Check for updates

15206688, 2020, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/pam.22206 by Ecole Normale Supérieure de Paris, Wiley Online Library on [08/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Methods for Policy Analysis

Burt Barnow, Editor

General comment: not a very good paper, not useful

Does not agree / know of Kloßner !

Cherry Picking with Synthetic Controls

Bruno Ferman, Cristine Pinto, and Vitor Possebom

Abstract

We evaluate whether a lack of guidance on how to choose the matching variables used in the Synthetic Control (SC) estimator creates specification-searching opportunities. We provide theoretical results showing that specification-searching opportunities are asymptotically irrelevant if we restrict to a subset of SC specifications. However, based on Monte Carlo simulations and simulations with real datasets, we show significant room for specification searching when the number of pre-treatment periods is in line with common SC applications, and when alternative specifications commonly used in SC applications are also considered. This suggests that such lack of guidance generates a substantial level of discretion in the choice of the comparison units in SC applications, undermining one of the advantages of the method. We provide recommendations to limit the possibilities for specification searching in the SC method. Finally, we analyze the possibilities for specification searching and provide our recommendations in a series of empirical applications. © *2020 by the Association for Public Policy Analysis and Management*

INTRODUCTION

The synthetic control (SC) method has been recently proposed in a series of seminal papers by Abadie and Gardeazabal (2003) and Abadie, Diamond, and Hainmueller (2010, 2015) as an alternative method to estimate treatment effects in comparative case studies. Despite being relatively new, this method has been used in a wide range of applications in Political Science, Economics, and other Social

Journal of Policy Analysis and Management, Vol. 39, No. 2, 510–532 (2020)
© 2020 by the Association for Public Policy Analysis and Management
Published by Wiley Periodicals, Inc. View this article online at wileyonlinelibrary.com/journal/pam
DOI:10.1002/pam.22206

15206688, 2020, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/pam.22206 by Ecole Normale Supérieure de Paris, Wiley Online Library on [08/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Sciences.[1] Athey and Imbens (2017) describe the SC method as arguably the most important innovation in the policy evaluation literature in the last fifteen years.

Abadie, Diamond, and Hainmueller (2010, 2015) describe many advantages of the SC estimator over techniques traditionally used in comparative studies. Among them, one important feature of the SC method is that it provides a transparent way to choose comparison units. In the SC method, a data-driven process is used to choose the weights that build the weighted-average of the controls' outcomes that estimates the counterfactual for the treated unit. Also, since the estimation of the SC weights does not require access to post-intervention outcomes, researchers could decide on the study design without knowing how those decisions would affect the conclusions of their studies. Taken together, these features potentially make the SC method less susceptible to specification searching relative to alternative methods for comparative case studies. This could be an important advantage of the SC method, given the growing debate about transparency in social science research (e.g., Miguel et al., 2014).[2]

An important limitation of the SC method, however, is that there is no consensus on the choice of predictor variables and covariates that should be used to estimate the SC weights.[3] Although Abadie, Diamond, and Hainmueller (2010) define vectors of linear combinations of pre-intervention outcomes that could be used as predictors, there is no specific recommendation about which variables should be used. Such lack of guidance on how to choose the predictors when implementing the synthetic control method translates into a wide variety of different specifications in empirical applications. If different specifications result in widely different choices of the SC unit, then a researcher would have relevant opportunities to select "statistically significant" specifications even when there is no effect. This flexibility may undermine one of the potential advantages of the SC method, as it essentially implies some discretionary power for the researcher to construct the counterfactual for the treated unit—and, therefore, the estimated treatment effects—by choosing which predictors to include, rather than having a purely data-driven process.[4]

In this paper, we investigate these opportunities for specification searching by considering only one particular step of the method: the choice of pre-treatment outcome lags used in the estimation of the SC weights.[5] In the following section,

[1] SC has been used to analyze terrorism (Abadie & Gardeazabal, 2003; Montalvo, 2011), political and economic reforms (Billmeier & Nannicini, 2011; Billmeier & Nannicini, 2013); crime and police (Cunningham & Shah, 2018; DeAngelo & Hansen, 2014; Donohue, Aneja, & Weber, 2018; Pinotti, 2013); natural resources and disasters (Barone & Mocetti, 2014; Cavallo et al., 2013; Smith, 2015); immigration (Bohn, Lofstrom, & Raphael, 2014; Dustmann, Schonberg, & Stuhler, 2017); education (Belot & Vandenberghe, 2014; Hinrichs, 2012); pregnancy and parental leave (Bartel et al., 2018; Lindo & Packham, 2017); taxation (Baccini, Li, & Mirkina, 2014; Kleven, Landais, & Saez, 2013); social connections (Acemoglu et al., 2016); local development (Gobillon & Magnac, 2016; Zou, 2018).

[2] See Christensen and Miguel (2018) for an extensive literature review on research transparency and reproducibility.

[3] Dube and Zipperer (2015) and Kaul et al. (2018) point out that there is little explicit guidance in the SC literature on how to choose predictors. However, they do not explore the implications of such lack of consensus on the possibilities for specification searching in SC applications.

[4] Olken (2015) and Coffman and Niederle (2015) evaluate the use of pre-analysis plans in social sciences. However, in many synthetic control applications both pre- and post-intervention information would be available to the researcher before the possibility of registering the study, implying that committing to a particular specification is infeasible. An alternative solution to this problem would be splitting samples (Fafchamps & Labonne, 2017). Once again, this solution is infeasible in SC applications because most of them have only one treated unit.

[5] There may be other important dimensions in the implementation of the SC method that provide discretionary choices for the researcher, such as the choice of which covariates to include as predictor

we first provide conditions under which different SC specifications lead to asymptotically equivalent estimators when the number of pre-treatment periods (T_0) goes to infinity and we restrict to specifications whose number of pre-treatment outcome lags used as predictors goes to infinity with T_0. This equivalence result is true whether or not covariates are included as predictors. Under these conditions, we also show that the placebo test suggested by Abadie, Diamond, and Hainmueller (2010) asymptotically leads to the same conclusion regardless of the chosen specification. On the one hand, these results show that the SC method is robust to specification searching, *provided we have a large number of pre-treatment periods, and we restrict to a specific subset of specifications*. This is an important feature of the SC estimator that is not generally shared by other methods. On the other hand, these results point out exactly when specification searching should be a problem in SC applications. First, many SC applications do not have a large number of pre-treatment periods to justify large-T_0 asymptotics, as argued by Doudchenko and Imbens (2016), possibly leaving room to specification searching even if we restrict to this specific class of SC specifications. Moreover, there are common SC specifications whose number of included pre-treatment periods does not go to infinity, possibly leading to specification-searching opportunities even when the number of pre-treatment periods is large.

Guided by our theoretical results, we then measure the specification-searching opportunities in SC applications using Monte Carlo (MC) simulations in the third section, and placebo simulations with the Current Population Survey (CPS) in Appendix E.[6] We calculate the probability that a researcher could find at least one specification such that she would reject the null using the test procedure proposed by Abadie, Diamond, and Hainmueller (2010), when the actual effect of the intervention is zero. If different SC specifications lead to similar SC estimators, then this probability would be close to 5 percent for a 5 percent significance level test, while it may be much higher than 5 percent if different SC specifications lead to wildly different estimates, implying that there is room for specification searching. We consider seven different specifications commonly used in SC applications.[7]

We find that the probability of detecting a false positive in at least one specification for a 5 percent significance test can be as high as 14 percent when there are 12 pre-treatment periods. The possibilities for specification searching remain high even when the number of pre-treatment periods is large. For example, with 400 pre-treatment periods—which is much longer than the usual SC application—we still find a probability of around 13 percent that at least one specification is significant at 5 percent. These results suggest that, even with a large number of pre-treatment periods, different specifications that are commonly used in SC applications can still lead to significantly different synthetic control units, generating substantial opportunities for specification searching. Given our theoretical results,

variables, the choice of how to split the pre-treatment periods into training and validation periods, and even the choice of software and data-sorting criteria (see Klößner et al., 2017, for details on this last point). Therefore, our results should be seen as a lower bound on the possibilities for specification searching in SC applications. We focus on the choice of pre-treatment outcome lags, rather than on the inclusion of covariates, because it is possible to systematically analyze the inclusion of pre-intervention outcomes lags in a way that encompasses all applications, while covariates may differ in complex ways from one application to another. We consider the possibility of specification searching in the decision to include covariates in our empirical applications.

[6] Appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's website and use the search engine to locate the article at http://onlinelibrary.wiley.com.

[7] In Appendix B, we also consider specifications that use time-invariant covariates as predictors, in addition to functions of the pre-treatment outcomes. All results remain similar.

15206688, 2020, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/pam.22206 by Ecole Normale Supérieure de Paris, Wiley Online Library on [08/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

15206688, 2020, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/pam.22206 by Ecole Normale Supérieure de Paris, Wiley Online Library on [08/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

it is expected that the significant specification-searching possibilities with a large T_0 are driven by specifications that do not increase the number of pre-treatment lags used as predictors when the number of pre-treatment periods goes to infinity. Indeed, we find that excluding those specifications from the set of options strongly attenuates the specification-searching problem when T_0 is large. However, we still find significant possibilities for specification searching for values of T_0 commonly considered in SC applications, suggesting that reliable asymptotic approximations may require unrealistically long time series. We also show that specification searching may remain a problem even when we restrict the set of options to specifications with a good pre-treatment fit.

Since transparency in the choice of comparison units is one of the often-advocated advantages of the method (Abadie, Diamond, & Hainmueller, 2010, p. 494), our main conclusion is that such an advantage is weakened by a lack of consensus on which variables should be chosen as predictors to estimate the SC weights. If there were a consensus on how the SC specification should be selected, then the risk of p-hacking (at least in this dimension) would be limited. For this reason, we specifically recommend focusing on the specification that uses all the pre-treatment outcome lags as matching variables, unless there is a strong prior belief that it is crucial to balance on a specific set of covariates. We discuss this and other recommendations in the fourth section.

Finally, we also consider, in the fifth section, the possibilities for specification searching and the implementability of the above recommendations in two empirical applications based on Abadie, Diamond, and Hainmueller (2015) and Bartel et al. (2018). We find that different specifications can reach either significant or non-significant results, showing the potential for specification searching with synthetic controls. In Appendix F, we consider three more examples, two based on Smith (2015) and one based on Abadie, Diamond, and Hainmueller (2010).[8] In the first example, the conclusions are robust to specification searching; in the second example, most specifications show insignificant effects, but it would be possible to find a few "statistically significant" specifications; and, in the third example, all results are significant, but at different significance levels. While in some of these empirical applications conclusions vary depending on the SC specification, we show that applying our recommendations from the fourth section to these empirical applications provides clear conclusions about the significance of these estimates.

Appendix A presents the formal theoretical results and proofs that guide our investigation about specification-searching opportunities. The code for all our simulations and empirical examples was made available by Ferman, Pinto, and Possebom (2020).

SYNTHETIC CONTROLS AND SPECIFICATION SEARCHING

Abadie and Gardeazabal (2003) and Abadie, Diamond, and Hainmueller (2010, 2015) have developed the Synthetic Control Method (SCM) in order to address counterfactual questions involving only one treated unit. This method uses a weighted average of control units and flexibly estimates treatment effects for each post-treatment period. Below, we explain the SCM following Abadie, Diamond, and Hainmueller (2010).

[8] Appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's website and use the search engine to locate the article at http://onlinelibrary.wiley.com.

Suppose we observe data for $(J+1) \in \mathbb{N}$ units during $T \in \mathbb{N}$ time periods and a treatment that affects only unit 1 from period T_0+1 to period T uninterruptedly. Let $Y^0_{j,t}$ be the potential outcome that would be observed for unit j in period t if there were no treatment for $j \in \{1,\ldots,J+1\}$ and $t \in \{1,\ldots,T\}$. Let $Y^1_{j,t}$ be the potential outcome under treatment. Define $\alpha_{j,t} := Y^1_{j,t} - Y^0_{j,t}$ as the treatment effect and $Y_{j,t}$ as the observed outcome.

We aim to identify $(\alpha_{1,T_0+1},\ldots,\alpha_{1,T})$. Since $Y^1_{1,t}$ is observable for $t > T_0$, we only need to estimate the counterfactual $Y^0_{1,t}$ to accomplish this goal.

Let $\mathbf{Y_j} := [Y_{j,1}\ldots Y_{j,T_0}]'$ be the vector of observed outcomes for unit $j \in \{1,\ldots,J+1\}$ in the pre-treatment period and $\mathbf{X_j}$ be a $(F \times 1)$-vector of predictors of $\mathbf{Y_j}$. Those predictors can be not only covariates that explain the outcome variable, but also linear combinations of the variables in $\mathbf{Y_j}$. Let also $\mathbf{Y_0} = [\mathbf{Y_2}|\ldots|\mathbf{Y_{J+1}}]$ be a $(T_0 \times J)$-matrix and $\mathbf{X_0} = [\mathbf{X_2}|\ldots|\mathbf{X_{J+1}}]$ be a $(F \times J)$-matrix.

Given the choice of predictors in matrix $\mathbf{X_j}$, the idea of the SC method is to construct the counterfactual for the treated unit using a weighted average of the control units, $\hat{Y}^0_{1,t} := \sum_{j=2}^{J+1} \hat{w}_j Y_{j,t}$.

The weights $\hat{\mathbf{W}} = [\hat{w}_2 \ldots \hat{w}_{j+1}]' := \hat{\mathbf{W}}(\hat{\mathbf{V}}) \in \mathbb{R}^J$ are given by the solution to a nested minimization problem:

$$\hat{\mathbf{W}}(\mathbf{V}) := \arg\min_{\mathbf{W}\in\mathcal{W}} (\mathbf{X_1} - \mathbf{X_0}\mathbf{W})'\mathbf{V}(\mathbf{X_1} - \mathbf{X_0}\mathbf{W}) \tag{1}$$

where $\mathcal{W} := \{\mathbf{W} = [w_2 \ldots w_{J+1}]' \in \mathbb{R}^J : w_j \geq 0 \text{ for each } j \in \{2,\ldots,J+1\} \text{ and } \sum_{j=2}^{J+1} w_j = 1\}$ and $\mathbf{V}$ is a diagonal positive semidefinite matrix of dimension $(F \times F)$. Moreover,

$$\hat{\mathbf{V}} := \arg\min_{\mathbf{V}} (\mathbf{Y_1} - \mathbf{Y_0}\hat{\mathbf{W}}(\mathbf{V}))'(\mathbf{Y_1} - \mathbf{Y_0}\hat{\mathbf{W}}(\mathbf{V})). \tag{2}$$

Intuitively, $\hat{\mathbf{W}}$ is a weighting vector that measures the relative importance of each unit in the synthetic control of unit 1, while $\hat{\mathbf{V}}$ measures the relative importance of each one of the F predictors. The relative importance of each predictor is estimated in a data-driven optimization problem presented in equation (2). We define the Synthetic Control Estimator of $\alpha_{1,t}$ (or the estimated gap) as $\hat{\alpha}_{1,t} := Y_{1,t} - \hat{Y}^0_{1,t}$ for each $t \in \{1,\ldots,T\}$, where $\hat{Y}^0_{1,t}$ is constructed using weights $\hat{\mathbf{W}}(\hat{\mathbf{V}})$.

Even though a crucial part in the implementation of the SC method is the choice of predictors, there is no consensus on which variables to include in matrix $\mathbf{X_j}$. This lack of guidance can create an opportunity for the researcher to look for specifications that yield "better" results by including or excluding some pre-treatment outcome values from its specification. This risk is even greater when we consider that there is no consensus about which functions of the outcome values should be included in $\mathbf{X_j}$.

To illustrate this lack of consensus, we present in Table 1 a list with all papers that use the SC method published in the American Economic Review, American Economic Journal–Economic Policy, American Economic Journal–Applied Economics, Quarterly Journal of Economics, Review of Economic Studies, Review of Economics and Statistics, Journal of Development Economics, Journal of Labor Economics, and Journal of Policy Analysis and Management, including information on the specifications used in the implementation of the method. Abadie and Gardeazabal (2003), Abadie, Diamond, and Hainmueller (2015), Kleven, Landais, and Saez (2013), Baccini, Li, and Mirkina (2014), and DeAngelo and Hansen (2014) use the mean of all pre-treatment outcome values and additional covariates; Cunningham and Shah (2018) pick Y_{j,T_0}, Y_{j,T_0-1}, Y_{j,T_0-2}, Y_{j,T_0-7}, Y_{j,T_0-8}, Y_{j,T_0-9}, Y_{j,T_0-11}, Y_{j,T_0-14}, Y_{j,T_0-15}, $Y_{j.T_0-16}$ and additional covariates; Smith (2015) uses Y_{j,T_0}, Y_{j,T_0-2},

15206688, 2020, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/pam.22206 by Ecole Normale Supérieure de Paris, Wiley Online Library on [08/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 1. Published articles using the SCM.

Authors	Journal	Pre-treatment Periods	Post-treatment periods	Number of Covariates[a]	Outcome Lags[b]	Number of Control Units
Abadie and Gardeazabal (2003)	AER	10	30	11	Mean	16
Kleven et al. (2013)	AER	11	5	3	Mean	14
DeAngelo and Hansen (2014)	AEJ:EP	37	35	14	Mean	46
Lindo and Packham (2017)	AEJ:EP	6	5	0	−1, −3, −5	38
Dustmann et al. (2017)	QJE	6	5	5	All	85
Cunningham and Shah (2018)	RESTUD	18	6	5	0, −1, −2, −7, −8, −9, −11, −14,−15, −16	50
Montalvo (2011)	RESTAT	4	1	2	0, −1	32
Hinrichs (2012)	RESTAT	9	6	0	All	3-7
Billmeier and Nannicini (2013	RESTAT	2-32	10	5	All	4-62
Cavallo et al. (2013)	RESTAT	11	10	7	First Half	53
Bohn et al. (2014)	RESTAT	9	3	42	All	45
Gobillon and Magnac (2016)	RESTAT	8	13	0	All	135
Smith (2015)	JDE	10-43	16-49	2	0, −2, −4, −6	7-32
Zou (2018)	JLE	2	1	6	All	2429
Baccini et al. (2014)	JPAM	7	5	0	Mean	36
Eren and Ozbeklik (2016)	JPAM	19	6	7	Even Lags	28
Bartel et al. (2018)	JPAM	5	9	11	All, Mean	49

[a]Number of Covariates included in matrix **Xj** besides the ones related to the outcome variable.

[b]Outcome Lags included in matrix **Xj** . The last pre-treatment period ($T0$) is denoted by the number 0.

Notes: List of articles using the SC method published at American Economic Review, American Economic Journal–Economic Policy, American Economic Journal–Applied Economics, Quarterly Journal of Economics, Review of Economic Studies, Review of Economics and Statistics, Journal of Development Economics, Journal of Labor Economics, and Journal of Policy Analysis and Management. We did not find any articles using the SC method published at Econometrica or the Journal of Political Economy.

Journal of Policy Analysis and Management DOI: 10.1002/pam
Published on behalf of the Association for Public Policy Analysis and Management

15206688, 2020, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/pam.22206 by École Normale Supérieure de Paris, Wiley Online Library on [08/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Y_{j,T_0-4}, Y_{j,T_0-6} and additional covariates; Abadie, Diamond, and Hainmueller (2010) pick Y_{j,T_0}, Y_{j,T_0-8}, Y_{j,T_0-13} and additional covariates; Lindo and Packham (2017) pick Y_{j,T_0-1}, Y_{j,T_0-3}, Y_{j,T_0-5}; Billmeier and Nannicini (2013), Bohn et al. (2014), Gobillon and Magnac (2016), Hinrichs (2012), Dustmann, Schonberg, and Stuhler (2017), Zou (2018) and Bartel et al. (2018) use all pre-treatment outcome values; Cavallo et al. (2013) use the first half of the pre-treatment outcome values and additional covariates; Eren and Ozbeklik (2016) use the even-numbered pre-treatment lags and additional covariates; and Montalvo (2011) uses only the last two pre-treatment outcome values and additional covariates.[9]

A key question, therefore, is whether different specifications may lead to substantially different SC estimators. We consider the asymptotic behavior of different SC specifications when $T_0 \to \infty$. We define a specification s by the set of predictors $\mathbf{X}_j(s, T_0)$ that are used when there are T_0 pre-treatment periods, which may include pre-treatment outcome lags, functions of pre-treatment outcome lags, or other observed covariates. Let $L(s, T_0)$ be the number of pre-treatment periods t such that $Y_{j,t}$ is included as a predictor when there are T_0 pre-treatment periods. For example, consider a specification s to be such that R covariates and the first half of the pre-treatment outcome lags $\{1, 2, \ldots, T_0/2\}$ are used as predictors. Then, $L(s, T_0) = \frac{T_0}{2}$. Note that, in this case, the dimension of $\mathbf{X}_j(s, T_0)$ would be $F = R + \frac{T_0}{2}$.

Let $\hat{\mathbf{W}}(s, T_0)$ be the SC weights using specification s when there are T_0 pre-intervention periods. We want to understand under which conditions $\hat{\mathbf{W}}(s, T_0)$ converges in probability to the same $\bar{\mathbf{W}}$ for any specification s when $T_0 \to \infty$. We show in Proposition 2 (see Appendix A) that this is the case when we consider specifications such that the number of pre-treatment outcomes used as predictors increases with T_0 (i.e., $L(s, T_0) \to \infty$ when $T_0 \to \infty$). The only assumption we need is that pre-treatment averages for subsequences of $(Y^0_{1,t}, \ldots, Y^0_{J+1,t})$ converge to the same value.[10] Given that, the difference between two SC estimators using specifications s and s' converge in probability to zero if $L(s, T_0) \to \infty$ and $L(s', T_0) \to \infty$ when $T_0 \to \infty$ (see Corollary 3 in Appendix A).

The intuition for these results is that, when $T_0 \to \infty$, the minimization problem (2) that chooses the matrix $\hat{\mathbf{V}}$ will only assign positive weights to the pre-treatment outcome lags if $L(s, T_0) \to \infty$, even when other covariates are included. Therefore, asymptotically, all such specifications will choose the SC weights by minimizing an average of a function of the pre-treatment outcomes that are included as predictors. A formal proof is presented in the Appendix.

While different SC specifications may generate different SC estimates, our theoretical results show that, under some conditions, different specifications will lead to asymptotically equivalent SC estimators, as long as the number of pre-treatment lags used as predictors goes to infinity with T_0. However, our results do not guarantee that different SC specifications would lead to similar SC estimates when T_0 is finite, nor determine a value of T_0 that is large enough to ensure that the asymptotic approximation is reliable. Moreover, there are common specifications used in

[9] By no means do we imply that those authors have engaged in specification searching. We have only listed them as prominent examples of different choices regarding predictor variables. Given that this is a relatively new method, there are not enough papers to formally test for specification searching (Brodeur et al., 2016; Simonsohn, Nelson, & Simmons, 2014). Also, specification searching is, of course, not something specific to the SC method, and our results do not imply that this problem is more relevant for the SC method when compared to alternative methods (Gardeazabal & Vega-Bayo, 2016).

[10] In Appendix A, we present in detail sufficient conditions for this assumption. Appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's website and use the search engine to locate the article at http://onlinelibrary.wiley.com.

15206688, 2020, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/pam.22206 by Ecole Normale Supérieure de Paris, Wiley Online Library on [08/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

15206688, 2020, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/pam.22206 by Ecole Normale Supérieure de Paris, Wiley Online Library on [08/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

SC applications that do not satisfy the condition that $L(s, T_0) \to \infty$ when $T_0 \to \infty$. For example, in roughly a third of the published papers that use the SC method in Table 1, the authors consider the use of the mean of all pre-treatment outcome values in addition to other covariates as predictors. These alternative specifications would generally lead to SC weights that will not converge to $\bar{\mathbf{W}}$, so there may still be significant variation in the SC estimates even when T_0 is large.

We also consider the implications of our results on the asymptotic equivalence of different SC specifications to the inference method proposed by Abadie, Diamond, and Hainmueller (2015). They permute which unit is assumed to be treated and estimate, for each $j \in \{2, \ldots, J+1\}$ and $t \in \{1, \ldots, T\}$, $\hat{\alpha}_{j,t}$ as described above. Then, they compute the *ratio of the mean squared prediction errors* as a test statistic:

$$RMSPE_j := \frac{\frac{\sum_{t=T_0+1}^{T} \left(Y_{j,t} - \widehat{Y_{j,t}^N}\right)^2}{(T-T_0)}}{\frac{\sum_{t=1}^{T_0} \left(Y_{j,t} - \widehat{Y_{j,t}^N}\right)^2}{T_0}}. \tag{3}$$

Moreover, they propose to calculate a p-value, $p := \frac{\sum_{j=1}^{J+1} 1[RMSPE_j \geq RMSPE_1]}{J+1}$, and reject the null hypothesis of no effect if p is less than some pre-specified significance level. Abadie, Diamond, and Hainmueller (2010) recognize that the randomization inference assumptions are very restrictive for the SC set-up, as treatment is not, in general, randomly assigned.[11] In the absence of random assignment, they interpret the p-value as the probability of obtaining an estimate value for the test statistics at least as large as the value obtained using the treated case as if the intervention were randomly assigned among the data. Although the p-value from this placebo test lacks a clear statistical interpretation, this test is commonly used in SC application. Therefore, our simulation exercises can be seen as the probability that a researcher applying the SC method would find a test statistic that is in the top 5 percent of the distribution of test statistics in the placebo runs, which is how researchers applying the SC method usually assess whether their estimates are significant. Moreover, note that, in our simulations, the placebo test considering a single SC specification would have a rejection rate under the null of 5 percent by construction. In Appendix G, we also consider as a robustness check an *infeasible* test based on the actual distribution of the test statistic in our MC simulations to assess the statistical significance of the results.[12]

Given our results that the difference between two SC estimators using specifications s and s' converge in probability to zero when both $L(s, T_0) \to \infty$ and $L(s', T_0) \to \infty$ when $T_0 \to \infty$, we also show that the ranking of $RMSPE_j$ will remain asymptotically invariant to changes in the SC specification when $T_0 \to \infty$, whenever we consider only specifications whose number of pre-treatment outcome lags goes to infinity with T_0 (see Corollary 4 in Appendix A). As a consequence, the test decision in the placebo test is asymptotically invariant to the specification choice when $T_0 \to \infty$, provided we restrain to such set of SC specifications. Therefore, in this case, the possibilities for specification searching are asymptotically irrelevant.

[11] Firpo and Possebom (2018) discuss a sensitivity mechanism analysis for this test, while Ferman and Pinto (2017) analyze the statistical properties of this placebo test when treatment is not randomly assigned. Hahn and Shi (2017) also consider the properties of a placebo test in the SC setting. For our purposes in this paper, we consider Abadie, Diamond, and Hainmueller's (2010) interpretation of the placebo test's p-value.

[12] Appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's website and use the search engine to locate the article at http://onlinelibrary.wiley.com.

This is a feature of the SC method that is not generally shared by other methods and is valid even when covariates are included.[13]

In addition to showing that the SC estimator is robust to specification searching when T_0 is large and when we restrict attention to a subset of specifications, these theoretical results provide guidance on the conditions in which specification searching might be relevant in SC applications: (i) when T_0 is not large enough to ensure a reliable asymptotic approximation or (ii) when one considers specifications with few pre-treatment outcomes as predictors.

Monte Carlo Simulations

We design an MC simulation guided by our results presented in the previous section. We evaluate whether values of T_0 commonly used in SC applications are large enough so that our asymptotic results provide a reliable approximation, and whether alternative specifications commonly used in SC applications, but that do not satisfy the conditions in our theoretical results, can imply significant specification-searching possibilities even when T_0 is large.

We generate 10,000 datasets and, for each one of them, test the null hypothesis of no effect whatsoever adopting several different specifications. Conditional on a given specification, in our simulations, this placebo test should provide a rejection rate of α percent under the null for a α percent significance test by construction. We are interested, however, in the probability of rejecting the null hypothesis at the α-percent-significance level for at least one specification. If different specifications result in wildly different SC estimators, then the probability of finding one specification that rejects the null at α percent can be significantly higher than α percent. In the extreme case, in which we have S different specifications and these specifications lead to independent estimators, this probability would be given by $1-(1-\alpha)^S$.[14] In this case, lack of guidance about specification choice could generate substantial opportunities for specification searching. In contrast, if different SC specifications lead to similar SC weights, then this rejection rate will be close to α percent and the risk of specification searching would be very low. We consider two data-generating processes (DGP).

In the first DGP, we consider a linear factor model in which all units are divided into groups that follow different stationary time trends.

$$Y_{j,t}^0 = \delta_t + \lambda_t^k + \epsilon_{j,t} \tag{4}$$

for some $k = 1, \ldots, K$. We consider the case in which $J+1 = 20$ and $K = 10$. Therefore, units 1 and 2 follow the trend λ_t^1, units 3 and 4 follow the trend λ_t^2, and so on. We consider that λ_t^k is normally distributed following an AR(1) process with 0.5 serial correlation parameter, $\delta_t \sim N(0,\ 1)$ and $\epsilon_{j,t} \sim N(0,\ 0.1)$.

[13] For example, consider a field experiment in which the researcher may decide which set of covariates to include. Given random assignment, we have that all covariates are uncorrelated with the treatment variable. If covariates are relevant in explaining the outcome, there would still be room for specification searching in the choice of which covariates to include, even when the number of observations goes to infinity. Our theoretical results show that this is not the case in the SC method.

[14] Lovell (1983) provides a similar formula but considering the decision on which variables to include in a regression model.

15206688, 2020, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/pam.22206 by Ecole Normale Supérieure de Paris, Wiley Online Library on [08/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

15206688, 2020, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/pam.22206 by Ecole Normale Supérieure de Paris, Wiley Online Library on [08/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

In our second DGP, we modify the linear factor model such that a subset of the common factors is non-stationary. In this case, we consider a DGP that includes a non-stationary trend ϕ_t^r that follows a random walk,

$$Y_{j,t}^0 = \delta_t + \lambda_t^k + \phi_t^r + \epsilon_{jt} \tag{5}$$

for some $k = 1, \ldots, K$ and $r = 1, \ldots, R$. We consider in our simulations $K = 10$ and $R = 2$. Therefore, units $j = 2, \ldots, 10$ follow the same non-stationary path ϕ_t^1 as the treated unit, although only unit $j = 2$ also follows the same stationary path λ_t^1 as the treated unit.

We fix the number of post-treatment periods $T - T_0 = 10$ and we vary the number of pre-intervention periods in the DGPs, $T_0 \in \{12,\ 32,\ 100,\ 400\}$. Note that seven papers in Table 1 use a number of pre-treatment periods around 12 (i.e., between eight and 16). Moreover, the longest pre-treatment period is 43. Therefore, setting $T_0 = 400$ in our Monte Carlo is useful to test the reliability of the asymptotic approximations described in the previous section, but we should bear in mind that this is an extreme setting that is unlikely to hold in common SC applications. In both models, we impose that there is no treatment effect, i.e., $Y_{j,t} = Y_{j,t}^0 = Y_{j,t}^1$ for each time period $t \in \{1, \ldots, T\}$.

In Appendix G, we consider variations in our stationary model (4) by setting (i) $\epsilon_{j,t} \sim N(0, 1)$, and (ii) $K = 2$. In Appendix B, we consider a DGP with time-invariant covariates. Moreover, in Appendix E, we consider placebo simulations with the CPS.[15] In all cases, we find similar results as the ones presented in the main text, showing that our conclusions are not restricted to the particular DGP we present in this section.

We calculate the SC estimator using the following seven specifications that differ only in the linear combinations of pre-treatment outcome values used as predictors:[16]

1. All pre-treatment outcome values: $\mathbf{X}_j = [Y_{j,1} \cdots Y_{j,T_0}]'$
2. The first three-fourths of the pre-treatment outcome values: $\mathbf{X}_j = [Y_{j,1} \cdots Y_{j,\frac{3T_0}{4}}]'$
3. The first half of the pre-treatment outcome values: $\mathbf{X}_j = [Y_{j,1} \cdots Y_{j,\frac{T_0}{2}}]'$
4. Odd pre-treatment outcome values: $\mathbf{X}_j = [Y_{j,1}\ Y_{j,3} \cdots Y_{j,(T_0-3)}\ Y_{j,(T_0-1)}]'$
5. Even pre-treatment outcome values: $\mathbf{X}_j = [Y_{j,2}\ Y_{j,4} \cdots Y_{j,(T_0-2)}\ Y_{j,T_0}]'$
6. Pre-treatment outcome mean: $\mathbf{X}_j = [\frac{\sum_{t=1}^{T_0} Y_{j,t}}{T_0}]$
7. Three outcome values (the first one, the middle one, and the last one): $\mathbf{X}_j = [Y_{j,1}\ Y_{j,\frac{T_0}{2}}\ Y_{j,T_0}]'$

Observe that specifications 1 through 5 satisfy the conditions for the asymptotic equivalence results we present in the previous section, while specifications 6 and 7 do not. In order to simplify the presentation of our results, we do not consider in our MC simulations the use of time-invariant covariates, as is commonly used in specifications that rely on the pre-treatment outcome mean. In Appendix B, we show

[15] Appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's website and use the search engine to locate the article at http://onlinelibrary.wiley.com.

[16] In order to compute the SC estimator, we use the *Synth* package in *R*. (See Abadie, Diamond, & Hainmueller, 2011, for details.) This package solves the nested minimization problem described by equations (1) and (2). We specify the optimization method to be *BFGS* only and use optimization routine *Low Rank Quadratic Programming* when *Interior Point* optimization routine does not converge.

Table 2. Specification searching.

	Stationary model		Non-stationary model	
	5% test (1)	10% test (2)	5% test (3)	10% test (4)
Panel A: specifications 1 to 7				
$T_0 = 12$	0.143	0.250	0.142	0.254
	(0.003)	(0.004)	(0.004)	(0.004)
$T_0 = 32$	0.146	0.255	0.158	0.275
	(0.003)	(0.004)	(0.004)	(0.005)
$T_0 = 100$	0.143	0.254	0.152	0.264
	(0.003)	(0.004)	(0.004)	(0.004)
$T_0 = 400$	0.134	0.241	0.145	0.255
	(0.003)	(0.004)	(0.004)	(0.005)
Panel B: specifications 1 to 5				
$T_0 = 12$	0.106	0.19	0.110	0.198
	(0.003)	(0.004)	(0.003)	(0.004)
$T_0 = 32$	0.100	0.179	0.109	0.191
	(0.003)	(0.004)	(0.004)	(0.005)
$T_0 = 100$	0.090	0.157	0.094	0.162
	(0.003)	(0.004)	(0.003)	(0.004)
$T_0 = 400$	0.077	0.138	0.081	0.142
	(0.003)	(0.004)	(0.004)	(0.005)

Notes: Rejection rates are estimated based on 10,000 observations and on seven specifications: (1) all pre-treatment outcome values, (2) the first three-quarters of the pre-treatment outcome values, (3) the first half of the pre-treatment outcome values, (4) odd pre-treatment outcome values, (5) even pre-treatment outcome values, (6) the mean of all pre-treatment outcome values, and (7) three outcome values. z% test indicates that the nominal size of the analyzed test is z percent and T_0 is the number of pre-treatment periods.

that our results remain valid if we consider specifications that use time-invariant covariates as predictors in addition to functions of the pre-treatment outcomes.[17]

For each specification, we run a placebo test using the root mean squared prediction error (RMSPE) test statistic proposed in Abadie, Diamond, and Hainmueller (2010) and reject the null at 5 percent significance level if the treated unit has the largest RMSPE among the 20 units. We are interested in the probability that we would reject the null at the 5 percent significance level in at least one specification. This is the probability that a researcher would be able to report a significant result even when there is no effect if she were to engage in specification searching. If all different specifications result in the same synthetic control unit, then we would find that the probability of rejecting the null in at least one specification would be equal to 5 percent as well. However, this probability may be higher if the synthetic estimator depends on specification choices, which may be the case in finite samples or for specifications 6 and 7.

We present in columns 1 and 2 of Table 2, panel A, the probability of rejecting the null at 5 percent and at 10 percent significance levels in at least one of our seven specifications for the stationary model. Columns 3 and 4 present the same results for the non-stationary model.[18] With $T_0 = 12$, a researcher considering these seven

[17] Appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's website and use the search engine to locate the article at http://onlinelibrary.wiley.com.
[18] See Table G1 in Appendix G for results using different data-generating processes.

15206688, 2020, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/pam.22206 by Ecole Normale Supérieure de Paris, Wiley Online Library on [08/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

15206688, 2020, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/pam.22206 by Ecole Normale Supérieure de Paris, Wiley Online Library on [08/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

different specifications would be able to report a specification with statistically significant results at the 5 percent (10 percent) level with probability 14.3 percent (25.0 percent) for the stationary model and 14.2 percent (25.4 percent) for the non-stationary model.[19] Therefore, with few pre-treatment periods, a researcher would have substantial opportunities to select statistically significant specifications even when the null hypothesis is true. Importantly, Table 1 shows that SC applications with around 12 pre-treatment periods are common.

If the variation in the SC weights across different specifications vanishes when the number of pre-treatment periods goes to infinity, then we would expect the rejection rate to get closer to 5 percent once the number of pre-treatment periods gets large. In this case, all different specifications would provide roughly the same SC unit and, therefore, the same treatment effect estimate. The results in Table 2 show that the probabilities of rejecting the null are still significantly higher than the test size even when the number of pre-intervention periods is large. In a scenario with 400 pre-intervention periods in the non-stationary model, it would be possible to reject the null in at least one specification 14.5 percent (25.5 percent) of the time for a 5 percent (10 percent) significance test.[20]

These results suggest that, when we include specifications that violate the conditions for the asymptotic equivalence results from the previous section, specification searching remains a problem for the SC method, even when the number of pre-intervention periods is remarkably large for empirical applications. Therefore, we present in panel B of Table 2 the same results excluding specifications 6 and 7. As expected, based on our theoretical results presented in the previous section, excluding specifications 6 and 7 significantly attenuates the specification-searching problem, especially when the number of pre-treatment periods is large.[21] However, it does not completely solve the problem even when T_0 is relatively large in comparison to usual dataset sizes in SC applications. Given that our theoretical results suggest that specification-searching possibilities within a well-defined class of specifications should be very small asymptotically, this result suggests that asymptotic results may not provide reliable approximations in most SC applications.

The results in Table 2 are driven by the fact that the weights of specifications 1 through 5 converge to the same set of weights when $T_0 \to \infty$, while weights of specifications 6 and 7 may converge to different points according to the theoretical discussion presented in the previous section. Moreover, for the DGP we consider in our simulation exercise, we can evaluate the proportion of weights that are misallocated to control units that do not follow the same trends as the treated unit. The proportion of misallocated weights is much larger for specifications 6 and 7,

[19] As a robustness check, we take advantage of the fact that the DGP is known in our MC simulations, and we replicate our results using an infeasible test based on the actual distributions of the test statistics to determine whether the SC estimator for a given specification is statistically significant. The results based on this infeasible test, presented in Table G3 in Appendix G, corroborate the results above, showing that they are not driven by potential distortions of the placebo test used in the SC inference.

[20] Note that the probability of specification searching is not monotonic in T_0. This happens because, with a very small T_0, the chance that a pre-treatment MSPE is close to zero is very high. Since there is a high correlation of pre-treatment MSPE across specifications, it is likely that one unit will have a pre-treatment MSPE close to zero for many specifications. This implies that this unit will have a large test statistic for all these specifications, so the placebo test will reject the null for these specifications most of the time. As T_0 increases, the probability of having a pre-treatment MSPE close to zero will be small.

[21] The attenuation in the specification-searching problem after excluding specifications 6 and 7 is not simply because we are considering five specifications instead of seven. If we exclude, for example, specifications 2 and 3 instead of specifications 6 and 7, then there is virtually no change in the specification-search problem relative to the case that we consider all seven specifications (Table G2 in Appendix G).

and it does not decrease with T_0. In contrast, for specifications 1 to 5, the proportion of misallocated weights is much smaller and decreasing with the number of pre-treatment periods. We present these results in detail in Appendix C.[22]

Finally, one important feature of the SC method emphasized by Abadie, Diamond, and Hainmueller (2010, 2015) is that the method should only be used in situations with good pre-treatment fit. Therefore, if the specification-searching problem documented in Table 2 came from specifications with a particularly poor pre-treatment fit, then this phenomenon would not be a crucial problem for the method, as those specifications should not be chosen by applied researchers. However, in Appendix D, we show that the probability of rejecting the null in at least one SC specifications remains substantially higher than the significance level of the test even when we restrict to specifications that have a good fit. Therefore, our main conclusion—that there can be substantial opportunities for specification searching in the SC method because there are commonly used specifications that do not satisfy the conditions for the asymptotic equivalence results seen in the second previous section or T_0 is usually not large enough to provide reliable asymptotic approximations—remains valid even when we restrict to specifications with a good pre-treatment fit. As detailed in Appendix D, this phenomenon is explained by the impact of conditioning on a good pre-treatment fit on the number of "acceptable" specifications and on the denominator of the test statistic.[23] On the one hand, if conditioning on a good fit does not actually restrict the set of options a researcher has, then we have the same results as in the unconditional case. This is generally what happens when data is non-stationary. On the other hand, if conditioning severely restricts the set of options, then we have over-rejection because the test statistic for the treated unit is conditional on a denominator that is close to zero, while the test statistics for the placebo units are unconditional (Ferman & Pinto, 2017).

Recommendations

The specification-searching problem we identify arises from a lack of consensus about which specifications should be used in SC applications. If there are no covariates, the specification including all pre-treatment periods should be used. This specification is the one that minimizes the RMSE in the pre-treatment period, and it is not subject to arbitrary decisions regarding which pre-treatment outcome lags are included as predictors.

The only reason not to use all pre-treatment periods is when the researcher believes that the SC unit must also balance a specific set of covariates. In this case, the researcher would have to use a specification that does not include all pre-treatment lags, otherwise all covariates would be rendered irrelevant in the estimation of weights, as documented by Kaul et al. (2018). In those situations, we first recommend considering only specifications that satisfy the conditions given earlier in the second section. Both our theoretical and simulation results show that the specification-searching problem is attenuated by focusing only on the specifications with those properties. This is especially true when we have a large number of pre-treatment periods, even though it does not solve the problem completely when we consider T_0 in line with common SC applications.

Since there is more than one possible specification that satisfies the conditions above, we recommend presenting results for many different specifications. In par-

[22] Appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's website and use the search engine to locate the article at http://onlinelibrary.wiley.com.
[23] Appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's website and use the search engine to locate the article at http://onlinelibrary.wiley.com.

15206688, 2020, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/pam.22206 by Ecole Normale Supérieure de Paris, Wiley Online Library on [08/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

ticular, we recommend that specification 1 is always included as a benchmark. However, even if we present results for all possible SC specifications with a hypothesis test for each specification, this would not provide a valid hypothesis test. If the decision rule is to reject the null if the test rejects in all specifications, then we could end up with a very conservative test (Romano & Wolf, 2015).[24] If the decision rule is to reject the null if the test rejects in at least one specification, then we would be back in the situation where we over-reject the null.

One possible solution is to base the inference procedure on a new test statistic that is a function that combines all the test statistics for the individual specifications (Imbens & Rubin, 2015). The drawback of this solution is that it does not provide an obvious point-estimator. There are two possible ways to handle this disadvantage. First, if the test function is simply a weighted average of the test statistics for individual specifications, then Christensen and Miguel (2008) and Cohen-Cole et al. (2009) suggest using the same weights to compute a weighted average of the point-estimator of each specification, generating an estimate that incorporates model uncertainty. As another alternative, we can focus on set identification, as suggested by Firpo and Possebom (2018). In this case, we would invert this combination of test statistics to compute a confidence set that contains all treatment effect functions within a pre-specified class that is not rejected by the inference procedure.

Another possibility is to consider a criterion for choosing among all possible specifications. If one restricts attention to only one specification that is chosen based on an objective criterion, without the need of subjective decisions by the researcher, then the possibility for specification searching would be limited, at least in this dimension. For example, Donohue, Aneja, and Weber (2018) report that they considered different specifications, and eventually chose the one that minimized the mean squared prediction error (MSPE) during the validation period. While this is a reasonable and interesting idea, it potentially allows for specification searching in other dimensions, such as the decision on how to split the pre-treatment periods into training and validation periods. Dube and Zipperer (2015) provide a similar idea but they consider the specification that minimizes the MSPE in the post-intervention periods for the placebo estimates.

Empirical Application

Example 1: German Reunification (Abadie, Diamond, & Hainmueller, 2015)

Abadie, Diamond, and Hainmueller (2015) evaluate the impact of the German Reunification in 1991 on GDP per capita.[25] The pre- and post-treatment periods are 1960 through 1990 and 1991 through 2003, respectively, with a training period of 1971 through 1980 and a validating period of 1981 through 1990. The donor pool consists of 16 Organisation for Economic Co-operation and Development (OECD) countries.

We reestimate the impact of the German reunification on GDP per capita using the synthetic control method with 14 different specifications. Specifically, we test the same seven specifications from the third section of the paper and, for each one

[24] When we adopt this decision rule in our MC simulations, then the probability of rejecting the null at 5 percent for all specifications is lower than 1 percent in all scenarios. If we discard specifications 6 and 7, then this rejection rate ranges from 1 percent when $T_0 = 12$ to 2.8 percent when $T_0 = 400$.

[25] Following the best practices in terms of transparency and replicability, Hainmueller (2014) made their dataset and replication files available online.

15206688, 2020, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/pam.22206 by Ecole Normale Supérieure de Paris, Wiley Online Library on [08/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 3. Specification searching—database from Abadie et al. (2015).

Specification	(1a)	(1b)	(2a)	(2b)	(3a)	(3b)	(4a)	(4b)
p-value	0.059	0.059	0.059	0.118	0.118	0.059	0.059	0.059
Specification	(5a)	(5b)	(6a)	(6b)	(7a)	(7b)		
p-value	0.118	0.059	0.588	0.059	0.353	0.059		

Notes: We analyze 14 different specifications. The number of the specifications refers to: (1) all pre-treatment outcome values, (2) the first three-fourths of the pre-treatment outcome values, (3) the first half of the pre-treatment outcome values, (4) odd pre-treatment outcome values, (5) even pre-treatment outcome values, (6) pre-treatment outcome mean (original specification by Abadie, Diamond, & Hainmueller, 2010), and (7) three outcome values. Specifications that end with an *a* do not include covariates, while specifications that end with a *b* include the covariates trade openness, inflation rate, industry share, schooling levels, and investment rate.

of them, we either include five covariates or not.[26, 27] Specifications ending with *a* do not include covariates, while those ending with *b* include them. Specification 6b is the original one in Abadie, Diamond, and Hainmueller (2015).

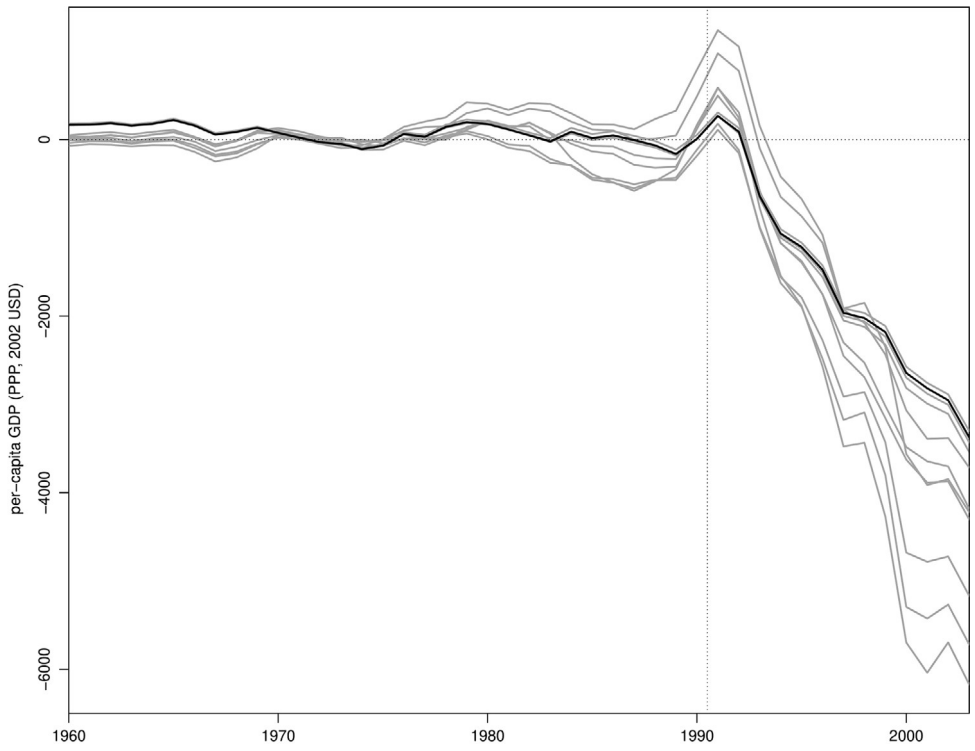
Table 3 shows the p-value for each specification. The results show that the researcher could try different specifications and pick one whose result is significant.[28] In particular, only nine of them are significant at the 10 percent significance level, while four of them are not, implying that different specifications could lead to different conclusions.

If we believe that covariates are not relevant to explain the German GDP per capita, the recommended specification uses all pre-treatment outcome lags. Note that specification 1a indicates that the treated unit has the largest RMSPE, suggesting that our treatment has a statistically significant effect.

However, if we believe that the SC unit should also match the covariates, then we should focus only on the specifications that satisfy the conditions outlined in the second section by dropping specifications 6 and 7. By looking at Table 3, we note that the significance of the treatment effect is not straightforward. By looking at Figure 1, we find that specifications 1 through 5 point to a treatment effect that is negative in the long run. However, the magnitude of this effect varies across specifications. The next step is to test the null hypothesis using a test statistic that combines the test statistics of specifications 1 through 5. We find that the p-value of a test that uses the mean of the RMSPE statistic across specifications (Imbens & Rubin, 2015), is equal to 0.059, suggesting that the German Reunification had a statistically significant impact on West Germany's per-capita GDP. In order to present point-estimates associated with this test, we follow Christensen and Miguel (2018) and Cohen-Cole et al. (2009) and show, in Figure 2, the average treatment effects across specifications 1 through 5 as the black line. This average treatment effect suggests a strongly negative effect in the long run. We also follow Firpo and Possebom (2018) to compute a confidence set (Figure 2) that includes all treatment effect functions that we fail to reject using this combined test statistic, considering functions that are deviations from the average treatment effect across specifications

[26] We follow Abadie, Diamond, and Hainmueller (2015) and consider for this exercise different specifications using only the training period in the first minimization problem (equation 1) and the validating period in the second minimization problem (equation 2).
[27] The included covariates are trade openness, inflation rate, industry share, schooling levels, and investment rate.
[28] All 14 specifications have a good pre-treatment fit according to the measure proposed in Appendix D. Appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's website and use the search engine to locate the article at http://onlinelibrary.wiley.com.

15206688, 2020, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/pam.22206 by Ecole Normale Supérieure de Paris, Wiley Online Library on [08/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License



Notes: The solid black line is the original specification by Abadie, Diamond, and Hainmueller (2015) and gray lines are specifications 1 through 5. The vertical line denotes the beginning of the post-treatment period.

Figure 1. Treatment Effects for Specifications 1 through 5 and the Original Specification—Database from Abadie, Diamond, and Hainmueller (2015).

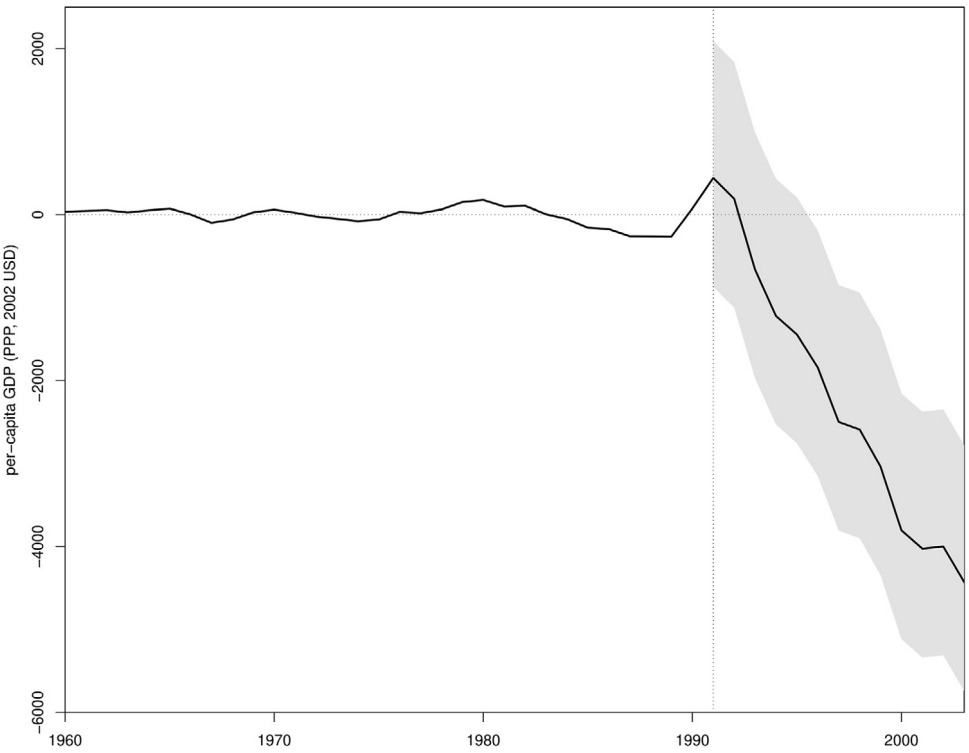
by an additive and constant factor. We find that, although we cannot reject treatment effect functions that are initially positive, all treatment effect functions in our confidence set are negative in the long run. Finally, we apply the choice criteria suggested by Dube and Zipperer (2015) and Donohue et al. (2018), restricting ourselves to specifications 1 through 5. The first criterion picks specification 1a (in this case, we would reject the null with a p-value of 0.059), while the second one picks specification 2b (in this case, we would marginally reject the null, with a p-value of 0.118).

After this analysis, a reasonable conclusion would be that there is a significant and negative treatment effect in the long run.

Example 2: Paid Family Leave (Bartel et al., 2018)

Bartel et al. (2018) evaluate the impact of the California's Paid Family Leave (CA-PFL) program on fathers' leave-taking. The pre- and post-treatment periods are 2000 through 2004 and 2005 through 2013 using data from the American Community Survey (ACS). The donor pool consists of the District of Columbia and all American states, excluding New Jersey, because it also implemented a similar program in 2008.

15206688, 2020, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/pam.22206 by Ecole Normale Supérieure de Paris, Wiley Online Library on [08/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License



Notes: We compute confidence sets by inverting the average test statistic across specifications. Our confidence sets include all treatment effect functions that we fail to reject using this combined test statistic, considering functions that are deviations from the average treatment effect across specifications by an additive and constant factor. The black line is the average treatment effect of West Germany and the gray area is the confidence set. The vertical lines denote the beginning of the post-treatment period.

Figure 2. Ninety Percent Confidence Sets Around the Average Across Specifications 1 through 5—Database from Abadie, Diamond, and Hainmueller (2015).

We reestimate the impact of the CA-PFL program on fathers' leave-taking using the synthetic control method with 14 specifications. Specifically, we test the same seven specifications from the third section of the paper and, for each one of them, we either include 11 covariates or not.[29] Specifications ending with *a* do not include covariates, while those ending with *b* include them. Similarly to our recommendations, Bartel et al. (2018) analyze and report results for many different specifications: our specifications 1b and 6b are their specifications 7 and 6, respectively (Bartel et al., 2018, Table 6).

Table 4 shows the p-value for the specifications with a good pre-treatment fit.[30] The results show that the researcher could try different specifications and pick one whose result is significant: specifications 1a, 3b, 5b, 6b, and 7b are significant at the

[29] The included covariates are related to racial composition, educational attainment, employment, and labor force participation.

[30] A good pre-treatment fit is defined according to the measure proposed in Appendix D. Appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's website and use the search engine to locate the article at http://onlinelibrary.wiley.com.

15206688, 2020, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/pam.22206 by Ecole Normale Supérieure de Paris, Wiley Online Library on [08/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

15206688, 2020, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/pam.22206 by Ecole Normale Supérieure de Paris, Wiley Online Library on [08/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 4. Specification searching—database from Bartel et al. (2018).

Specification	(1a)	(1b)	(2b)	(3b)	(4b)	(5b)	(6b)	(7b)
p-value	0.02	0.12	0.06	0.02	0.125	0.04	0.021	0.021

Notes: We analyze 14 different specifications and only report the ones with good pre-treatment fit according to the measure proposed in Appendix D. The number of the specifications refers to: (1) all pre-treatment outcome values (specification 7 by Bartel et al., 2018), (2) the first three-fourths of the pre-treatment outcome values, (3) the first half of the pre-treatment outcome values, (4) odd pre-treatment outcome values, (5) even pre-treatment outcome values, (6) pre-treatment outcome mean (specification 6 by Bartel et al., 2018), and (7) three outcome values. Specifications that end with an *a* do not include covariates, while specifications that end with a *b* include the covariates related to racial composition, educational attainment, employment, and labor force participation.

5 percent level; specification 2b is significant at the 10 percent level; and specifications 1b and 4b are not significant. As a consequence, different specifications could lead to different conclusions.

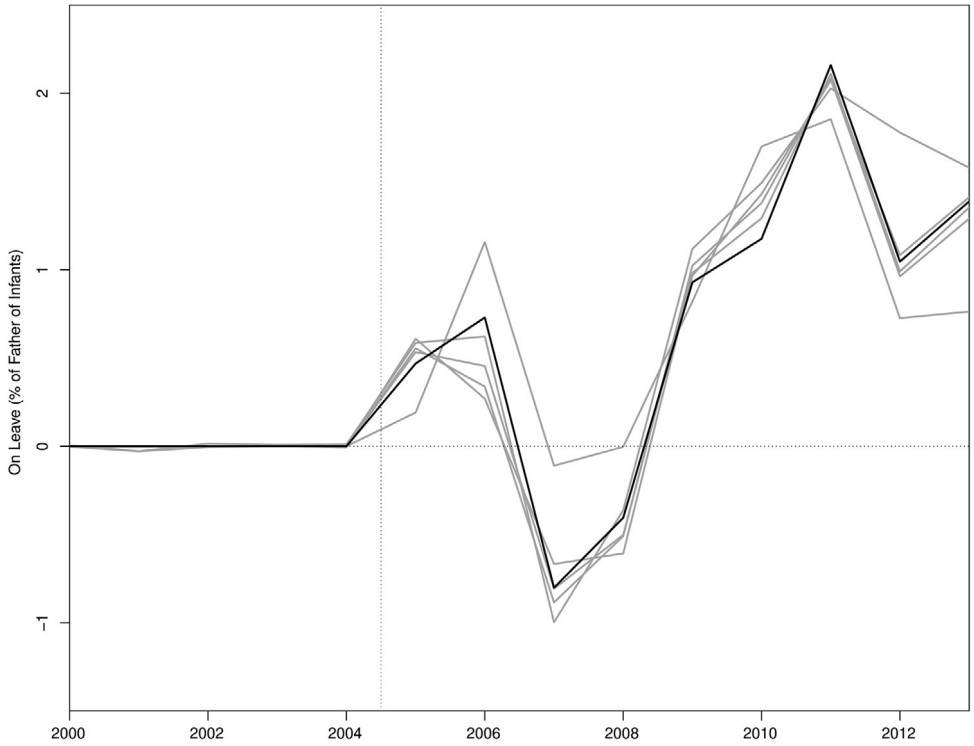
If we believe that covariates are not relevant to explain fathers' leave-taking, the recommended specification uses all pre-treatment outcome lags. Note that specification 1a indicates that the treated unit has the largest RMSPE, suggesting that our treatment has a statistically significant effect.

However, if we believe that the SC unit should also directly match the covariates, then we should focus only on the specifications that satisfy the conditions outlined in the second section by dropping specifications 6 and 7. By looking at Table 4, we note that the significance of the treatment effect is not straightforward.[31] By looking at Figure 3, we find that specifications 1 through 5 point to a treatment effect of similar magnitude and positive in the long run. The next step is to test the null hypothesis using a test statistic that combines the test statistics of specifications 1 through 5. We find that the p-value of a test that uses the mean of the RMSPE statistic across specifications (Imbens & Rubin, 2015) is equal to 0.021, suggesting that the CA-PFL program had an impact on fathers' leave-taking behavior. In order to present point-estimates associated with this test, we follow Christensen and Miguel (2018) and Cohen-Cole et al. (2009), and show, in Figure 4, the average treatment effects across specifications 1 through 5 as a black line, suggesting a positive effect in the long run. We also follow Firpo and Possebom (2018) to compute the confidence set (Figure 4) that includes all treatment effect functions that we fail to reject using this combined test statistic, considering functions that are deviations from the average treatment effect across specifications by an additive and constant factor. We find that, although we cannot reject treatment effect functions that are initially negative, all treatment effect functions in our confidence set are positive in the long run. Finally, we apply the choice criterion suggested by Dube and Zipperer (2015), restricting ourselves to specifications 1 through 5. The choice criterion picks specification 5b (in this case, we would reject the null with a p-value of 0.040).

After this analysis, a reasonable conclusion would be that there is a significant and positive treatment effect in the long run.

In Appendix F, we consider other empirical applications. In particular, we present an empirical application based on Smith (2015) in which we can find a few "statistically significant" specifications although most specifications show insignificant

[31] Note that the p-values of specifications 1a and 1b are different in Table 4. Although Kaul et al. (2018) show that the same weights solve the minimization problem for these specifications, the solution may not be unique when the number of control units is larger than the number of pre-treatment periods, as is the case in this empirical example. As a consequence, the command *synth* in *R* picks different solutions for specifications 1a and 1b.



Notes: The solid black line is the specification 7 by Bartel et al. (2018); gray lines are the other specifications in Table 4 that satisfy the conditions outlined in the second section. The vertical line denotes the beginning of the post-treatment period.

Figure 3. Treatment Effects for Specifications 1 through 5—Database from Bartel et al. (2018).

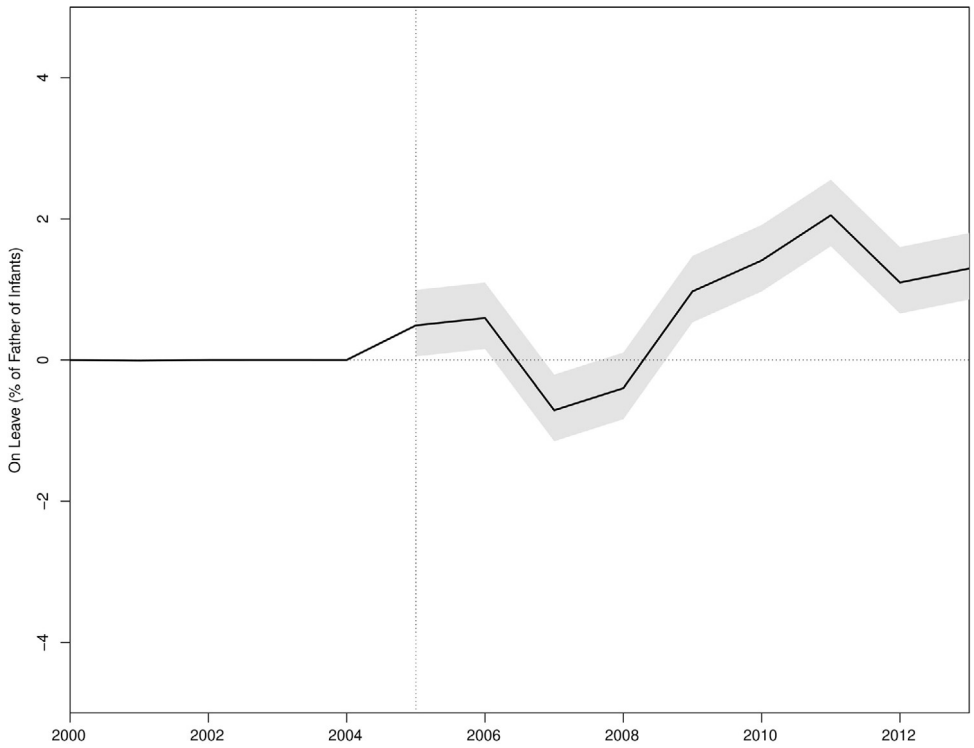
effects, illustrating the potential for specification searching in SC applications.[32] Following our recommendations, we provide clear evidence that the effects are not significant in this application.[33]

CONCLUSION

We analyze whether a lack of guidance on how to choose among different SC specifications creates the potential for specification searching with synthetic controls. We first provide theoretical results showing that the possibility for specification searching becomes asymptotically irrelevant if the number of pre-treatment outcome lags used as predictors goes to infinity when the number of pre-treatment periods goes to infinity. However, guided by our theoretical results, we provide evidence from simulations that specification searching may be a relevant problem in real SC applications for at least two reasons. First, many SC applications do not have a large number of pre-treatment periods to guarantee that our asymptotic re-

[32] Appendices are available at the end of this article as it appears in JPAM online. Go to the publisher's website and use the search engine to locate the article at http://onlinelibrary.wiley.com.
[33] Smith (2015) did *not* choose one of these "statistically significant" specifications.

15206688, 2020, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/pam.22206 by Ecole Normale Supérieure de Paris, Wiley Online Library on [08/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License



Notes: We compute confidence sets by inverting the average test statistic across specifications. Our confidence set includes all treatment effect functions that we fail to reject using this combined test statistic, considering functions that are deviations from the average treatment effect across specifications by an additive and constant factor. The black line is the average treatment effect of CA-PFL and the gray area is the confidence set. The vertical lines denote the beginning of the post-treatment period.

Figure 4. Ninety Percent Confidence Sets Around the Average Across Specifications 1 through 5—Database from Bartel et al. (2018).

sults are approximately valid. Second, many SC applications rely on specifications that do not satisfy the conditions in our theoretical results. We provide a series of recommendations to limit the scope for specification searching in SC applications.

BRUNO FERMAN is an Associate Professor of Economics at the Sao Paulo School of Economics - FGV, Rua Itapeva 474, São Paulo, Brazil 01332-000 (e-mail: bruno.ferman@fgv.br).

CRISTINE PINTO is an Associate Professor of Economics at the Sao Paulo School of Economics -FGV, Rua Itapeva, 474, 12o andar, São Paulo, Brazil CEP 01332-000 (e-mail: cristine.pinto@fgv.br).

VITOR POSSEBOM is a PhD Candidate in the Department of Economics at Yale University, New Haven, CT 06511 (e-mail: vitoraugusto.possebom@yale.edu).

15206688, 2020, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/pam.22206 by Ecole Normale Supérieure de Paris, Wiley Online Library on [08/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

15206688, 2020, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/pam.22206 by Ecole Normale Supérieure de Paris, Wiley Online Library on [08/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

ACKNOWLEDGMENTS

We would like to thank Juan Camilo Castillo, Sergio Firpo, Ricardo Masini, Masayuki Sawada, and participants at the Sao School of Economics seminar, Yale Econometrics Lunch, African Meeting of the Econometric Society, the 2016 Meeting of the Brazilian Econometric Society, and the Young Economists Symposium 2018 for the excellent comments and suggestions. Deivis Angeli and Murilo S. Cardoso provided outstanding research assistance. Bruno Ferman gratefully acknowledges financial support from CNPq.

REFERENCES

Abadie, A., Diamond, A., & Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. Journal of the American Statistical Association, 105, 493–505.

Abadie, A., Diamond, A., & Hainmueller, J. (2011). Synth: An R package for synthetic control methods in comparative case studies. Journal of Statistical Software, 42, 1–17.

Abadie, A., Diamond, A., & Hainmueller, J. (2015). Comparative politics and the synthetic control method. American Journal of Political Science, 59, 495–510.

Abadie, A., & Gardeazabal, J. (2003). The economic costs of conflict: A case study of the Basque country. American Economic Review, 93, 113–132.

Acemoglu, D., Johnson, S., Kermani, A., Kwak, J., & Mitton, T. (2016). The value of connections in turbulent times: Evidence from the United States. Journal of Financial Economics, 121, 368–391.

Athey, S., & Imbens, G. W. (2017). The state of applied econometrics: Causality and policy evaluation. Journal of Economic Perspectives, 31, 3–32.

Baccini, L., Li, Q., & Mirkina, I. (2014). Corporate tax cuts and foreign direct investment. Journal of Policy Analysis and Management, 33, 977–1006.

Barone, G., & Mocetti, S. (2014). Natural disasters, growth and institutions: A tale of two earthquakes. Journal of Urban Economics, 84, 52–66.

Bartel, A. P., Rossin-Slater, M., Ruhm, C. J., Stearns, J., & Waldfogel, J. (2018). Paid family leave, fathers' leave-taking, and leave-sharing in dual-earner households. Journal of Policy Analysis and Management, 37, 10–37.

Belot, M., & Vandenberghe, V. (2014). Evaluating the threat effects of grade repetition: Exploiting the 2001 reform by the French-speaking community of Belgium. Education Economics, 22, 73–89.

Billmeier, A., & Nannicini, T. (2011). Trade openness and growth: Pursuing empirical glasnost. Oxford Bulletin of Economics and Statistics, 73, 287–314.

Billmeier, A., & Nannicini, T. (2013). Assessing economic liberalization episodes: A synthetic control approach. The Review of Economics and Statistics, 95, 983–1001.

Bohn, S., Lofstrom, M., & Raphael, S. (2014). Did the 2007 Legal Arizona Workers Act reduce the state's unauthorized immigrant population? The Review of Economics and Statistics, 96, 258–269.

Brodeur, A., Le, M., Sangnier, M., & Zylberberg, Y. (2016). Star wars: The empirics strike back. American Economic Journal: Applied Economics, 8, 1–32.

Cavallo, E., Galiani, S., Noy, I., & Pantano, J. (2013). Catastrophic natural disasters and economic growth. The Review of Economics and Statistics, 95, 1549–1561.

Christensen, G., & Miguel, E. (2018). Transparency, reproducibility, and the credibility of economics research. Journal of Economic Literature, 56, 920–980.

Coffman, L. C., & Niederle, M. (2015). Pre-analysis plans have limited upside, especially where replications are feasible. Journal of Economic Perspectives, 29, 81–98.

Cohen-Cole, E., Durlauf, S., Fagan, J., & Nagin, D. (2009). Model uncertainty and the deterrent effect of capital punishment. American Law and Economics Review, 11, 335–369.

15206688, 2020, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/pam.22206 by Ecole Normale Supérieure de Paris, Wiley Online Library on [08/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Cunningham, S., & Shah, M. (2018). Decriminalizing indoor prostitution: Implications for sexual violence and public health. The Review of Economic Studies, 85, 1683–1715.

DeAngelo, G., & Hansen, B. (2014). Life and death in the fast lane: Police enforcement and traffic fatalities. American Economic Journal: Economic Policy, 6, 231–257.

Donohue, J. J., Aneja, A., & Weber, K. D. (2018). Right-to-Carry laws and violent crime: A comprehensive assessment using panel data, the LASSO, and a state-level synthetic controls analysis. NBER Working Paper No. 23510. Cambridge, MA: National Bureau of Economic Research.

Doudchenko, N., & Imbens, G. (2016). Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. NBER Working Paper No. 22791. Cambridge, MA: National Bureau of Economic Research.

Dube, A., & Zipperer, B. (2015). Pooling multiple case studies using synthetic controls: An application to minimum wage policies. IZA Discussion Papers 8944. Bonn, Germany: Institute of Labor Economics.

Dustmann, C., Schonberg, U., & Stuhler, J. (2017). Labor supply shocks, native wages, and the adjustment of local employment. The Quarterly Journal of Economics, 132, 435–483.

Eren, O., & Ozbeklik, S. (2016). What do Right-to-Work laws do? Evidence from a synthetic control method analysis. Journal of Policy Analysis and Management, 35, 173–194.

Fafchamps, M., & Labonne, J. (2017). Using split samples to improve inference on causal effects. Political Analysis, 25, 465–482.

Ferman, B., & Pinto, C. (2017). Placebo tests for synthetic controls. MPRA Paper 78079. Germany: University Library of Munich.

Ferman, B., Pinto, C., & Possebom, V. (2020) Replication data for: Cherry picking with synthetic controls. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor]. Retrieved January 24, 2020, from https://doi.org/10.3886/E117261V2.

Firpo, S., & Possebom, V. (2018). Synthetic control method: Inference, sensitivity analysis and confidence sets. Journal of Causal Inference, 6, 1–26.

Gardeazabal, J., & Vega-Bayo, A. (2016). An empirical comparison between the synthetic control method and Hsiao et al.'s panel data approach to program evaluation. Journal of Applied Econometrics, 32, 983–1002.

Gobillon, L., & Magnac, T. (2016). Regional policy evaluation: Interactive fixed effects and synthetic controls. Review of Economics and Statistics, 98, 535–551.

Hahn, J., & Shi, R. (2017). Synthetic control and inference. Econometrics, 5.

Hainmueller, J. (2014), Replication data for: Comparative politics and the Synthetic Control Method. Retrieved January 24, 2020, from https://doi.org/10.7910/DVN/24714. Harvard Dataverse.

Hinrichs, P. (2012). The effects of affirmative action bans on college enrollment, educational attainment, and the demographic composition of universities. Review of Economics and Statistics, 94, 712–722.

Imbens, G. W., & Rubin, D. B. (2015). Causal inference for statistics, social and biomedical sciences: An introduction. Cambridge, UK: Cambridge University Press.

Kaul, A., Klößner, S., Pfeifer, G., & Schieler, M. (2018). Synthetic control methods: Never use all pre-intervention outcomes together with covariates. Working Paper. Retrieved January 24, 2020, from https://goo.gl/KztyPq.

Kleven, H. J., Landais, C., & Saez, E. (2013). Taxation and international migration of superstars: Evidence from European football market. American Economic Review, 103, 1892–1924.

Klößner, S., Kaul, A., Pfeifer, G., & Schieler, M. (2017). Comparative politics and the synthetic control method revisited: A note on Abadie et al. (2015). Swiss Journal of Economics and Statistics, 154.

Lindo, J. M., & Packham, A. (2017). How much can expanding access to long-acting reversible contraceptives reduce teen birth rates? American Economic Journal: Economic Policy, 9, 348–376.

Lovell, M. (1983). Data mining. The Review of Economics and Statistics, 65, 1–12.

Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K. M., Gerber, A., ... Van der Laan, M. (2014). Promoting transparency in social science research. Science, 343, 30–31.

Montalvo, J. G. (2011). Voting after the bombings: A natural experiment on the effect of terrorist attacks on democratic elections. Review of Economics and Statistics, 93, 1146–1154.

Olken, B. A. (2015). Promises and perils of pre-analysis plans. Journal of Economic Perspectives, 29, 61–80.

Pinotti, P. (2013). Organized crime, violence, and the quality of politicians: Evidence from Southern Italy. In P. J. Cook, S. Machin, O. Marie, & G. Mastrobuoni (Eds.), Lessons from the economics of crime: What reduces offending? Cambridge, MA: The MIT Press.

Romano, J. P., & Wolf, M. (2005). Stepwise multiple testing as formalized data snooping. Econometrica, 73, 1237–1282.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. Journal of Experimental Psychology: General, 143, 534–547.

Smith, B. (2015). The resource curse exorcised: Evidence from a panel of countries. Journal of Development Economics, 116, 57–73.

Zou, B. (2018). The local economic impacts of military personnel. Journal of Labor Economics, 36, 589–621.

15206688, 2020, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/pam.22206 by Ecole Normale Supérieure de Paris, Wiley Online Library on [08/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

15206688, 2020, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/pam.22206 by Ecole Normale Supérieure de Paris, Wiley Online Library on [08/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

APPENDIX A: THEORETICAL RESULTS

Main Theoretical Results

Here, we formalize the theoretical results presented in the second section in the main paper. We consider a sufficient assumption to guarantee that a broad set of SC specifications will be asymptotically equivalent when $T_0 \to \infty$.

Assumption 1. *For any sequence of integers* $\{t_k\}_{k\in\mathbb{N}}$ *with* $t_k > t_{k-1}$, *and for any* $j \in \{1, \ldots, J+1\}$, *we have that*

$$\sup_{\mathbf{W}\in\mathcal{W}} \left| \frac{1}{K} \sum_{k=1}^{K} \left(Y^0_{j,t_k} - \mathbf{y}'^0_{-j,t_k} \mathbf{W} \right)^2 - Q_j(\mathbf{W}) \right| \underset{p}{\to} 0 \text{ when } K \to \infty,$$

where $Q_j(\mathbf{W})$ *is a continuous and strictly convex function.*

Assumption 1 implies that pre-treatment averages of the second moments of every subsequence of $(Y^0_{1,t}, \ldots, Y^0_{J+1,t})$ converge to the same value. We show below that this assumption is satisfied if, for example, we assume that $\{\mathbf{y}^0_t \mathbf{y}'^0_t\}_{t=1}^{T_0}$ has weak stationarity, each element of $\{\mathbf{y}^0_t \mathbf{y}'^0_t\}$ has absolutely summable covariances, and $\mathbb{E}[\mathbf{y}^0_t \mathbf{y}'^0_t]$ is non-singular, where $\mathbf{y}^0_t = (Y^0_{1,t}, \ldots, Y^0_{J+1,t})'$.

Define $\hat{\alpha}_{1t}(s, T_0)$ as the estimated gap when specification s is used, and consider the definitions of $\hat{\mathbf{W}}(s, T_0)$ and $L(s, T_0)$ given in the second section of the main paper. Then, we have the following results (see Proposition 2 below).

Proposition 2. *Let* $\hat{\mathbf{W}}(s, T_0)$ *be the SC weights using specification s when there are* T_0 *pre-intervention periods. If* $L(s, T_0) \to \infty$ *when* $T_0 \to \infty$, *then, under Assumption* 1, $\hat{\mathbf{W}}(s, T_0) \to_p \bar{\mathbf{W}} = argmin_{\mathbf{W}\in\mathcal{W}} Q_1(\mathbf{W})$. (*See details below for Proof of Proposition 2.*)

Corollary 3. *Let* $\hat{\alpha}_{1t}(s, T_0)$ *and* $\hat{\alpha}_{1t}(s', T_0)$ *be two SC estimators for the treatment effect at time* $t > T_0$ *using specifications s and s′ such that* $L(s, T_0) \to \infty$ *and* $L(s', T_0) \to \infty$ *when* $T_0 \to \infty$. *Then, under Assumption* 1, $|\hat{\alpha}_{1t}(s, T_0) - \hat{\alpha}_{1t}(s', T_0)| = o_p(1)$. (*See details below for Proof of Corollary 3.*)

Therefore, while different SC specifications may generate different SC estimates, our result from Proposition 2 and Corollary 3 show that, under some conditions, different specifications will lead to asymptotically equivalent SC estimators, as long as the number of pre-treatment lags used as predictors goes to infinity with T_0.

Our results are valid irrespective of whether the SC estimator is unbiased, as we are only comparing the asymptotic behavior of the SC estimator under different specifications. For a thorough analysis on the asymptotic bias of the SC estimator when $T_0 \to \infty$, see Ferman and Pinto (2019). In our Monte Carlo simulations in the third section of the paper and in Appendix E, the conditions in which the SC estimator is unbiased are satisfied. Also, our results are related to the results from Kaul et al. (2018), who show that covariates would become irrelevant in the minimization problem (1) if all pre-treatment outcome lags are included as predictors. Since our theoretical results hold whether or not other covariates are included as predictors, this implies that covariates would also become asymptotically irrelevant in the minimization problem (1) whenever we consider specifications such that $L(s, T_0) \to \infty$ when $T_0 \to \infty$, even if we do not include all pre-treatment outcome lags. This, however, does not necessarily imply that the SC weights will not attempt to match the covariates of the treated unit, nor that the SC estimator will be asymptotically biased, as explained by Botosaru and Ferman (2019).

15206688, 2020, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/pam.22206 by Ecole Normale Supérieure de Paris, Wiley Online Library on [08/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

As a corollary from both results, we show that the ranking of $RMSPE_j$ (defined in equation 3) will remain asymptotically invariant to changes in the SC specification when $T_0 \to \infty$ whenever we consider only specifications whose number of pre-treatment outcome lags goes to infinity with T_0.

Corollary 4. *Under Assumption 1 and assuming that Y_{jt} is continuous, with probability approaching one when $T_0 \to \infty$ and $T - T_0$ is fixed, the ordering of $\{RMSPE_1, \ldots, RMSPE_{J+1}\}$ is invariant to SC specifications such that $L(s, T_0) \to \infty$ when $T_0 \to \infty$. (See details below for Proof of Corollary 4.)*

As a consequence of Corollary 4, the test decision in the placebo test is asymptotically invariant to the specification choice when $T_0 \to \infty$, provided that we restrain to SC specifications whose number of pre-treatment outcome lags goes to infinity with T_0.

Proof of Proposition 2. Let $\tilde{\mathcal{W}} = \{\hat{\mathbf{W}} \in \mathcal{W} | \hat{\mathbf{W}} \in \text{argmin}_{\mathbf{W} \in \mathcal{W}} (\mathbf{X}_1 - \mathbf{X}_0\mathbf{W})'\mathbf{V}(\mathbf{X}_1 - \mathbf{X}_0\mathbf{W})$ for some $\mathbf{V} \in \mathcal{V}\}$, and $\hat{Q}_{T_0}(\mathbf{W}) = \frac{1}{T_0}(\mathbf{Y}_1 - \mathbf{Y}_0\mathbf{W})'(\mathbf{Y}_1 - \mathbf{Y}_0\mathbf{W})$. Also, let $\hat{f}^s_{T_0}(\mathbf{W}, \mathbf{V}) = (\mathbf{X}_1 - \mathbf{X}_0\mathbf{W})'\mathbf{V}(\mathbf{X}_1 - \mathbf{X}_0\mathbf{W})$, where $\mathbf{X_j}$ includes the predictors used in specification s when there are T_0 pre-treatment periods.

The SC weights computed from the nested optimization problem proposed in Abadie, Diamond, and Hainmueller (2010) can be defined by:

$$\hat{\mathbf{W}}(s, T_0) = \arg\min_{\mathbf{W} \in \tilde{\mathcal{W}}} \hat{Q}_{T_0}(\mathbf{W}).$$

We want to show that $\hat{\mathbf{W}}(s, T_0) \to_p \bar{\mathbf{W}}$. First, let $\mathbf{V}^*(s, T_0)$ be a diagonal matrix with diagonal entries equal to one for pre-treatment outcome lags and zero for other predictors when we consider the predictors used in specification s with T_0 pre-treatment periods. Then we have that $\frac{1}{L(s,T_0)} \hat{f}^s_{T_0}(\mathbf{W}, \mathbf{V}^*(s, T_0)) = \frac{1}{L(s,T_0)} \sum_{t \in \mathcal{I}(s,T_0)} (Y^0_{1,t} - \mathbf{y}'^0_{-1,t}\mathbf{W})^2$. By Assumption 1 and by the fact that $L(s, T_0) \to \infty$ when $T_0 \to \infty$, $\frac{1}{L(s,T_0)} \hat{f}^s_{T_0}(\mathbf{W}, \mathbf{V}^*(s, T_0))$ converges uniformly in probability to $Q_1(\mathbf{W})$, which is uniquely minimized at $\bar{\mathbf{W}}$. Let $\hat{\mathbf{W}}(s, \mathbf{V}^*(s, T_0), T_0) = \text{argmin}_{\mathbf{W} \in \mathcal{W}} \frac{1}{L(s,T_0)} \hat{f}^s_{T_0}(\mathbf{W}, \mathbf{V}^*(s, T_0))$. Since $\mathcal{W}$ is compact, we have that $\hat{\mathbf{W}}(s, \mathbf{V}^*(s, T_0), T_0) \to_p \bar{\mathbf{W}}$ when $T_0 \to \infty$ (Theorem 2.1 of Newey & McFadden, 1994).

We now show that the solution to the nested problem proposed in Abadie, Diamond, and Hainmueller (2010) will also converge in probability to $\bar{\mathbf{W}}$. First, note that $\hat{\mathbf{W}}(s, T_0)$ always exists. According to Berge's Maximum Theorem (Ok, 2007, p. 306), $\hat{\mathbf{W}}(\mathrm{V})$ is a compact-value, upper hemicontinuous and closed correspondence. As a consequence, $\tilde{\mathcal{W}}$ is a compact set. To see that, take any sequence $\{\tilde{\mathrm{W}}_n\}_{n \in \mathbb{N}}$ such that $\tilde{\mathrm{W}}_n \in \tilde{\mathcal{W}}$ for any $n \in \mathbb{N}$. Since $\tilde{\mathcal{W}} = \cup_{V \in \mathcal{V}} \hat{\mathbf{W}}(\mathrm{V})$ by its definition, there exists $\mathrm{V}_n \in \mathcal{V}$ for each $n \in \mathbb{N}$ such that $\tilde{\mathrm{W}}_n \in \hat{\mathbf{W}}(\mathrm{V}_n)$. We also know that there exists a convergent subsequence $\{\mathrm{V}_{n_m}\}_{m \in \mathbb{N}}$ such that $\lim_{m \to +\infty} \mathrm{V}_{n_m} =: \overline{\overline{\mathrm{V}}} \in \mathcal{V}$ because $\mathcal{V}$ is a compact set. By the definition of upper hemicontinuity (Stokey & Lucas, 1989, p. 56), there exists a convergent subsequence $\{\tilde{\mathrm{W}}_{n_{m_l}}\}_{l \in \mathbb{N}}$ such that $\lim_{l \to +\infty} \tilde{\mathrm{W}}_{n_{m_l}} =: \overline{\overline{\mathrm{W}}} \in \hat{\mathrm{W}}(\overline{\overline{\mathrm{V}}}) \subset \cup_{V \in \mathcal{V}} \hat{\boldsymbol{W}}(\mathrm{V}) = \tilde{\mathcal{W}}$, proving that $\tilde{\mathcal{W}}$ is a compact set. Consequently, Weierstrass' Extreme Value Theorem guarantees that $\hat{\mathbf{W}}(s, T_0)$ exists.

From Assumption 1, we have that $\hat{Q}_{T_0}(\mathbf{W})$ converges uniformly to $Q_1(\mathbf{W})$ over $\mathcal{W}$. Therefore, for any $\epsilon > 0$, (i) uniform convergence of $\hat{Q}_{T_0}(\mathbf{W})$ implies that $\hat{Q}_{T_0}(\hat{\mathbf{W}}(s, \mathbf{V}^*(s, T_0), T_0)) < Q_1(\hat{\mathbf{W}}(s, \mathbf{V}^*(s, T_0), T_0)) + \frac{\epsilon}{3}$ and $Q_1(\bar{\mathbf{W}}) < \hat{Q}_{T_0}(\bar{\mathbf{W}}) + \frac{\epsilon}{3}$ with probability approaching to one (w.p.a.1), and (ii) convergence in probability of

15206688, 2020, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/pam.22206 by Ecole Normale Supérieure de Paris, Wiley Online Library on [08/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

$\hat{\mathbf{W}}(s, \mathbf{V}^*(s, T_0), T_0)$ and continuity of $Q_1(\mathbf{W})$ implies that $Q_1(\hat{\mathbf{W}}(s, \mathbf{V}^*(s, T_0), T_0)) < Q_1(\bar{\mathbf{W}})) + \frac{\epsilon}{3}$ w.p.a.1. Therefore, $\hat{Q}_{T_0}(\hat{\mathbf{W}}(s, \mathbf{V}^*(s, T_0), T_0)) < \hat{Q}_{T_0}(\bar{\mathbf{W}}) + \epsilon$ w.p.a.1.

Suppose now that $\hat{\mathbf{W}}(s, T_0)$ does not converge in probability to $\bar{\mathbf{W}}$. Then $\exists \tilde{\epsilon} > 0$ such that $\lim Pr(|\hat{\mathbf{W}}(s, T_0) - \bar{\mathbf{W}}| > \tilde{\epsilon}) \neq 0$ when $T_0 \to \infty$. Since $\mathcal{W}$ is compact and $Q_1(\mathbf{W})$ is uniquely minimized at $\bar{\mathbf{W}}$, then $|\hat{\mathbf{W}}(s, T_0) - \bar{\mathbf{W}}| > \tilde{\epsilon}$ implies that $\exists \eta > 0$ such that $Q_1(\hat{\mathbf{W}}(s, T_0)) > Q_1(\bar{\mathbf{W}}) + 3\eta$. Uniform convergence of $\hat{Q}_{T_0}(\mathbf{W})$ implies that $\hat{Q}_{T_0}(\hat{\mathbf{W}}(s, T_0)) > Q_1(\hat{\mathbf{W}}(s, T_0)) - \eta$ and $Q_1(\bar{\mathbf{W}}) > \hat{Q}_{T_0}(\bar{\mathbf{W}}) - \eta$ w.p.a.1. Therefore, $\hat{Q}_{T_0}(\hat{\mathbf{W}}(s, T_0)) > \hat{Q}_{T_0}(\bar{\mathbf{W}}) + \eta$ w.p.a.1.

However, if we set $\epsilon = \eta$, then we have $\hat{Q}_{T_0}(\hat{\mathbf{W}}(s, \mathbf{V}^*(s, T_0), T_0)) < \hat{Q}_{T_0}(\hat{\mathbf{W}}(s, T_0))$ w.p.a.1, which contradicts the fact that for all $\tilde{T}_0$ we can always find $T_0 > \tilde{T}_0$ such that $\hat{\mathbf{W}}(s, T_0) \in \tilde{\mathcal{W}}$ with $|\hat{\mathbf{W}}(s, T_0) - \bar{\mathbf{W}}| > \tilde{\epsilon}$ minimizes $\hat{Q}_{T_0}(\mathbf{W})$ with positive probability. Therefore, it must be that $\hat{\mathbf{W}}(s, T_0)$ converges in probability to $\bar{\mathbf{W}}$.

Proof of Corollary 3. Notice that we can write each estimator as:

$$\hat{\alpha}_{1t}(s, T_0) = Y_{1t} - \sum_{j=2}^{J+1} \hat{w}_j(s, T_0) Y_{j,t} \text{ for any } s.$$

Using the result of Proposition 2, under Assumption 1:

$$\hat{\alpha}_{1t}(s, T_0) \underset{p}{\to} Y_{1t} - \sum_{j=2}^{J+1} \bar{w}_j Y_{j,t} \text{ as } T_0 \to \infty \text{ for any } s.$$

Hence, for any s and s' such that $L(s, T_0) \to \infty$ and $L(s', T_0) \to \infty$ when $T_0 \to \infty$:

$$\left|\hat{\alpha}_{1t}(s, T_0) - \hat{\alpha}_{1t}(s', T_0)\right| \underset{p}{\to} 0.$$

Proof of Corollary 4. Let $\mathrm{y}_{-j,t}$ be the vector of outcomes at time t excluding unit j, $\hat{\mathbf{W}}_j$ be the SC weights when unit j is used as treated, and $\bar{\mathbf{W}}_j := \text{argmin}_{\mathbf{W} \in \tilde{\mathcal{W}}} Q_j(\mathbf{W})$.

If the outcomes are conditions and conditioning on the realization of the random variables $\{Y_{1,t}, \ldots, Y_{J+1,t}\}_{t=T_0+1}^{T} = \{y_{1,t}, \ldots, y_{J+1,t}\}_{t=T_0+1}^{T}$, we can define $\{(1), \ldots, (J+1)\}$ such that, with probability one:[34]

$$\frac{\frac{1}{T-T_0}\sum_{t=T_0}^{T}\left(y_{(1),t} - \mathrm{y}_{-(1),t}{}'\bar{\mathbf{W}}_{(1)}\right)^2}{Q_{(1)}\left(\bar{\mathbf{W}}_{(1)}\right)} > \cdots > \frac{\frac{1}{T-T_0}\sum_{t=T_0}^{T}\left(y_{(J+1),t} - \mathrm{y}_{-(J+1),t}{}'\bar{\mathbf{W}}_{(J+1)}\right)^2}{Q_{(J+1)}\left(\bar{\mathbf{W}}_{(J+1)}\right)} \tag{A.1}$$

From Proposition 2, we know that $\hat{\mathbf{W}}_j \to_p \bar{\mathbf{W}}_j$ and $\frac{1}{T_0}\sum_{t=1}^{T_0}(y_{j,t} - \mathrm{y}_{-j,t}{}'\hat{\mathbf{W}}_j)^2 \to {}_pQ_j(\bar{\mathbf{W}}_j)$. Therefore, the inequalities in equation (A.1) will remain valid w.p.a.1 when we consider the test statistics for the placebo runs.

Sufficient Conditions for Assumption 1

Let $\mathbf{y}_t^0 = (Y_{1,t}^0, \ldots, Y_{J+1,t}^0)'$. We show that the following assumption is sufficient for Assumption 1.

[34] Continuous outcomes guarantee that ties will happen with probability zero.

15206688, 2020, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/pam.22206 by Ecole Normale Supérieure de Paris, Wiley Online Library on [08/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Assumption 5. $\{\mathbf{y}_t^0\mathbf{y}_t'^0\}$ *has weak stationarity, each element of* $\{\mathbf{y}_t^0\mathbf{y}_t'^0\}$ *has absolutely summable covariances, and* $\mathbb{E}[\mathbf{y}_t^0\mathbf{y}_t'^0]$ *is non-singular.*

Let A_t be one element of $\{\mathbf{y}_t^0\mathbf{y}_t'^0\}$. Under Assumption 5, we can define $\mathbb{E}[A_t] = \mu$ and $\mathbb{E}[(A_t - \mu)(A_{t-j} - \mu)] = \gamma_j$, where $\sum_{j=0}^{\infty} |\gamma_j| < \infty$. Consider a subsequence $\{t_k\}_{k\in\mathbb{N}}$ with $t_k > t_{k-1}$. Note that $\mathbb{E}[\frac{1}{K}\sum_{k=1}^{K} A_{t_k}] = \mu$. We want to show that $\mathbb{E}[\frac{1}{K}\sum_{k=1}^{K}(A_{t_k} - \mu)]^2 \to 0$ when $K \to \infty$. Note that:

$$\begin{aligned} K^2\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}(A_{t_k}-\mu)\right]^2 &= (\gamma_0 + \gamma_{t_1-t_2} + \cdots + \gamma_{t_1-t_K}) + (\gamma_{t_2-t_1} + \gamma_0 + \cdots + \gamma_{t_2-t_K}) \\ &\quad + \cdots + \left(\gamma_{t_K-t_1} + \gamma_{t_{K-1}-t_1} + \cdots + \gamma_0\right) \\ &= K\gamma_0 + \sum_{k=1}^{K-1}\left[\sum_{l=k+1}^{K} 2\gamma_{t_l-t_k}\right] \\ &\leq K|\gamma_0| + \sum_{k=1}^{K-1}\left[\sum_{l=k+1}^{K} 2\left|\gamma_{t_l-t_k}\right|\right]. \end{aligned}$$

Let $lim\sum_{l=0}^{T}|\gamma_l| = C$. Now note that, for each k, $\sum_{l=k+1}^{K} 2|\gamma_{|t_l-t_k|}|$ is the sum of a subsequence of $\{|\gamma_l|\}$. Therefore, for any k, we have that $\sum_{l=k+1}^{K} 2|\gamma_{|t_l-t_k|}| \leq \sum_{l=1}^{t_K} 2|\gamma_l| \leq C$. Therefore:

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}(A_{t_k}-\mu)\right]^2 \leq \frac{1}{K}|\gamma_0| + \frac{K-1}{K^2}C$$

which implies that $\mathbb{E}[\frac{1}{K}\sum_{k=1}^{K}(A_{t_k} - \mu)]^2 \to 0$ when $K \to \infty$. Therefore, we have that all elements of the pre-treatment averages of $\{\mathbf{y}_t^0\mathbf{y}_t'^0\}$ for any subsequence $\{t_k\}_{k\in\mathbb{N}}$ converge in probability to their corresponding expected values.

Since $\frac{1}{K}\sum_{k=1}^{K}(Y_{j,t_k}^0 - \mathbf{y}'^0_{-j,t_k}\mathbf{W})^2$ is a linear combination of pre-treatment averages of elements of $\{\mathbf{y}_t^0\mathbf{y}_t'^0\}$ for a given subsequence $\{t_k\}_{k\in\mathbb{N}}$, for any $\mathbf{W} \in \mathcal{W}$, we have that:

$$\tilde{Q}_K(\mathbf{W}) \equiv \frac{1}{K}\sum_{k=1}^{K}\left(Y_{j,t_k}^0 - \mathbf{y}'^0_{-j,t_k}\mathbf{W}\right)^2 \underset{p}{\to} \mathbb{E}\left[\left(Y_{j,t}^0 - \mathbf{y}'^0_{-j,t}\mathbf{W}\right)^2\right]$$

where $\mathbb{E}[(Y_{j,t}^0 - \mathbf{y}'^0_{-j,t}\mathbf{W})^2]$ is continuous and strictly convex.

Finally, we show that this convergence in probability is uniform. For any $\mathbf{W}', \mathbf{W} \in \mathcal{W}$, using the mean value theorem, we can find $\tilde{\mathbf{W}} \in \mathcal{W}$ such that:

$$\begin{aligned} \left|\tilde{Q}_K(\mathbf{W}') - \tilde{Q}_K(\mathbf{W})\right| &= \left|2\left(\frac{1}{K}\sum_{k=1}^{K}\mathbf{y}_{-j,t_k}^0 Y_{j,t_k}^0 - \frac{1}{K}\sum_{k=1}^{K}\mathbf{y}_{-j,t_k}^0\mathbf{y}_{-j,t_k}^{0\prime}\tilde{\mathbf{W}}\right)\cdot(\mathbf{W}' - \mathbf{W})\right| \\ &\leq \left[\left(2\frac{1}{K}\sum_{k=1}^{K}\mathbf{y}_{-j,t_k}^0 Y_{j,t_k}^0 + \frac{1}{K}\sum_{k=1}^{K}\mathbf{y}_{-j,t_k}^0\mathbf{y}_{-j,t_k}^{0\prime}\cdot\tilde{\mathbf{W}}\right)\mathbf{W}' - \mathbf{W}\right]^{\frac{1}{2}}. \end{aligned}$$

Define $B_K = 2\frac{1}{K}\sum_{k=1}^{K} \mathbf{y}^0_{-j,t_k} Y^0_{j,t_k} + \frac{1}{K}\sum_{k=1}^{K} \mathbf{y}^0_{-j,t_k} \mathbf{y}'^0_{-j,t_k} \cdot C$. Since $\mathcal{W}$ is compact, $\tilde{\mathbf{W}}$ is bounded, so we can find a constant C such that $|\tilde{Q}_K(\mathbf{W}') - \tilde{Q}_K(\mathbf{W})| \leq B_K(\mathbf{W}' - \mathbf{W})^{\frac{1}{2}}$. From Assumption 5, B_K converges in probability to a positive constant, so $B_K = O_p(1)$. Note also that $\mathbb{E}[(Y^0_{j,t} - \mathbf{y}'^0_{-j,t}\mathbf{W})^2]$ is uniformly continuous on $\mathcal{W}$. Therefore, from Corollary 2.2 of Newey (1991), we have that $\tilde{Q}_K$ converges uniformly in probability to $\mathbb{E}[(Y^0_{j,t} - \mathbf{y}'^0_{-j,t}\mathbf{W})^2]$ for any subsequence $\{t_k\}_{k\in\mathbb{N}}$.

15206688, 2020, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/pam.22206 by Ecole Normale Supérieure de Paris, Wiley Online Library on [08/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

15206688, 2020, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/pam.22206 by Ecole Normale Supérieure de Paris, Wiley Online Library on [08/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

APPENDIX B: MODEL WITH TIME-INVARIANT COVARIATES

In the paper's third section, we provide evidence that specifications 6 (pre-treatment outcome mean as economic predictor) and 7 (initial, middle, and final years of the pre-intervention period as economic predictors) fail to take into account the time-series dynamics of the data, which implies that the SC estimator using these specifications does not converge to the SC estimators using the other specifications, which satisfy the conditions outlined in the second section. As a consequence, the possibilities for specification searching do not vanish even when the number of pre-treatment periods is large in contrast to the behavior of the specifications within the scope of our theoretical results. However, in most applications that use specifications 6 and 7, other time-invariant covariates are also considered as economic predictors. Here, we consider an alternative MC simulation where we include time-invariant covariates and we show that the same pattern observed in the third section can arise even when we consider specifications that also include time-invariant covariates as economic predictors.

The alternative DGP is given by:

$$Y_{j,t}^{0} = \delta_t + \lambda_t^k + \theta_t Z_i + \epsilon_{jt}$$

where $Z_i = 1$ for $i = 1, \ldots, 10$ and $Z_i = 0$ for $i = 11, \ldots, 20$. As in our DGP from the main paper, we consider $K = 10$.[35] We consider that λ_t^k is normally distributed following an AR(1) process with 0.5 serial correlation parameter, $\delta_t \sim N(0,\ 1)$, $\epsilon_{j,t} \sim N(0,\ 0.1)$, and $\theta_t \sim N(0,\ 1)$. We consider the same seven specifications as in the third section of the main paper, except that we also include Z_i as economic predictor.

In columns (1) and (2) of Table B1, we present the probability of rejecting the null in at least one of our seven specifications at, respectively, 5 percent and 10 percent significance levels. The possibilities for specification searching remain high for large T_0 because specifications 6 and 7 remain poorly behaved in comparison to the other specifications. This result is similar to our findings in the main paper.

Table B1. Specification searching—model with time-invariant covariates.

	All Specifications		Excluding 6 and 7	
	5% test (1)	10% test (2)	5% test (3)	10% test (4)
$T_0 = 12$	0.142	0.232	0.107	0.196
	(0.003)	(0.004)	(0.004)	(0.005)
$T_0 = 32$	0.141	0.224	0.101	0.175
	(0.003)	(0.004)	(0.004)	(0.005)
$T_0 = 100$	0.136	0.215	0.089	0.158
	(0.003)	(0.004)	(0.003)	(0.004)
$T_0 = 400$	0.125	0.200	0.078	0.138
	(0.003)	(0.004)	(0.003)	(0.004)

Notes: This table presents results based on 10,000 observations of the MC simulations described in Appendix B. Columns (1) and (2) present the probability of rejecting the null in at least one specification at, respectively, 5 percent and 10 percent significance levels. Columns (3) and (4) present the probability of rejecting the null in at least one specification at, respectively, 5 percent and 10 percent significance levels when we exclude specifications 6 and 7.

[35] Therefore, units 1 and 2 follow the trend λ_t^1, units 3 and 4 follow the trend λ_t^2, and so on.

In columns (3) and (4) of Table B1, we present the probability of rejecting the null at the 5 percent and 10 percent significance levels in at least one of the five specifications that satisfy the conditions outlined in the second section of the main paper. We stress that the possibilities for specification searching decrease a lot for each T_0 and, most importantly, the rejection rate decreases when the pre-treatment period gets larger. Once more, this result is similar to our findings in the main paper.

15206688, 2020, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/pam.22206 by Ecole Normale Supérieure de Paris, Wiley Online Library on [08/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

15206688, 2020, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/pam.22206 by Ecole Normale Supérieure de Paris, Wiley Online Library on [08/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

APPENDIX C: VARIABILITY AND MISALLOCATION OF WEIGHTS

Based on Proposition 2 (see Appendix A), specifications 1 through 5 should provide similar SC weights, while specifications 6 and 7 could potentially provide SC weights that differ wildly. To analyze this possibility, we calculate a measure of variability of weights in comparison to specification 1. For each specification $s \in \{2, \ldots, 7\}$, we compute the difference between the weight allocated by specification 1 and specification s for each unit in the donor pool. Then, we take the maximum value of this difference across units in the donor pool. We present this measure for specifications 2 through 7 in Table C1. On the one hand, analyzing specifications 2 through 5, we find that the variability of weights between specifications is small (even when T_0 is small) and, most importantly, decreasing when the pre-intervention period gets large, as expected given our theoretical results. On the other hand, for specifications 6 and 7, we find strikingly different results: their weights differ substantially from the weights of specification 1 and this difference does not decrease when the pre-intervention period gets large.

Table C1. Variability of weights.

Distance between weights of specification 1 vs. specification s:						
	2	3	4	5	6	7
Panel A: Stationary model						
$T_0 = 12$	0.156	0.210	0.137	0.137	0.631	0.337
	(0.001)	(0.002)	(0.001)	(0.001)	(0.003)	(0.003)
$T_0 = 32$	0.085	0.134	0.073	0.074	0.693	0.370
	(0.001)	(0.001)	(0.000)	(0.000)	(0.003)	(0.003)
$T_0 = 100$	0.055	0.080	0.051	0.051	0.724	0.381
	(0.000)	(0.000)	(0.000)	(0.000)	(0.003)	(0.004)
$T_0 = 400$	0.032	0.048	0.032	0.032	0.740	0.391
	(0.000)	(0.000)	(0.000)	(0.000)	(0.003)	(0.004)
Panel B: Non-stationary model						
$T_0 = 12$	0.137	0.185	0.114	0.115	0.661	0.295
	(0.001)	(0.002)	(0.001)	(0.001)	(0.003)	(0.003)
$T_0 = 32$	0.071	0.115	0.067	0.066	0.723	0.312
	(0.001)	(0.001)	(0.001)	(0.001)	(0.004)	(0.004)
$T_0 = 100$	0.049	0.070	0.049	0.049	0.756	0.313
	(0.000)	(0.000)	(0.000)	(0.000)	(0.003)	(0.003)
$T_0 = 400$	0.034	0.046	0.036	0.036	0.769	0.318
	(0.000)	(0.000)	(0.000)	(0.000)	(0.004)	(0.004)

Notes: The average variability of weights is based on 10,000 observations and captures the average maximum difference of allocated weights between specifications s and 1. Specification s is one of the specifications used to compute the synthetic control unit: (2) the first three-quarters of the pre-treatment outcome values, (3) the first half of the pre-treatment outcome values, (4) odd pre-treatment outcome values, (5) even pre-treatment outcome values, (6) the mean of all pre-treatment outcome values, and (7) three outcome values. T_0 is the number of pre-treatment periods.

Beyond the variability of weights between specifications, an interesting feature of our MC simulations is that the SC estimator should assign positive weights only for unit 2 (which has the same factor loadings of unit 1), so we can actually calculate the proportion of weights that are misallocated for each specification. We present in columns 1 to 7 of Table C2 the proportion of misallocated weights for each specification using both of our DGPs. Interestingly, specifications 6 and 7 misallocate substantially more weights relative to the other specifications. For the stationary model (panel A), with $T_0 = 12$, specifications 6 and 7 misallocate more than 80 percent and 45 percent of the weights, while the misallocation share for

15206688, 2020, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/pam.22206 by Ecole Normale Supérieure de Paris, Wiley Online Library on [08/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table C2. Misallocation of weights.

	Specification						
	1	2	3	4	5	6	7
Panel A: Stationary model							
$T_0 = 12$	0.225	0.278	0.315	0.249	0.248	0.813	0.474
	(0.001)	(0.002)	(0.003)	(0.002)	(0.002)	(0.003)	(0.004)
$T_0 = 32$	0.148	0.163	0.193	0.143	0.143	0.811	0.459
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.003)	(0.004)
$T_0 = 100$	0.110	0.115	0.119	0.099	0.099	0.811	0.450
	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)	(0.003)	(0.004)
$T_0 = 400$	0.091	0.092	0.094	0.086	0.085	0.812	0.451
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.003)	(0.004)
Panel B: Non-stationary model							
$T_0 = 12$	0.187	0.233	0.267	0.204	0.203	0.805	0.401
	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)	(0.004)	(0.004)
$T_0 = 32$	0.116	0.125	0.159	0.119	0.120	0.807	0.373
	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.004)	(0.005)
$T_0 = 100$	0.085	0.087	0.097	0.080	0.080	0.815	0.357
	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)	(0.004)	(0.004)
$T_0 = 400$	0.072	0.072	0.075	0.070	0.069	0.819	0.355
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.005)	(0.005)

Notes: The average of misallocated weights is based on 10,000 observations. The reasoning behind this variable is the following: Since, in our DGP, we divide units into groups whose trends are parallel only when compared to units in the same group, the sum of the weights allocated to the units in the other groups is a measure of the relevance given by the synthetic control method to units whose true potential outcome follows a different trajectory than the one followed by the unit chosen to be the treated one. Specification s is one of the specifications used to compute the synthetic control unit: (1) all pre-treatment outcome values, (2) the first three-quarters of the pre-treatment outcome values, (3) the first half of the pre-treatment outcome values, (4) odd pre-treatment outcome values, (5) even pre-treatment outcome values, (6) the mean of all pre-treatment outcome values, and (7) three outcome values. T_0 is the number of pre-treatment periods.

other specifications ranges from 23 to 32 percent. Most importantly, the misallocation of weights decreases with T_0 for all specifications, except for specifications 6 and 7. Results are qualitatively the same for the non-stationary model (panel B). These results suggest that specifications outside the scope of Proposition 2, such as specifications 6 and 7, behave poorly because they do not capture the time-series dynamics of the units, which is the main goal of the SC method.[36,37]

[36] Although any specification could potentially take into account the time-series dynamics of the outcome variable because the matrix **V** is chosen to minimize the pre-treatment MSPE in the second step of the optimization process, this process is very limited because the first minimization problem can severely restrict the set of possible weights W*(**V**) that may be chosen in the second step, as suggested in Ferman and Pinto (2019).

[37] In Appendix B, we show that specifications 6 and 7 can fail to properly exploit the time-series dynamics of the data even if we also include time-invariant covariates as economic predictors. Therefore, our result that the possibilities of specification searching may not diminish with the number of pre-treatment periods when we consider specifications outside the scope of Proposition 2 remains valid.

15206688, 2020, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/pam.22206 by Ecole Normale Supérieure de Paris, Wiley Online Library on [08/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

APPENDIX D: CONDITIONING ON A GOOD PRE-TREATMENT FIT

In the exercise presented in Table 2, we assumed that the researcher would be able to choose any of the seven specifications we considered in our MC simulations. However, Abadie, Diamond, and Hainmueller (2010, 2015) emphasize that the SC control estimator should only be used in the situations with good pre-treatment fit. Therefore, we check whether the specification-searching problem we identified in the SC method arises because we allow the researcher to choose specifications that provide a poor pre-treatment fit. We consider a pre-treatment normalized mean squared error index to determine whether a specification provides a good pre-treatment fit:[38]

$$\tilde{R}^2 = 1 - \frac{\sum_{t=1}^{T_0} \left(Y_{1,t} - \hat{Y}_{1,t}^N\right)^2}{\sum_{t=1}^{T_0} \left(Y_{1,t} - \bar{Y}_1\right)^2} \tag{D.1}$$

where $\bar{Y}_1 = \frac{\sum_{t=1}^{T_0} Y_{1,t}}{T_0}$. If this measure is one, then we have a perfect fit.[39]

In order to capture a good fit, we consider two thresholds for $\tilde{R}^2$, $\tilde{R}^2 > 0.80$ and $\tilde{R}^2 > 0.95$. Considering these two thresholds, panel A of Table D1 shows the probability of finding a good pre-treatment fit for at least one of the seven specifications. The probability of finding specifications with a good pre-treatment fit depends crucially on how we define whether a specification provided a good fit and on whether we consider a stationary or a non-stationary model. We present in columns 1 and 2 the results for the stationary model. With a moderate T_0, the probability of finding at least one specification with good fit is close to one when we consider the weaker definition of good fit, and close to zero when we consider the more stringent definition. We highlight that, according to panels B and C, the specifications that do not satisfy the conditions outlined in the second section of the main paper have a relatively small chance of providing a good pre-intervention fit, even under the weaker definition of good fit.

We present, in columns 3 and 4, the results for the non-stationary model. In this case, the probability of having at least one specification with a good fit is close to one even when we consider the more stringent definition of good fit. Also, there is a high probability that all specifications (including specifications 6 and 7) provide a good fit, especially when T_0 is large. This happens because, with large T_0, the non-stationary factors dominate the variance of $Y_{1,t}$. Since the SC estimator is extremely effective in controlling for the non-stationary factors (Ferman & Pinto, 2019), it will usually provide a good pre-treatment fit.

Given these definitions of good fit, we present in Table D2 the probabilities of rejecting the null in at least one specification when we restrict the researcher to consider only specifications that provide a good pre-treatment fit. The possibilities for specification searching in the non-stationary model (columns 3 and 4) are virtually the same as when we do not restrict to specifications with a good pre-treatment fit, especially when T_0 is large (columns 3 and 4 of Table 2). This is not surprising, given that all specifications will usually provide a good pre-treatment fit in this model. For the stationary model (columns 1 and 2 of Table D2), the specification-search problem is attenuated when we restrict to specifications with a good fit if

[38] This measure is very similar to the "pre-treatment fit index" proposed by Adhikari and Alm (2016). Differently from their suggestion, our measure is invariant to linearly additive changes. Dube and Zipperer (2015) also propose a pre-treatment fit criterion that is equal to the numerator of our measure. Differently from our suggestion, their measure is not scale-invariant.

[39] Note that, differently from the standard R^2 measure, $\tilde{R}^2$ can be negative.

15206688, 2020, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/pam.22206 by Ecole Normale Supérieure de Paris, Wiley Online Library on [08/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table D1. Probability of good pre-treatment fit.

	Stationary model		Non-stationary model	
	$\tilde{R}^2 > 0.80$ (1)	$\tilde{R}^2 > 0.95$ (2)	$\tilde{R}^2 > 0.80$ (3)	$\tilde{R}^2 > 0.95$ (4)
Panel A: At least one specification with good fit				
$T_0 = 12$	0.947	0.271	0.990	0.642
	(0.001)	(0.003)	(0.001)	(0.003)
$T_0 = 32$	0.993	0.085	1.000	0.857
	(0.001)	(0.003)	(0.001)	(0.004)
$T_0 = 100$	1.000	0.002	1.000	0.993
	(0.001)	(0.003)	(0.001)	(0.003)
$T_0 = 400$	1.000	0.000	1.000	1.000
	(0.001)	(0.003)	(0.001)	(0.004)
Panel B: Specification 6 has a good fit				
$T_0 = 12$	0.163	0.015	0.323	0.082
	(0.004)	(0.001)	(0.004)	(0.004)
$T_0 = 32$	0.164	0.004	0.456	0.145
	(0.004)	(0.001)	(0.005)	(0.005)
$T_0 = 100$	0.170	0.000	0.757	0.242
	(0.004)	(0.001)	(0.004)	(0.004)
$T_0 = 400$	0.168	0.000	0.994	0.667
	(0.004)	(0.001)	(0.005)	(0.005)
Panel C: Specification 7 has a good fit				
$T_0 = 12$	0.579	0.092	0.779	0.350
	(0.005)	(0.002)	(0.003)	(0.004)
$T_0 = 32$	0.576	0.024	0.837	0.525
	(0.005)	(0.002)	(0.004)	(0.005)
$T_0 = 100$	0.590	0.001	0.931	0.718
	(0.005)	(0.002)	(0.003)	(0.004)
$T_0 = 400$	0.585	0.000	0.994	0.898
	(0.005)	(0.002)	(0.004)	(0.005)

Notes: Results are based on 10,000 observations and on seven specifications: (1) all pre-treatment outcome values, (2) the first three-quarters of the pre-treatment outcome values, (3) the first half of the pre-treatment outcome values, (4) odd pre-treatment outcome values, (5) even pre-treatment outcome values, (6) the mean of all pre-treatment outcome values, and (7) three outcome values. T_0 is the number of pre-treatment periods. Our measure of goodness of fit is defined by equation (D.1). We consider two definitions of good fit: $\tilde{R}^2 > 0.80$ and $\tilde{R}^2 > 0.95$.

we use the more lenient definition of good fit (panel A). In practice, in this case the restriction of considering only specifications with a good fit prevents the researcher from choosing specifications 6 and 7, whose weights, as we show in Appendix C, are very different from the ones chosen by specifications that satisfy the conditions of Proposition 2 and Corollaries 3 and 4. If we consider the more stringent definition of good fit, however, then the probability of rejecting the null in at least one specification is substantially higher (panel B). This happens because, if we consider that the SC method should only be used when the pre-treatment fit is good (as suggested in Abadie, Diamond, & Hainmueller, 2010, 2015), then there is a low probability of finding a good fit for at least one specification and we would only consider specifications such that the denominator of the test statistic for the treated unit is close to zero. Since the test statistics for the placebo units are not conditional on a good pre-treatment, this leads to over-rejection, as shown in Ferman and Pinto (2017).

Table D2. Specification searching conditional on a good pre-treatment fit.

	Stationary model		Non-stationary model	
	5% test (1)	10% test (2)	5% test (3)	10% test (4)
Panel A: $\tilde{R}^2 > 0.80$				
$T_0 = 12$	0.119	0.205	0.124	0.218
	(0.003)	(0.004)	(0.003)	(0.004)
$T_0 = 32$	0.110	0.193	0.138	0.240
	(0.003)	(0.004)	(0.004)	(0.005)
$T_0 = 100$	0.101	0.174	0.141	0.243
	(0.003)	(0.004)	(0.003)	(0.004)
$T_0 = 400$	0.093	0.163	0.145	0.255
	(0.003)	(0.004)	(0.004)	(0.005)
Panel B: $\tilde{R}^2 > 0.95$				
$T_0 = 12$	0.199	0.323	0.129	0.222
	(0.008)	(0.009)	(0.004)	(0.005)
$T_0 = 32$	0.218	0.348	0.123	0.210
	(0.014)	(0.016)	(0.004)	(0.005)
$T_0 = 100$	0.130	0.217	0.114	0.193
	(0.084)	(0.098)	(0.003)	(0.004)
$T_0 = 400$	-	-	0.130	0.227
	-	-	(0.004)	(0.005)

Notes: Rejection rates are estimated based on 10,000 observations and on seven specifications: (1) all pre-treatment outcome values, (2) the first three-quarters of the pre-treatment outcome values, (3) the first half of the pre-treatment outcome values, (4) odd pre-treatment outcome values, (5) even pre-treatment outcome values, (6) the mean of all pre-treatment outcome values, and (7) three outcome values. *z% test* indicates that the nominal size of the analyzed test is z percent and T_0 is the number of pre-treatment periods. Our measure of goodness of fit is defined by equation (D.1). We consider two definitions of good fit: $\tilde{R}^2 > 0.80$ and $\tilde{R}^2 > 0.95$. In panel B, there is no information on specification searching probabilities for $T_0 = 400$ in the stationary model because all specifications fail to provide a good fit given this definition in all simulations.

Overall, these results suggest that restricting the researcher to consider only specifications with a good fit does not necessarily attenuate the specification-searching problem. On the one hand, if conditioning on a good fit does not actually restrict the set of options a researcher has (as happens with our non-stationary model), then we have the same results as in the unconditional case. On the other hand, if conditioning severely restricts the set of options, then we have over-rejection because the test statistic for the treated unit is conditional on a denominator that is close to zero, while the test statistics for the placebo units are unconditional.

We also present in Table D3 the same results excluding specifications 6 and 7.

15206688, 2020, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/pam.22206 by Ecole Normale Supérieure de Paris, Wiley Online Library on [08/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table D3. Specification searching—excluding specifications 6 and 7.

	Stationary model		Non-stationary model	
	5% test (1)	10% test (2)	5% test (3)	10% test (4)
Panel A: Conditional on $\tilde{R}^2 > 0.80$				
$T_0 = 12$	0.104	0.184	0.107	0.192
	(0.003)	(0.004)	(0.003)	(0.004)
$T_0 = 32$	0.099	0.177	0.108	0.191
	(0.003)	(0.004)	(0.004)	(0.005)
$T_0 = 100$	0.090	0.157	0.094	0.162
	(0.003)	(0.004)	(0.003)	(0.004)
$T_0 = 400$	0.077	0.138	0.081	0.142
	(0.003)	(0.004)	(0.004)	(0.005)
Panel B: Conditional on $\tilde{R}^2 > 0.95$				
$T_0 = 12$	0.183	0.183	0.120	0.210
	(0.008)	(0.008)	(0.004)	(0.005)
$T_0 = 32$	0.208	0.208	0.113	0.195
	(0.013)	(0.013)	(0.004)	(0.005)
$T_0 = 100$	0.130	0.130	0.094	0.162
	(0.082)	(0.082)	(0.003)	(0.004)
$T_0 = 400$	-	-	0.081	0.142
	-	-	(0.004)	(0.005)

Notes: Rejection rates are estimated based on 10,000 observations and on five specifications: (1) all pre-treatment outcome values, (2) the first three-quarters of the pre-treatment outcome values, (3) the first half of the pre-treatment outcome values, (4) odd pre-treatment outcome values, and (5) even pre-treatment outcome values. *z% test* indicates that the nominal size of the analyzed test is z percent and T_0 is the number of pre-treatment periods. Our measure of goodness of fit is defined by equation (D.1). We consider two definitions of good fit: $\tilde{R}^2 > 0.80$ and $\tilde{R}^2 > 0.95$. In panel B, there is no information on specification searching probabilities for $T_0 = 400$ in the stationary model because all specifications fail to provide a good fit given this definition in all simulations.

15206688, 2020, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/pam.22206 by Ecole Normale Supérieure de Paris, Wiley Online Library on [08/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

15206688, 2020, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/pam.22206 by Ecole Normale Supérieure de Paris, Wiley Online Library on [08/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

APPENDIX E: SIMULATIONS WITH REAL DATA

The results presented in the main paper suggest that different specifications of the SC method can generate significant specification-searching opportunities in samples of sizes commonly used in SC applications. In particular, we also find that using only specifications that satisfy the conditions outlined in the second section of the paper alleviate this problem even though it does not solve it completely. We now check whether the results we find in our MC simulations are also relevant when we consider real datasets by conducting simulations of placebo interventions with the Current Population Survey (CPS). We use the CPS Merged Outgoing Rotation Groups for the years 1979 to 2014. Following Bertrand, Duflo, and Mullainathan (2004), we extract information on employment status and earnings for women between ages 25 and 50. We also consider in a separate set of simulations information on men in the same age range.

Before we proceed to the placebo simulations, we briefly discuss the raw data for these outcome variables. There are important distinctions in the time series characteristics when we consider information for men versus women and when we consider log wages versus employment. Figures E1a and E1b present the time series of log wages for all U.S. states, respectively for men and women. As expected, the time series of log wages is non-stationary and increasing for both men and women. These graphs suggest that there is a strong non-stationary factor that affects all states in the same way. Figures E1c and E1d present the time series of employment for all U.S. states, respectively for men and women. In this case, the time series for men should be closer to our stationary model from the third section of the main paper, while the time series for women has an increasing trend in the 80s and 90s.

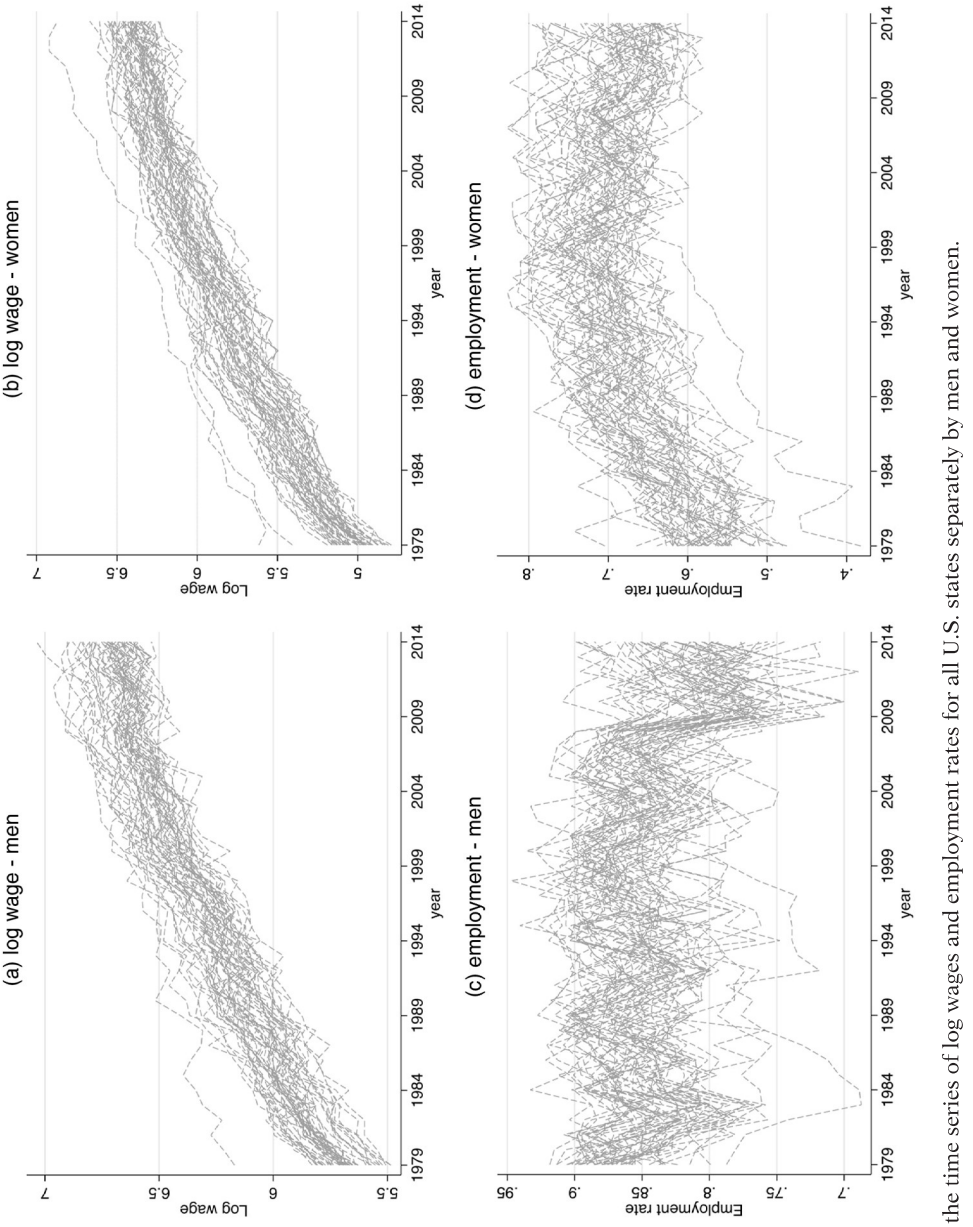
We first consider simulations with 12 pre-intervention periods, four post-intervention periods, and 20 states. In each simulation, we randomly select one treated and 19 control states out of the 51 states (including Washington, DC) and then we randomly select the first period between 1979 and 1999. Then we consider simulations with 32 pre-intervention periods, four post-intervention periods, and 20 states. In this case, we randomly select 20 states and use the entire 36 years of data. In each scenario, we run 5,000 simulations using either employment or log wages as the dependent variable and test the null hypothesis using the same seven specifications from the third section of the main paper.[40]

We start presenting the probability of finding specifications with a good fit in Table E1. When the outcome variable is log wages, the probability of having at least one specification with a good fit is close to one, especially when we consider $T_0 = 32$ (columns 1 to 4, panel A). Most importantly, when we consider $T_0 = 32$, specifications 6 and 7 have a high probability of fitting the data closely. These results are consistent with our MC simulations considering that the log wages series appear to have important non-stationary common factors. The probability of finding specifications with a good fit is lower when we consider employment instead of log wages as the outcome variable, and even lower when we consider men relative to women. This is consistent with the employment time series for men being closer to a stationary process.

We present in Table E2 the probabilities of rejecting the null in at least one specification.[41] In panel A, we present the specification-search probabilities including any of the seven specifications that provide a good fit, i.e., $\tilde{R}^2 > 0.80$. The results

[40] Standard errors are clustered at the level of the treated state when we calculate the probability of having a good fit and when we calculate rejection rates.

[41] Standard errors for these simulation results are clustered at the treated-state level, in order to take into account that the simulations are not independent.



Note: We present the time series of log wages and employment rates for all U.S. states separately by men and women.

Figure E1. Outcome Trajectories in the CPS Data.

15206688, 2020, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/pam.22206 by Ecole Normale Supérieure de Paris, Wiley Online Library on [08/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table E1. Probability of good pre-treatment fit—CPS.

	Log wages				Employment			
	Women		Men		Women		Men	
	$\tilde{R}^2 > 0.80$ (1)	$\tilde{R}^2 > 0.95$ (2)	$\tilde{R}^2 > 0.80$ (3)	$\tilde{R}^2 > 0.95$ (4)	$\tilde{R}^2 > 0.80$ (5)	$\tilde{R}^2 > 0.95$ (6)	$\tilde{R}^2 > 0.80$ (7)	$\tilde{R}^2 > 0.95$ (8)
Panel A: At least one specification								
$T_0 = 12$	0.914	0.573	0.876	0.413	0.276	0.031	0.153	0.017
	(0.028)	(0.043)	(0.031)	(0.044)	(0.03)	(0.011)	(0.031)	(0.008)
$T_0 = 32$	0.963	0.949	0.983	0.906	0.653	0.042	0.066	0.000
	(0.026)	(0.029)	(0.017)	(0.032)	(0.057)	(0.023)	(0.03)	-
Panel B: Specification 6 has a good fit								
$T_0 = 12$	0.846	0.224	0.719	0.087	0.069	0.000	0.008	0.000
	(0.033)	(0.035)	(0.038)	(0.023)	(0.015)	-	(0.003)	-
$T_0 = 32$	0.959	0.914	0.981	0.777	0.343	0.000	0.002	0.000
	(0.029)	(0.03)	(0.017)	(0.043)	(0.056)	-	(0.001)	-
Panel C: Specification 7 has a good fit								
$T_0 = 12$	0.874	0.317	0.790	0.168	0.107	0.001	0.020	0.001
	(0.031)	(0.036)	(0.033)	(0.031)	(0.015)	(0.001)	(0.007)	(0.001)
$T_0 = 32$	0.963	0.934	0.983	0.860	0.359	0.008	0.003	0.000
	(0.026)	(0.031)	(0.017)	(0.037)	(0.053)	(0.007)	(0.002)	-

Notes: Results are based on seven specifications: (1) all pre-treatment outcome values, (2) the first three-quarters of the pre-treatment outcome values, (3) the first half of the pre-treatment outcome values, (4) odd pre-treatment outcome values, (5) even pre-treatment outcome values, (6) the mean of all pre-treatment outcome values, and (7) three outcome values—and on 5,000 observations for each outcome variable (employment and log wages), for each sample (men and women), and number of pre-treatment periods ($T_0 \in \{12,\ 32\}$). Our measure of goodness of fit is defined by equation (D.1). We consider two definitions of good fit: $\tilde{R}^2 > 0.80$ and $\tilde{R}^2 > 0.95$.

are very similar to our findings in the MC simulations. With $T_0 = 12$, depending on the sample and outcome variable, there is 13 to 26 percent probability of finding a specification with statistically significant results at 5 percent and a 21 to 41 percent probability of finding a specification with statistically significant results at 10 percent. With $T_0 = 32$, these probabilities are slightly lower, but still significantly higher than the test nominal size for all cases but men's employment rates. In panel B, we present the results searching only specifications that satisfy the conditions outlined in the second section of the main paper, i.e., we exclude specifications 6 and 7. As in our MC simulations, restricting to specifications 1 through 5 reduces the specification-searching problem but does not solve it entirely. In particular, for $T_0 = 32$, we cannot reject the null hypothesis that the rejection rate is equal to the nominal level for all but one case. We stress that this reduction is not a mechanical consequence of searching five instead of seven specifications. If we exclude specifications 2 and 3, we find rejection rates that are very similar to the ones including all seven specifications.[42] In general, these results suggest that specification-searching possibilities in SC applications can be relevant in real applications of the SC method even when we restrict ourselves to specifications that satisfy the conditions outlined in the second section of the main paper.

[42] Detailed results are available upon request.

15206688, 2020, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/pam.22206 by Ecole Normale Supérieure de Paris, Wiley Online Library on [08/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table E2. Specification searching—CPS simulations.

	Log wages				Employment			
	Women		Men		Women		Men	
	5% test (1)	10% test (2)	5% test (3)	10% test (4)	5% test (5)	10% test (6)	5% test (7)	10% test (8)
Panel A: Conditional on $\tilde{R}^2 > 0.80$ - All Specifications								
$T_0 = 12$	0.137	0.234	0.130	0.218	0.217	0.351	0.262	0.415
	(0.013)	(0.019)	(0.013)	(0.018)	(0.025)	(0.026)	(0.027)	(0.029)
$T_0 = 32$	0.123	0.215	0.117	0.203	0.141	0.228	0.151	0.242
	(0.029)	(0.039)	(0.029)	(0.04)	(0.045)	(0.056)	(0.08)	(0.108)
Panel B: Conditional on $\tilde{R}^2 > 0.80$ - Excluding Specifications 6 and 7								
$T_0 = 12$	0.108	0.192	0.106	0.183	0.201	0.325	0.253	0.405
	(0.012)	(0.018)	(0.011)	(0.016)	(0.024)	(0.027)	(0.027)	(0.029)
$T_0 = 32$	0.082	0.149	0.071	0.138	0.105	0.186	0.151	0.242
	(0.023)	(0.033)	(0.021)	(0.033)	(0.036)	(0.049)	(0.08)	(0.108)

Notes: Rejection rates are estimated based on seven specifications: (1) all pre-treatment outcome values, (2) the first three-quarters of the pre-treatment outcome values, (3) the first half of the pre-treatment outcome values, (4) odd pre-treatment outcome values, (5) even pre-treatment outcome values, (6) the mean of all pre-treatment outcome values, and (7) three outcome values—and on 5,000 observations for each outcome variable (employment and log wages), for each sample (men and women), and number of pre-treatment periods ($T_0 \in \{12,\ 32\}$). *z% test* indicates that the nominal size of the analyzed test is z percent. Our measure of goodness of fit is defined by equation (D.1). Here, we consider one definition of good fit: $\tilde{R}^2 > 0.80$. Unconditional results and conditional results imposing $\tilde{R}^2 > 0.95$ are available upon request. The numbers in panels A and B for male employment levels when $T_0 = 32$ are the same because there are only 21 observations whose specifications 6 and 7 provide a good pre-treatment fit and, in all these cases, they do not change the test decision based only on specifications 1 through 5.

15206688, 2020, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/pam.22206 by Ecole Normale Supérieure de Paris, Wiley Online Library on [08/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

15206688, 2020, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/pam.22206 by Ecole Normale Supérieure de Paris, Wiley Online Library on [08/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

APPENDIX F: SUPPLEMENTARY EMPIRICAL APPLICATIONS

Empirical Application: Resource Curse

Smith (2015) evaluates the impact of major natural resource discoveries since 1950 on GDP per capita using different methods, including the synthetic control method.[43] Major oil and gas discoveries happened in Equatorial Guinea and Ecuador in 1992 and 1972, respectively, implying that pre- and post-treatment periods are 1950 through 1991 and 1992 through 2008 for the first country and 1950 through 1971 and 1972 through 2008 for the second one. While the donor pool for Equatorial Guinea consists of Sub-Saharan African Countries, the donor pool for Ecuador consists of Latin American and Caribbean countries.

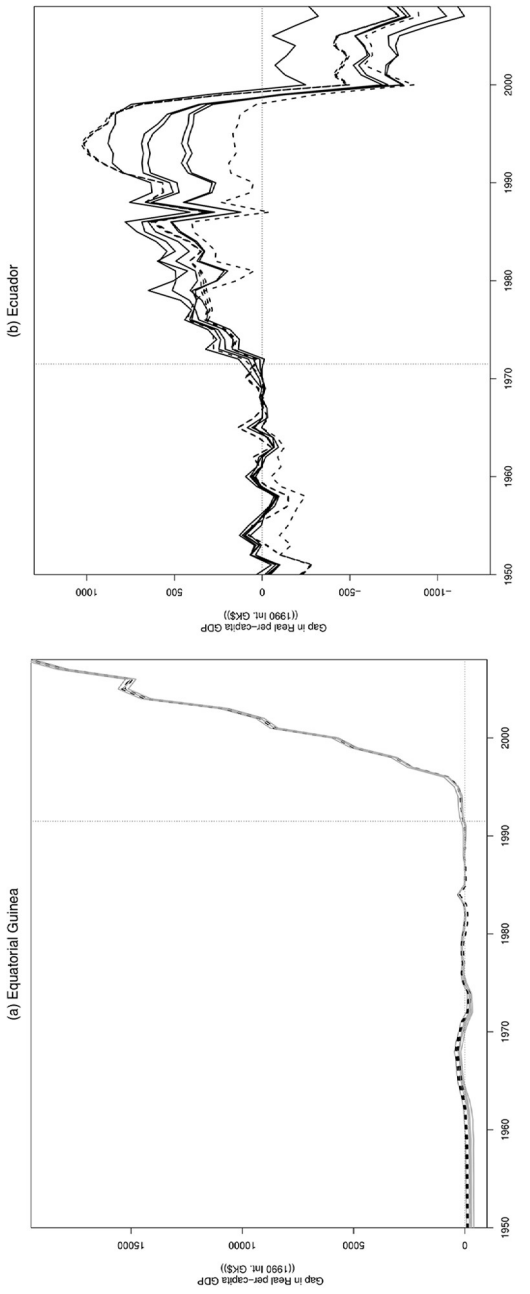
We estimate the impact of major oil and gas discoveries on GDP per capita using the synthetic control method with 14 different specifications. Specifically, we test the same seven specifications from the main paper and, for each one of them, we either include two covariates or not.[44,45]

Table F1 shows the p-value and our goodness of fit measure for each specification and each country. On the one hand, the results for Equatorial Guinea are robust to specification searching, since all specifications provide treatment effect estimates that are significant at the 5 percent level. On the other hand, the results for Ecuador show that the researcher could try different specifications and pick one whose result is significant. In particular, all 14 specifications have a good fit ($\tilde{R}^2 > 0.80$), but only two of them are significant (specifications 4a and 6a), implying that the researcher could, potentially, report a false-positive result.[46]

If we believe that covariates are not relevant to explain GDP per capita, the recommended specification uses all pre-treatment outcome lags. Note that specification 1a indicates that the treatment is significant for Equatorial Guinea, but not significant for Ecuador.

However, if we believe that the SC unit should also match the covariates, then we should focus only on specifications that satisfy the conditions outlined in Appendix D by dropping specifications 6 and 7. By looking at Table F1, a sensible conclusion would be that major oil and gas discoveries had a significant effect on Equatorial Guinea's GDP per capita even though there is no evidence of such effect on Ecuador's GDP per capital. Figure F1 shows that this conclusion is reasonable since, in the case of Equatorial Guinea, we find that all specifications with a good fit have estimates of similar magnitude while, in the case of Ecuador, our results vary widely across specifications. The next step is to test the null hypothesis using a test statistic that combines the test statistics of specifications 1 through 5. Restricting ourselves to specifications with good fit ($\tilde{R}^2 > 0.80$), we find that the p-value of a test that uses the mean of the RMSPE statistic across specifications (Imbens & Rubin, 2015), is equal to 0.031 and 0.308 for Equatorial Guinea and Ecuador, corroborating our conclusion that the treatment effect is positive in the first case and zero in the second

[43] Following the best practices in terms of transparency and replicability, Smith (2015) made his dataset and replication files available online (http://www.brockdsmith.com/research.html).
[44] We follow Smith (2015) and consider for this exercise different specifications using only seven years of pre-treatment data in the first minimization problem (equation 1) while accounting for the entire pre-treatment period in the second minimization problem (equation 2). Had we considered only seven years of pre-treatment data in the second step, we would reach similar conclusions to the ones in the main text. Had we considered the same specifications using the full pre-treatment data in the first step, then we would fail to reject the null for all specifications. This is consistent with our result that the variation between specifications that use pre-treatment outcome lags as economic predictors diminishes when the number of pre-treatment periods increases. Results are available upon request.
[45] The included covariates are ethnic fragmentation and population size one year before the discovery.
[46] The specification considered by Smith (2015) does not reject the null.



Notes: Gray lines have $\tilde{R}^2 \leq 0.80$, dashed lines have $0.80 < \tilde{R}^2 \leq 0.95$, and solid black lines have $\tilde{R}^2 > 0.95$, where $\tilde{R}^2$ is defined by equation (D.1). The vertical lines denote the beginning of the post-treatment period.

Figure F1. Treatment Effects for Specifications 1 through 5—Database from Smith (2015).

15206688, 2020, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/pam.22206 by Ecole Normale Supérieure de Paris, Wiley Online Library on [08/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table F1. Specification searching—database from Smith (2015).

	Equatorial Guinea		Ecuador	
	p-value (1)	$\tilde{R}^2$ (2)	p-value (3)	$\tilde{R}^2$ (4)
(1a)	0.031	0.797	0.385	0.975
(1b)	0.031	0.866	0.538	0.881
(2a)	0.031	0.832	0.308	0.975
(2b)	0.031	0.777	0.538	0.881
(3a)	0.031	0.790	0.231	0.972
(3b)	0.031	0.809	0.615	0.880
(4a)	0.031	0.536	0.077	0.970
(4b)	0.031	0.891	0.308	0.969
(5a)	0.031	0.744	0.769	0.804
(5b)	0.031	0.828	0.538	0.881
(6a)	0.031	0.657	0.077	0.972
(6b)	0.031	0.848	0.538	0.804
(7a)	0.031	0.671	0.231	0.955
(7b)	0.031	0.849	0.692	0.838
# of Countries	33		13	

Notes: We analyze 14 different specifications. The number of the specifications refers to: (1) all pre-treatment outcome values, (2) the first three-quarters of the pre-treatment outcome values, (3) the first half of the pre-treatment outcome values, (4) odd pre-treatment outcome values, (5) even pre-treatment outcome values (original specification by Smith, 2015), (6) the mean of all pre-treatment outcome values, and (7) three outcome values. Specifications that end with an *a* do not include covariates, while specifications that end with a *b* include the covariates ethnic fragmentation and population size one year before the discovery. Our measure of goodness of fit is defined by equation (D.1).

one. Now, following Christensen and Miguel (2018) and Cohen-Cole et al. (2009), Figure F2 shows the average treatment effects across specifications 1 through 5 as a black line, suggesting a strongly positive effect for Equatorial Guinea and a zero effect for Ecuador. Now, following Firpo and Possebom (2018), we invert tests based on the mean of the RMSPE statistic across specifications 1 through 5 to compute confidence sets for the treatment effect over time. Our confidence sets (Figure F2) include all treatment effect functions that we fail to reject using this combined test statistic, considering functions that are deviations from the average treatment effect across specifications by an additive and constant factor. Analyzing Sub figure F2a, we see that, although we cannot reject treatment effect functions that are initially negative, all treatment effect functions in our confidence sets increase very fast, becoming positive after a few years of treatment. For Ecuador (Sub figure F2b), we find that our confidence set includes a zero effect for almost all years after the beginning of treatment, suggesting that the discovery of oil and gas in Ecuador had almost no impact on per-capita GDP. Finally, we apply the choice criterion suggested by Dube and Zipperer (2015) and Donohue et al. (2018), restricting ourselves to specifications 1 through 5 that have a good pre-treatment fit. The first criterion picks specification 2a for Equatorial Guinea (in this case, we would reject the null with a p-value of 0.031) and specification 5a for Ecuador (in this case, we would not reject the null), while the second one picks specification 4b for Equatorial Guinea (in this case, we would reject the null with a p-value of 0.031) and specification 1a for Ecuador (in this case, we would not reject the null).

After this analysis, a reasonable conclusion would be that there is a significant and positive effect for Equatorial Guinea and a null effect for Ecuador.

15206688, 2020, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/pam.22206 by Ecole Normale Supérieure de Paris, Wiley Online Library on [08/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Notes: We compute confidence sets by inverting the average test statistic across specifications that satisfy $\tilde{R}^2 > 0.80$, where $\tilde{R}^2$ is defined by equation (D.1). Our confidence sets include all treatment effect functions that we fail to reject using this combined test statistic, considering functions that are deviations from the average treatment effect across specifications by an additive and constant factor. The black line is the average treatment effect of the treated country and the gray area is the confidence set. The vertical lines denote the beginning of the post-treatment period.

Figure F2. Ninety Percent Confidence Sets Around the Average Across Specifications 1 through 5—Database from Smith (2015).

15206688, 2020, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/pam.22206 by Ecole Normale Supérieure de Paris, Wiley Online Library on [08/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

15206688, 2020, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/pam.22206 by Ecole Normale Supérieure de Paris, Wiley Online Library on [08/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table F2. Specification searching—database from Abadie, Diamond, and Hainmueller (2010).

Specification	(1a)	(1b)	(2a)	(2b)	(3a)	(3b)	(4a)	(4b)
p-value	0.077	0.077	0.077	0.077	0.051	0.026	0.051	0.026
$\tilde{R}^2$	0.979	0.979	0.969	0.974	0.978	0.978		
Specification	(5a)	(5b)	(6a)	(6b)	(7a)	(7b)		
p-value	0.077	0.077	0.077	0.077	0.077	0.026		
$\tilde{R}^2$	0.979	0.979	0.525	0.828	0.909	0.975		

Notes: We analyze 14 different specifications. The number of the specifications refers to: (1) all pre-treatment outcome values, (2) the first three-fourths of the pre-treatment outcome values, (3) the first half of the pre-treatment outcome values, (4) odd pre-treatment outcome values, (5) even pre-treatment outcome values, (6) pre-treatment outcome mean, and (7) three outcome values (original specification by Abadie, Diamond, & Hainmueller, 2010). Specifications that end with an *a* do not include covariates, while specifications that end with a *b* include the covariates average retail price of cigarettes, per capita state personal income (logged), percentage of the population ages 15 through 24, and per capita beer consumption. Our measure of goodness of fit is defined by equation (D.1).

Empirical Application: Tobacco Control (Abadie, Diamond, & Hainmueller, 2010)

Abadie, Diamond, and Hainmueller (2010) evaluate the effect of Proposition 99, a large-scale tobacco control program that California implemented in 1988, on annual per-capita cigarette sales.[47] The pre- and post-treatment periods are 1970 through 1988 and 1989 through 2000. The donor pool includes 38 American states.

We reestimate the impact of Proposition 99 on California's annual per-capita cigarette sales using the synthetic control method with 14 different specifications. Specifically, we test the same seven specifications from the main paper and, for each one of them, we either include five covariates or not.[48] Specifications ending with *a* do not include covariates, while those ending with *b* include them. Specification 7b is the original one in Abadie, Diamond, and Hainmueller (2010).

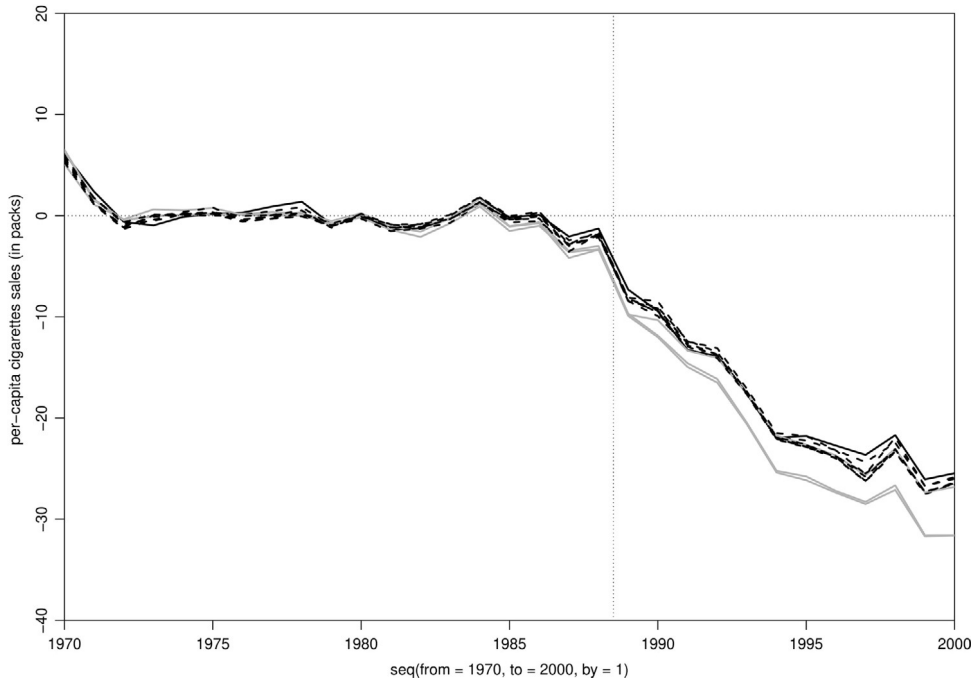
Table F2 shows the p-value and our goodness of fit measure for each specification. Note that quality of the fit varies widely across specifications: eight of them fit the data very closely ($\tilde{R}^2 \geq 0.975$), five of them have an intermediate value for our measure of goodness of fit ($0.80 < \tilde{R}^2 < 0.975$), and one of them fits the data very poorly ($\tilde{R}^2 \leq 0.80$). Most importantly, all specifications with good fit have significant estimates whose magnitude is similar according to Figure F3, although p-values vary from 0.026 (the p-value in the specification considered in Abadie, Diamond, & Hainmueller, 2010) to 0.077 depending on the specification.

If we believe that covariates are not relevant to explain GDP per capita, the recommended specification uses all pre-treatment outcome lags. Note that specification 1a indicated that the treatment effect is significant at the 10 percent significance level but not at the 5 percent significance level.

However, if we believe that the SC unit should also match the covariates, then we should focus only on specifications that satisfy the conditions outlined in the second section of the main paper by dropping specifications 6 and 7. By looking at Table F2, a sensible conclusion would be that the treatment is significant at least at the 10 percent level. To have a better understanding of the significance of the treatment effect, we test the null hypothesis using a test statistic that combines the

[47] Following the best practices in terms of transparency and replicability, Abadie, Diamond, and Hainmueller (2010) made their dataset and replication files available through the command synth in the software Stata.

[48] The included covariates are average retail price of cigarettes, per capita state personal income (logged), percentage of the population ages 15 through 24, and per capita beer consumption.

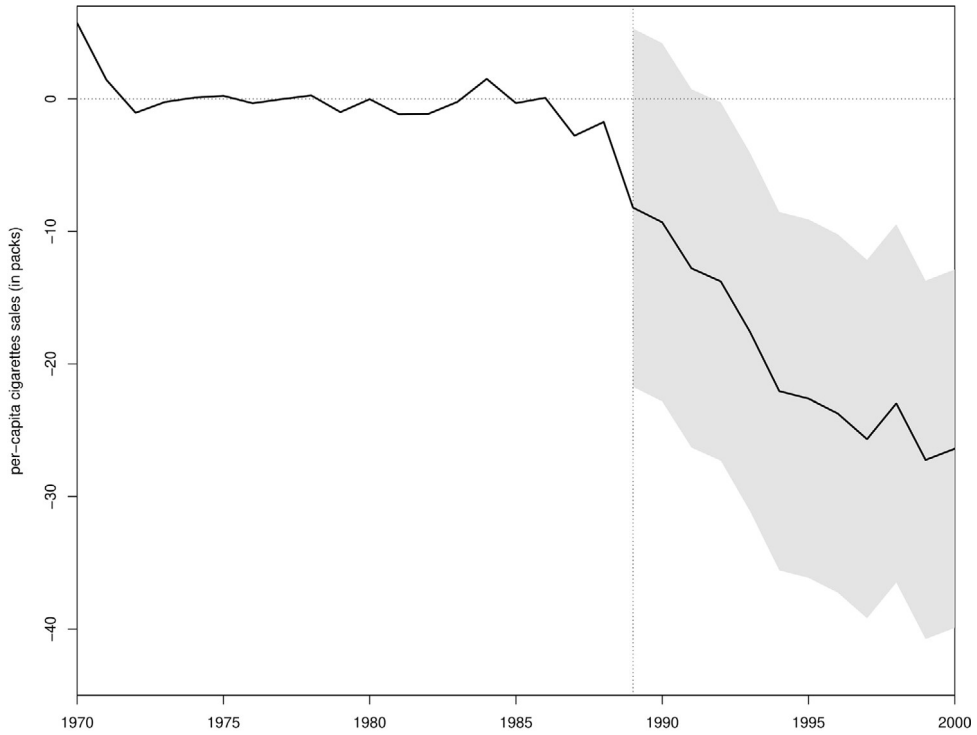


Notes: The solid black line is the original specification by Abadie, Diamond, and Hainmueller (2010), whose measure of goodness of fit is $\tilde{R}^2 = 0.975$, where $\tilde{R}^2$ is defined by equation (D.1). Gray lines have $\tilde{R}^2 \leq 0.975$ and dashed lines have $\tilde{R}^2 > 0.975$. The vertical line denotes the beginning of the post-treatment period.

Figure F3. Treatment Effects for Specifications and the Original Specification—Database from Abadie, Diamond, and Hainmueller (2010).

test statistics of all specifications. Restricting ourselves to specifications with a fit as good as the original specification ($\tilde{R}^2 \geq 0.975$), we find that the p-value of a test that uses the mean of the RMSPE statistic across specifications 1 through 5, as suggested by Imbens and Rubin (2015), is equal to 0.077, which is larger than the p-value of the original specification (0.026). Hence, the treatment effect is still significant even though the test statistic for California does not stand out as the largest one among all placebo runs as it does when we consider the original specification. Now, following Christensen and Miguel (2018) and Cohen-Cole et al. (2009), Figure F4 shows the average treatment effects across specifications 1 through 5 as a black line, suggesting a strongly negative effect in the long run. Now, following Firpo and Possebom (2018), we invert tests based on the mean of the RMSPE statistic across specifications to compute confidence sets for the treatment effect over time. Our confidence set includes all treatment effect functions that we fail to reject using this test, considering functions that are deviations from the average treatment effect across specifications by an additive and constant factor. Analyzing Figure F4, we see that, although we cannot reject treatment effect functions that are initially positive, all treatment effect functions in our confidence sets become negative after a few years of treatment, suggesting Proposition 99 eventually reduced tobacco consumption in California. Finally, we apply the choice criterion suggested by Dube and Zipperer (2015), restricting ourselves to specifications 1 through 5 that have a

15206688, 2020, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/pam.22206 by Ecole Normale Superieure de Paris, Wiley Online Library on [08/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License



Notes: We compute confidence sets by inverting the average test statistic across specifications that satisfy $\bar{R}^2 > 0.80$, where $\bar{R}^2$ is defined by equation (D.1). Our confidence sets include all treatment effect functions that we fail to reject using this combined test statistic, considering functions that are deviations from the average treatment effect across specifications by an additive and constant factor. The black line is the average treatment effect of California and the gray area is the confidence set. The vertical lines denote the beginning of the post-treatment period.

Figure F4. Ninety Percent Confidence Sets Around the Average Across Specifications 1 through 5—Database from Abadie, Diamond, and Hainmueller (2010).

good pre-treatment fit ($\tilde{R} \geq 0.975$). The choice criterion picks specification 4b (in this case, we would reject the null with a p-value of 0.026).

After this analysis, a reasonable conclusion is that the effect of California's tobacco control program is significantly different from zero, although the test statistic for California is not always the largest one among all placebo runs when we consider different specifications, even if we consider only specifications that provide a good pre-treatment fit.

15206688, 2020, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/pam.22206 by Ecole Normale Supérieure de Paris, Wiley Online Library on [08/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

15206688, 2020, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/pam.22206 by Ecole Normale Supérieure de Paris, Wiley Online Library on [08/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

APPENDIX G: SUPPLEMENTARY TABLES

Table G1. Specification searching—alternative models.

	Model (4) with $\epsilon_{j,t} \sim N(0,\ 1)$		Model (4) with K = 2	
	5% test (1)	10% test (2)	5% test (3)	10% test (4)
$T_0 = 12$	0.139	0.246	0.142	0.25
	(0.003)	(0.004)	(0.003)	(0.004)
$T_0 = 32$	0.132	0.235	0.147	0.247
	(0.003)	(0.004)	(0.003)	(0.004)
$T_0 = 100$	0.130	0.235	0.133	0.243
	(0.003)	(0.004)	(0.003)	(0.004)
$T_0 = 400$	0.119	0.218	0.129	0.230
	(0.003)	(0.004)	(0.003)	(0.004)

Notes: Rejection rates are estimated based on 10,000 observations and on seven specifications: (1) all pre-treatment outcome values, (2) the first three-quarters of the pre-treatment outcome values, (3) the first half of the pre-treatment outcome values, (4) odd pre-treatment outcome values, (5) even pre-treatment outcome values, (6) the mean of all pre-treatment outcome values, and (7) three outcome values. *z% test* indicates that the nominal size of the analyzed test is z percent and T_0 is the number of pre-treatment periods.

Table G2. Specification searching—excluding specifications 2 and 3.

	Stationary model		Non-stationary model	
	5% test (1)	10% test (2)	5% test (3)	10% test (4)
$T_0 = 12$	0.125	0.225	0.123	0.224
	(0.003)	(0.004)	(0.003)	(0.004)
$T_0 = 32$	0.131	0.232	0.138	0.251
	(0.003)	(0.004)	(0.004)	(0.005)
$T_0 = 100$	0.131	0.237	0.139	0.248
	(0.003)	(0.004)	(0.003)	(0.004)
$T_0 = 400$	0.127	0.23	0.138	0.245
	(0.003)	(0.004)	(0.004)	(0.005)

Notes: Rejection rates are estimated based on 10,000 observations and on five specifications: (1) all pre-treatment outcome values, (4) odd pre-treatment outcome values, (5) even pre-treatment outcome values, (6) the mean of all pre-treatment outcome values, and (7) three outcome values. *z% test* indicates that the nominal size of the analyzed test is z percent and T_0 is the number of pre-treatment periods.

Table G3. Infeasible test.

	Stationary model		Non-stationary model	
	5% test (1)	10% test (2)	5% test (3)	10% test (4)
Panel A: Including All Specifications				
$T_0 = 12$	0.201	0.344	0.192	0.330
	(0.004)	(0.005)	(0.004)	(0.005)
$T_0 = 32$	0.176	0.308	0.185	0.320
	(0.004)	(0.005)	(0.005)	(0.006)
$T_0 = 100$	0.155	0.274	0.167	0.291
	(0.004)	(0.005)	(0.004)	(0.005)
$T_0 = 400$	0.134	0.240	0.152	0.266
	(0.004)	(0.005)	(0.005)	(0.006)
Panel B: Excluding Specifications 6 and 7				
$T_0 = 12$	0.152	0.266	0.146	0.259
	(0.003)	(0.004)	(0.003)	(0.004)
$T_0 = 32$	0.130	0.231	0.132	0.234
	(0.003)	(0.004)	(0.004)	(0.005)
$T_0 = 100$	0.102	0.184	0.105	0.191
	(0.003)	(0.004)	(0.003)	(0.004)
$T_0 = 400$	0.078	0.148	0.083	0.154
	(0.003)	(0.004)	(0.004)	(0.005)

Notes: This table presents results for the infeasible test. This test is based on the true distribution of the test statistics in our Monte Carlos. In panel A, rejection rates are estimated based on 10,000 observations and on seven specifications: (1) all pre-treatment outcome values, (2) the first three-quarters of the pre-treatment outcome values, (3) the first half of the pre-treatment outcome values, (4) odd pre-treatment outcome values, (5) even pre-treatment outcome values, (6) the mean of all pre-treatment outcome values, and (7) three outcome values. In panel B, rejection rates are estimated based on 10,000 observations and excluding the last two specifications. *z% test* indicates that the nominal size of the analyzed test is z percent and T_0 is the number of pre-treatment periods.

APPENDIX REFERENCES

Adhikari, B., & Alm, J. (2016). Evaluating the economic effects of flat tax reforms using synthetic control methods. Southern Economic Journal, 83, 437–463.

Bertrand, M., Duflo, E., & Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? Quarterly Journal of Economics, 119, 249–275.

Botosaru, I., & Ferman, B. (2019). On the role of covariates in the synthetic control method. The Econometrics Journal, 22, 117–130.

Ferman, B., & Pinto, C. (2019). Synthetic controls with imperfect pre-treatment fit. See arXiv preprint arXiv:1911.08521, 2019.

Newey, W. K. (1991). Uniform convergence in probability and stochastic equicontinuity. Econometrica, 59, 1161–1167.

Newey, W. K., & McFadden, D. (1994). Large sample estimation and hypothesis testing (Chapter 36). In Handbook of Econometrics, 4, 2111–2245. Amsterdam, Netherlands: Elsevier.

Ok, E. A. (2007). Real analysis with economic applications. Princeton, NJ: Princeton University Press.

Stokey, N. L., & Lucas, R. E. (1989). Recursive methods in economic dynamics. Cambridge, MA: Harvard University Press.

15206688, 2020, 2, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/pam.22206 by Ecole Normale Supérieure de Paris, Wiley Online Library on [08/03/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License