Cherry Picking with Synthetic Controls

Ferman, Bruno and Pinto, Cristine and Possebom, Vitor

Sao Paulo School of Economics - FGV, Sao Paulo School of Economics - FGV, Yale

8 April 2017

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

First Draft: June 2016
This Draft: April 2017

Please click here for the most recent version

Abstract

We show that a lack of guidance on how to choose the matching variables used in the Synthetic Control (SC) estimator creates specification-searching opportunities in SC applications. This undermines one of the potential advantages of the method, which is providing a transparent way of choosing comparison units and, therefore, being less susceptible to specification searching than alternative methods. To address this problem, we provide recommendations to limit the possibilities for specification searching in the SC method. Finally, we analyze the possibilities for specification searching and our recommendations in two empirical applications.

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

JEL Codes: C12; C21; C33

[*] We would like to thank Sergio Firpo, Masayuki Sawada, and participants at the Sao School of Economics seminar, Yale Econometrics Lunch, African Meeting of the Econometric Society, and the 2016 meeting of the Brazilian Econometric Society for the excellent comments and suggestions. Deivis Angeli provided outstanding research assistance.

[†] *Corresponding Author:* (email) bruno.ferman@fgv.br, (tel) +55 11 3799-3350, (fax) + 55 11 3799-3357

[‡] 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 & Gardeazabal (2003), Abadie et al. (2010), and Abadie et al. (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, 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] Athey & Imbens (2016) describe the SC method as arguably the most important innovation in the evaluation literature in the last fifteen years.

Abadie et al. (2010) and Abadie et al. (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 will build the weighted-average of the controls' outcomes that will represent 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 (Miguel et al. (2014)).[2]

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

[2] See Christensen & Miguel (2016) for an extensive literature review on research transparency and reproducibility both in economics and other fields.

An important limitation of the SC method, however, is that it does not provide clear guidance on the choice of predictor variables that should be used to estimate the SC weights.[3] Although Abadie et al. (2010) define vectors of linear combinations of pre-intervention outcomes that could be used as predictors, there is no specific recommendation about which linear combinations should be used. Such lack of guidance on how to choose the economic predictors when implementing the synthetic control method translates into a wide variety of different specifications in empirical applications of this method. For example, some applied papers use all pre-treatment outcome lags as economic predictors, other papers select a subset of the pre-treatment outcome lags as economic predictors, while other papers use the mean of all pre-treatment outcome lags and other covariates as economic predictors.[4] 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 particular specification before knowing how these decisions would affect the conclusion of her study, this flexibility may undermine one of the main advantages of the SC method.[5]

In this paper, we evaluate the extent to which this variety of options in the synthetic control method creates opportunities for specification searching considering only one particular step of the method: the choice of which pre-treatment outcome values to include in the estimation of the SC weights.[6] Using Monte Carlo (MC) simulations and placebo simulations with

[3] To the best of our knowledge, Dube & Zipperer (2015) and Kaul et al. (2015) are the only other authors to point out that there is little explicit guidance in the SC literature to determine the choice of predictors. However, they do not explore the implications of such lack of specific guidance on the possibilities for specification searching in SC applications.

[4] For example, Abadie & Gardeazabal (2003), Abadie et al. (2015) and Kleven et al. (2013) use the mean of all pre-treatment outcome values and other covariates as predictors; Billmeier & Nannicini (2013), Bohn et al. (2014), Gobillon & 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.

[5] Olken (2015) and Coffman & 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.

[6] There may be other dimensions in the implementation of the SC method that provide discretionary choices for the researcher. For example, Klöbner et al. (2016) show that different SC estimators are obtained depending on the software used or on how the dataset is sorted, and Dube & Zipperer (2015) mention the addition of

the Current Population Survey (CPS), we calculate the probability that a researcher would find at least one specification that would lead him to reject the null at 5%. If different SC specifications lead to wildly different estimates, then the probability that a researcher would be able to find a specification that rejects the null at 5% can be much higher than 5%, implying significant room for specification searching. We consider six different specifications commonly used in SC applications: (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.[7]

We find that 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 (22% if we consider a 10% significance test). The possibilities for specification searching remain high even when the number of pre-treatment periods is large. With 400 pre-treatment periods, we still find a probability of around 11% that at least one specification is significant at 5% (21% 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. This is true both in data generating processes with stationary and non-stationary common factors. We also find similar results in placebo simulations using the CPS.

Abadie et al. (2010) and Abadie et al. (2015) emphasize that the SC method should only be used if there is a vector of weights such that the weighted average of the pre-intervention outcomes of the controls approximates well the pre-intervention outcomes of the treated unit. Since it is expected that a researcher applying the SC method show the pre-intervention fit of the chosen SC specification, this could potentially help reduce the scope for specification searching, as a researcher would only be able to select among specifications that yield a good

covariates. We focus only on the choice of pre-treatment outcome values to include in the estimation of the SC weights.

[7] In order to simplify the presentation of our results, we do not consider in our simulations the use of time-invariant covariates, as is commonly used in specifications that rely on the pre-treatment outcome mean. In Appendix A we show that our results remain valid if we consider specifications that use time-invariant covariates as economic predictors in addition to functions of the pre-treatment outcomes. Note also that these six specifications do not exhaust all specification options that have been considered in SC applications.

pre-intervention fit. We still find, however, that the probability of rejecting the null in at least one specification can be significantly higher than the nominal test size even when we restrict the set of choices to specifications with a good pre-treatment fit. There are at least two possible explanations for these results. First, in many SC applications, including those in Abadie & Gardeazabal (2003), Abadie et al. (2010), and Abadie et al. (2015), the outcome variable is non-stationary. In this case, most SC specifications will provide a good pre-treatment fit, as it will provide a good approximation to the non-stationary trend, as shown in Ferman & Pinto (2016). Our results suggest that, in this scenario, different SC specifications can still yield substantially different estimators even if most specifications provide a good approximation to the non-stationary trend.[8] Second, as shown in Ferman & Pinto (2017), the SC permutation test can lead to over-rejection if we consider the SC estimator conditional on a good pre-treatment fit.[9] This explains why we may still have significant over-rejection even when the researcher has only a few (or even just one) specifications with a good pre-intervention fit to choose from.[10]

The data-generating process (DGP) in our MC simulations 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 treated one. The specification that uses the mean of the 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.[11] Still, this result reinforces the argument that using only

[8] This is the case, for example, when we consider placebo simulations with the CPS using log wages as outcome variable.

[9] This happens because the test statistic for the treated unit is conditional on a good pre-treatment fit (that is, the denominator is close to zero), while the test statistics for the placebo units are unconditional. The over-rejection is decreasing in the probability that the SC estimator provides a good pre-intervention fit.

[10] This is the case, for example, when we consider placebo simulations with the CPS using male employment rate as outcome variable.

[11] In their Appendix, Ferman & Pinto (2016) analyze the asymptotic properties of the SC estimator 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. Importantly, we find that excluding this specification strongly attenuates the specification-searching problem, especially when the number of pre-treatment periods is large, even though it does not solve the problem completely.

It is important to note that our results by no means imply that researchers that have implemented the SC method did engage in specification searching. Given that this is a relatively new method, there would not be enough papers to formally test for specification searching.[12] However, given the mounting evidence that there is a high return for reporting "significant" results and that scientists tend to engage in p-hacking, our findings raise important concerns about the synthetic control method.[13] Also, while we find room for specification searching in the SC method, it does not imply that this problem is more relevant for the SC method when compared to alternatives methods.[14] The main conclusion of our paper is that, despite providing a data-driven method to construct the counterfactual unit, the SC method does not completely solve the specification-searching problem due to a lack of consensus on how the SC weights should be estimated.

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. Our results on the specification that uses the average of the pre-treatment outcome as economic predictor suggest that this specification-searching problem in the SC method is magnified by specifications with undesirable properties. More generally, our results suggest that restricting the set of options for researchers can

the average of the pre-treatment outcomes when the number of pre-treatment periods goes to infinity. They show that, in this case, there is no guarantee that the SC weights will converge to weights that reconstruct the factor loadings of the treated.

[12] 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 a 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 would not be enough 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.

[13] See Rosenthal (1979), Lovell (1983), De Long & Lang (1992), Simmons et al. (2011), Simonsohn et al. (2014), and Brodeur et al. (2016).

[14] For example, Gardeazabal & Vega-Bayo (2016) compare the synthetic control method with a panel data approach developed in Hsiao et al. (2012), and conclude that the SC estimator is more robust to changes in the donor pool.

strongly attenuate this problem. Another possible solution 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)). One alternative is to consider a test statistic for the permutation test that combines the test statistics for all individual specifications, as suggested in Imbens & Rubin (2015). Finally, another alternative would be to have a data-driven rule to determine which specification should be used. As an example, Dube & Zipperer (2015) propose a mean squared prediction error (MSPE) criterion based on the estimated post-treatment effects in placebo estimations whose minimizer could be the focus of an analysis that uses the synthetic control method.

Finally, we also consider the possibilities for specification searching and the implementability of the above recommendations in two empirical applications, based on Smith (2015) and Abadie et al. (2010). In our empirical examples, we analyze three cases: one whose conclusion is robust to specification searching, one where different specifications can reach either significant and non-significant results (clearly showing the potential for specification searching in the synthetic control framework), and one where all results are significant, but at different significance levels. Moreover, after applying our recommendations, we show that one can reach a clear conclusion about the significance of the results in all three examples.

The remainder of this paper proceeds as follows. In Section 2, we provide a brief overview of the SC estimation. We highlight the optimization problem used to find the weights. Then, we provide Monte Carlo simulations in Section 3 and simulations with real data in Section 4. We present our main recommendations in Section 5, and we discuss three empirical examples in Section 6. We conclude in Section 7.

2 Synthetic Control Method and Specification Searching

Abadie & 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 variables and covariates. For this reason, this weighted average of control units is known as the synthetic control unit 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 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 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 the counterfactual $Y^0_{1,t}$ 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 $\mathbf{Y_j}$.[15] 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.

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:

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

The weights $\widehat{\mathbf{W}} = [\widehat{w}_2...\widehat{w}_{j+1}]' := \widehat{\mathbf{W}}(\widehat{\mathbf{V}}) \in \mathbb{R}^J$ are 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.

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

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. Abadie et al. (2010) discuss alternative ways to choose the matrix $\widehat{\mathbf{V}}$. We focus

[15] 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.

our attention on the most common method of choosing $\widehat{\mathbf{V}}$, which involves solving the nested minimization problem given by equations (3) and (4).

Even though a crucial part in the implementation of the SC method is the choice of economic predictors, there is little guidance about which variables should be included in matrix $\mathbf{X_j}$. This lack of guidance can create an opportunity for the researcher to 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 functions of the outcome values should be included in $\mathbf{X_j}$: Abadie & Gardeazabal (2003), Abadie et al. (2015) and Kleven et al. (2013) use the mean of all pre-treatment outcome values and additional covariates; Smith (2015) uses Y_{j,T_0}, Y_{j,T_0-2}, Y_{j,T_0-4} and Y_{j,T_0-6}; Abadie et al. (2010) picks Y_{j,T_0}, Y_{j,T_0-8} and Y_{j,T_0-13}; Billmeier & Nannicini (2013), Bohn et al. (2014), Gobillon & Magnac (2016), Hinrichs (2012) use all pre-treatment outcome values; and Montalvo (2011) uses only the last two pre-treatment outcome values.[16]

Abadie et al. (2015) propose an inference procedure that consists in a straightforward placebo 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

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

where the acronym RMSPE stands for *ratio of the mean squared prediction errors*. 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 0.05. Abadie et al. (2010) recognize that the randomization inference assumptions are very restrictive for

[16] By no means we imply that those authors have engaged in specification searching. We have only listed them as prominent examples of different choices regarding predictor variables.

the SC set-up. However, 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 was randomly assigned among the data.[17]

3 Monte Carlo Simulations

In order to verify the possibility of specification searching, we elaborate a Monte Carlo exercise in which we generate 5,000 data sets and, for each one of them, test the null hypothesis of no effect whatsoever adopting several different specifications. Conditional on a given specification, this placebo test should provide a rejection rate of $\alpha\%$ under the null for a $\alpha\%$ significance test by construction. We are interested, however, in the probability of rejecting the null hypothesis at the 5%-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 $\alpha\%$ can be significantly higher than $\alpha\%$. In the extreme case in which we have K different specifications and these specifications lead to independent estimators, this probability would be given by $1-(1-\alpha)^K$, where K is the number of different specifications.[18] In this case, such lack of guidance about which specifications should be used 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 $\alpha\%$ and the risk of specification searching would be very low. We consider two data generating processes. In Section 4 we consider placebo simulations with the CPS.

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

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

[17] Firpo & Possebom (2016) discuss this inference procedure in the case of random assignment, while Ferman & Pinto (2017) analyze the statistical properties of this placebo test when treatment is not randomly assigned. For our purposes in this paper, we consider Abadie et al. (2010) interpretation of the placebo test p-value.

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

for some $k = 1, ..., K$. 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 $\epsilon_{j,t} \sim N(0, 0.1)$.

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

$$Y_{j,t}^0 = \delta_t + \lambda_t^k + \phi_t^r + \epsilon_{jt} \tag{8}$$

for some $k = 1, ..., K$ and $r = 1, ..., R$. We consider in our simulations $K = 10$ and $R = 2$. Therefore, units $j = 2, ..., 10$ 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.

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 $t \in \{1, ..., T_0\}$. We fix the number of post-treatment periods $T - T_0 = 10$ and we vary the number of pre-intervention periods in the DGPs, $T_0 \in \{12, 32, 100, 400\}$. In the Appendix, we consider variations in our stationary model (7) by setting (i) $\epsilon_{j,t} \sim N(0, 1)$, (ii) $K = 2$, or (iii) including time-invariant covariates. We find similar results as the ones presented in the main text.

We calculate the SC estimator using the following six specifications that differ only in the linear combinations of pre-treatment outcome values used as predictors:[19]

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: $\mathbf{X}_j = \left[Y_{j,1} \cdots Y_{j,3T_0/4}\right]'$

[19] In order to compute the SC estimator, we use the *Synth* package in *R*. (See Abadie et al. (2011) for details.) This package solves the nested minimization problem described by equations (3) and (4). 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.

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: $\mathbf{X}_j = \left[Y_{j,2} \; Y_{j,4} \cdots Y_{j,(T_0-2)} \; Y_{j,T_0}\right]'$

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

For each specification, we run a permutation test using the RMSPE test statistic proposed in Abadie et al. (2010) 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 the 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 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% as well. However, this probability may be higher if the synthetic control weights depend on specification choices.

We present in columns 1 and 2 of Table 1 the probability of rejecting the null at 5% and at 10% significance levels in at least one specification for the stationary model. Columns 3 and 4 present the same results for the non-stationary model.[20] With $T_0 = 12$, a researcher considering these six different specifications would be able to report a specification with statistically significant results at the 5% level with probability 12.7% for the stationary model and 12.4% for the non-stationary. If we consider 10% significance tests, then the probability of rejecting the null in at least one specification would be up to 22.5% and 22.1%, respectively for the stationary and the non-stationary models. 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, note that it is not unusual to have

[20] See table A.1 for results using different data generating processes.

SC applications with as few as 12 pre-intervention periods.[21]

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 closer 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. The results in Table 1 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 11.8% (21.4%) of the time for a 5% (10%) significance test.[22] These results suggest that specification searching remains a problem for the SC method even when the number of pre-intervention periods is remarkably large for empirical applications.

In the previous exercise, we assumed that the researcher would be able to choose any of the 6 specifications we considered in our MC simulations. However, Abadie et al. (2010) and Abadie et al. (2015) emphasize that the SC control estimator should only be used in the situations with good pre-treatment fit, i.e., in situations in which the weighted average of the controls' pre-treatment outcomes is a good approximation for the treated pre-treatment outcome. It is important, therefore, to 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:[23]

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

[22] 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.

[23] This measure is very similar to the "pretreatment fit index" proposed by Adhikari & Alm (2016). These authors propose a measure that is the ratio between the squared root of the mean squared predicted error (the numerator of $1 - \tilde{R}^2$) and $\sqrt{\frac{\sum_{t=1}^{T_0} Y_{1t}^2}{T_0}}$. The advantage of our measure relative to the one proposed by Adhikari & Alm (2016) is that our measure is invariant to linearly additive changes. Dube & Zipperer (2015)

$$\tilde{R}^2 = 1 - \frac{\sum_{t=1}^{T_0} \left(Y_{1,t} - \widehat{Y}_{1,t}^N \right)^2}{\sum_{t=1}^{T_0} \left(Y_{1,t} - \overline{Y}_1 \right)^2} \tag{9}$$

where $\overline{Y}_1 = \frac{\sum_{t=1}^{T_0} Y_{1,t}}{T_0}$. If this measure is one, then we have a perfect fit.[24]

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 2 shows the probability of finding a good pre-treatment fit for at least one of the six specifications.[25] 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. Even when we consider the weaker definition of good fit, it is interesting to note that the average number of specifications with good fit is close to 5 (panel B of Table 2). This happens because the probability of having a good fit for the specification that uses the pre-treatment mean as economic predictor is relatively low (panel C of Table 2). 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 the specification that uses the pre-treatment mean as economic predictor) 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 efficient in controlling for the non-stationary factors (see Ferman & Pinto (2016)), it will usually provide a good pre-treatment fit.

Given these definitions of good fit, we present in Table 3 the probabilities of rejecting the

also propose a pre-treatment fit criterion that is equal to the numerator of our measure, the root of the mean squared error predictor between the synthetic and the actual outcomes in the pre-treatment period. However, differently from our suggestion, their measure is not scale invariant.

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

[25] See table A.2 for results using different data generating processes.

null in at least one specification when we restrict the researcher to consider only specifications that provide a good pre-treatment fit.[26] Note that the possibilities for specification searching in the non-stationary model (columns 3 and 4) are virtually the same as when we do not restrict for specifications with a good pre-treatment fit, especially when T_0 is large (columns 3 and 4 of Table 1). 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 3), the specification-search problem is attenuated when we restrict to specifications with a good fit if 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 the specification that uses the pre-treatment mean as economic predictor, whose weights, as we show below, are very different from the ones chosen by the other specifications. If we consider the more stringent definition of good fit, however, then the probability of rejecting the null in at least one specifications 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 et al. (2010) and Abadie et al. (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 statistic for the placebo units are not conditional on a good pre-treatment, this leads to over-rejection, as shown in Ferman & Pinto (2017).

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.

The results so far indicate that different specifications can provide substantially different

[26] See table A.3 for results using different data generating processes.

SC estimators. An interesting feature of our MC simulations is that the SC estimator should assigned 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 is misallocated for each specification. We present in columns 1 to 6 of Table 4 the proportion of misallocated weights for each specification in different scenarios.[27] Interestingly, specification 1 (which uses the pre-treatment mean as economic predictor) misallocates substantially more weights relative to the other specifications. For the stationary model (panel A), with $T_0 = 12$, specification 1 misallocates more than 80% of the weights, while the misallocation for other specifications ranges from 23% to 29%. The misallocation of weights decreases with T_0 for all specifications, except for specification 1. Results are qualitatively the same for the non-stationary model (panel B). These results suggest that using the pre-treatment outcome mean might not capture the time-series dynamics of the units, which is the main goal of the SC method.[28]

We also calculate a measure of variability of weights. 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 present this measure in column 7 of Table 4.[29] Interestingly, this measure does not decrease in T_0, suggesting that increasing T_0 does not imply that different SC specifications will lead to similar SC estimators. If we consider this measure excluding specification 1, however, then it decreases in T_0 (column 8 of Table 4). These results indicate that, as T_0 increases, the SC estimators using specifications 2 to 6 become more similar. However, the SC estimator using specification 1 can be considerably different from the SC estimators using the other specifications even when T_0 is large. This result is intuitive, given that specifications 2 to 6 exploit the time-series dynamics of the data, while specification 1 does not.[30]

[27] See table A.4 for results using different data generating processes.

[28] In specifications that use other covariates in addition to the pre-treatment mean, the matrix V would be chosen to minimize the pre-treatment MSPE in the second step of the optimization process, so this estimator would somewhat take the time-series dynamics of the outcome into account. However, this would be very limited because the first minimization problem can severely restrict the set of possible weights $\mathbf{W}^*(V)$ that may be chosen in the second step, as suggested in Ferman & Pinto (2016).

[29] See table A.4 for results using different data generating processes.

[30] In Appendix A we show that specification 1 can fail to properly exploit the time-series dynamics of the data even if we also include time-invariant covariates as economic predictors. In this case, it will still remain different from the specifications that use pre-treatment outcome lags as economic predictors, even when the number

Given that the specification that uses the pre-treatment mean as the economic predictor stands out by misallocating significantly more weights, in Table 5 we consider the specification-searching possibilities excluding specification 1.[31] Excluding specification 1 significantly attenuates the specification-searching problem, especially when the number of pre-treatment periods is large, although it does not completely solve the problem.[32] This attenuation in the specification-searching problem is not simply because we are considering five specifications instead of six. If we exclude, for example, specification 2 instead of specification 1, then there is virtually no change in the specification-search problem relative to the case that we consider six specifications (Appendix Table A.6).

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. Following Bertrand et al. (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 1a and 1b present the time series of log wages for all US states, respectively for men and women. As expected, the time series of log wages is

of pre-treatment periods is large. Therefore, our result that the possibilities of specification searching may not diminish with the number of pre-treatment periods when we consider the specification that uses the pre-treatment outcome mean as economic predictor remains valid even if we consider the addition of time-invariant variables as economic predictors.

[31] See table A.5 for results using different data generating processes.

[32] The only exception is when we consider the stationary model conditional on a good fit with $\tilde{R}^2 > 0.95$. This happens because, in this case, there is a low probability that we find at least one specification with a good fit.

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 1c and 1d present the time series of employment for all US states, respectively for men and women. In this case, the time series for men should be closer to our stationary model from Section 3, while the time series for women has an increasing trend in the 80s and 90s.

We first consider simulations with 12 pre-intervention periods, 4 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, D.C.) and then we randomly select the first period between 1979 and 1999. 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 5,000 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 start presenting the probability of finding specifications with a good fit in Table 6. 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). Also, when we consider $T_0 = 32$, the number of specifications with a good fit is close to 6, which suggests that there is a high probability that all specifications will provide a good fit (columns 1 to 4, panel B). 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 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 7 the probabilities of rejecting the null in at least one specification.[33] In panel A, we present the unconditional specification-searching probabilities. The results are very similar to our findings in the MC simulations. With $T_0 = 12$, depending on the sample and outcome variable, there is 10-12% probability of finding a specification with sta-

[33] 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.

tistically significant results at 5% and a 19-21% probability of finding a specification with statistically significant results at 10%. With $T_0 = 32$ these probabilities are slightly lower, but still significantly higher than the test nominal size. In panels B and C we present results restricting the choices to specifications with a good pre-treatment fit. As in our MC simulations, conditioning on a good fit does not attenuate the specification-searching problem. When we consider log wages as outcome variable, conditioning does not affect much these probabilities, as most specifications provide a good fit, especially when $T_0 = 32$. When we consider employment as outcome variable, conditioning leads to over-rejection because there is a high probability that no specification provides a good fit and the surviving specifications present a pre-treatment fit (the denominator of the test statistic) that is much lower than the pre-treatment fit of the control units, that are not restricted to have a good fit. These results suggest that specification-searching possibilities in SC applications can be relevant in real applications of the SC method.

We also consider in Table 8 the specification-searching probabilities excluding the specification that uses the pre-treatment mean as economic predictor. Similar to our MC simulations, excluding this specification attenuates the specification-searching problem, although it does not completely solve the problem. With $T_0 = 32$, the probability of rejecting the null at 5% in at least one specification ranges from 7% to 7.9% depending on the sample and outcome variable, and in no case we can reject that this probability is different from the nominal test size. Again, this attenuation is not a simple mechanical effect due to the fact that we are considering fewer specifications. In Appendix Table A.7, we show that there is virtually no change in the probabilities of rejecting the null in at least one specification when we exclude specification 2 (instead of specification 1) relative to the case where we consider all six specifications.

5 Recommendations

The specification-searching problem we identify arises from a lack of consensus about which specifications should be used in SC applications. Our first recommendation is that researchers should only consider specifications that capture the dynamic of the outcome of the treated unit in the pre-treatment period, because our results suggest that the specification-searching problem is magnified by specifications with undesirable properties, such as the specification that uses only the mean pre-treatment outcome as economic predictor. If we discard this specification, then the specification-searching problem is attenuated, especially if we have a large number of pre-treatment periods, even though it does not solve the problem completely.

We also recommend 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 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 (2005)).[34] 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, as suggested by Imbens & Rubin (2015). Although this function can be non-linear, if it is simply a weighted average of the test statistics for individual specifications, then Christensen & Miguel (2016) and Cohen-Cole et al. (2009) suggest using the same weights to compute a weighted average of the point-estimator of each specification and using this weighted average as an estimate that incorporates model uncertainty.

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 criterium, without the need of subjective decisions by the researcher, then the possibility for specification searching would be limited, at least in this dimension. One such possibility is to follow Dube & Zipperer (2015) and choose the specification that minimizes the mean squared prediction error (MSPE) for the post-intervention periods for the placebo estimates.

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

6 Empirical Applications

We analyze the possibilities for specification searching and the implementability of our recommendations in two empirical examples.

6.1 The resource curse exorcised: Evidence from a panel of countries (Smith (2015))

Smith (2015) evaluates the impact of major natural resource discoveries since 1950 on GDP per capita using different methods, including the synthetic control method.[35] Major oil and gas discoveries happened in Equatorial Guine and Equator in 1992 and 1972 respectively, implying that pre and post-treatment periods are 1950-1991 and 1992-2008 for the first country and 1950-1971 and 1972-2008 for the second one. While the donor pool for Equatorial Guine consists of Sub-Saharan African Countries (Benin, Burkina Faso, Burundi, Cameroon, Cape Verde, Central African Republic, Chad, Cote d'Ivoire, Gambia, Ghana, Guinea, Kenya, Lesotho, Liberia, Madagascar, Malawi, Mali, Mauritania, Mauritius, Mozambique, Namibia, Niger, Rwanda, Senegal, Somalia, Sudan, Swaziland, Tanzania, Togo, Uganda, Zambia, Zimbabwe), the donor pool for Ecuador consists of Latin American and Caribbean countries (Costa Rica, Cuba, Dominican Republic, El Salvador, Guatemala, Honduras, Jamaica, Nicaragua, Panama, Paraguay, Puerto Rico, Uruguay).

We estimate the impact of major oil and gas discoveries on GDP per capita using the synthetic control method with twelve different specifications. Specifically, we test six different specifications that differ in which functions of the pre-treatment periods are included and, for each one of them, we either include the covariates *ethnic fragmentation* and *population size one year before the discovery* or not. Our the six basic specifications are:[36]

[35] Following the best practices in terms of transparency and replicability, he made his dataset and replication files available online (http://www.brockdsmith.com/research.html).

[36] Although the number of pre-treatment years is larger than seven, we followed Smith (2015) and considered for this exercise different specifications using only seven years of pre-treatment data in the first minimization problem (equation (3)) while accounting for the entire pre-treatment period in the second minimization problem (equation (4)). 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

1. Original Specification (Even pre-treatment outcome values): $\mathbf{X}_j = \left[Y_{j,(T_0-6)} \;\; Y_{j,(T_0-4)} \;\; Y_{j,(T_0-2)} \;\; Y_{j,T_0}\right]'$

2. Pre-treatment outcome mean: $\mathbf{X}_j = \left[\sum_{t=T_0-6}^{T_0} Y_{j,t}/7\right]$

3. All pre-treatment outcome values: $\mathbf{X}_j = \left[Y_{j,(T_0-6)} \cdots Y_{j,T_0}\right]'$

4. The first half of the pre-treatment outcome values: $\mathbf{X}_j = \left[Y_{j,(T_0-6)} \cdots Y_{j,(T_0-4)}\right]'$

5. The first three fourths of the pre-treatment outcome values: $\mathbf{X}_j = \left[Y_{j,(T_0-6)} \cdots Y_{j,(T_0-2)}\right]'$

6. Odd pre-treatment outcome values: $\mathbf{X}_j = \left[Y_{j,(T_0-5)} \;\; Y_{j,(T_0-3)} \;\; Y_{j,(T_0-1)}\right]'$

where $T_0 = 1991$ for Equatorial Guine and $T_0 = 1971$ for Ecuador.

Table 9 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%-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 twelve specifications have a good fit ($\tilde{R}^2 > 0.80$), but only two of them are significant (specifications (2b) and (6b)), implying that the researcher could, potentially, report a false-positive result.[37]

We now test our recommendations in these particular applications. First of all, by presenting results for more than one specification as we do in Table 9, 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 2 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 statistics that combine the test statistics of all specifications. 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, as suggested by Imbens & Rubin

predictor diminishes when the number of pre-treatment periods increases. Results are available upon request.

[37] We stress that the specification considered by Smith (2015) is not one of these two that would have led him to conclude that there is a significant effect.

(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 one. Now, following the suggestion of Christensen & Miguel (2016) and Cohen-Cole et al. (2009), figure 3 shows the average treatment effect across specifications with good fit as a black line and the associated placebo effects as gray lines. Clearly, the effects for Equatorial Guinea and Ecuador are, respectively, large and small when compared to their empirical distributions. Finally, we apply the MSPE criterion suggested by Dube & Zipperer (2015) to select only one specification. For Equatorial Guinea, we find that specification 5b (first three-fourths of pre-treatment outcome values without covariates) minimizes the MSPE criterion, and the the p-value for this specification is 0.031 . For Ecuador, we find that specification 5b (first three-fourths pre-treatment outcome values without covariates) minimizes the MSPE criterion, and the p-value for this specification is 0.308. Figure 4 shows the treatment effect and the placebo effects for specifications 5b for Equatorial Guinea and Ecuador, respectively.

The results based on the recommendations by Imbens & Rubin (2015), Christensen & Miguel (2016) and Cohen-Cole et al. (2009) point all to the same direction. Therefore, a reasonable conclusion would be that the treatment effect is significant in the case of Equatorial Guinea and statistically zero in the case of Ecuador. Importantly, without following these recommendations, the results for Ecuador point out that it would be possible to find particular specifications with a good pre-treatment fit that would lead the researcher to conclude that the effect for Ecuador is statistically significant.

6.2 Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program (Abadie et al. (2010))

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

[38] Following the best practices in terms of transparency and replicability, they made their dataset and replication files available through the command *synth* in the software *Stata*.

eight American states (Alabama, Arkansas, Colorado, Connecticut, Delaware, Georgia, Idaho, Illinois, Indiana, Iowa, Kansas, Kentucky, Louisiana, Maine, Minnesota, Mississippi, Missouri, Montana, Nebraska, Nevada, New Hampshire, New Mexico, North Carolina, North Dakota, Ohio, Oklahoma, Pennsylvania, Rhode Island, South Carolina, South Dakota, Tennessee, Texas, Utah, Vermont, Virginia, West Virginia, Wisconsin, Wyoming).

We estimate the impact of Proposition 99 on California's annual per-capita cigarette sales using the synthetic control method with fourteen different specifications. Specifically, we test seven different specifications that differ in which functions of the pre-treatment periods are included and, for each one of them, we either include the covariates *average retail price of cigarettes*, *per capita state personal income (logged)*, *percentage of the population age 15–24*, and *per capita beer consumption* or not. The seven basic specification are (1) original specification by Abadie et al. (2010) (outcome values for 1975, 1980 and 1988), (2) pre-treatment outcome mean, (3) all pre-treatment outcome values, (4) the first half of the pre-treatment outcome values, (5) the first three fourths of the pre-treatment outcome values, (6) odd pre-treatment outcome values, (7) even pre-treatment outcome values.

Table 10 shows the p-value and our goodness of fit measure for each of the 14 specifications we considered. 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 fit 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 5, although p-values vary from 0.026 (the p-value in the specification considered in Abadie et al. (2010)) to 0.077 depending on the specification.

Now, we test the null hypothesis using a test statistic that combine the test statistics of all specifications. Restricting ourselves to specifications with a fit as good as the original specification ($\tilde{R}^2 > 0.975$), we find that the p-value of a test that uses the mean of the RMSPE statistic across specifications, as suggested by Imbens & 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 stands out as the largest one among all placebo runs as it does when we consider the original specification. Additionally, figure 3 shows the average treatment effect across specifications with good fit as a black line and the associated placebo effects as gray lines following the suggestion of Christensen & Miguel (2016) and Cohen-Cole et al. (2009). Note that the treatment effects for California seem to be larger (or, at least, more stable) than the placebo effects.

Finally, we apply the MSPE criterion suggested by Dube & Zipperer (2015) to select only one specification. We find that specification 6a (odd pre-treatment outcome values) minimizes the MSPE criterion. The p-value for this specification is 0.026. This result is consistent with the one reached by the method suggested by Imbens & Rubin (2015), although we would reject the null at a lower significance level.

Overall, our results suggest that the effect of the 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.

7 Conclusion

We show that a lack of specific guidance on how to choose among different SC specifications creates the potential for specification searching with synthetic controls. We also show that restricting the set of options a researcher has when applying the SC method can substantially attenuate this specification-searching problem. We move in this direction by showing that the specification that uses the average of the pre-treatment outcome may fail to exploit the dynamics of the time series, which is the main goal of the SC method. Discarding this specification significantly reduces the room for specification searching when the number of pre-treatment periods is large, even though it does not completely solve the problem. However, further research is necessary to determine in which circumstances one should use all pre-treatment lags as economic predictors or only a subset of the pre-treatment outcome lags (and, in this case, which subset should be used). Consequently, additional restrictions on the set of specifications applied researchers can use when employing the SC method in a given application can further reduce the scope for specification searching with synthetic controls. Furthermore, we also recommend that researchers report results using different specifications, and we suggest alternatives to take into account the fact that the treatment effect can be estimated using different specifications. Finally, we show that these recommendations can easily be implemented in practice, providing clear conclusions about the significance of an estimate.

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 Statiscal Association* **105**(490), 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**(13), 1–17.

Abadie, A., Diamond, A. & Hainmueller, J. (2015), 'Comparative Politics and the Synthetic Control Method', *American Journal of Political Science* **59**(2), 495–510.

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

Acemoglu, D., Johnson, S., Kermani, A., Kwak, J. & Mitton, T. (2013), 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.

Adhikari, B. & Alm, J. (2016), 'Evaluating the economic effects of flat tax reforms using synthetic control methods', *Southern Economic Journal* **83**(2), 437–463.
URL: *http://dx.doi.org/10.1002/soej.12152*

Ando, M. (2015), '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.

Athey, S. & Imbens, G. W. (2016), 'The state of applied econometrics - causality and policy evaluation', *mimeo* .

Barone, G. & Mocetti, S. (2014), 'Natural Disasters, Growth and Institutions: a Tale of Two Earthquakes', *Journal of Urban Economics* pp. 52–66.

Bauhoff, S. (2014), 'The Effect of School Nutrition Policies on Dietary Intake and Overweight: a Synthetic Control Approach', *Economics and Human Biology* pp. 45–55.

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**(1), 73–89.

Bertrand, M., Duflo, E. & Mullainathan, S. (2004), 'How much should we trust differences-in-differences estimates?', *Quarterly Journal of Economics* p. 24975.

Billmeier, A. & Nannicini, T. (2009), 'Trade Openness and Growth: Pursuing Empirical Glasnost', *IMF Staff Papers* **56**(3), 447–475.

Billmeier, A. & Nannicini, T. (2013), 'Assessing Economic Liberalization Episodes: A Synthetic Control Approach', *The Review of Economics and Statistics* **95**(3), 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**(2), 258–269.

Bove, V., Elia, L. & Smith, R. P. (2014), 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.

Brodeur, A., Lé, M., Sangnier, M. & Zylberberg, Y. (2016), 'Star Wars: The Empirics Strike Back', *American Economic Journal: Applied Economics* **8**(1), 1–32.

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

Carrasco, V., de Mello, J. M. P. & Duarte, I. (2014), A Década Perdida: 2003 – 2012. Texto para Discussão, http://www.econ.puc-rio.br/uploads/adm/trabalhos/files/td626.pdf.

Cavallo, E., Galiani, S., Noy, I. & Pantano, J. (2013), 'Catastrophic Natural Disasters and Economic Growth', *The Review of Economics and Statistics* **95**(5), 1549–1561.

Chan, H. F., Frey, B. S., Gallus, J. & Torgler, B. (2014), 'Academic Honors and Performance', *Labour Economics* **31**, 188–204.

Christensen, G. & Miguel, E. (2016), Transparency, reproducibility, and the credibility of economics research, Technical report.

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

Coffman, M. & Noy, I. (2011), 'Hurricane Iniki: Measuring the Long-Term Economic Impact of Natural Disaster Using Synthetic Control', *Environment and Development Economics* **17**, 187–205.

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**(2), 335–369.

De Long, J. B. & Lang, K. (1992), 'Are all economic hypotheses false?', *Journal of Political Economy* pp. 1257–1272.

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

Dhungana, S. (2011), 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.

Dube, A. & Zipperer, B. (2015), Pooling Multiple Case Studies Using Synthetic Controls: An Application to Minimum Wage Policies, IZA Discussion Papers 8944, Institute for the Study of Labor (IZA).
URL: *https://ideas.repec.org/p/iza/izadps/dp8944.html*

DuPont, W. & Noy, I. (2012), 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.

Ferman, B. & Pinto, C. (2016), Revisiting the synthetic control estimator.

Ferman, B. & Pinto, C. (2017), Placebo Tests for Synthetic Controls.

Firpo, S. & Possebom, V. (2016), Synthetic Control Estimator: A Generalized Inference Procedure and Confidence Sets. Working Paper, https://goo.gl/oQTX9c.

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* pp. n/a–n/a. jae.2557.
URL: *http://dx.doi.org/10.1002/jae.2557*

Gathani, S., Santini, M. & Stoelinga, D. (2013), 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.

Gobillon, L. & Magnac, T. (2016), 'Regional Policy Evaluation: Interative Fixed Effects and Synthetic Controls', *Review of Economics and Statistics* . Forthcoming.

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**(3), 712–722.

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

Hsiao, C., Steve Ching, H. & Ki Wan, S. (2012), 'A panel data approach for program evaluation: Measuring the benefits of political and economic integration of hong kong with mainland china', *Journal of Applied Econometrics* **27**(5), 705–740.
URL: *http://dx.doi.org/10.1002/jae.1230*

Imbens, G. W. & Rubin, D. B. (2015), *Causal Inference for Statistics, Social and Biomedical Sciences: An Introduction*, 1st edn, Cambridge University Press, United Kingdom.

Jinjarak, Y., Noy, I. & Zheng, H. (2013), 'Capital Controls in Brazil — Stemming a Tide with a Signal?', *Journal of Banking & Finance* **37**, 2938–2952.

Kaul, A., Klöbner, S., Pfeifer, G. & Schieler, M. (2015), 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.

Kirkpatrick, A. J. & Bennear, L. S. (2014), 'Promoting Clean Enery Investment: an Empirical Analysis of Property Assessed Clean Energy', *Journal of Environmental Economics and Management* **68**, 357–375.

Kleven, H. J., Landais, C. & Saez, E. (2013), 'Taxation and International Migration of Superstars: Evidence from European Football Market', *American Economic Review* **103**(5), 1892–1924.

Klöbner, S., Kaul, A., Pfeifer, G. & Schieler, M. (2016), Comparative Politics and the Synthetic Control Method Reviseted: A Note on Abadie et al. (2015).

Kreif, N., Grieve, R., Hangartner, D., Turner, A. J., Nikolova, S. & Sutton, M. (2015), 'Examination of the Synthetic Control Method for Evaluating Health Policies with Multiple Treated Units', *Health Economics* .

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

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

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

Mideksa, T. K. (2013), 'The Economic Impact of Natural Resources', *Journal of Environmental Economics and Management* **65**, 277–289.

Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K. M., Gerber, A., Glennerster, R., Green, D. P., Humphreys, M., Imbens, G., Laitin, D., Madon, T., Nelson, L., Nosek, B. A., Petersen, M., Sedlmayr, R., Simmons, J. P., Simonsohn, U. & Van der Laan, M. (2014), 'Promoting transparency in social science research', *Science* **343**(6166), 30–31.
URL: *http://science.sciencemag.org/content/343/6166/30*

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**(4), 1146–1154.

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

Pinotti, P. (2012*a*), Organized Crime, Violence and the Quality of Politicians: Evidence from Southern Italy. Available at: http://dx.doi.org/10.2139/ssrn.2144121.

Pinotti, P. (2012*b*), 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.

Ribeiro, F., Stein, G. & Kang, T. (2013), 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.

Romano, J. P. & Wolf, M. (2005), 'Stepwise multiple testing as formalized data snooping', *Econometrica* **73**(4), 1237–1282.

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

Sanso-Navarro, M. (2011), '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.

Saunders, J., Lundberg, R., Braga, A. A., Ridgeway, G. & Miles, J. (2014), 'A Synthetic Control Approach to Evaluating Place-Based Crime Interventions', *Journal of Quantitative Criminology* .

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

Sills, E. O., Herrera, D., Kirkpatrick, A. J., Brandao, A., Dickson, R., Hall, S., Pattanayak, S., Shoch, D., Vedoveto, M., Young, L. & Pfaff, A. (2015), 'Estimating the Impact of a Local Policy Innovation: The Synthetic Control Method Applied to Tropica Desforestation', *PLOS One* .

Simmons, J. P., Nelson, L. D. & Simonsohn, U. (2011), 'False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant', *Psychological science* p. 0956797611417632.

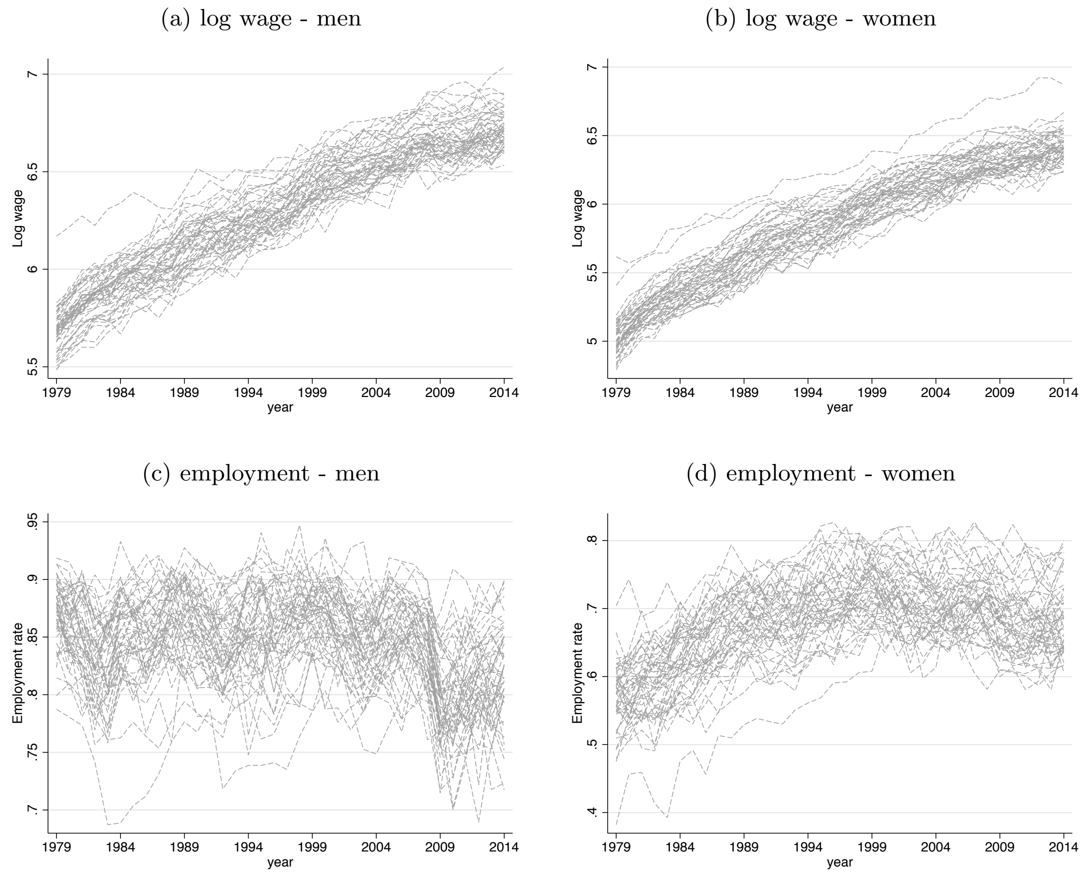
Simonsohn, U., Nelson, L. D. & Simmons, J. P. (2014), 'P-curve: A key to the file-drawer.', *Journal of Experimental Psychology: General* **143**(2), 534–547.
URL: *http://dx.doi.org/10.1037/a0033242*

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

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

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

Figure 1: **Outcome trajectories in the CPS data**



(a) log wage - men

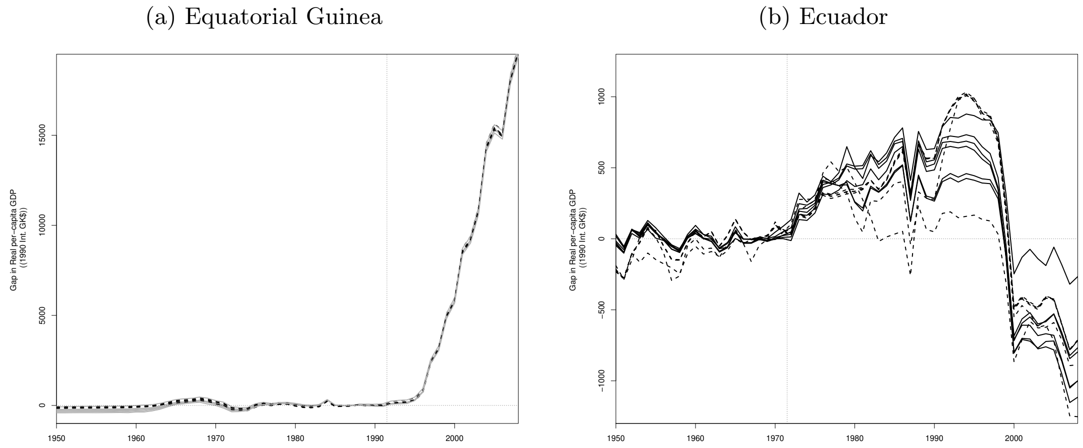
(b) log wage - women

(c) employment - men

(d) employment - women

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

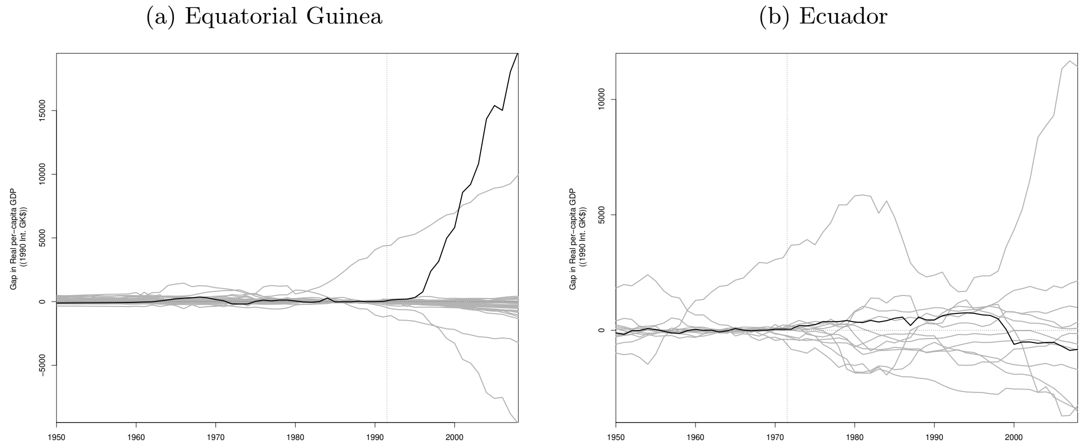
Figure 2: **Treatment Effects for All Specifications - Database from Smith (2015)**


(a) Equatorial Guinea

(b) Ecuador

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 (9). The vertical lines denote the beginning of the post-treatment period.

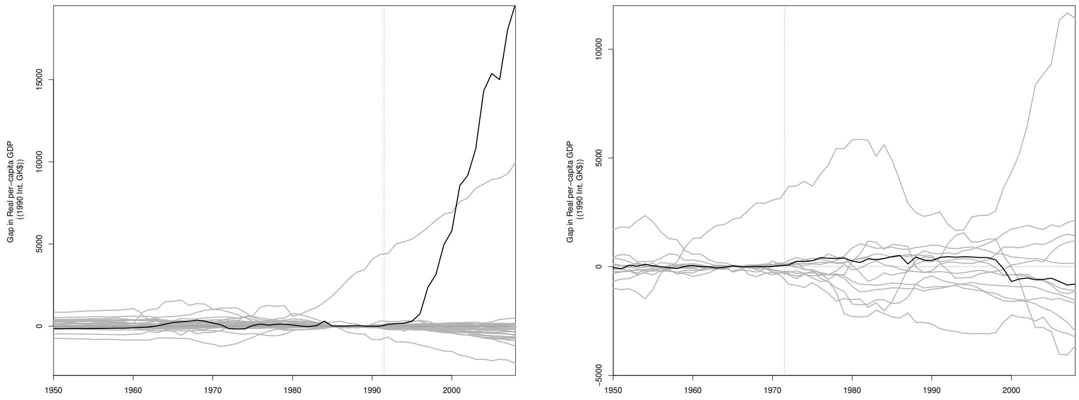
Figure 3: **Placebo Effects Using the Average Across Specifications - Database from Smith (2015)**


(a) Equatorial Guinea

(b) Ecuador

Notes: We only consider specifications that satisfy $\tilde{R}^2 > 0.80$ to compute the average treatment effect across specification, where $\tilde{R}^2$ is defined by equation (9). Gray lines are the placebo effects of the control countries and the black line is the treatment effect of the treated country. The vertical lines denote the beginning of the post-treatment period.

Figure 4: **Placebo Effects Using the MSPE Criterion - Database from Smith (2015)**

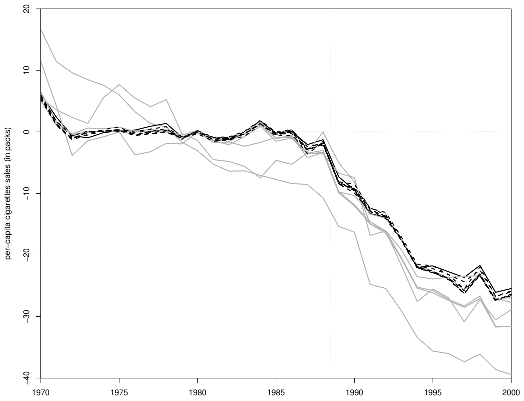


(a) Equatorial Guinea - Specification (6b)

(b) Ecuador - Specification (2b)

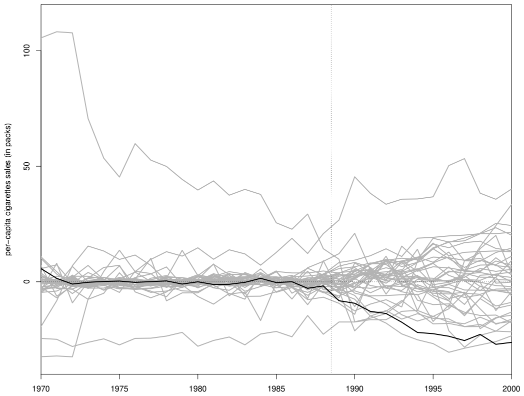
Notes: We only consider specifications that satisfy $\tilde{R}^2 > 0.80$ when minimizing the MSPE criterion (Dube & Zipperer (2015)) across specifications, where $\tilde{R}^2$ is defined by equation (9). Gray lines are the placebo effects of the control countries and the black line is the treatment effect of the treated country. The vertical lines denote the beginning of the post-treatment period.

Figure 5: **Treatment Effects for All Specifications - Database from Abadie et al. (2010)**



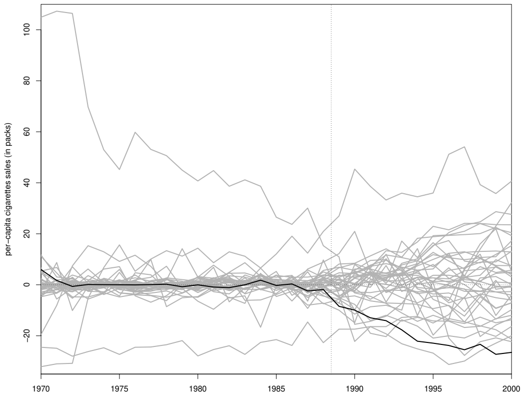
Notes: The solid black line is the original specification by Abadie et al. (2010), whose measure of goodness of fit is $\tilde{R}^2 = 0.0975$, where $\tilde{R}^2$ is defined by equation (9). Gray lines have $\tilde{R}^2 \leq 0.975$ and dashed lines have $\tilde{R}^2 > 0.975$. The vertical line denotes the beginning of the post-treatment period.

Figure 6: **Placebo Effects Using the Average Across Specifications - Database from Abadie et al. (2010)**



Notes: We only consider specifications that satisfy $\tilde{R}^2 > 0.0975$ to compute the average treatment effect across specification, where $\tilde{R}^2$ is defined by equation (9) Gray lines are the placebo effects of the control state and the black line is the treatment effect of California. The vertical line denotes the beginning of the post-treatment period.

Figure 7: **Placebo Effects Using the MSPE Criterion (Specification 5a) - Database from Abadie et al. (2010)**



Notes: We only consider specifications that satisfy $\tilde{R}^2 > 0.975$ when minimizing the MSPE criterion (Dube & Zipperer (2015)) across specifications, where $\tilde{R}^2$ is defined by equation (9). Gray lines are the placebo effects of the control states and the black line is the treatment effect of California. The vertical line denotes the beginning of the post-treatment period.

Table 1: **Specification searching**

	Stationary Model		Non-stationary Model	
	5% test	10% test	5% test	10% test
	(1)	(2)	(3)	(4)
$T_0 = 12$	0.127	0.225	0.124	0.221
	(0.005)	(0.006)	(0.005)	(0.006)
$T_0 = 32$	0.128	0.234	0.137	0.240
	(0.005)	(0.006)	(0.005)	(0.006)
$T_0 = 100$	0.122	0.222	0.130	0.236
	(0.005)	(0.006)	(0.005)	(0.006)
$T_0 = 400$	0.114	0.212	0.118	0.214
	(0.005)	(0.006)	(0.005)	(0.006)

Note: Rejection rates are estimated based on 5,000 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. *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: **Probability of good pre-treatment fit**

	Stationary model		Non-stationary model	
	$\tilde{R}^2 > 0.80$	$\tilde{R}^2 > 0.95$	$\tilde{R}^2 > 0.80$	$\tilde{R}^2 > 0.95$
	(1)	(2)	(3)	(4)
Panel A: At least one specification with good fit				
$T_0 = 12$	0.942	0.262	0.988	0.609
	(0.002)	(0.004)	(0.001)	(0.004)
$T_0 = 32$	0.994	0.091	1.000	0.818
	(0.002)	(0.004)	(0.001)	(0.004)
$T_0 = 100$	1.000	0.003	1.000	0.997
	(0.002)	(0.004)	(0.001)	(0.004)
$T_0 = 400$	1.000	0.000	1.000	1.000
	(0.002)	(0.004)	(0.001)	(0.004)
Panel B: # of specifications with good fit				
$T_0 = 12$	4.522	0.923	5.128	2.604
	(0.012)	(0.015)	(0.008)	(0.022)
$T_0 = 32$	5.068	0.339	5.389	4.002
	(0.012)	(0.015)	(0.008)	(0.022)
$T_0 = 100$	5.169	0.007	5.753	5.246
	(0.012)	(0.015)	(0.008)	(0.022)
$T_0 = 400$	5.166	0.000	5.991	5.681
	(0.012)	(0.015)	(0.008)	(0.022)
Panel C: Specification 1 has a good fit				
$T_0 = 12$	0.162	0.015	0.300	0.073
	(0.005)	(0.001)	(0.006)	(0.006)
$T_0 = 32$	0.164	0.005	0.394	0.131
	(0.005)	(0.001)	(0.006)	(0.005)
$T_0 = 100$	0.169	0.000	0.753	0.268
	(0.005)	(0.001)	(0.006)	(0.005)
$T_0 = 400$	0.166	0.000	0.991	0.681
	(0.005)	(0.001)	(0.006)	(0.006)

Note: Descriptive statistics are estimated based on 5,000 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. T_0 is the number of pre-treatment periods. Our measure of goodness of fit is defined by equation (9). We consider two definitions of good fit; $\tilde{R}^2 > 0.80$ and $\tilde{R}^2 > 0.95$.

Table 3: **Specification searching conditional on a good pre-treatment fit**

	Stationary Model		Non-stationary Model	
	5% test	10% test	5% test	10% test
	(1)	(2)	(3)	(4)
	Panel A: $\tilde{R}^2 > 0.80$			
$T_0 = 12$	0.106	0.188	0.111	0.194
	(0.004)	(0.005)	(0.005)	(0.006)
$T_0 = 32$	0.100	0.182	0.119	0.205
	(0.004)	(0.005)	(0.005)	(0.006)
$T_0 = 100$	0.090	0.157	0.124	0.221
	(0.004)	(0.005)	(0.005)	(0.006)
$T_0 = 400$	0.079	0.143	0.118	0.214
	(0.004)	(0.005)	(0.005)	(0.006)
	Panel B: $\tilde{R}^2 > 0.95$			
$T_0 = 12$	0.197	0.321	0.125	0.215
	(0.011)	(0.013)	(0.006)	(0.007)
$T_0 = 32$	0.192	0.328	0.113	0.191
	(0.019)	(0.022)	(0.005)	(0.006)
$T_0 = 100$	0.154	0.308	0.103	0.179
	(0.110)	(0.130)	(0.004)	(0.006)
$T_0 = 400$	-	-	0.107	0.191
	-	-	(0.004)	(0.006)

Note: Rejection rates are estimated based on 5,000 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. *z% test* indicates that the nominal size of the analyzed test is z% and T_0 is the number of pre-treatment periods. Our measure of goodness of fit is defined by equation (9). We consider two definitions of good fit; $\tilde{R}^2 > 0.80$ and $\tilde{R}^2 > 0.95$.

Table 4: **Variability and Misallocation of weights**

	Misallocation of weights in specification:						Variability of weights	
	1	2	3	4	5	6	All	Exclude 1
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
				Panel A: Stationary Model				
$T_0 = 12$	0.811	0.226	0.260	0.290	0.243	0.247	0.736	0.317
	(0.005)	(0.002)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
$T_0 = 32$	0.810	0.147	0.143	0.180	0.141	0.142	0.763	0.181
	(0.005)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.004)	(0.001)
$T_0 = 100$	0.812	0.110	0.099	0.124	0.099	0.099	0.774	0.115
	(0.005)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.004)	(0.001)
$T_0 = 400$	0.813	0.091	0.086	0.096	0.086	0.085	0.769	0.069
	(0.005)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	(0.004)	(0.000)
				Panel B: Non-stationary Model				
$T_0 = 12$	0.807	0.192	0.219	0.249	0.209	0.212	0.753	0.287
	(0.005)	(0.002)	(0.003)	(0.003)	(0.003)	(0.003)	(0.004)	(0.003)
$T_0 = 32$	0.812	0.117	0.122	0.151	0.120	0.119	0.784	0.165
	(0.005)	(0.001)	(0.001)	(0.002)	(0.001)	(0.001)	(0.004)	(0.001)
$T_0 = 100$	0.814	0.086	0.081	0.099	0.081	0.082	0.794	0.107
	(0.005)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.004)	(0.001)
$T_0 = 400$	0.818	0.073	0.070	0.077	0.070	0.071	0.792	0.070
	(0.005)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	(0.005)	(0.000)

Note: The average of misallocated weights is based on 5,000 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. The average of variability of weights is based on 5,000 observations and captures the average maximum difference of allocated weights across specifications. Specification 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. T_0 is the number of pre-treatment periods.

Table 5: **Specification searching - Excluding specification 1**

	Stationary Model		Non-stationary Model	
	5% test	10% test	5% test	10% test
	(1)	(2)	(3)	(4)
Panel A: Unconditional				
$T_0 = 12$	0.105	0.189	0.106	0.187
	(0.004)	(0.005)	(0.004)	(0.005)
$T_0 = 32$	0.099	0.180	0.110	0.186
	(0.004)	(0.005)	(0.004)	(0.005)
$T_0 = 100$	0.087	0.152	0.098	0.167
	(0.004)	(0.005)	(0.004)	(0.005)
$T_0 = 400$	0.077	0.140	0.080	0.141
	(0.004)	(0.005)	(0.004)	(0.005)
Panel B: Conditional on $\tilde{R}^2 > 0.80$				
$T_0 = 12$	0.101	0.182	0.104	0.184
	(0.004)	(0.005)	(0.004)	(0.005)
$T_0 = 32$	0.098	0.178	0.110	0.186
	(0.004)	(0.005)	(0.004)	(0.005)
$T_0 = 100$	0.087	0.152	0.098	0.167
	(0.004)	(0.005)	(0.004)	(0.005)
$T_0 = 400$	0.077	0.140	0.080	0.141
	(0.004)	(0.005)	(0.004)	(0.005)
Panel C: Conditional on $\tilde{R}^2 > 0.95$				
$T_0 = 12$	0.192	0.316	0.124	0.214
	(0.011)	(0.013)	(0.005)	(0.007)
$T_0 = 32$	0.187	0.326	0.111	0.188
	(0.018)	(0.022)	(0.005)	(0.006)
$T_0 = 100$	0.154	0.308	0.098	0.168
	(0.109)	(0.129)	(0.004)	(0.005)
$T_0 = 400$	-	-	0.080	0.141
	-	-	(0.004)	(0.005)

Note: Rejection rates are estimated based on 5,000 observations and on five specifications — (1) all pre-treatment outcome values, (2) the first half of the pre-treatment outcome values, (3) the first three quarters 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% and T_0 is the number of pre-treatment periods. Our measure of goodness of fit is defined by equation (9). We consider two definitions of good fit; $\tilde{R}^2 > 0.80$ and $\tilde{R}^2 > 0.95$.

Table 6: **Probability of good pre-treatment fit - CPS**

	Log wages				Employment			
	Women		Men		Women		Men	
	$\tilde{R}^2 > 0.80$	$\tilde{R}^2 > 0.95$	$\tilde{R}^2 > 0.80$	$\tilde{R}^2 > 0.95$	$\tilde{R}^2 > 0.80$	$\tilde{R}^2 > 0.95$	$\tilde{R}^2 > 0.80$	$\tilde{R}^2 > 0.95$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Panel A: At least one specification							
$T_0 = 12$	0.913	0.573	0.875	0.410	0.284	0.033	0.156	0.017
	(0.028)	(0.043)	(0.031)	(0.044)	(0.030)	(0.011)	(0.032)	(0.008)
$T_0 = 32$	0.963	0.950	0.982	0.906	0.655	0.042	0.066	0.000
	(0.026)	(0.028)	(0.018)	(0.032)	(0.057)	(0.024)	(0.030)	-
	Panel B: # of specifications with good fit							
$T_0 = 12$	5.368	2.713	5.000	1.741	1.160	0.093	0.496	0.034
	(0.176)	(0.233)	(0.194)	(0.214)	(0.135)	(0.037)	(0.115)	(0.018)
$T_0 = 32$	5.771	5.657	5.886	5.279	3.542	0.193	0.303	0.000
	(0.160)	(0.172)	(0.112)	(0.202)	(0.327)	(0.111)	(0.143)	-

Note: Descriptive statistics are estimated based 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 — 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 (9). We consider two definitions of good fit; $\tilde{R}^2 > 0.80$ and $\tilde{R}^2 > 0.95$.

Table 7: **Specification searching - CPS simulations**

	Log wages				Employment			
	Women		Men		Women		Men	
	5% test	10% test	5% test	10% test	5% test	10% test	5% test	10% test
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Panel A: Unconditional							
$T_0 = 12$	0.114***	0.204***	0.109***	0.188***	0.122***	0.209***	0.112***	0.205***
	(0.013)	(0.019)	(0.012)	(0.016)	(0.012)	(0.016)	(0.013)	(0.018)
$T_0 = 32$	0.102**	0.184**	0.097*	0.175**	0.092	0.166*	0.100**	0.191***
	(0.026)	(0.035)	(0.026)	(0.038)	(0.030)	(0.039)	(0.023)	(0.035)
	Panel B: Conditional on $\tilde{R}^2 > 0.80$							
$T_0 = 12$	0.123***	0.219***	0.119***	0.201***	0.201***	0.331***	0.253***	0.408***
	(0.013)	(0.019)	(0.012)	(0.017)	(0.023)	(0.026)	(0.028)	(0.029)
$T_0 = 32$	0.104**	0.187**	0.099*	0.178**	0.126*	0.210**	0.151	0.241
	(0.027)	(0.037)	(0.027)	(0.038)	(0.042)	(0.054)	(0.080)	(0.108)
	Panel C: Conditional on $\tilde{R}^2 > 0.95$							
$T_0 = 12$	0.163***	0.275***	0.171***	0.280***	0.403***	0.522***	0.417***	0.595***
	(0.016)	(0.023)	(0.017)	(0.022)	(0.052)	(0.045)	(0.046)	(0.055)
$T_0 = 32$	0.106**	0.189**	0.094*	0.178**	0.185***	0.379***	-	-
	(0.027)	(0.037)	(0.025)	(0.037)	(0.043)	(0.057)	-	-

Note: Rejection rates are estimated based 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 — 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%. Our measure of goodness of fit is defined by equation (9). We consider two definitions of good fit; $\tilde{R}^2 > 0.80$ and $\tilde{R}^2 > 0.95$. * 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%.

Table 8: **Specification searching excluding specification 1 - CPS simulations**

	Log wages				Employment			
	Women		Men		Women		Men	
	5% test	10% test	5% test	10% test	5% test	10% test	5% test	10% test
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Panel A: Unconditional							
$T_0 = 12$	0.100***	0.178***	0.096***	0.165***	0.105***	0.183***	0.100***	0.181***
	(0.012)	(0.018)	(0.011)	(0.016)	(0.011)	(0.016)	(0.012)	(0.017)
$T_0 = 32$	0.079	0.145	0.071	0.135	0.070	0.133	0.077	0.149
	(0.022)	(0.032)	(0.021)	(0.033)	(0.025)	(0.034)	(0.019)	(0.030)
	Panel B: Conditional on $\tilde{R}^2 > 0.80$							
$T_0 = 12$	0.108***	0.193***	0.106***	0.183***	0.198***	0.322***	0.251***	0.406***
	(0.012)	(0.018)	(0.011)	(0.016)	(0.023)	(0.026)	(0.027)	(0.029)
$T_0 = 32$	0.082	0.149	0.072	0.137	0.105	0.185*	0.151	0.241
	(0.023)	(0.033)	(0.021)	(0.033)	(0.036)	(0.049)	(0.080)	(0.108)
	Panel C: Conditional on $\tilde{R}^2 > 0.95$							
$T_0 = 12$	0.154***	0.264***	0.166***	0.273***	0.403***	0.522***	0.417***	0.595***
	(0.016)	(0.022)	(0.016)	(0.022)	(0.052)	(0.045)	(0.046)	(0.055)
$T_0 = 32$	0.083	0.151	0.076	0.142	0.185***	0.379***	-	-
	(0.024)	(0.033)	(0.022)	(0.034)	(0.043)	(0.057)	-	-

Note: Rejection rates are estimated based on five specifications — (1) all pre-treatment outcome values, (2) the first half of the pre-treatment outcome values, (3) the first three quarters of the pre-treatment outcome values, (4) odd pre-treatment outcome values, and (5) even pre-treatment 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%. Our measure of goodness of fit is defined by equation (9). We consider two definitions of good fit; $\tilde{R}^2 > 0.80$ and $\tilde{R}^2 > 0.95$. * 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%.

Table 9: **Specification Searching - Database from Smith (2015)**

	Equatorial Guinea		Ecuador	
	p-value	$\tilde{R}^2$	p-value	$\tilde{R}^2$
	(1)	(2)	(3)	(4)
(1a)	0.031	0.828	0.538	0.881
(1b)	0.031	0.744	0.769	0.804
(2a)	0.031	0.848	0.538	0.804
(2b)	0.031	0.657	0.077	0.972
(3a)	0.031	0.866	0.538	0.881
(3b)	0.031	0.797	0.385	0.975
(4a)	0.031	0.809	0.615	0.880
(4b)	0.031	0.790	0.231	0.972
(5a)	0.031	0.777	0.538	0.881
(5b)	0.031	0.832	0.308	0.975
(6a)	0.031	0.891	0.308	0.969
(6b)	0.031	0.536	0.077	0.970
# of Permutations	33		14	

Note: We analyze twelve different specifications. The number of the specifications refer to: (1) even pre-treatment outcome values (original specification by Smith (2015)), (2) the mean of all pre-treatment outcome values, (3) all pre-treatment outcome values, (4) the first half of the pre-treatment outcome values, (5) the first three quarters of the pre-treatment outcome values and (6) odd pre-treatment outcome values. Specifications that end with an *a* include the covariates *ethnic fragmentation* and *population size one year before the discovery*, while specifications that end with an *b* do not include covariates. Our measure of goodness of fit is defined by equation (9).

Table 10: **Specification Searching - Database from Abadie et al. (2010)**

Specification	(1a)	(1b)	(2a)	(2b)	(3a)	(3b)	(4a)	(4b)
p-value	0.026	0.077	0.077	0.077	0.077	0.077	0.026	0.051
$\tilde{R}^2$	0.975	0.909	0.828	0.525	0.979	0.979	0.968	0.969

Specification	(5a)	(5b)	(6a)	(6b)	(7a)	(7b)
p-value	0.077	0.077	0.026	0.051	0.077	0.077
$\tilde{R}^2$	0.976	0.974	0.978	0.978	0.979	0.978

Note: We analyze fourteen different specifications. The number of the specifications refer to: (1) original specification by Abadie et al. (2010)), (2) the mean of all pre-treatment outcome values, (3) all pre-treatment outcome values, (4) the first half of the pre-treatment outcome values, (5) the first three quarters of the pre-treatment outcome values, (6) odd pre-treatment outcome values and (7) even pre-treatment outcome values. Specifications that end with an *a* include the covariates *average retail price of cigarettes*, *per capita state personal income (logged)*, *percentage of the population age 15–24*, and *per capita beer consumption*, while specifications that end with an *b* do not include covariates. Our measure of goodness of fit is defined by equation (9).

ONLINE APPENDIX (NOT FOR PUBLICATION)

A Model with time-invariant covariates

In Section 3 we provide evidence that specification 1 (pre-treatment outcome mean as economic predictor) fails to take into account the time-series dynamics of the data, which implies that the SC estimator using this specification does not converge to the SC estimators using the other specifications. As a consequence, the possibilities for specification searching do not vanish even when the number of pre-treatment periods is large. However, in most applications that use the pre-treatment outcome mean as economic predictor, 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 Section 3 can arise even when we consider specifications that also include time-invariant covariates as economic predictors.

The alternative DGP is given by:

$$Y^0_{j,t} = \delta_t + \lambda^k_t + \theta_t Z_i + \epsilon_{jt} \tag{10}$$

where $Z_i = 1$ for $i = 1, ..., 10$ and $Z_i = 0$ for $i = 11, ..., 20$. As in our DGP from Section 3, we consider $K = 10$.[39] We consider that λ^k_t 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 six specifications as in Section 3, except that we also include Z_i as economic predictor.

In column 1 of Table A.8 we present the proportion of misallocated weights for specification 1. Similarly to our findings in Section 3, specification 1 misallocates significantly more weight relative to the other specifications, and, importantly, the misallocation of weights remains constant when T_0 increases.[40] In column 2 of Table A.8 we show that our measure of variability

[39] Therefore, units 1 and 2 follow the trend λ^1_t, units 3 and 4 follow the trend λ^2_t, and so on.

[40] The misallocation for the other specifications is similar to the stationary model considered in Section 3. Results available upon request.

of weights also remain constant with T_0, which implies that there is still substantial differences in the SC estimators when we consider different specifications, even when T_0 is large. Given that specification 1 remains very different from the other specification even with large T_0, the possibilities for specification searching remain high for large T_0, as presented in columns 3 and 4 of Table A.8. This is similar to our findings in Section 3. The intuition is that including Z_i as an economic predictor helps prevent that the SC estimator will allocate positive weights to units $i = 11, ..., 20$. However, this specification still fails to capture the time-series dynamics when allocating weights among units $i = 2, ..., 10$.

B Appendix Tables

Table A.1: **Specification searching - Alternative Models**

	Model (7) with $\epsilon_{j,t} \sim N(0,1)$		Model (7) with $K = 2$	
	5% test	10% test	5% test	10% test
	(1)	(2)	(3)	(4)
$T_0 = 12$	0.121	0.218	0.124	0.225
	(0.004)	(0.006)	(0.005)	(0.006)
$T_0 = 32$	0.122	0.211	0.130	0.229
	(0.004)	(0.006)	(0.005)	(0.006)
$T_0 = 100$	0.113	0.208	0.118	0.217
	(0.004)	(0.006)	(0.005)	(0.006)
$T_0 = 400$	0.100	0.187	0.113	0.202
	(0.004)	(0.006)	(0.005)	(0.006)

Note: Rejection rates are estimated based on 5,000 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. *z% test* indicates that the nominal size of the analyzed test is z% and T_0 is the number of pre-treatment periods.

Table A.2: **Probability of good pre-treatment fit - Alternative Models**

	Model (7) with $\epsilon_{j,t} \sim N(0,1)$		Model (7) with $K=2$	
	$\tilde{R}^2 > 0.80$	$\tilde{R}^2 > 0.95$	$\tilde{R}^2 > 0.80$	$\tilde{R}^2 > 0.95$
	(1)	(2)	(3)	(4)
Panel A: At least one specification with good fit				
$T_0 = 12$	0.234	0.005	0.991	0.671
	(0.003)	(0.001)	(0.001)	(0.007)
$T_0 = 32$	0.008	0.000	1.000	0.642
	(0.003)	(0.001)	(0.001)	(0.007)
$T_0 = 100$	0.000	0.000	1.000	0.493
	(0.003)	(0.001)	(0.001)	(0.007)
$T_0 = 400$	0.000	0.000	1.000	0.349
	(0.003)	(0.001)	(0.001)	(0.007)
Panel B: # of specifications with good fit				
$T_0 = 12$	0.615	0.008	5.606	2.764
	(0.010)	(0.001)	(0.008)	(0.031)
$T_0 = 32$	0.018	0.000	5.794	2.636
	(0.010)	(0.001)	(0.008)	(0.031)
$T_0 = 100$	0.000	0.000	5.823	1.801
	(0.010)	(0.001)	(0.008)	(0.031)
$T_0 = 400$	0.000	0.000	5.801	1.322
	(0.010)	(0.001)	(0.008)	(0.031)
Panel C: Specification 1 has a good fit				
$T_0 = 12$	0.003	0	0.732	0.098
	(0.000)	-	(0.006)	(0.003)
$T_0 = 32$	0.000	0	0.796	0.048
	(0.000)	-	(0.006)	(0.003)
$T_0 = 100$	0.000	0	0.823	0.002
	(0.000)	-	(0.006)	(0.003)
$T_0 = 400$	0.000	0	0.801	0.000
	(0.000)	-	(0.006)	(0.003)

Note: Descriptive statistics are estimated based on 5,000 observations and on the six specifications defined in Section 3. T_0 is the number of pre-treatment periods. Our measure of goodness of fit is defined by equation (9). We consider two definitions of good fit; $\tilde{R}^2 > 0.80$ and $\tilde{R}^2 > 0.95$.

Table A.3: **Specification searching conditional on a good pre-treatment fit - Alternative Models**

	Model (7) with $\epsilon_{j,t} \sim N(0,1)$		Model (7) with $K = 2$	
	5% test	10% test	5% test	10% test
	(1)	(2)	(3)	(4)
		Panel A: $\tilde{R}^2 > 0.80$		
$T_0 = 12$	0.235	0.379	0.122	0.222
	(0.012)	(0.014)	(0.005)	(0.006)
$T_0 = 32$	0.286	0.524	0.125	0.222
	(0.066)	(0.075)	(0.005)	(0.006)
$T_0 = 100$	-	-	0.113	0.210
	-	-	(0.005)	(0.006)
$T_0 = 400$	-	-	0.109	0.194
	-	-	(0.005)	(0.006)
		Panel B: $\tilde{R}^2 > 0.95$		
$T_0 = 12$	0.808	0.808	0.144	0.256
	(0.079)	(0.079)	(0.006)	(0.007)
$T_0 = 32$	-	-	0.135	0.229
	-	-	(0.006)	(0.007)
$T_0 = 100$	-	-	0.103	0.190
	-	-	(0.007)	(0.008)
$T_0 = 400$	-	-	0.087	0.152
	-	-	(0.008)	(0.01)

Note: Rejection rates are estimated based on 5,000 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. *z% test* indicates that the nominal size of the analyzed test is z% and T_0 is the number of pre-treatment periods. Our measure of goodness of fit is defined by equation (9). We consider two definitions of good fit; $\tilde{R}^2 > 0.80$ and $\tilde{R}^2 > 0.95$.

Table A.4: **Variability and Misallocation of weights - Alternative Models**

	Misallocation of weights in specification:						Variability of weights	
	1	2	3	4	5	6	All	Exclude 1
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Panel A: Model (7) with $\epsilon_{j,t} \sim N(0,1)$							
$T_0 = 12$	0.883	0.703	0.734	0.748	0.719	0.725	0.677	0.554
	(0.003)	(0.003)	(0.004)	(0.004)	(0.004)	(0.004)	(0.003)	(0.003)
$T_0 = 32$	0.872	0.576	0.571	0.618	0.564	0.570	0.631	0.43
	(0.004)	(0.002)	(0.003)	(0.004)	(0.003)	(0.003)	(0.002)	(0.002)
$T_0 = 100$	0.878	0.508	0.480	0.531	0.477	0.479	0.605	0.295
	(0.003)	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.001)
$T_0 = 400$	0.877	0.472	0.461	0.480	0.462	0.459	0.574	0.170
	(0.003)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)
	Panel B: Model (7) with $K = 2$							
$T_0 = 12$	0.207	0.044	0.049	0.06	0.048	0.048	0.737	0.586
	(0.003)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.003)
$T_0 = 32$	0.219	0.027	0.029	0.041	0.028	0.029	0.673	0.497
	(0.003)	(0.000)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.002)
$T_0 = 100$	0.204	0.016	0.018	0.025	0.018	0.018	0.612	0.372
	(0.003)	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	(0.003)	(0.001)
$T_0 = 400$	0.219	0.008	0.010	0.013	0.010	0.010	0.540	0.217
	(0.003)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.003)	(0.001)

Note: The average of misallocated weights is based on 5,000 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. The average of variability of weights is based on 5,000 observations and captures the average maximum difference of allocated weights across specifications. Specification 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. T_0 is the number of pre-treatment periods.

Table A.5: **Specification searching - Excluding specification 1 - Alternative Models**

	Model (7) with $\epsilon_{j,t} \sim N(0,1)$		Model (7) with $K = 2$	
	5% test	10% test	5% test	10% test
	(1)	(2)	(3)	(4)
	Panel A: Unconditional			
$T_0 = 12$	0.102	0.186	0.098	0.186
	(0.004)	(0.005)	(0.004)	(0.005)
$T_0 = 32$	0.106	0.184	0.104	0.186
	(0.004)	(0.005)	(0.004)	(0.005)
$T_0 = 100$	0.092	0.172	0.09	0.172
	(0.004)	(0.005)	(0.004)	(0.005)
$T_0 = 400$	0.079	0.148	0.081	0.148
	(0.004)	(0.005)	(0.004)	(0.005)
	Panel B: Conditional on $\tilde{R}^2 > 0.80$			
$T_0 = 12$	0.234	0.377	0.099	0.187
	(0.012)	(0.014)	(0.004)	(0.005)
$T_0 = 32$	0.286	0.524	0.104	0.186
	(0.066)	(0.075)	(0.004)	(0.005)
$T_0 = 100$	-	-	0.09	0.172
	-	-	(0.004)	(0.005)
$T_0 = 400$	-	-	0.081	0.148
	-	-	(0.004)	(0.005)
	Panel C: Conditional on $\tilde{R}^2 > 0.95$			
$T_0 = 12$	0.808	0.808	0.133	0.243
	(0.079)	(0.079)	(0.006)	(0.007)
$T_0 = 32$	-	-	0.13	0.223
	-	-	(0.006)	(0.007)
$T_0 = 100$	-	-	0.103	0.190
	-	-	(0.006)	(0.008)
$T_0 = 400$	-	-	0.087	0.152
	-	-	(0.008)	(0.010)

Note: Rejection rates are estimated based on 5,000 observations and on specifications (2)-(6), defined in Section 3. *z% test* indicates that the nominal size of the analyzed test is z% and T_0 is the number of pre-treatment periods.Our measure of goodness of fit is defined by equation (9). We consider two definitions of good fit; $\tilde{R}^2 > 0.80$ and $\tilde{R}^2 > 0.95$.

Table A.6: **Specification searching - Excluding specification 2**

	Stationary Model		Non-stationary Model	
	5% test	10% test	5% test	10% test
	(1)	(2)	(3)	(4)
	Panel A: Unconditional			
$T_0 = 12$	0.125	0.222	0.122	0.218
	(0.005)	(0.006)	(0.005)	(0.006)
$T_0 = 32$	0.126	0.231	0.135	0.239
	(0.005)	(0.006)	(0.005)	(0.006)
$T_0 = 100$	0.121	0.220	0.128	0.233
	(0.005)	(0.006)	(0.005)	(0.006)
$T_0 = 400$	0.114	0.211	0.117	0.213
	(0.005)	(0.006)	(0.005)	(0.006)
	Panel B: Conditional on $\tilde{R}^2 > 0.80$			
$T_0 = 12$	0.103	0.185	0.109	0.190
	(0.004)	(0.005)	(0.005)	(0.006)
$T_0 = 32$	0.098	0.178	0.117	0.203
	(0.004)	(0.005)	(0.005)	(0.006)
$T_0 = 100$	0.089	0.154	0.122	0.219
	(0.004)	(0.005)	(0.005)	(0.006)
$T_0 = 400$	0.079	0.142	0.117	0.213
	(0.004)	(0.005)	(0.005)	(0.006)
	Panel C: Conditional on $\tilde{R}^2 > 0.95$			
$T_0 = 12$	0.208	0.336	0.121	0.210
	(0.012)	(0.014)	(0.006)	(0.007)
$T_0 = 32$	0.188	0.328	0.110	0.189
	(0.020)	(0.024)	(0.005)	(0.006)
$T_0 = 100$	0.286	0.429	0.101	0.176
	(0.152)	(0.178)	(0.004)	(0.006)
$T_0 = 400$	-	-	0.106	0.189
	-	-	(0.004)	(0.006)

Note: Rejection rates are estimated based on 5,000 observations and on specifications (1) and (3)-(6), defined on Section 3. *z% test* indicates that the nominal size of the analyzed test is z% and T_0 is the number of pre-treatment periods. Our measure of goodness of fit is defined by equation (9). We consider two definitions of good fit; $\tilde{R}^2 > 0.80$ and $\tilde{R}^2 > 0.95$.

Table A.7: **Specification searching excluding specification 2 - CPS simulations**

	Log wages				Employment			
	Women		Men		Women		Men	
	5% test	10% test	5% test	10% test	5% test	10% test	5% test	10% test
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Unconditional								
$T_0 = 12$	0.110***	0.198***	0.106***	0.184***	0.119***	0.205***	0.110***	0.199***
	(0.012)	(0.018)	(0.011)	(0.016)	(0.011)	(0.016)	(0.013)	(0.017)
$T_0 = 32$	0.101**	0.183**	0.096*	0.173*	0.092	0.166*	0.099**	0.190**
	(0.025)	(0.035)	(0.026)	(0.037)	(0.030)	(0.039)	(0.023)	(0.035)
Panel B: Conditional on $\tilde{R}^2 > 0.80$								
$T_0 = 12$	0.120***	0.213***	0.116***	0.198***	0.201***	0.329***	0.256***	0.407***
	(0.012)	(0.018)	(0.012)	(0.016)	(0.023)	(0.025)	(0.027)	(0.028)
$T_0 = 32$	0.103**	0.186**	0.098*	0.176**	0.126*	0.210**	0.151	0.246
	(0.026)	(0.036)	(0.027)	(0.038)	(0.042)	(0.054)	(0.082)	(0.109)
Panel C: Conditional on $\tilde{R}^2 > 0.95$								
$T_0 = 12$	0.163***	0.274***	0.173***	0.283***	0.431***	0.528***	0.409***	0.545***
	(0.015)	(0.022)	(0.016)	(0.021)	(0.070)	(0.059)	(0.066)	(0.058)
$T_0 = 32$	0.105**	0.188**	0.094*	0.176**	0.188***	0.375***	-	-
	(0.027)	(0.037)	(0.025)	(0.037)	(0.043)	(0.056)	-	-

Note: Rejection rates are estimated based on five specifications — (1) the mean of all pre-treatment outcome values, (2) the first half of the pre-treatment outcome values, (3) the first three quarters of the pre-treatment outcome values, (4) odd pre-treatment outcome values, and (5) even pre-treatment 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%. Our measure of goodness of fit is defined by equation (9). We consider two definitions of good fit; $\tilde{R}^2 > 0.80$ and $\tilde{R}^2 > 0.95$. * 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%.

Table A.8: **Model with time-invariant covariates**

	Misallocation of Weights	Variability of Weights	Specification Searching	
			5% test	10% test
	(1)	(2)	(3)	(4)
$T_0 = 12$	0.624	0.611	0.126	0.230
	(0.005)	(0.004)	(0.005)	(0.006)
$T_0 = 32$	0.615	0.599	0.123	0.224
	(0.005)	(0.004)	(0.005)	(0.006)
$T_0 = 100$	0.613	0.593	0.121	0.212
	(0.005)	(0.005)	(0.005)	(0.006)
$T_0 = 400$	0.609	0.578	0.108	0.197
	(0.005)	(0.005)	(0.005)	(0.006)

Note: This table presents results based on 5,000 observations of the MC simulations described in Appendix A. Column 1 presents the misallocation of weights for specification 1 (pre-treatment outcome mean and Z_i as economic predictors). Column 2 presents the variability of weights considering all six specifications including Z_i as economic predictor. Columns (3) and (4) present the probability of rejecting the null in at least one specification at, respectively, 5% and 10% significance level.