Working Paper | 420 Junho de 2016

FGV | SAO PAULO SCHOOL OF ECONOMICS

Cherry Picking with Synthetic Controls

Bruno Ferman
Cristine Pinto
Vitor Possebom

Escola de Economia de São Paulo da Fundação Getulio Vargas FGV-EESP
www.eesp.fgv.br

Cherry Picking with Synthetic Controls

Bruno Ferman[*]
Sao Paulo School of Economics - FGV

Cristine Pinto[†]
Sao Paulo School of Economics - FGV

Vitor Possebom[‡]
Yale University

June, 2016

Abstract

The synthetic control (SC) method has been recently proposed as an alternative method to estimate treatment effects in comparative case studies. Abadie et al. [2010] and Abadie et al. [2015] argue that one of the advantages of the SC method is that it imposes a data-driven process to select the comparison units, providing more transparency and less discretionary power to the researcher. However, an important limitation of the SC method is that it does not provide clear guidance on the choice of predictor variables used to estimate the SC weights. We show that such lack of specific guidances provides significant opportunities for the researcher to search for specifications with statistically significant results, undermining one of the main advantages of the method. Considering six alternative specifications commonly used in SC applications, we calculate in Monte Carlo simulations the probability of finding a statistically significant result at 5% in at least one specification. We find that this probability can be as high as 13% (23% for a 10% significance test) when there are 12 pre-intervention periods and decay slowly with the number of pre-intervention periods. With 230 pre-intervention periods, this probability is still around 10% (18% for a 10% significance test). We show that the specification that uses the average pre-treatment outcome values to estimate the weights performed particularly bad in our simulations. However, the specification-searching problem remains relevant even when we do not consider this specification. We also show that this specification-searching problem is relevant in simulations with real datasets looking at placebo interventions in the Current Population Survey (CPS). In order to mitigate this problem, we propose a criterion to select among SC different specifications based on the prediction error of each specifications in placebo estimations.

Keywords: inference; synthetic control; p-hacking; specification searching; publication bias

JEL Codes: C12; C21; C33

[*] bruno.ferman@fgv.br
[†] cristine.pinto@fgv.br
[‡] vitoraugusto.possebom@yale.edu

1 Introduction

The synthetic control (SC) method has been recently proposed in a series of seminal papers by Abadie and Gardeazabal [2003], Abadie et al. [2010], and Abadie et al. [2015] as an alternative method to estimate treatment effects in comparative case studies. In situations in which there is only one treated unit and many control units, SC provides a way to estimate the counterfactual for the treated unit using a data-driven weighted average of the outcomes for the control units. Since then, SC has been used in a wide range of applications, including the evaluation of the impact of terrorism, civil wars and political risk, natural resources and disasters, international finance, education and research policy, health policy, economic and trade liberalization, political reforms, labor, taxation, crime, social connections, and local development.[1]

Abadie et al. [2010] and Abadie et al. [2015] describe many advantages of the synthetic control estimator over techniques traditionally used in comparative studies. According to them, an important advantage of this method is that it provides a transparent way to choose comparison units. It reduces researchers' discretionary power by imposing a data-driven process that computes the control unit that is most similar to the treated unit. They also argue that, since estimation of the synthetic control 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. Consequently, the synthetic control method would be less susceptible to specification searching. Given the growing debate on the importance of transparency in social science research (Miguel et al. [2014]), this could be an important advantage of the SC method.

However, an important limitation of the synthetic control method is that it does not provide clear guidance on the choice of predictor variables should be used to estimate the synthetic control weights. Abadie et al. [2010] define vectors of linear combinations of pre-intervention outcomes that could be used as predictors. They give examples where these linear combinations yield the value of the outcome variable in the period immediately prior to the intervention or the average of the outcome variable in the pre-treatment period. However, there is no guidance about which linear combinations should be used. They also suggest that, if the number of pre-intervention periods is large enough, researchers may divide them into an initial training period and a subsequent validation period. However, there is again no guidance determining when

[1] SC has been used in the evaluation of the impact of terrorism, civil wars and political risk (Abadie and Gardeazabal [2003], Bove et al. [2014], Li [2012], Montalvo [2011], Yu and Wang [2013]), natural resources and disasters (Barone and Mocetti [2014], Cavallo et al. [2013], Coffman and Noy [2011], DuPont and Noy [2012], Mideksa [2013], Sills et al. [2015], Smith [2015]), international finance (Jinjarak et al. [2013], Sanso-Navarro [2011]), education and research policy (Belot and Vandenberghe [2014], Chan et al. [2014], Hinrichs [2012]), health policy (Bauhoff [2014], Kreif et al. [2015]), economic and trade liberalization (Billmeier and Nannicini [2013], Gathani et al. [2013], Hosny [2012]), political reforms (Billmeier and Nannicini [2009], Carrasco et al. [2014], Dhungana [2011] Ribeiro et al. [2013]), labor (Bohn et al. [2014], Calderon [2014]), taxation (Kleven et al. [2013], de Souza [2014]), crime (Pinotti [2012a], Pinotti [2012b], Saunders et al. [2014]), social connections (Acemoglu et al. [2013]), and local development (Ando [2015], Gobillon and Magnac [2016], Kirkpatrick and Bennear [2014], Liu [2015], Severnini [2014]).

a researcher should follow this strategy and on how the pre-intervention periods should be divided into these two periods[2]. Such lack of consensus on how to implement the synthetic control method translates into a wide variety of specification choices in empirical applications of this method.[3] If different specifications result in widely different choices of the synthetic control unit, then a researcher would have relevant opportunities to select "statistically significant" specifications even when there is no effect. Since a researcher would usually not be able to commit to a specific specification before knowing how these decisions would affect the conclusion of his study, this flexibility may undermine one of the main advantages of the SC method.[4]

In this paper, we evaluate the extent to which this variety of options in the synthetic control method creates opportunities for specification searching considering one particular dimension of this problem: the choice of which pre-treatment outcome values to include in the estimation of the synthetic control unit. Using Monte Carlo simulations, we calculate the probability that a researcher would find at least one specification that would lead him to reject the null at 5%. Considering six different specifications commonly used in SC applications[5], the probability of detecting a false positive in at least one specification can be as high as 13% when there are 12 pre-treatment periods (23% if we consider a 10% significance test). If the variation in the synthetic control weights across different specifications vanishes when the number of pre-treatment periods is large, then we would expect this probability to get close to 5% in applications with a large number of pre-treatment periods.[6] We do find that the possibility of specification searching decreases with the number of pre-treatment periods. However, even with 230 pre-treatment periods, we still find a probability of around 10% that at least one specification is significant at 5% (18% if we consider a 10% significance test). These results suggest that even with a large number of pre-treatment periods, different specifications can lead to significantly different synthetic control units, generating substantial opportunities for specification searching. We find simular results in placebo simulations using the Current Population Survey (CPS).

[2] Moreover, Abadie et al. [2015] are not clear on how to compute their recommended test statistic when the pre-intervention period is divided in a training period and in a validation period.

[3] For example, Abadie and Gardeazabal [2003], Abadie et al. [2015] and Kleven et al. [2013] use the mean of all pre-treatment outcome values as predictor; Billmeier and Nannicini [2013], Bohn et al. [2014], Gobillon and Magnac [2016], Hinrichs [2012] use all the pre-treatment outcome values; Smith [2015] selects 4 out of 10 pre-treatment periods; Abadie et al. [2010] select 3 out of 19 pre-treatment periods; and Montalvo [2011] uses only the last two pre-treatment outcome values.

[4] Olken [2015] and Coffman and Niederle [2015] evaluate the use of pre-analysis plans in social sciences. For randomized control trials (RCT), the American Economic Association (AEA) launched a site to register experimental designs. However, there is no site where one would be able to register a prospective synthetic control study. Moreover, in many synthetic control applications both pre- and post-intervention information would be available to the researcher before the possibility of registering the study. In this case, it would be unfeasible to commit to a particular specification.

[5] We consider (1) the mean of all pre-treatment outcome values, (2) all pre-treatment outcome values, (3) the first half of the pre-treatment outcome values, (4) the first three quarters of the pre-treatment outcome values, (5) odd pre-treatment outcome values, and (6) even pre-treatment outcome values. Note that these six specifications do not exhaust all specification options that have been considered in SC applications.

[6] In this case, all different specifications would provide roughly the same synthetic control unit and, therefore, the same treatment effect estimate.

The data-generating process (DGP) in our MC slmulations also provides a way to measure the extent to which different specifications assign positive weight to control units that should not be considered in the synthetic control unit. 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 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. The specification that uses the mean of all pre-treatment outcome values as predictor misallocates remarkably more weight when compared to alternative specifications. This result is not surprising given that, in our DGP, the expected value of the outcome variable is the same for all groups. Still, this result highlights that, when no covariates are used to compute the synthetic control unit, using the average of the pre-treatment outcome values might not capture the time-series dynamics of the groups, which is the main goal of the SC method. Excluding this specification, however, there is only a mechanical and marginal decrease in the probability of rejecting the null in at least one specification.

It is important to note that our results by no means imply that researchers that have implemented the synthetic control method did incur in specification searching. Given that SC is a relatively new method, there would not be enough papers to formally test for specification searching.[7] However, given the evidence that there is a high return for reporting "significant" results and that scientists tend to incur in p-hacking (Rosenthal [1979], Lovell [1983], De Long and Lang [1992], Simmons et al. [2011] and Simonsohn et al. [2014]), our findings raise important concerns about the synthetic control method. One possibility to mitigate this problem would be to require researchers applying the SC method to report results for different specifications. However, it is important to note that testing all the possible SC specifications separately would not provide a valid hypothesis test since there would not be a defined decision rule (see White [2000]). Our preferred solution is to adopt a mean squared prediction error (MSPE) criterion based on the estimated post-treatment effects in placebo estimations and, then, focus all the analysis on the specification that minimizes this criterion. If there is an agreement on how the SC specification should be selected, then the scope for p-hacking (in this dimension) would be limited.

Finally, we find some implementation issues when using the *Synth* package for R (Abadie et al. [2011a])

[7] Brodeur et al. [2016] analyzes 641 articles (providing more than 50,000 tests) published in the *American Economic Review*, the *Journal of Political Economy*, and the *Quarterly Journal of Economics*. They identify a residual in the distribution of tests that cannot be explained solely by journals favoring rejection of the null hypothesis. Simonsohn et al. [2014] suggest the use of the P-curve as a way to distinguish between selective reporting findings and true effects. One of the requirements to the inference from p-curve to be valid is that we have great pool of studies from which we can select studies and p-values that test similar hypothesis. Given that the synthetic control estimator is a relatively recent method, there are less than 100 published papers that used this method even if we consider a wide range of journals. Therefore, it would be unfeasible to replicate these methodologies for synthetic control applications.

and the *synth* command for Stata, that are available on of the author's webpage (Abadie et al. [2011b]). When implementing the synthetic control method by solving a nested minimization problem (R's default and Stata's *nested* option), all of our placebo simulations using the CPS have not found a synthetic control unit for at least one of the observed states.[8] When using Stata's regression-based version of the Synthetic Control Method, we find that this procedure pays attention only to a small subset of the predictor variables. In particular, it only considers at most $J + 1$ predictor variables, where J is the number of control units, reducing the accuracy of the Synthetic Control Estimator.

2 Synthetic Control Method and Specification Search

Abadie and Gardeazabal [2003], Abadie et al. [2010] and Abadie et al. [2015] have recently developed the Synthetic Control Method in order to address counterfactual questions involving only one treated unit and a few control units. Intuitively, this method estimates the potential outcome of the treated unit if there were no treatment by constructing a weighted average of control units that is as similar as possible to the treated unit regarding the pre-treatment outcome variable and covariates. For this reason, this weighted average of control units is known as the synthetic control and treatment effects can be flexibly estimated for each post-treatment period. Below, we follow Abadie et al. [2010], explaining their estimator.

Suppose that we observe data for $(J + 1) \in \mathbb{N}$ units during $T \in \mathbb{N}$ time periods. Additionally, assume that there is a treatment that affects only unit 1[9,10] from period $T_0 + 1$ to period T uninterruptedly, where $T_0 \in (1, T) \cap \mathbb{N}$. Let the scalar $Y^0_{j,t}$ be the potential outcome that would be observed for unit j in period t if there were no treatment for $j \in \{1, ..., J + 1\}$ and $t \in \{1, ..., T\}$. Let the scalar $Y^1_{j,t}$ be the potential outcome that would be observed for unit j in period t if unit j received the treatment from period $T_0 + 1$ to T. Define

$$\alpha_{j,t} := Y^1_{j,t} - Y^0_{j,t} \tag{1}$$

as the treatment effect for unit j in period t and $D_{j,t}$ as a dummy variable that assumes value 1 if unit j is treated in period t and value 0 otherwise. With this notation, we have that the observed outcome for unit

[8] This issue may be caused by two different problems: (i) the theoretical literature have not yet elaborated a sufficient condition for the existence of the synthetic control estimator — in parallel with the rank condition for the OLS estimator — and there may be no solution for some datasets; and/or (ii) SC objective function may be approximately flat in some regions of the parameter space, making it hard to find a solution using iterated process.

[9] We can assume that there is only one treated unit without loss of generality, because, if there is more than one treated unit, we can look to their weighted average, where the weights are given, for example, by their population share as suggested by Imbens and Wooldridge [2009].

[10] The control units $j \in \{2, \cdots, J + 1\}$ constitute the donor pool.

j in period t is given by

$$Y_{j,t} := Y^0_{j,t}\left(1 - D_{j,t}\right) + Y^1_{j,t} D_{j,t}.$$

Since only the first unit receives the treatment from period $T_0 + 1$ to T, we have that:

$$D_{j,t} := \begin{cases} 1 & \text{if } j = 1 \text{ and } t > T_0 \\ 0 & \text{otherwise} \end{cases}$$

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

Let $\mathbf{Y_j} := [Y_{j,1}...Y_{j,T_0}]'$ be the vector of observed outcomes for unit $j \in \{1, ..., J+1\}$ in the pre-treatment period and $\mathbf{X_j}$ 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 $\mathbf{Y_j}$.[11] Let also $\mathbf{Y_0} = [\mathbf{Y_2}...\mathbf{Y_{J+1}}]$ be a $(T_0 \times J)$-matrix and $\mathbf{X_0} = [\mathbf{X_2}...\mathbf{X_{J+1}}]$ be a $(F \times J)$-matrix.

Since we want to make unit 1's synthetic control as similar as possible to the actual unit 1, the Synthetic Control Estimator of $Y^0_{1,t}$ is given, for each $t \in \{1, ..., T\}$, by

$$\widehat{Y}^0_{1,t} := \sum_{j=2}^{J+1} \widehat{w}_j Y_{j,t}, \tag{2}$$

where $\widehat{\mathbf{W}} = [\widehat{w}_2...\widehat{w}_{j+1}]' := \widehat{\mathbf{W}}(\widehat{\mathbf{V}}) \in \mathbb{R}^J$ is given by the solution to a nested minimization problem:

$$\widehat{\mathbf{W}}(\mathbf{V}) := \arg\min_{\mathbf{W} \in \mathcal{W}} (\mathbf{X_1} - \mathbf{X_0}\mathbf{W})'\mathbf{V}(\mathbf{X_1} - \mathbf{X_0}\mathbf{W}) \tag{3}$$

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

$$\widehat{\mathbf{V}} := \arg\min_{\mathbf{V} \in \mathcal{V}} (\mathbf{Y_1} - \mathbf{Y_0}\widehat{\mathbf{W}}(\mathbf{V}))'(\mathbf{Y_1} - \mathbf{Y_0}\widehat{\mathbf{W}}(\mathbf{V})) \tag{4}$$

where $\mathcal{V}$ is the set of diagonal positive semidefinite matrix of dimension $(F \times F)$ whose trace equals one.

Intuitively, $\widehat{\mathbf{W}}$ is a weighting vector that measures the relative importance of each unit in the synthetic control of unit 1 and $\widehat{\mathbf{V}}$ measures the relative importance of each one of the F predictors. Consequently, this

[11] For example, if the outcome variable is a country's per capita GDP and $T_0 = 12$, $\mathbf{X_j}$ may contain the investment rate, some measures of human capital and institutional quality, population, and the average per capita GDP from 1 to 4, from 5 to 8 and from 9 to 12.

technique makes the synthetic control of unit 1 as similar as possible with the actual unit 1 considering the F predictors and the pre-intervention values of the outcome variable when we choose the Euclidean metric (or a reweighed version of it) to evaluate the distance between the observed variables for unit 1 and the values predicted by the Synthetic Control Method.

Abadie and Gardeazabal [2003], Abadie et al. [2010] and Abadie et al. [2015] propose two other ways to choose $\widehat{\mathbf{V}}$. The first and most simple one is to use subjective and previous knowledge about the relative importance of each predictor. Since, according to the authors, one of the advantages of the Synthetic Control Method is to make the choice of comparison groups in comparative case studies more objective, this method of choosing $\mathbf{V}$ is discouraged by those authors. Another choice method for $\widehat{\mathbf{V}}$ is to divide the pre-intervention period in two sub-periods: one training period and one validation period. While data from the training period are used to solve problem (3), data for the validation period are used to solve problem (4). Intuitively, this technique of cross-validation chooses matrix $\widehat{\mathbf{W}}(\widehat{\mathbf{V}})$ to minimize the out-of-sample prediction errors, an advantage when compared to the method described above. However, the cost of this improvement is the need of a longer pre-intervention period. Moreover, the Stata command made available by those authors also allows the researcher to use a regression-based method in order to compute matrix $\widehat{\mathbf{V}}$.[12] According to Kaul et al. [2015], this choice method estimates, for each time period $t \in \{1, ..., T_0\}$, a cross-section regression of $y_{j,t}$ on all F predictor variables, yielding regression coefficients $\widehat{\beta}_{t,f}$ for each $t \in \{1, ..., T_0\}$ and for each $f \in \{1, ..., F\}$. It then populates the diagonal of matrix $\widehat{\mathbf{V}}$ by imposing $v_{f,f} := \frac{\sum_{t=1}^{T_0} \widehat{\beta}_{t,f}^2}{\sum_{f=1}^{F} \sum_{t=1}^{T_0} \widehat{\beta}_{t,f}^2}$ for each $f \in \{1, ..., F\}$. A natural problem of this choice method is that it will allocate strictly positive weights $v_{f,f}$ for only $\widetilde{f} := \min\{F, J+1\}$ due to a degrees of freedom restriction, actually making it impossible to use many predictor variables — i.e., $F > J+1$.[13]

Finally, we define the Synthetic Control Estimator of $\alpha_{1,t}$ (or the estimated gap) as

$$\widehat{\alpha}_{1,t} := Y_{1,t} - \widehat{Y}_{1,t}^N \tag{5}$$

for each $t \in \{1, ..., T\}$.

Abadie et al. [2015] propose an inference procedure that consists in a straightforward permutation test. They permute which unit is assumed to be treated and estimate, for each $j \in \{2, ..., J+1\}$ and $t \in \{1, ..., T\}$, $\widehat{\alpha}_{j,t}$ as described above. Then, they compute the test statistic

[12] This method is not covered in Abadie and Gardeazabal [2003], Abadie et al. [2010] and Abadie et al. [2015].

[13] This situation arises in the empirical literature when $T_0 > J+1$ and the empirical researcher wants to use all pre-treatment outcome values as predictor variables, following, for example, Billmeier and Nannicini [2013].

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

where the acronym RMSPE stands for *ratio of the mean squared prediction errors*[14]. Moreover, they propose to calculate a p-value

$$p := \frac{\sum_{j=1}^{J+1} \mathbb{1}\left[RMSPE_j \geq RMSPE_1\right]}{J+1}, \tag{6}$$

where $\mathbb{1}[\diamond]$ is the indicator function of event $\diamond$, and reject the null hypothesis of no effect if p is less than some pre-specified significance level, such as the traditional value of 0.1.[15] Ferman and Pinto [2015] show that a test statistic that includes in the denominator only the pre-treatment periods not used in the estimation of the SC weights provides a better heteroskedasticity correction, although they also argue that such test could have low power if there are only few pre-treatment periods not included in the estimation of the SC weights.

Abadie et al. [2010] and Abadie et al. [2015] describes many advantages of the synthetic control estimator over techniques traditionally used in comparative studies. According to them, this method is a transparent way to choose comparison units, reducing researchers' discretionary power by imposing a data-driven process that computes the control unit that is most similar to the treated unit. They also argue that, since estimation of the synthetic control weights does not require access to post-intervention outcomes, researchers could decide on study design without knowing how those decisions would affect the conclusions of their studies. Consequently, the synthetic control method would be robust against specification searching, a common problem in many empirical areas as pointed out by Brodeur et al. [2016]. However, researchers frequently have access to post-intervention outcomes in the empirical contexts most common for synthetic control applications, creating an opportunity to search for significant results if there is publication bias.

Abadie et al. [2010] and Abadie et al. [2015] recommend several robustness checks for synthetic control applications: in-time placebo tests[16], leave-one-out tests in the comparison units dimension[17] and restricted synthetic control unit[18]. Reducing the donor pool size is another robustness check recommend by them that

[14] When dividing the pre-treatment period in a training period and in a validation period, the authors are not clear about whether the denominator of this test statistic should include all pre-treatment values or only the validation period outcome values.

[15] Firpo and Possebom [2016] detailedly discuss this inference procedure, formalizing and generalizing it.

[16] This test consists in assigning the beginning of the post-treatment period to a earlier period $t^* < T_0$ and look for a treatment effect before T_0. If there is one, there is evidence that estimated effect is not actually caused by the investigated treatment.

[17] Generally, the synthetic control method allocates positive weights to a small number of control units. This test consists in dropping one of the units that received a positive weight from the donor pool and reestimating the treatment effect. It aims to verify the influence of a particular unit in the estimated result.

[18] In order to avoid over-fitting, this test force the synthetic control method to allocate positive weights only to a fixed number of control units, mimicking the n nearest neighbors matching estimator.

many authors follow, e.g.: Ando [2015], Barone and Mocetti [2014], Kreif et al. [2015] and Mideksa [2013]. Despite all the cautionary notes by Abadie et al. [2010] and Abadie et al. [2015] about the importance of robustness checks in synthetic control applications, the literature is mute, to the best of our knowledge, about the specification-search possibilities generated by the choice of predictors variables[19]. Since any linear combination of pre-treatment outcome values can be an element of $\mathbf{X_j}$, a researcher may look for a significant estimate 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 outcome values should be included in $\mathbf{X_j}$: Abadie and Gardeazabal [2003], Abadie et al. [2015] and Kleven et al. [2013] use the mean of all pre-treatment outcome values; Smith [2015] uses Y_{j,T_0}, Y_{j,T_0-3}, Y_{j,T_0-5} and Y_{j,T_0-7}; Abadie et al. [2010] picks Y_{j,T_0}, Y_{j,T_0-8} and Y_{j,T_0-13}; Billmeier and Nannicini [2013], Bohn et al. [2014], Gobillon and Magnac [2016], Hinrichs [2012] use all pre-treatment outcome values; and Montalvo [2011] uses only the last two pre-treatment outcome values.[20] If different specifications result in wildly different SC estimators, then it would be possible to search for specifications that reject then null hypothesis.

3 Monte Carlo Simulations

In order to verify the possibility of specification search, we elaborate a Monte Carlo Experiment in which we generate 20,000 data sets and, for each one of them, test the null hypothesis of no effect adopting several different specifications. We compute the rejection rate of this simulation by counting how many data sets are associated to a rejected null hypothesis at the 5%-significance level for at least one specification. If different SC specifications lead to similar SC weights, then this rejection rate would be close to 5% and the risk of specification searching would be low.

Our data generating process (DGP) is given by:

$$Y_{j,t}^0 = \delta_t + \lambda_t \mu_j + \epsilon_{j,t} \tag{7}$$

We set the total number of units $J = 20$; λ_t is a $(K, 1)$ vector with $\lambda_t^k \sim N(0, 1)$; μ_j is such that the 20 units are divided into K groups where the units in each group assign $\mu_j^k = 1$ for one k and zero for the others[21]; $\epsilon_{j,t} \sim N(0, 1)$. Finally, we impose that there is no treatment effect, i.e., $Y_{j,t} = Y_{j,t}^0 = Y_{j,t}^1$ for each

[19] From now on, we call the choice of predictors variables of the synthetic control estimator as specification search because it mimics the choice of functional form in a linear model.

[20] By no means, we imply that those authors have incurred in specification search. We have only listed them as prominent examples of different choices regarding predictor variables.

[21] For example, with $K = 2$ we have $\lambda_t = (\lambda_t^1, \lambda_t^2)$, $\mu_j = (1, 0)$ for $j = 1, ..., 10$ and $\mu_j = (0, 1)$ for $j = 11, ..., 20$.

time period $t \in \{1, ..., T_0\}$.

We consider variations in the DGP along two dimensions:

- The number of pre-intervention periods: $T_0 \in \{12, 32, 100, 230\}$.
- The number of different groups whose units follow parallel trends: $K \in \{2, 10\}$.

We test the null hypothesis of no effect at the 5%-significance level adopting the following six specifications that differ only in the linear combinations of pre-treatment outcome values used as predictors:

1. Pre-treatment outcome mean: $\mathbf{X}_j = \left[\sum_{t=1}^{T_0} Y_{j,t}/T_0\right]$
2. All pre-treatment outcome values: $\mathbf{X}_j = [Y_{j,1}, \cdots, Y_{j,T_0}]'$
3. The first half of the pre-treatment outcome values: $\mathbf{X}_j = \left[Y_{j,1}, \cdots, Y_{j,T_0/2}\right]'$
4. The first three fourths of the pre-treatment outcome values[22]: $\mathbf{X}_j = \left[Y_{j,1}, \cdots, Y_{j,3T_0/4}\right]'$
5. Odd pre-treatment outcome values: $\mathbf{X}_j = \left[Y_{j,1}, Y_{j,3}, \cdots, Y_{j,(T_0-3)}, Y_{j,(T_0-1)}\right]'$
6. Even pre-treatment outcome values: $\mathbf{X}_j = \left[Y_{j,2}, Y_{j,4}, \cdots, Y_{j,(T_0-2)}, Y_{j,T_0}\right]'$

We run our simulations in Stata using the *synth* module (Abadie et al. [2011b]). We do not use the *nested* option and we set the diagonal matrix $\mathbf{V}$ as the identity matrix. All of our MC results are very similar if we compare to simulations using the command in R (Abadie et al. [2011a]), which uses a fully nested optimization procedure to determine the optimal $\mathbf{V}$.[23] We focus on the non-nested minimization problem because in the placebo simulations using CPS that we run in Stata with the *nested* option, the nested minimization problem failed to converge. More importantly, the default option in the Stata command without the *nested* option uses a data-driven regression based method to obtain the variable weights contained in the matrix $\mathbf{V}$. However, we found that the number of non-zero entries in the diagonal matrix $\mathbf{V}$ that results from this procedure is always equal to $min\{F, J+1\}$. This implies that, when T_0 grows, this method essentially discards part of the information that should be used to estimate the SC weights. For example, with $J+1 = 20$ and considering the specification that uses all pre-treatment outcome values, this command would solve essentially the same minimization problem whether T_0 is equal to 32, 100 or 230. In all these cases, the weights would be based on the fit of only 20 out of the T_0 pre-treatment periods, so more pre-treatment periods would not translate

[22] When $T_0 = 230$, we use the smallest integer that is larger than $3T_0/4$ as the last period included in matrix $\mathbf{X}_j$

[23] Results available upon request.

into more accurate weights. Since all our specifications include pre-treatment lags of the outcome variable (except for specification 1 in which the choice of $\mathbf{V}$ is not relevant), we follow Bohn et al. [2014] and estimate the SC estimators setting the diagonal matrix $\mathbf{V}$ as the identity matrix, which implies that we consider equal weights for all lags.

For each specification, we run a permutation test using the Abadie et al. [2015] RMSPE test statistic and reject the null at 5%-significance level if the treated unit has the largest RMSPE among the 20 units. By construction, this leads to a 5% rejection rate when we look at each specification separately. We are interested, however, in the probability that we would reject the null at 5%-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 he incurs 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% as well. However, this number may be higher if the synthetic control weights depend on specification choices.

We present in column 1 of Table 1 the probability of rejecting the null at 5% in at least one specification when there are 10 groups of 2 units each following different parallel trends. With $T_0 = 12$, a researcher considering these six different specification would be able to report a specification with statistically significant results at the 5% level with probability 13.3%. If we consider 10% significance tests, then the probability of rejecting the null in at least one specification would be up to 23.5%. Therefore, with few pre-treatment periods, a researcher would have substantial opportunities to select statistically significant specifications even when the null hypothesis is true. Note that it is not unusual to have SC applications with as few as 12 pre-intervention periods[24].

If the variation in the synthetic control weights across different specifications vanishes when the number of pre-treatment periods goes to infinity, then we would expect this probability to get close to 5% once the number of pre-treatment periods gets large. In this case, all different specifications would provide roughly the same synthetic control unit and, therefore, the same treatment effect estimate. We show in columns 1 and 2 of Table 1 the probabilities of rejecting the null in at least one specification is, as expected, decreasing with the number of pre-treatment periods. However, these probabilities decrease very slowly. Even if we consider a scenario with 230 pre-intervention periods, it would still be possible to reject the null in at least one specification 10% (18%) of the time for a 5% (10%) significance test. Therefore, specification searching

[24] See, for example, Abadie and Gardeazabal [2003], Kleven et al. [2013], Kreif et al. [2015], Smith [2015], Ando [2015], Liu [2015], Sills et al. [2015], Billmeier and Nannicini [2013], Bohn et al. [2014], Cavallo et al. [2013], Gobillon and Magnac [2016], Hinrichs [2012], Montalvo [2011], Li [2012] and Hosny [2012].

would remain a problem for the SC method unless the number of pre-intervention periods is remarkably large.[25] In columns 3 and 4 of Table 1 we present the results when there are 2 groups of 10 units each following different parallel trends. The probabilities of rejecting the null in at least one specification are similar to the previous case.

We present in Table 2 a measure of the variability in the allocation of weights across specifications. For each unit in the donor pool we look for the specifications that allocate the most and the least weight for this unit. Then we take the maximum value of this difference across units in the donor pool. With 12 pre-treatment periods, we find, on average, 0.67, which suggests that different specifications allocate wildly different weights across units in the donor pool. These numbers decreases very slowly with the number of pre-treatment periods, which justifies the fact that there are still significant possibilities for specification searching even when T_0 is large.

Our data-generating process (DGP) also provides a way to measure the extent to which different specifications assign positive weight to units in the donor pool that should not be considered in the synthetic control unit. 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. In Panel A of Table 3, we present the proportion of misallocated weights by the SC estimator for each specification when there are 10 groups of 2 units each following different parallel trends. Note that, in this case, there is only one unit that follows the same parallel trend as the treated unit, while all the other units follow different parallel trends. Therefore, the SC control method should allocate 100% of the weight to the unit that follows a parallel trend. The results presented in column (1) shows that the specification that uses the mean of all pre-treatment outcome values as predictor misallocates remarkably more weight when compared to alternative specifications. More specifically, almost 91% of the weights used to construct the SC unit is allocated to units that follow different trends relative to the treated unit. Also, this number does not decrease with the number of pre-intervention periods. This result is not surprising given that, in our DGP, the expected value of the outcome variable is the same for all groups. Still, this result highlights that, when there is no covariates, using the average of the pre-treatment outcome values might not capture the time-series dynamics of the groups, which is the main goal of the SC method. The other specifications do a slightly better job in assigning weights to the correct

[25] We stress that, in our review of the empirical literature, Saunders et al. [2014] use the largest pre-treatment period that we have found: 36 months.

unit in the donor pool. However, the share of misallocated weights is still higher than 50% even when there are 230 pre-intervention periods. In Panel B we present the results when there are 2 groups of 10 units each following different parallel trends. We find the same patterns, although the numbers are mechanically lower because now there are only 10 units (instead of 18) that should not receive positive weights.

Given that the specification that uses the average of the pre-intervention outcomes stands out by misallocating significantly more weights, we replicate our results on specification-searching possibilities and on the variability in the allocation of weights in Tables 4 and 5. The probability of rejecting the null in at least one specification decreases mechanically relative to the previous case, since we are considering only 5 instead of 6 specifications. We find that the probability of rejecting the null in at least one specification decreases more rapidly with T_0 once we discard specification 1, although there is still a probability of 7.6% of rejecting the null at 5% even when we have 230 pre-intervention periods (14.7% if we consider a 10% test). With $T_0 = 32$, which is more pre-treatment periods than available for most SC application, the probability of rejecting the null in at least one of the five specifications at 5% is around 10% (around 18% if we consider a 10% significance level). The variability in the allocation of weights also decreases substantially once we discard specification 1.

Finally, in Table 6 we present specification-searching possibilities using an alternative test statistic that uses in the denominator only the MSPE of lags not used in the estimation of the SC weights. As suggested in Ferman and Pinto [2015], this alternative test statistic provides a better heteroskedasticity correction with finite T_0 in MC simulations. The specification-searching possibilities are higher than with the original test statistic. The reason is that using only lags not used in the estimation of the SC weights induces more variability in the denominator, which increases the probability that different test statistics would provide different test results.

4 Simulations with Real Data

The results presented in Section 3 suggest that different specifications of the SC method can generate significant specification-searching opportunities. 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, and extract information on employment status and earnings for women between ages 25 and 50, following Bertrand et al. [2004]. We first consider simulations with 12 pre-intervention periods, 4 post-intervention periods, and 20 states. In each simulation, we randomly select 20 out of the 51 states (including Washington, D.C.) and then we randomly select the first periods between 1979 and 1999 and we extract 16 consecutive years of data. Then we consider simulations with 32 pre-intervention periods, 4 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 20,000 simulations using either employment or log wages as the dependent variable and test the null hypothesis using the same six specifications of section 3.

We present the probabilities of rejecting the null in at least one specification in Table 7. We present in Panel A results considering all 6 specifications defined in Section 3. In Panel B we exclude specification 1, while in Panel C we use the test statistic suggested in Ferman and Pinto [2015]. We use standard errors clustered at the level of the treated state in order to take into account that the simulations are not independent. The results are very similar to our finding in the MC simulations. In particular, with 32 pre-intervention periods, we find that the a researcher would have a probability higher than 10% of finding at least one specification that is statistically significant at 5% even excluding specification 1 (the probability of rejecting the null in at least one specification is around 20% if we consider 10% significance level tests). These results suggest that specification-searching possibilities in SC applications can be relevant in real applications of the method.

5 Recommendations

Our first recommendation is that researchers applying the SC should report results for different specifications. However, even if a researcher present results for all possible SC specifications with an hypothesis test for each specification, this would not provide a valid 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 and Wolf [2005]).[26] 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. Also, we would not have objectively a point estimate for the policy effect, as different specifications would yield different results. Therefore, it is important to provide a valid hypothesis testing procedure combined with a decision rule that determines which specification we should look at as suggested by White [2000].

We need to consider a criterion to choose among all possible specifications. One possibility is to choose the specification that minimizes the mean squared prediction error (MSPE) for the post-intervention period.

[26] When we adopt this decision rule, the rejection rate in our simulations for a 5%-significance level test in section 3 considering the six specifications is around 0.6% when $T_0 = 12$ and around 1.6% when $T_0 = 230$.

For each specification, we compute the SC estimator considering each unit j in the donor pool as the treated. Then we calculate for each specification s, unit $j \in \{2, .., J+1\}$ and time $t \in \{T_0 + 1, ..., T\}$ the predicted error:

$$Y_{j,t} - \widehat{Y}_{j,t}^{s} = Y_{j,t} - \sum_{l \neq j} \widehat{w}_l^{j,s} Y_{l,t} \tag{8}$$

where $l \in \{1, ..., J+1\} \setminus \{j\}$ indexes units in the control group that are used to construct the synthetic control and $\widehat{w}_l^{j,s}$ represents the SC weights using specification s when unit j is used as treated. Then we can calculate the MSPE of each specification by averaging across time and units in the donor pool:

$$MSPE(s) = \frac{1}{(T - T_0)J} \sum_{j=2}^{J+1} \sum_{t=T_0+1}^{T} \left(Y_{j,t} - \widehat{Y}_{j,t}^{s} \right)^2 \tag{9}$$

If the specification provides an accurate synthetic control estimator, we expect that the mean squared error of prediction should be close to zero, since the units in the control group were not affect by the intervention in the post-treatment period.

Our recommendation for the researcher is to choose the specification that minimizes the MSPE. If we have $S \in \mathbb{N}$ specifications, we choose the specification s that solves the following minimization problem:

$$min_{s \in S} \left[\frac{1}{(T - T_0)J} \sum_{j=2}^{J+1} \sum_{t=T_0+1}^{T} \left(Y_{j,t} - \widehat{Y}_{j,t}^{s} \right)^2 \right]$$

The idea of using the mean squared error criterion to choose the model specification is not new in the literature. In the time series literature, the mean squared error of forecast has been used to evaluate post-sample prediction performance (Tsurumi and Wago [1991]). while the out-of-sample mean squared prediction error has been used to evaluate the predictability of the models (Clark and West [2006]). In addition, the mean squared error of prediction is also used to compare alternative models in ecological and agronomic systems (Wallach and Goffinet [1989]).

Another possible solution to the potential specification-searching problem is suggested by Imbens and Rubin [2015]. After reporting results for all specifications, the researcher could construct a new test statistic as a function that combines all the test statistics and base his or her inference procedure on this new test statistic. One limitation of this approach is that it would not determine a point estimate for the effect.

6 Conclusion

We show that the lack of specific guidance on how to choose among different SC specification generates scope for specification searching. While the possibility for specification searching diminishes when the number of pre-intervention periods increases, the probability of rejecting the null hypothesis in at least one specification diminishes very slowly. The results from our MC simulations and simulations with the CPS suggest significant possibilities for specification searching even when one has information on 230 pre-intervention periods, which would imply almost 20 years of monthly data. Therefore, common SC applications should be susceptible to this specification-searching problem.

Our results point out that using the average of the pre-intervention average, which is a common specification in SC applications, does a poor job in capturing the time series dynamics of the data. However, there are still plenty of alternative specifications that a researcher could choose in SC applications even when we discard this specification. This lack of specific guidance in determining which economic predictors should be used is an important limitation of the SC method, and further research is needed in order to determine which specification should be chosen. We work in this direction by proposing a MSPE criterion based on the estimated post-treatment effects in placebo estimations.

Finally, we show that the choice of the weighting matrix V can be important, and that the commonly used Stata command without the *nested* option provides a choice for this matrix that should not be used in applications with a large number of pre-intervention periods. On the other hand, there is a high chance that the optimization program will not converge if one uses the *nested* command in Stata or the command in R. Therefore, it is also is important that applied researchers present all details in the implementation of the SC method, including which software was used and what was the choice for matrix V.

References

Alberto Abadie and Javier Gardeazabal. The Economic Costs of Conflict: A Case Study of the Basque Country. *American Economic Review*, 93(1):113–132, 2003.

Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program. *Journal of the American Statiscal Association*, 105(490):493–505, 2010.

Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Synth: An R Package for Synthetic Control Methods in Comparative Case Studies. *Journal of Statistical Software*, 42(13):1–17, 2011a.

Alberto Abadie, Alexis Diamond, and Jens Hainmueller. SYNTH: Stata module to implement Synthetic Control Methods for Comparative Case Studies. Statistical Software Components, Boston College Department of Economics, October 2011b. URL https://ideas.repec.org/c/boc/bocode/s457334.html.

Alberto Abadie, Alexis Diamond, and Jens Hainmueller. Comparative Politics and the Synthetic Control Method. *American Journal of Political Science*, 59(2):495–510, 2015.

Daron Acemoglu, Simon Johnson, Amir Kermani, James Kwak, and Todd Mitton. The Value of Connections in Turbulent Times: Evidence from the United States. NBER Working Paper 19701. Available at: http://www.nber.org/papers/w19701.pdf, December 2013.

Michihito Ando. Dreams of Urbanization: Quantitative Case Studies on the Local Impacts of Nuclear Power Facilities using the Synthetic Control Method. *Journal of Urban Economics*, 85:68–85, June 2015.

Guglielmo Barone and Sauro Mocetti. Natural Disasters, Growth and Institutions: a Tale of Two Earthquakes. *Journal of Urban Economics*, pages 52–66, 2014.

Sebastian Bauhoff. The Effect of School Nutrition Policies on Dietary Intake and Overweight: a Synthetic Control Approach. *Economics and Human Biology*, pages 45–55, 2014.

Michele Belot and Vincent Vandenberghe. Evaluating the Threat Effects of Grade Repetition: Exploiting the 2001 Reform by the French-Speaking Community of Belgium. *Education Economics*, 22(1):73–89, 2014.

Marianne Bertrand, Esther Duflo, and Sendhil Mullainathan. How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics*, page 24975, 2004.

Andreas Billmeier and Tommaso Nannicini. Trade Openness and Growth: Pursuing Empirical Glasnost. *IMF Staff Papers*, 56(3):447–475, 2009.

Andreas Billmeier and Tommaso Nannicini. Assessing Economic Liberalization Episodes: A Synthetic Control Approach. *The Review of Economics and Statistics*, 95(3):983–1001, 2013.

Sarah Bohn, Magnus Lofstrom, and Steven Raphael. Did the 2007 Legal Arizona Workers Act Reduce the State's Unauthorized Immigrant Population? *The Review of Economics and Statistics*, 96(2):258–269, 2014.

Vincenzo Bove, Leandro Elia, and Ron P. Smith. The Relationship between Panel and Synthetic Control Estimators on the Effect of Civil War. Working Paper, http://www.bbk.ac.uk/ems/research/BirkCAM/working-papers/BCAM1406.pdf, October 2014.

Abel Brodeur, Mathias Lé, Marc Sangnier, and Yanos Zylberberg. Star Wars: The Empirics Strike Back. *American Economic Journal: Applied Economics*, 8(1):1–32, 2016.

Gabriela Calderon. The Effects of Child Care Provision in Mexico. Working paper, http://goo.gl/YSEs9B., July 2014.

Vinicius Carrasco, Joao M. P. de Mello, and Isabel Duarte. A Década Perdida: 2003 – 2012. Texto para Discussão, http://www.econ.puc-rio.br/uploads/adm/trabalhos/files/td626.pdf, 2014.

Eduardo Cavallo, Sebastian Galiani, Ilan Noy, and Juan Pantano. Catastrophic Natural Disasters and Economic Growth. *The Review of Economics and Statistics*, 95(5):1549–1561, 2013.

Ho Fai Chan, Bruno S. Frey, Jana Gallus, and Benno Torgler. Academic Honors and Performance. *Labour Economics*, 31:188–204, 2014.

Todd E. Clark and Kenneth D. West. Using out-of-sample mean squared prediction errors to test the martingale difference hypothesis. *Journal of Econometrics*, 135(1-2):155 – 186, 2006. ISSN 0304-4076. doi: http://dx.doi.org/10.1016/j.jeconom.2005.07.014. URL http://www.sciencedirect.com/science/article/pii/S0304407605001648.

Lucas C. Coffman and Muriel Niederle. Pre-analysis plans have limited upside, especially where replications are feasible. *Journal of Economic Perspectives*, 29(3):81–98, 2015. doi: 10.1257/jep.29.3.81. URL http://www.aeaweb.org/articles.php?doi=10.1257/jep.29.3.81.

Makena Coffman and Ilan Noy. Hurricane Iniki: Measuring the Long-Term Economic Impact of Natural Disaster Using Synthetic Control. *Environment and Development Economics*, 17:187–205, 2011.

J Bradford De Long and Kevin Lang. Are all economic hypotheses false? *Journal of Political Economy*, pages 1257–1272, 1992.

Fernando Friaça Asmar de Souza. Tax Evasion and Inflation: Evidence from the Nota Fiscal Paulista Program. Master's thesis, Pontifícia Universidade Católica, March 2014. Available at http://www.dbd.puc-rio.br/pergamum/tesesabertas/1212327_2014_completo.pdf.

Sandesh Dhungana. Identifying and Evaluating Large Scale Policy Interventions: What Questions Can We Answer? Available at: https://openknowledge.worldbank.org/bitstream/handle/10986/3688/WPS5918.pdf?sequence=1., December 2011.

William DuPont and Ilan Noy. What Happened to Kobe? A Reassessment of the Impact of the 1995 Earthquake in Japan. Available at: http://www.economics.hawaii.edu/research/workingpapers/WP_12-4.pdf., March 2012.

Bruno Ferman and Cristine Pinto. Inference in Differences-in-Differences with Different Group Sizes. Working Paper, October 2015.

Sergio Firpo and Vitor Possebom. Synthetic Control Estimator: A Generalized Inference Procedure and Confidence Sets. Working Paper, https://goo.gl/oQTX9c, April 2016.

Sachin Gathani, Massimiliano Santini, and Dimitri Stoelinga. Innovative Techniques to Evaluate the Impacts of Private Sector Developments Reforms: An Application to Rwanda and 11 other Countries. Working Paper, https://blogs.worldbank.org/impactevaluations/files/impactevaluations/methods_for_impact_evaluations_feb06-final.pdf, February 2013.

Laurent Gobillon and Thierry Magnac. Regional Policy Evaluation: Interative Fixed Effects and Synthetic Controls. *Review of Economics and Statistics*, 2016. Forthcoming.

Peter Hinrichs. The Effects of Affirmative Action Bans on College Enrollment, Educational Attainment, and the Demographic Composition of Universities. *Review of Economics and Statistics*, 94(3):712–722, March 2012.

Amr Sadek Hosny. Algeria's Trade with GAFTA Countries: A Synthetic Control Approach. *Transition Studies Review*, 19:35–42, 2012.

Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social and Biomedical Sciences: An Introduction*. Cambridge University Press, United Kingdom, 1st edition, 2015.

Guido W. Imbens and Jeffrey M. Wooldridge. Recent Developments in the Econometrics of Program Evaluation. *Journal of Economic Literature*, 47(1):5–86, 2009.

Yothin Jinjarak, Ilan Noy, and Huanhuan Zheng. Capital Controls in Brazil — Stemming a Tide with a Signal? *Journal of Banking & Finance*, 37:2938–2952, 2013.

Ashok Kaul, Stefan Klöbner, Gregor Pfeifer, and Manuel Schieler. Synthetic Control Methods: Never Use All Pre-Intervention Outcomes as Economic Predictors. Working Paper. Available at: http://www.oekonometrie.uni-saarland.de/papers/SCM_Predictors.pdf., May 2015.

A. Justin Kirkpatrick and Lori S. Bennear. Promoting Clean Enery Investment: an Empirical Analysis of Property Assessed Clean Energy. *Journal of Environmental Economics and Management*, 68:357–375, 2014.

Henrik Jacobsen Kleven, Camille Landais, and Emmanuel Saez. Taxation and International Migration of Superstars: Evidence from European Football Market. *American Economic Review*, 103(5):1892–1924, 2013.

Noémi Kreif, Richard Grieve, Dominik Hangartner, Alex James Turner, Silviya Nikolova, and Matt Sutton. Examination of the Synthetic Control Method for Evaluating Health Policies with Multiple Treated Units. *Health Economics*, 2015.

Qi Li. Economics Consequences of Civil Wars in the Post-World War II Period. *The Macrotheme Review*, 1 (1):50–60, 2012.

Shimeng Liu. Spillovers from Universities: Evidence from the Land-Grant Program. *Journal of Urban Economics*, 87:25–41, 2015.

Michael Lovell. Data Mining. *The Review of Economics and Statistics*, 65(1):1–12, 1983.

Torben K. Mideksa. The Economic Impact of Natural Resources. *Journal of Environmental Economics and Management*, 65:277–289, 2013.

E. Miguel, C. Camerer, K. Casey, J. Cohen, K. M. Esterling, A. Gerber, R. Glennerster, D. P. Green, M. Humphreys, G. Imbens, D. Laitin, T. Madon, L. Nelson, B. A. Nosek, M. Petersen, R. Sedlmayr, J. P. Simmons, U. Simonsohn, and M. Van der Laan. Promoting transparency in social science research. *Science*, 343(6166):30–31, 2014. ISSN 0036-8075. doi: 10.1126/science.1245317. URL http://science.sciencemag.org/content/343/6166/30.

José G. Montalvo. Voting after the Bombings: A Natural Experiment on the Effect of Terrorist Attacks on Democratic Elections. *Review of Economics and Statistics*, 93(4):1146–1154, 2011.

Benjamin A. Olken. Promises and perils of pre-analysis plans. *Journal of Economic Perspectives*, 29(3): 61–80, 2015. doi: 10.1257/jep.29.3.61. URL http://www.aeaweb.org/articles.php?doi=10.1257/jep.29.3.61.

Paolo Pinotti. The Economic Costs of Organized Crime: Evidence from Southern Italy. Temi di Discussione (Working Papers), http://www.bancaditalia.it/pubblicazioni/temi-discussione/2012/2012-0868/en_tema_868.pdf, April 2012a.

Paolo Pinotti. Organized Crime, Violence and the Quality of Politicians: Evidence from Southern Italy. Available at: http://dx.doi.org/10.2139/ssrn.2144121., 2012b.

Felipe Ribeiro, Guilherme Stein, and Thomas Kang. The Cuban Experiment: Measuring the Role of the 1959 Revolution on Economic Performance using Synthetic Control. Available at: http://economics.ca/2013/papers/SG0030-1.pdf, May 2013.

Joseph P Romano and Michael Wolf. Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4):1237–1282, 2005.

Robert Rosenthal. The file drawer problem and tolerance for null results. *Psychological bulletin*, 86(3):638, 1979.

Marcos Sanso-Navarro. The effects on American Foreign Direct Investment in the United Kingdom from Not Adopting the Euro. *Journal of Common Markets Studies*, 49(2):463–483, 2011.

Jessica Saunders, Russel Lundberg, Anthony A. Braga, Greg Ridgeway, and Jeremy Miles. A Synthetic Control Approach to Evaluating Place-Based Crime Interventions. *Journal of Quantitative Criminology*, 2014.

Edson R. Severnini. The Power of Hydroelectric Dams: Agglomeration Spillovers. IZA Discussion Paper, No. 8082, http://ftp.iza.org/dp8082.pdf., March 2014.

Erin O. Sills, Diego Herrera, A. Justin Kirkpatrick, Amintas Brandao, Rebecca Dickson, Simon Hall, Subhrendu Pattanayak, David Shoch, Mariana Vedoveto, Luisa Young, and Alexander Pfaff. Estimating the Impact of a Local Policy Innovation: The Synthetic Control Method Applied to Tropica Desforestation. *PLOS One*, 2015.

Joseph P Simmons, Leif D Nelson, and Uri Simonsohn. False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, page 0956797611417632, 2011.

Uri Simonsohn, Leif D. Nelson, and Joseph P. Simmons. P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, 143(2):534–547, 2014. ISSN 1939-2222. doi: 10.1037/a0033242. URL http://dx.doi.org/10.1037/a0033242.

Brock Smith. The Resource Curse Exorcised: Evidence from a Panel of Countries. *Journal of Development Economics*, 116:57–73, 2015.

Hiroki Tsurumi and Hajime Wago. Mean squared errors of forecast for selecting nonnested linear models and comparison with other criteria. *Journal of Econometrics*, 48(1):215 – 240, 1991. ISSN 0304-4076. doi: http://dx.doi.org/10.1016/0304-4076(91)90039-G. URL http://www.sciencedirect.com/science/article/pii/030440769190039G.

D. Wallach and B. Goffinet. Mean squared error of prediction as a criterion for evaluating and comparing system models. *Ecological Modelling*, 44(3):299 – 306, 1989. ISSN 0304-3800. doi: http://dx.doi.org/10.1016/0304-3800(89)90035-5. URL http://www.sciencedirect.com/science/article/pii/0304380089900355.

Halbert White. A reality check for data snooping. *Econometrica*, 68(5):1097–1126, 2000. ISSN 1468-0262. doi: 10.1111/1468-0262.00152. URL http://dx.doi.org/10.1111/1468-0262.00152.

Jingwen Yu and Chunchao Wang. Political Risk and Economic Development: A Case Study of China. *Eknomska Istrazianja - Economic Research*, 26(2):35–50, 2013.

Table 1: **Probability of rejecting the null in at least one specification**

	$K = 10$		$K = 2$	
	5% test	10% test	5% test	10% test
	(1)	(2)	(3)	(4)
$T_0 = 12$	0.133	0.235	0.132	0.232
	(0.002)	(0.003)	(0.002)	(0.003)
$T_0 = 32$	0.117	0.211	0.115	0.211
	(0.002)	(0.003)	(0.002)	(0.003)
$T_0 = 100$	0.103	0.193	0.104	0.193
	(0.002)	(0.003)	(0.002)	(0.003)
$T_0 = 230$	0.097	0.184	0.097	0.179
	(0.002)	(0.003)	(0.002)	(0.003)

Source: Authors' own elaboration. Note: Rejection rates are estimated based on 20,000 observations and on six specifications — (1) the mean of all pre-treatment outcome values, (2) all pre-treatment outcome values, (3) the first half of the pre-treatment outcome values, (4) the first three quarters of the pre-treatment outcome values, (5) odd pre-treatment outcome values, and (6) even pre-treatment outcome values. $K \in \mathbb{N}$ indicates the number of groups whose units follow parallel trends, *z% test* indicates that the nominal size of the analyzed test is z% and T_0 is the number of pre-treatment periods.

Table 2: **Variability of Weights**

	$K = 10$	$K = 2$
	(1)	(2)
$T_0 = 12$	0.670	0.671
	(0.001)	(0.001)
$T_0 = 32$	0.580	0.565
	(0.001)	(0.001)
$T_0 = 100$	0.549	0.470
	(0.001)	(0.002)
$T_0 = 230$	0.545	0.422
	(0.001)	(0.002)

Source: Authors' own elaboration. Note: The average variability of weights is based on 20,000 observations. This measure is computed in the following way: for each unit in the donor pool, we look for the specifications that allocate the most and the least weight for this unit. Then we take the maximum value of this difference across units in the donor pool. We use six specifications — (1) the mean of all pre-treatment outcome values, (2) all pre-treatment outcome values, (3) the first half of the pre-treatment outcome values, (4) the first three quarters of the pre-treatment outcome values, (5) odd pre-treatment outcome values, and (6) even pre-treatment outcome values — to construct this variable. $K \in \mathbb{N}$ indicates the number of groups whose units follow parallel trends and T_0 is the number of pre-treatment periods.

Table 3: **Misallocation of weights**

	Specification					
	1	2	3	4	5	6
			Panel A: $K = 10$			
$T_0 = 12$	0.907	0.695	0.774	0.728	0.775	0.771
	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
$T_0 = 32$	0.909	0.591	0.659	0.616	0.658	0.660
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
$T_0 = 100$	0.910	0.532	0.564	0.543	0.563	0.562
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
$T_0 = 230$	0.911	0.513	0.529	0.518	0.528	0.529
	(0.001)	(0.000)	(0.001)	(0.000)	(0.001)	(0.001)
			Panel B: $K = 2$			
$T_0 = 12$	0.410	0.215	0.247	0.227	0.250	0.249
	(0.002)	(0.001)	(0.002)	(0.001)	(0.002)	(0.002)
$T_0 = 32$	0.418	0.171	0.197	0.183	0.199	0.198
	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
$T_0 = 100$	0.414	0.134	0.156	0.143	0.155	0.156
	(0.002)	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)
$T_0 = 230$	0.413	0.114	0.130	0.120	0.130	0.130
	(0.002)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)

Source: Authors' own elaboration. Note: The average of misallocated weights is based on 20,000 observations. The reasiong 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. Specifications s is one of the specifications used to compute the synthetic control unit: (1) the mean of all pre-treatment outcome values, (2) all pre-treatment outcome values, (3) the first half of the pre-treatment outcome values, (4) the first three quarters of the pre-treatment outcome values, (5) odd pre-treatment outcome values, and (6) even pre-treatment outcome values. $K \in \mathbb{N}$ indicates the number of groups whose units follow parallel trends and T_0 is the number of pre-treatment periods.

Table 4: **Probability of rejecting the null in at least one specification excluding specification 1**

	$K = 10$		$K = 2$	
	5% test	10% test	5% test	10% test
	(1)	(2)	(3)	(4)
$T_0 = 12$	0.116	0.209	0.115	0.206
	(0.002)	(0.003)	(0.002)	(0.003)
$T_0 = 32$	0.100	0.183	0.099	0.182
	(0.002)	(0.003)	(0.002)	(0.003)
$T_0 = 100$	0.085	0.160	0.087	0.162
	(0.002)	(0.003)	(0.002)	(0.003)
$T_0 = 230$	0.076	0.147	0.078	0.146
	(0.002)	(0.003)	(0.002)	(0.002)

Source: Authors' own elaboration. Note: Rejection rates are estimated based on 20,000 observations and on five specifications — (2) all pre-treatment outcome values, (3) the first half of the pre-treatment outcome values, (4) the first three quarters of the pre-treatment outcome values, (5) odd pre-treatment outcome values, and (6) even pre-treatment outcome values. $K \in \mathbb{N}$ indicates the number of groups whose units follow parallel trends, *z% test* indicates that the nominal size of the analyzed test is z% and T_0 is the number of pre-treatment periods.

Table 5: **Variability of weights excluding specification 1**

	$K = 10$	$K = 2$
	(1)	(2)
$T_0 = 12$	0.581	0.591
	(0.001)	(0.001)
$T_0 = 32$	0.401	0.438
	(0.001)	(0.001)
$T_0 = 100$	0.241	0.295
	(0.000)	(0.001)
$T_0 = 230$	0.171	0.216
	(0.000)	(0.000)

Source: Authors' own elaboration. Note: The average variability of weights is based on 20,000 observations. This measure is computed in the following way: for each unit in the donor pool, we look for the specifications that allocate the most and the least weight for this unit. Then we take the maximum value of this difference across units in the donor pool. We use six specifications — (1) the mean of all pre-treatment outcome values, (2) all pre-treatment outcome values, (3) the first half of the pre-treatment outcome values, (4) the first three quarters of the pre-treatment outcome values, (5) odd pre-treatment outcome values, and (6) even pre-treatment outcome values — to construct this variable. $K \in \mathbb{N}$ indicates the number of groups whose units follow parallel trends and T_0 is the number of pre-treatment periods.

Table 6: **Probability of rejecting the null in at least one specification using alternative test statistic**

	$K = 10$		$K = 2$	
	5% test	10% test	5% test	10% test
	(1)	(2)	(3)	(4)
$T_0 = 12$	0.155	0.272	0.154	0.272
	(0.003)	(0.003)	(0.003)	(0.003)
$T_0 = 32$	0.141	0.244	0.138	0.241
	(0.002)	(0.003)	(0.002)	(0.003)
$T_0 = 100$	0.115	0.203	0.116	0.206
	(0.002)	(0.003)	(0.002)	(0.003)
$T_0 = 230$	0.097	0.178	0.099	0.178
	(0.002)	(0.003)	(0.002)	(0.003)

Source: Authors' own elaboration. Note: Rejection rates are estimated based on 20,000 observations and on six specifications — (1) the mean of all pre-treatment outcome values, (2) all pre-treatment outcome values, (3) the first half of the pre-treatment outcome values, (4) the first three quarters of the pre-treatment outcome values, (5) odd pre-treatment outcome values, and (6) even pre-treatment outcome values. $K \in \mathbb{N}$ indicates the number of groups whose units follow parallel trends, *z% test* indicates that the nominal size of the analyzed test is z% and T_0 is the number of pre-treatment periods. Here, we consider the alternative test statistic that calculates the pre-intervention MSPE using only lags not used in the estimation of weights.

Table 7: **Probability of rejecting the null in at least one specification - CPS simulations**

	Pooled		Employment		Log wages	
	5% test	10% test	5% test	10% test	5% test	10% test
	(1)	(2)	(3)	(4)	(5)	(6)
			Panel A: all specifications			
$T_0 = 12$	0.136***	0.240***	0.141***	0.242***	0.132***	0.238***
	(0.010)	(0.014)	(0.013)	(0.017)	(0.012)	(0.017)
$T_0 = 32$	0.126***	0.219***	0.119**	0.209***	0.133***	0.228***
	(0.021)	(0.027)	(0.032)	(0.041)	(0.029)	(0.038)
			Panel B: exclude specification 1			
$T_0 = 12$	0.123***	0.219***	0.127***	0.221***	0.118***	0.216***
	(0.010)	(0.014)	(0.012)	(0.017)	(0.012)	(0.017)
$T_0 = 32$	0.111***	0.199***	0.107*	0.194**	0.116**	0.203***
	(0.019)	(0.026)	(0.029)	(0.039)	(0.027)	(0.036)
			Panel C: alternative test statistic			
$T_0 = 12$	0.146***	0.258***	0.148***	0.260***	0.144***	0.256***
	(0.007)	(0.009)	(0.009)	(0.012)	(0.008)	(0.012)
$T_0 = 32$	0.135***	0.235***	0.129***	0.223***	0.141***	0.247***
	(0.019)	(0.025)	(0.030)	(0.039)	(0.028)	(0.037)

Source: Authors' own elaboration. Note: Rejection rates are estimated based on 20,000 observations for each outcome variable (employment and log wages) and number of pre-treatment periods ($T_0 \in \{12, 32\}$). Columns 1 and 2 present results pooling results using employment and log wages as outcome variable. In Panel A we consider the probability of rejecting in at least one of the six specifications described in Section 3. In Panel B we present this probability excluding specification 1. In Panel C we consider the alternative test statistic that calculates the pre-intervention MSPE using only lags not used in the estimation of weights. *z% test* indicates that the nominal size of the analyzed test is z%. * means that we reject at 10% the null that the probability of rejecting at least one specification at *z%* is equal to *z%*. ** means that we reject at 5%, while *** means that we reject at 1%.