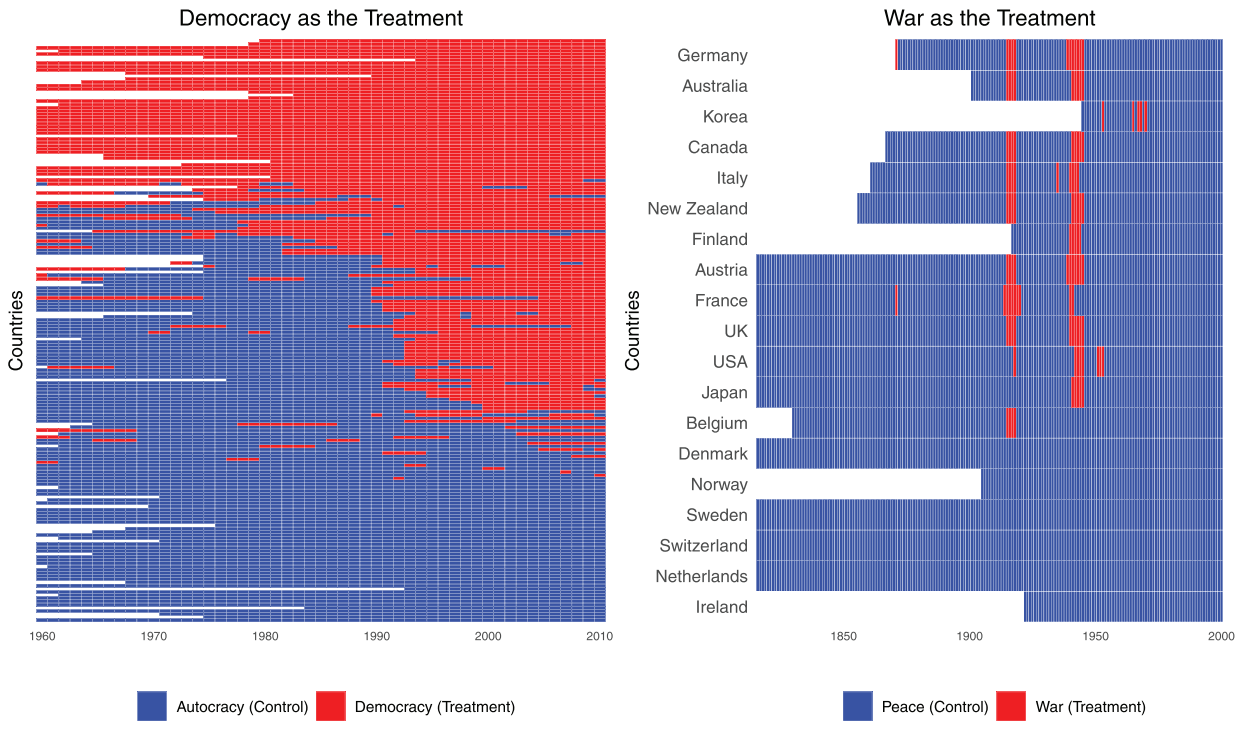
The last four columns of Table 1 present the results. Columns (5) and (7) report the results obtained using the static model given in Equation (4) without and with the time-varying covariates, respectively. Similarly, columns (6) and (8) are based on the dynamic model specified in Equation (6) without 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.

The Treatment Variation Plot

A variety of linear regression models with fixed effects used in these studies represent the most common methodological approaches to causal inference with TSCS data. However, a major drawback of these models is that it is difficult to understand how they use observed data to estimate relevant counterfactual quantities (Imai and Kim 2019, 2021).

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. 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

FIGURE 1 Treatment Variation Plots for Visualizing the Distribution of Treatment across Space and Time



Note: 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).

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 posttreatment bias. Alternatively, researchers can conduct comparison within the same year, which would correspond to 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 study, in which a treated (control) observation represents the time of interstate war (peace) indicated by a red (blue) rectangle. 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.[4] 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. Because 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.

[4]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).

The Proposed Methodology

In this section, we propose a general matching method for causal inference with TSCS data, which 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 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 DiD estimator in order to account for an underlying time trend. At the end of this section, we establish the connections to the linear fixed effects regression estimator and discuss covariate balance diagnostics and standard errors.

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.

Causal Quantity of Interest. The first step of the proposed methodology is to define a causal quantity by choosing a nonnegative 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 whereas $F = 2$ 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, our methodology requires researchers to select another nonnegative integer L as the number of *lags* to adjust for. Unlike the choice of leads, which should be primarily driven by researchers' substantive interests, selecting the number of lags is part of the

identification assumption. That is, researchers should evaluate the extent to which past treatment status could be a confounder affecting the current outcome as well as the current treatment (Imai and Kim 2019). As in the regression approach, the choice of L is important and faces a bias–variance tradeoff. Although 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, for example, $X_{i't'}$ with $i' \neq i$ and for any t', nor on the previous treatment status of the same unit after L time periods, that is, $\{X_{i,t-\ell}\}_{\ell=L+1}^{t-1}$. 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. Aronow and Samii 2017; Hudgens and Halloran 2008; Imai et al. 2021; Tchetgen Tchetgen and VanderWeele 2010). 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),

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

where the treated observations are those who experience the policy change, that is, $X_{i,t-1}=0$ and $X_{it}=1$. In our two applications, this quantity represents the average causal effect of democratization on economic growth and that of war initiation on inheritance taxation, respectively.

In this definition, $Y_{i,t+F}(X_{it}=1, X_{i,t-1}=0, \{X_{i,t-\ell}\}_{\ell=2}^{L})$ is the potential outcome under a policy change, whereas $Y_{i,t+F}(X_{it}=0, X_{i,t-1}=0, \{X_{i,t-\ell}\}_{\ell=2}^{L})$ represents the potential outcome without the policy change, that is, $X_{i,t-1}=X_{it}=0$. In both cases, the rest of the treatment history, that is, $\{X_{i,t-\ell}\}_{\ell=2}^{L} = \{X_{i,t-2}, \ldots, X_{i,t-L}\}$, 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.[5]

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, that is, $X_{i,t+\ell}=0$ 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.

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 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.

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 et al. 2000),

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

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

[5]One may be interested in the average treatment effect of *policy reversal* among the reversed (ART). This quantity corresponds to the effects of authoritarian reversal and is defined as,

$$\begin{aligned}\xi(F, L) = \mathbb{E}\big\{ & Y_{i,t+F}\big(X_{it}=0, X_{i,t-1}=1, \{X_{i,t-\ell}\}_{\ell=2}^{L}\big) \\ & - Y_{i,t+F}\big(X_{it}=1, X_{i,t-1}=1, \{X_{i,t-\ell}\}_{\ell=2}^{L}\big) \\ & \mid X_{it}=0, X_{i,t-1}=1\big\}. \end{aligned}$$

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 DiD design (e.g. Abadie 2005). Specifically, we make the following parallel trend assumption after conditioning on the treatment, outcome and covariate histories,

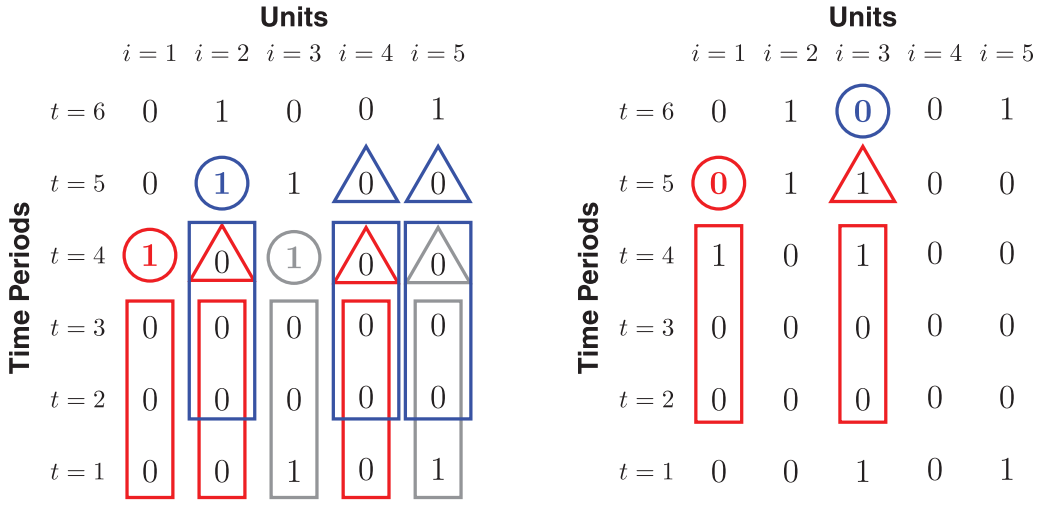
$$\begin{aligned}
&\mathbb{E}\big[Y_{i,t+F}\big(X_{it}=0, X_{i,t-1}=0, \{X_{i,t-\ell}\}_{\ell=2}^{L}\big) - Y_{i,t-1} \mid X_{it}=1,\\
&\quad X_{i,t-1}=0, \{X_{i,t-\ell}, Y_{i,t-\ell}\}_{\ell=2}^{L}, \{\mathbf{Z}_{i,t-\ell}\}_{\ell=0}^{L}\big]\\
&= \mathbb{E}\big[Y_{i,t+F}\big(X_{it}=0, X_{i,t-1}=0,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

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 pretreatment periods.

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 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 ART (see footnote 5) where the treated observation (red

FIGURE 2 An Example of Matched Sets with Five Units and Six Time Periods



(a) Matched Sets for ATT (b) Matched Sets for ART

Note: Panels (a) and (b) illustrate how matched sets are chosen for the ATT (as defined in Equation (11)) and the ART (see footnote 5), respectively, when $L = 3$. For each treated observation (coloured circles), we select a set of control observations from other units in the same time period (triangles with the same colour) that have an identical treatment history (rectangles with the same colour).

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,

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

for the treated observations with $X_{it} = 1$ and $X_{i,t-1} = 0$. For the ART, 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 et al. (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.

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 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, that is, $|\mathcal{M}_{it}| \leq J$. 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,

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

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 $\boldsymbol{\Sigma}_{it'}$ is the sample covariance matrix of $\mathbf{V}_{it'}$. That is, given a matched control unit, we compute the standardized distance using the time-varying covariates and average it across time periods.[6]

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 pretreatment 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 logistic regression model,

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

where $\mathbf{U}_{it'} = (X_{it'}, \mathbf{V}_{it'}^\top)^\top$.[7] In practice, researchers may assume a more parsimonious model, in which some elements of β are set to zero. For example, setting $\beta = 0$ 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 covariates by matching on the estimated propensity score, yielding the following distance measure,

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

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 to the treated unit in terms of the observed

[6] 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 DiD design.

[7] Note that because we only use the observations contained in the matched sets, this is equivalent to modelling the conditional probability of policy change (as opposed to no change).

confounders. Formally, the refined matched set for the treated observation (i, t) is given by,

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

where $S_{it}^{(J)}$ is 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 et al. 2003), based on the propensity score model given in Equation (13).[8] In this case, the weight for a matched control unit i' is defined as,

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

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 because the latter assigns an equal weight to each unit in the refined matched set $\mathcal{M}_{it}^{*}$,

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

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

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 DiD estimate of the ATT for each treated observation and then average it across all treated observations. Formally, our ATT estimator is given by,

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

[8] One can also use calibration weights instead of inverse propensity score weights.

where $D_{it} = X_{it}(1 - X_{i,t-1}) \cdot \mathbf{1}\{|\mathcal{M}_{it}| > 0\}$, and $w_{it}^{i'}$ represents the nonnegative 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.

When researchers are interested in a noncontemporaneous 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. Appendix A on p. 1 further discusses this alternative quantity of interest.

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 framework, 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 pretreatment time period, that is, $t' < t$. We further standardize this difference, at any given pretreatment 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 pretreatment time period $t - \ell$ as,

$$B_{it}(j, \ell) = \frac{V_{i,t-\ell,j} - \sum_{i' \in \mathcal{M}_{it}} w_{it}^{i'} V_{i',t-\ell,j}}{\sqrt{\frac{1}{N_1 - 1} \sum_{i'=1}^{N} \sum_{t'=L+1}^{T-F} D_{i't'} (V_{i',t'-\ell,j} - \overline{V}_{t'-\ell,j})^2}}, \quad (19)$$

where $N_1 = \sum_{i'=1}^{N} \sum_{t'=L+1}^{T-F} D_{i't'}$ is the total number of treated observations and $\overline{V}_{t-\ell,j} = \sum_{i=1}^{N} D_{i,t-\ell,j}/N$. We

then aggregate this covariate balance measure across all treated observations for each covariate and pretreatment time period.

$$\overline{B}(j,\ell) = \frac{1}{N_1}\sum_{i=1}^{N}\sum_{t=L+1}^{T-F} D_{it}\, B_{it}(j,\ell). \tag{20}$$

Finally, we emphasize that one must examine the balance of the lagged outcome variables over multiple pretreatment 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.

Relations with Linear Fixed Effects Regression Estimators

It is well known that the standard DiD estimator is 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 does not generalize to the multiperiod DiD design considered in this article, in which the number of time periods may exceed two and each unit may receive the treatment multiple times (see e.g. Abraham and Sun 2018; Athey and Imbens 2018; Chaisemartin and D'Haultfœuille 2018; Goodman-Bacon 2018; Imai and Kim 2011, 2021). 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 et al. (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 (1), in which the treatment will be in place at least for F time periods. This generalizes the result of Imai and Kim (2021). Specifically, we allow for estimating both short-term and long-term average treatment effects with nonparametric covariate adjustment.

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

$$\hat{\beta}_{DiD} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}\{(Y_{it} - \overline{Y}_i^* - \overline{Y}_t^* + \overline{Y}^*) - \beta(X_{it} - \overline{X}_i^* - \overline{X}_t^* + \overline{X}^*)\}^2. \tag{21}$$

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

$$W_{it} = \sum_{i'=1}^{N}\sum_{t'=1}^{T} D_{i't'} \cdot v_{it}^{i't'} \quad \text{and}$$

$$v_{it}^{i't'} = \begin{cases} 1 & \text{if } (i,t) = (i', t'+F) \\ 1 & \text{if } (i,t) = (i', t'-1) \\ w_{i't'}^{i} & \text{if } i \in \mathcal{M}_{i't'},\ t = t'+F \\ -w_{i't'}^{i} & \text{if } i \in \mathcal{M}_{i't'},\ t = t'-1 \\ 0 & \text{otherwise.} \end{cases} \tag{22}$$

Proof is in Appendix B on p. 1.

Importantly, the regression weight W_{it} can take a negative value in many cases, implying 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 modelling assumption is correct.

Standard Error Calculation

To compute the standard errors of the proposed estimator given in Equation (18), we condition on the weights implied by the matching procedure, which represents the number of times an observation is used for matching (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). For the proposed estimator, this observation-specific weight can be computed as follows,

$$W_{it}^{*} = \sum_{i'=1}^{N}\sum_{t'=1}^{T} D_{i't'} \cdot v_{it}^{i't'} \quad \text{and}$$

$$v_{it}^{i't'} = \begin{cases} 1 & \text{if } (i,t) = (i', t'+F) \\ -1 & \text{if } (i,t) = (i', t'-1) \\ -w_{i't'}^{i} & \text{if } i \in \mathcal{M}_{i't'},\ t = t'+F \\ w_{i't'}^{i} & \text{if } i \in \mathcal{M}_{i't'},\ t = t'-1 \\ 0 & \text{otherwise,} \end{cases} \quad (23)$$

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: $\hat{\delta}(F, L) = \sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}^{*} Y_{it} / \sum_{i=1}^{N}\sum_{t=1}^{T} D_{it}$.

We consider both conditional and unconditional standard errors. In both cases, we apply the strategy of matching as nonparametric preprocessing (Ho et al. 2007) and do not account for the uncertainty of the matching process. This results in the treatment of the weight W_{it} as an observed variable. Define $A = \sum_{i=1}^{N} A_i$ with $A_i = \sum_{t=1}^{T} W_{it}^{*} Y_{it}$ and $B = \sum_{i=1}^{N} B_i$ with $B_i = \sum_{t=1}^{T} D_{it}$. Then, for the conditional standard error, under the assumption of independence across units (but not across time periods), we have

$$\mathbb{V}(\hat{\delta}(F, L) \mid \mathbf{D}) = \frac{N^{*}\mathbb{V}(A_i)}{B^2},$$

where N^{*} represents the total number of units with at least one nonzero weight.

For the unconditional standard error, we use the first-order Taylor approximation for the asymptotic variance.

$$\mathbb{V}(\hat{\delta}(F, L)) = \mathbb{V}\left(\frac{A}{B}\right) \approx \frac{1}{\mathbb{E}(B)^2} \left\{\mathbb{V}(A) - 2\frac{\mathbb{E}(A)}{\mathbb{E}(B)}\mathrm{Cov}(A, B) + \frac{\mathbb{E}(A)^2}{\mathbb{E}(B)^2}\mathbb{V}(B)\right\},$$

where $\mathbb{E}(A) = N \cdot \mathbb{E}(A_i)$, $\mathbb{V}(A) = N \cdot \mathbb{V}(A_i)\mathbb{E}(B) = N \cdot \mathbb{E}(B_i)$, $\mathbb{V}(B) = N \cdot \mathbb{V}(B_i)$, $\mathrm{Cov}(A, B) = N \cdot \mathrm{Cov}(A_i, B_i)$. For unconditional standard error, it is also possible to 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. Abadie and Imbens (2008) show that a standard bootstrap procedure yields an invalid inference for matching estimators. However, we circumvent this problem by conditioning on the weights rather than recompute them for each bootstrapped sample (see also Otsu and Rai 2017).

A Simulation Study

We conduct simulations 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 choose a simulation setting that is favourable to OLS by generating the data from a linear model. We then introduce model misspecification 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 analysing TSCS data.

All the details and results of the simulation study are given in Appendix C on p. 3. Even in this simulation setting favourable to OLS, we find that the proposed matching estimator is much more robust to the omission of relevant lags than the linear regression estimator with fixed effects. However, this increased robustness of matching comes at the expense of statistical power. This finding reflects a fundamental tradeoff between bias and variance in statistics. In general, matching estimators tend to have less bias but also less efficient than regression estimators.

Empirical Analyses

We revisit the two motivating studies described earlier and reanalyse their data by applying the proposed methodology. 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.

Application of Matching Methods

For the Acemoglu et al. study, we estimate the two effects of democracy on economic growth, the effect of democratization and that of authoritarian reversal. Because 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,

$$\mathbb{E}\left[Y_{i,t+F}\left(X_{it} = 0, X_{i,t-1} = 1, \{X_{i,t-\ell}\}_{\ell=2}^{L}\right) - Y_{i,t+F}\left(X_{it} = 1, X_{i,t-1} = 1, \{X_{i,t-\ell}\}_{\ell=2}^{L}\right) \mid X_{it} = 0, X_{i,t-1} = 1\right]. \quad (24)$$

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 (1). We present the covariate balance for this alternative quantity of interest in Appendix D on p. 11.

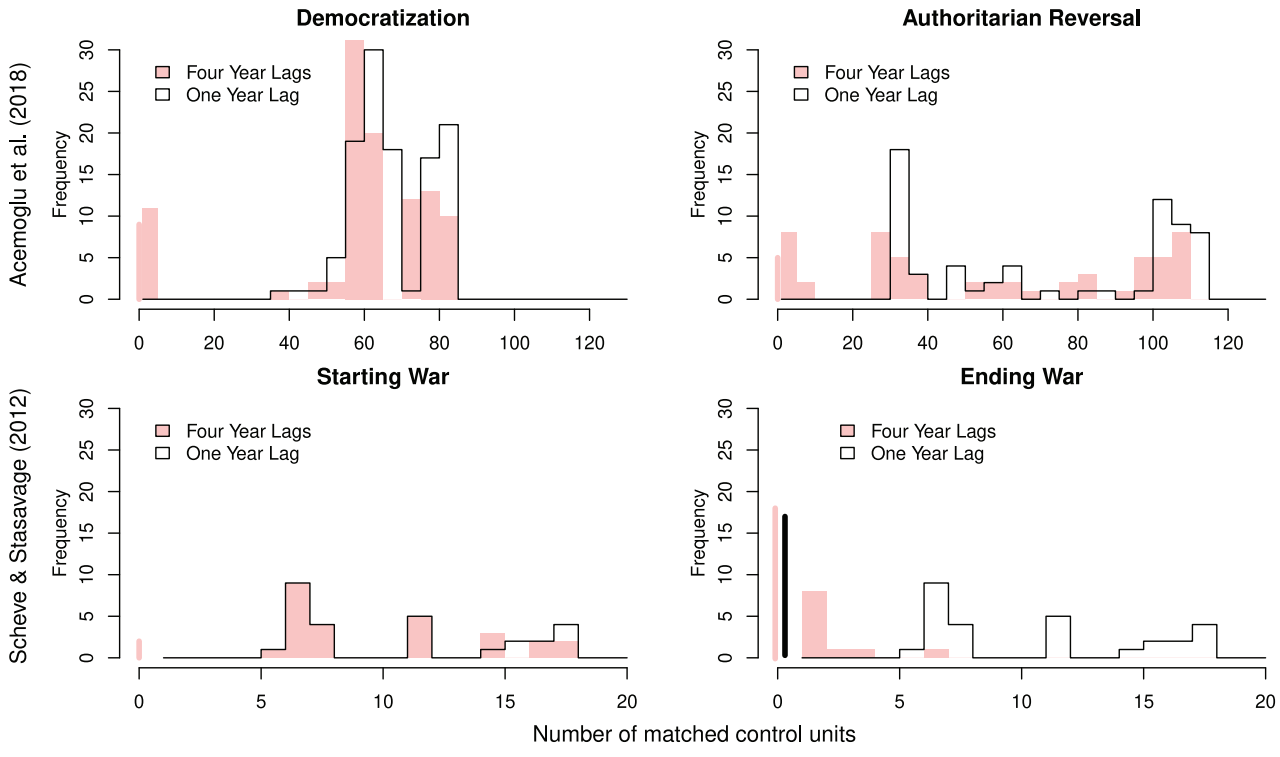
As shown in the left panel of Figure 1, although most countries transition from autocracy to democracy, we also observe enough cases of authoritarian reversal, suggesting that we may have sufficient data to estimate both effects. In contrast, for the Scheve and Stasavage study, we focus on the effect of involvement in a war on inheritance tax rather than the effect of ending a war because 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 D on p. 11.

We use the original studies to guide the specification of matching methods. In their regression models, Acemoglu et al. include 4 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 4 years of lag, that is, $L = 4$, and estimate the ATT up to 4 years after regime change, that is, $F = 1, 2, 3, 4$. In contrast, the dynamic model of Scheve and Stasavage adjusts only for 1-year lag of the outcome variable (see Equation (6)). Because 1-year lag may not be sufficient, we also conduct an analysis based on 4-year lags 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 1- and 4-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 4-year period rather than that of 1-year period.

For the Acemoglu et al. study in the upper panel, there are nine (five) treated observations for

FIGURE 3 Frequency Distribution of the Number of Matched Control Units



Note: The transparent (red) bar represents the number of matched control units that share the same treatment history as a treated observation for 1 year (4 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.

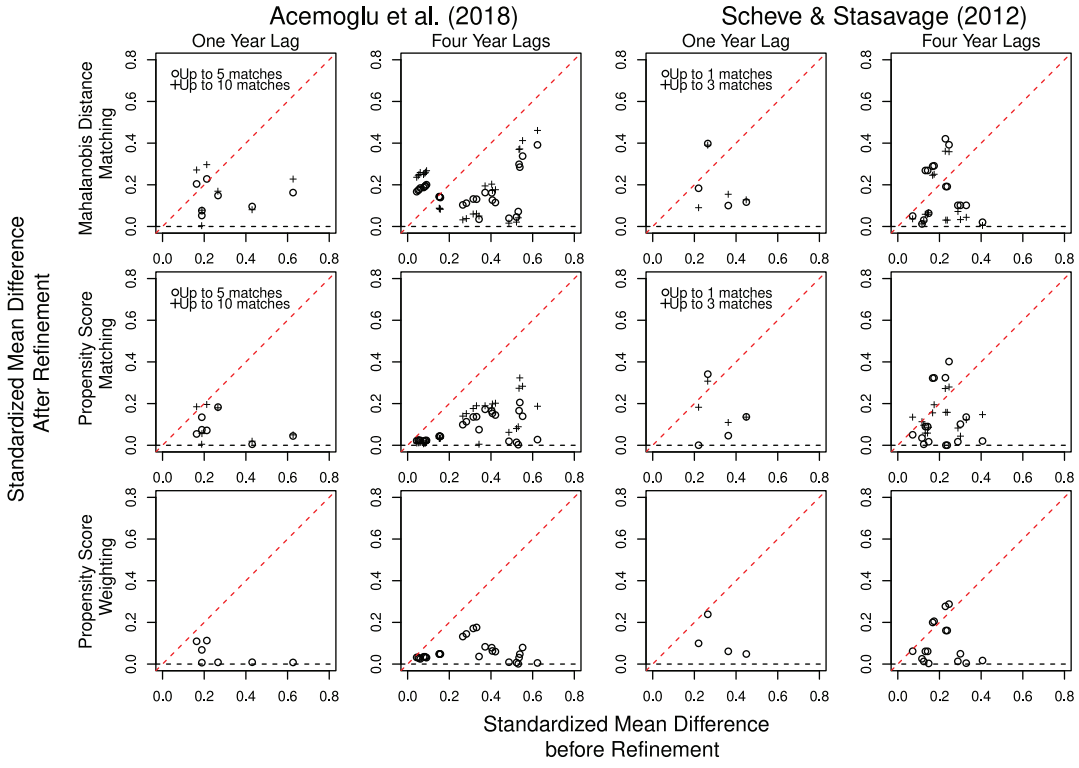
democratization (authoritarian reversal) that have no control unit with the same treatment history when the number of lags is four (represented by a thin red vertical bar at zero), whereas no such treated observation exists for the case of 1-year lag. We have enough matched control units for both democratization and authoritarian reversal: Most treated observations have more than 30 matched control units.

However, for the Scheve and Stasavage 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 two treated observations without any matched control units if we match on four lags, as represented by a thinner red vertical bar at zero. 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. Thus, causal inference is challenging especially when estimating the effects of ending war. Below, we do not estimate the effects of ending war because such estimates have low validity.

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. study to examine the sensitivity of empirical findings to the maximum number of matches. For the Scheve and Stasavage 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), whereas 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. 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

FIGURE 4 Improved Covariate Balance Due to Refinement of Matched Sets



Note: Each scatter plot compares the absolute value of standardized mean difference for each covariate j and lag year ℓ defined in Equation (20) before (horizontal axis) and after (vertical axis) the refinement of matched sets. Rows represents the results based on different matching and weighting methods whereas the columns represent the results using the adjustments for different lag lengths.

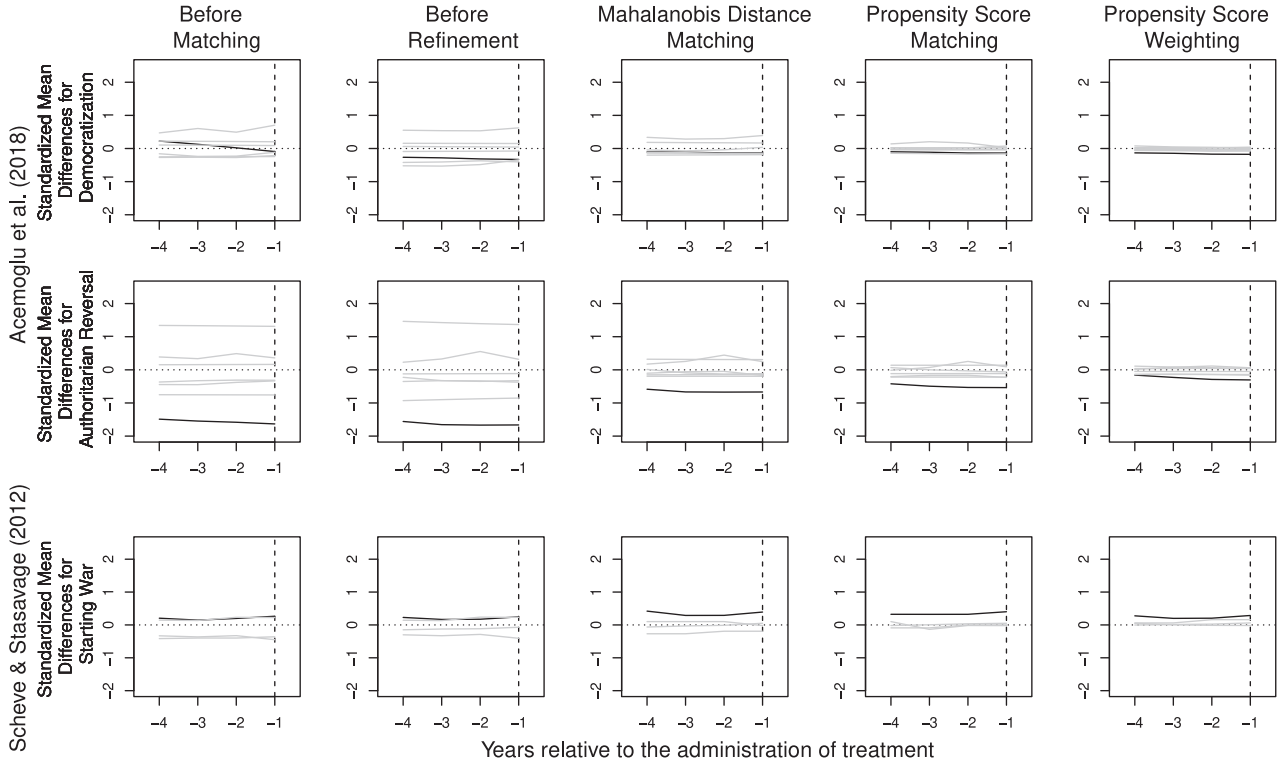
volume as a fraction of GDP and a dichotomous measure of social unrest (though the original authors do not include all variables at once in their regression model). Similarly, for the Scheve and Stasavage study, we use all available time-varying covariates, that is, an indicator variable for leftist 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 (20) 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 time-varying covariate. 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 pretreatment time period. We focus on the results for matching methods that adjust for time-varying covariates during the 4-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. study whereas the bottom row shows that for the treatment of starting war in the Scheve and Stasavage 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 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

FIGURE 5 Improved Covariate Balance Due to Matching over the Pre-Treatment Time Period



Note: Each plot plots the standardized mean difference defined in Equation (20) (vertical axis) over the pretreatment time period of 4 years (horizontal axis). The left column shows the balance before matching, whereas 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.

lagged outcome stays relatively constant over the entire pretreatment period. This suggests that the assumption of parallel trend for the proposed DiD estimator may be appropriate.

Empirical Findings

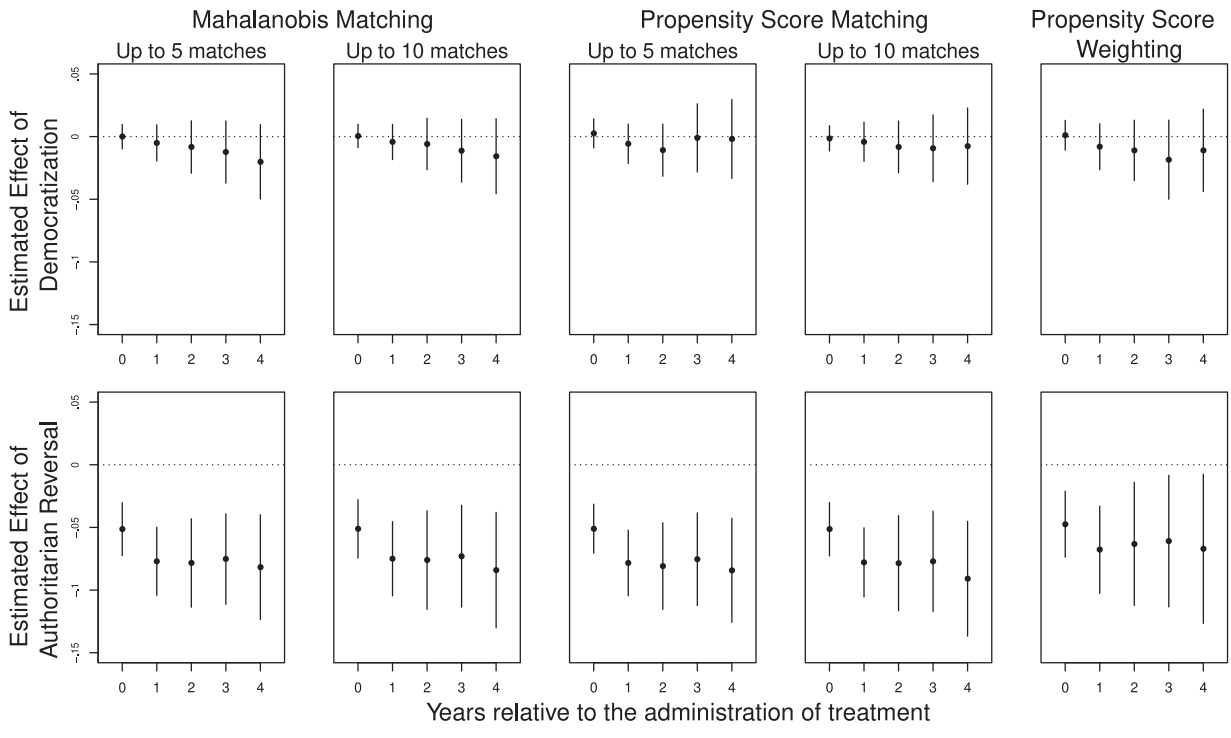
We now present the estimated ATTs based on the matching methods. Figure 6 shows the matching estimates of the effects of democratization (upper panel) and authoritarian reversal (lower panel) on logged GDP per capita for the period of 5 years after the transition, that is, $F = 0, 1, \ldots, 4$. Across all five methods (columns), we find that the point estimates of the effects for democratization are mostly close to zero over the 5-year time period. On the other hand, the estimated effects of authoritarian reversal are negative and statistically significant across all refinement methods during the year of transition and the 1 to 4 years immediately after the transition when the treatment reversal is allowed. The estimated effects are substantively large, indicating an approximately 5% to 8% reduction of GDP per capita. Although the confidence interval is wide, this effect size is greater than the estimated effect of 1% found in the original analysis (see Table 1). In Figure E.1 of Appendix E on p. 16, as a robustness check, we show that the same analysis with the refinement based on 1-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. 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.[9]

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

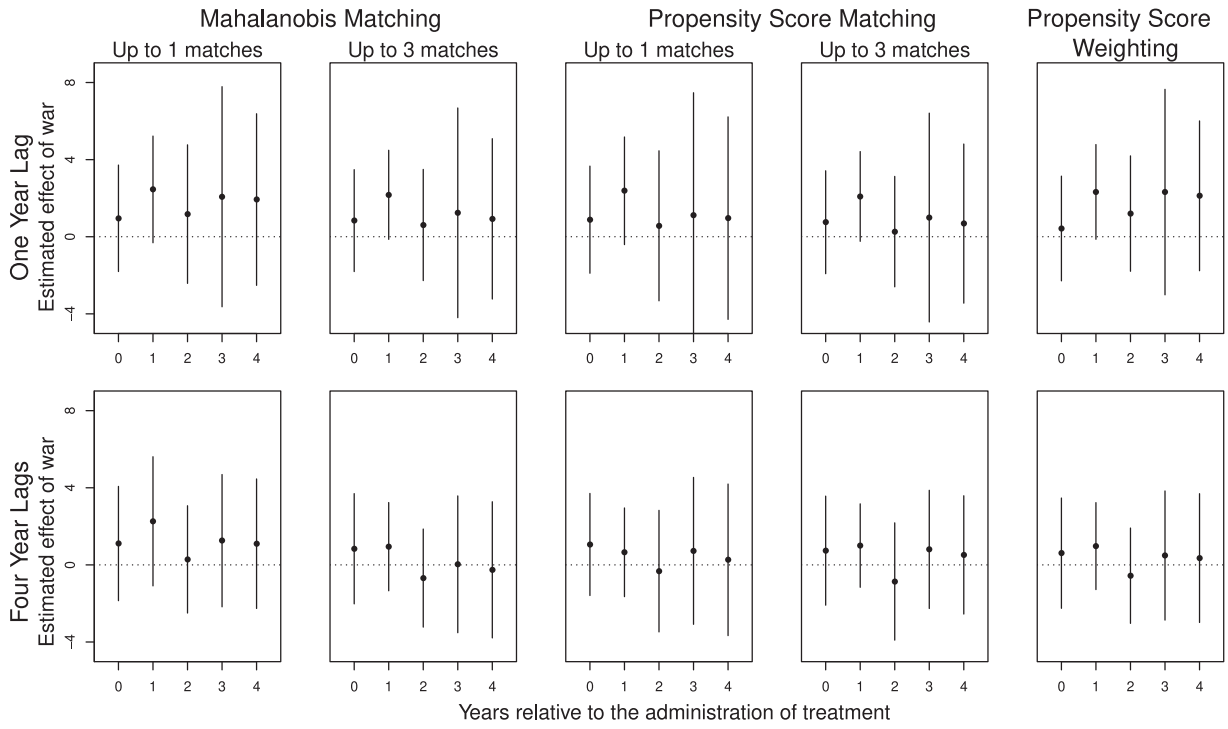
[9]The original authors also seek to separately estimate the effects of democratic transition and authoritarian reversal, using the linear regression models. Appendix F on p. 16 discusses this approach in detail. The empirical results obtained from this approach substantively differ from those presented here.

FIGURE 6 Estimated Average Effects of Democracy on Logged GDP per Capita



Note: The estimates are based on the matching method that adjusts for the treatment and covariate histories during the 4-year period prior to the treatment, that is, $L = 4$. The estimates for the average effects of democratization (upper panel) and authoritarian reversal (lower panel) are shown for the period of 5 years after the transition, that is, $F = 0, 1, \ldots, 4$, with 95% asymptotic confidence intervals as vertical bars. Five different refinement methods are considered and their results are presented in different columns.

FIGURE 7 Estimated Average Effects of Interstate War on Inheritance Tax Rate



Note: The matching method adjusts for the treatment and covariate histories during the 1- (upper panel) or 4 (lower panel) year period prior to the treatment. The estimated effects are shown for the period of 5 years after the war, that is, $F = 0, 1, \ldots, 4$, with 95% asymptotic confidence intervals as vertical bars. Five different matching/weighting methods are considered and their results are presented in different columns.

the treatment and covariates from 1 year period prior to the treatment. In contrast, the lower panel presents the estimates based on the adjustment for the 4-year pretreatment 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% asymptotic confidence intervals.

We find that if we refine the matched set using the 1-year pretreatment period, most of the estimated effects are not statistically significant. All of the estimated causal effects are not statistically significant if we refine the matched sets by adjusting for the 4-year pretreatment 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. Our analysis suggests that it is difficult to conclusively establish the positive effects of war on inheritance tax rate.

Concluding Remarks

Due to its simplicity and transparency, matching methods have become part of tool kit for empirical researchers who wish to estimate causal effects in observational studies. Yet, most matching methods have been developed for causal inference with cross-sectional data. We fill this gap by developing a methodological framework that enables the application of matching methods to causal inference with 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, our methods offer simple diagnostics through balance checking.

The proposed methodology can be extended in a number of ways. First, although we focus on the binary treatment variable in this article, the method can be extended to deal with a nonbinary (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. Although 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 neighbours as well as its own treatment history. We plan to explore such extensions of the proposed methods in our future research.

References

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

Abadie, A., A. Diamond, and J. 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(490): 493–505.

Abadie, A., and G. W. Imbens. 2008. "On the Failure of the Bootstrap for Matching Estimators." *Econometrica* 76(6): 1537–57.

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

Abraham, S., and L. Sun. 2018. "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects." Tech. rep., Department of Economics, Massachusetts Institute of Technology.

Acemoglu, D., S. Naidu, P. Restrepo, and J. A. Robinson. 2019. "Democracy Does Cause Growth." *Journal of Political Economy* 127(1): 47–100.

Angrist, J. D., and J. S. Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.

Arellano, M., and S. Bond. 1991. "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations." *Review of Economic Studies* 58(2): 277–97.

Aronow, P., and C. Samii. 2017. "Estimating Average Causal Effects under General Interference." *Annals of Applied Statistics* 11(4): 1912–47.

Athey, S., and G. Imbens. 2018. "Design-Based Analysis in Difference-in-differences Settings with Staggered Adoption." Tech. rep., Stanford Graduate School of Business, https://arxiv.org/abs/1808.05293.

Beck, N., and J. N. Katz. 1995. "What to Do (and Not to Do) with Time-Series Cross-Section Data." *American Political Science Review* 89(3): 634–47.

Ben-Michael, E., A. Feller, and J. Rothstein. 2019a. "The Augmented Synthetic Control Method." Tech. rep., arXiv:1811.04170.

Ben-Michael, E., A. Feller, and J. Rothstein. 2019b. "Synthetic Controls and Weighted Event Studies with Staggered Adoption." Tech. rep., arXiv:1912.03290.

Bertrand, M., E. Duflo, and S. Mullainathan. 2004. "How Much Should We Trust Differences-in-Differences Estimates?" *Quarterly Journal of Economics* 119(1): 249–75.

Blackwell, M., and A. Glynn. 2018. "How to Make Causal Inferences with Time-Series Cross-Sectional Data under Selection on Observables." *American Political Science Review* 112(4): 1067–82.

Chaisemartin, C. d., and X. D'Haultfœuille. 2018. "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects." Tech. rep., Department of Economics, University of California, Santa Barbara, https://arxiv.org/abs/1803.08807.

Diamond, A., and J. Sekhon. 2013. "Genetic Matching for Estimating Causal Effects: A New Method of Achieving Balance in Observational Studies." *Review of Economics and Statistics* 95(3): 932–45.

Doudchenko, N., and G. Imbens. 2017. "Balancing, Regression, Difference-in-Differences and Synthetic Control Methods: A Synthesis." Tech. rep., arXiv:1610.07748.

Goodman-Bacon, A. 2018. "Difference-in-differences with Variation in Treatment Timing." Working Paper 25018, National Bureau of Economic Research.

Hansen, B. B. 2004. "Full Matching in an Observational Study of Coaching for the SAT." *Journal of the American Statistical Association* 99(467): 609–18.

Hernán, M. A., and J. M. Robins. 2016. "Using Big Data to Emulate A Target Trial When a Randomized Trial Is Not Available." *American Journal of Epidemiology* 183(8): 758–64.

Hirano, K., G. Imbens, and G. Ridder. 2003. "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score." *Econometrica* 71(4): 1307–38.

Ho, D. E., K. Imai, G. King, and E. A. Stuart. 2007. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." *Political Analysis* 15(3): 199–236.

Hudgens, M. G., and E. Halloran. 2008. "Toward Causal Inference with Interference." *Journal of the American Statistical Association* 103(482): 832–42.

Iacus, S., G. King, and G. Porro. 2011. "Multivariate Matching Methods that are Monotonic Imbalance Bounding." *Journal of the American Statistical Association* 106(493): 345–61.

Imai, K., Z. Jiang, and A. Malai. 2021. "Causal Inference with Interference and Noncompliance in Two-Stage Randomized Experiments." *Journal of the American Statistical Association* 116(534): 632–44.

Imai, K., and I. S. Kim 2011. "On the Use of Linear Fixed Effects Regression Models for Causal Inference." Tech. rep., Princeton University.

Imai, K., and I. S. Kim. 2019. "When Should We Use Linear Unit Fixed Effects Regression Models for Causal Inference with Longitudinal Data?" *American Journal of Political Science* 63(2): 467–90.

Imai, K., and I. S. Kim. 2021. "On the Use of Two-Way Fixed Effects Regression Models for Causal Inference with Panel Data." *Political Analysis* 29(3): 405–15.

Imai, K., and M. Ratkovic. 2014. "Covariate Balancing Propensity Score." *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* 76(1): 243–63.

Imai, K., and M. Ratkovic. 2015. "Robust Estimation of Inverse Probability Weights for Marginal Structural Models." *Journal of the American Statistical Association* 110(511): 1013–23.

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

Li, Y. P., K. J. Propert, and P. R. Rosenbaum. 2001. "Balanced Risk Set Matching." *Journal of the American Statistical Association* 96(455): 870–82.

Nickell, S. 1981. "Biases in Dynamic Models with Fixed Effects." *Econometrica* 49(6): 1417–26.

Nielsen, R., and J. Sheffield. 2009. "Matching with Time-series Cross-sectional Data. Tech. rep., Harvard University." http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.510.7097&rep=rep1&type=pdf.

Otsu, T., and Y. Rai. 2017. "Bootstrap Inference of Matching Estimators for Average Treatment Effects." *Journal of the American Statistical Association* 112(520): 1720–32.

Robins, J. M. 1994. "Correcting for Non-Compliance in Randomized Trials Using Structural Nested Mean Models." *Communications in Statistics – Theory and Methods* 23(8): 2379–412.

Robins, J. M., M. A. Hernán, and B. Brumback. 2000. "Marginal Structural Models and Causal Inference in Epidemiology." *Epidemiology* 11(5): 550–60.

Rosenbaum, P. R., R. N. Ross, and J. 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(477): 75–83.

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: 212–18.

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

Scheve, K., and D. Stasavage. 2012. "Democracy, War, and Wealth: Lessons from Two Centuries of Inheritance Taxation." *American Political Science Review* 106(1): 81–102.

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

Tchetgen Tchetgen, E. J., and T. J. VanderWeele. 2010. "On Causal Inference in the Presence of Interference." *Statistical Methods in Medical Research* 21(1): 55–75.

Xu, Y. 2017. "Generalized Synthetic Control Method: Causal Inference with Interactive Fixed Effects Models." *Political Analysis* 25(1): 57–76.

Zubizarreta, J. R. 2012. "Using Mixed Integer Programming for Matching in an Observational Study of Kidney Failure after Surgery." *Journal of the American Statistical Association* 107(500): 1360–71.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Appendix A: Specifying the Future Treatment Sequence
Appendix B: Proof of Theorem 1
Appendix C: A Simulation Study
Appendix D: Covariate Balance when the Treatment Reversal is Not Allowed **Figure D.1**: Covariate Balance due to the Refinement of Matched Sets when Estimating the Average Effects of Stable Policy Change, with $F = 1$.
Appendix E: The Results based on One Year Lag
Appendix F: The Estimated Effects of Democratization and Authoritarian Reversal based on the Linear Regression Models.