SYNTHETIC CONTROL ESTIMATOR: A GENERALIZED INFERENCE PROCEDURE AND CONFIDENCE SETS*

SERGIO FIRPO† and VÍTOR POSSEBOM‡

APRIL 2ND, 2016

ABSTRACT

The Synthetic Control Method (SCM) was proposed to answer questions involving counterfactuals when only one treated unit and a few control units are observed. Although this method was applied in many empirical works, the formal theory behind its inference procedure is still an open question. In order to fulfill this lacuna, we make clear the sufficient hypotheses that guarantee the adequacy of Fisher's Exact Hypothesis Testing Procedure for panel data, allowing us to test any *sharp null hypothesis* and, consequently, to propose a new way to estimate Confidence Sets for the SCM by inverting a test statistic, the first confidence set when we have access only to finite sample, aggregate level data whose cross-sectional dimension may be larger than its time dimension. Moreover, we analyze the size and the power of the proposed test with a Monte Carlo experiment and find that test statistics that use the SCM outperforms test statistics commonly used in the evaluation literature. We also extend our framework for the cases when we observe more than one outcome of interest or more than one treated unit and when heteroskedasticity is present. Furthermore, we propose a sensitivity analysis that allows the researcher to verify the robustness of his or her empirical conclusions to one of our inference procedure's underlying assumptions. Finally, we apply our theoretical developments to reevaluate economic impact of ETA's terrorism on the Basque Country.

Key words: Synthetic Control Estimator, Hypothesis Testing, Confidence Sets

JEL Classification Number: C21, C23, C33

* We are grateful to *FAPESP*, that provided financial aid through grant number 2014/23731-3. We also thank useful suggestions by Ricardo Paes de Barros, Marinho Bertanha, Gabriel Cepaluni, Bruno Ferman, Brigham Frandsen, Dalia Ghanem, Federico Gutierrez, Hugo Jales, Ricardo Masini, Marcela Mello, Áureo de Paula, Cristine Pinto, Edson Severnini and seminar participants at EESP-FGV, the California Econometrics Conference 2015, the 37th Brazilian Meeting of Econometrics, the 2016 Latin American Workshop in Econometrics. All errors are our own.

† Insper Institute of Education and Research. firpo@insper.edu.br

‡ Sao Paulo School of Economics, EESP-FGV: vitorapossebom@gmail.com. *Corresponding author.*

1 INTRODUCTION

The Synthetic Control Method was proposed by Abadie and Gardeazabal [2003], Abadie et al. [2010] and Abadie et al. [2015] to address counterfactual questions involving only one treated unit and a few control units. Intuitively, this method constructs a weighted average of control units that is as similar as possible to the treated unit regarding the pre-treatment outcome variable and covariates. For this reason, this weighted average of control units is known as the synthetic control. Although the empirical literature applying the Synthetic Control Method is vast[1], this tool's theoretical foundation is still under development.

Our first contribution to this literature is to formalize the current existing inference procedure proposed by Abadie et al. [2010]. Adapting the permutation test framework described by Imbens and Rubin [2015] to a synthetic control context, we clearly state hypotheses that guarantee the validity of Fisher's Exact Hypothesis Testing Procedure, a method that compares an observed test statistic to its empirical distribution in order to verify whether there is enough evidence to reject the null hypothesis. Particularly, our framework allow us to test not only the null hypothesis of no effect whatsoever, but also any kind of *sharp null hypothesis*, generalizing the current existing inference procedure. The possibility of testing any *sharp null hypothesis* is relevant in order to approximate the intervention effect function by simpler functions that can be used to predict its future behavior. Most importantly, being able to test more flexible null hypothesis is fundamental to compare the costs and benefits of a policy. For example, one can interpret the intervention effect as the policy's benefit and test whether it is different than its costs. It also enables the empirical researcher to test theories related to the analyzed phenomenon, particularly the ones that predict some specific kind of intervention effect.

Based on our generalization of the current existing inference procedure, we propose a novel way to estimate Confidence Sets for the Synthetic Control Estimator by inverting a test statistic. We modify the method described by Imbens and Rubin [2015] to estimate confidence intervals based on Fisher's Exact Hypothesis Testing Procedure in order to apply it to a panel data framework, using test statistics generated by the Synthetic Control Method. To the best of our knowledge, this is the first work to propose Confidence Sets for the Synthetic Control Estimator when we observe aggregate level data for only one treated unit and a few control units (i.e., small finite samples) in a context whose cross-section dimension may be larger than its time

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

dimension. With our confidence sets, a researcher can quickly show, by using a graph, not only the significance of the estimated intervention effect, but also the precision of this point-estimate. This plot summarizes a large amount of information that is important to measure the strength of qualitative conclusions achieved after an econometric analysis.

Since this generalized inference method and the associated confidence sets can use many different test statistics, we verify, by a Monte Carlo experiment, the size and the power of five test statistics when they are used in this inference procedure. We choose them based on our review of the empirical literature that applies the Synthetic Control Method. More specifically, we compare test statistics that use the Synthetic Control Estimator to test statistics that use simpler methods (e.g.: difference in means and a permuted differences-in-differences test that are commonly used in the evaluation literature) and to the asymptotic inference procedure for the difference-in-differences estimator proposed by Conley and Taber [2011]. We find that a inference procedure based on a test statistic that uses the Synthetic Control Method performs much better than the ones that do not use this method when we compare their size and power.

We also extend our framework to cover hypothesis testing and confidence set estimation for a pooled effect among few treated units, as a formalization and a generalization of the test proposed by Cavallo et al. [2013], and to simultaneously test null hypotheses for different outcome variables. This last extension, that also expands the framework described by Anderson [2008] to a panel data context, is important, for example, to evaluate political reforms (Billmeier and Nannicini [2009], Billmeier and Nannicini [2013], Carrasco et al. [2014], Jinjarak et al. [2013], Sanso-Navarro [2011]) that generally affect multiple outcomes variables, such as income levels and investment. Moreover, we can also interpret each post-intervention time period as a different outcome variable, allowing us to investigate the timing of an intervention effect — a relevant possibility when the empirical researcher aims to uncover short and long term effects. As one last extension, we make some brief comments about cases in which heteroskedasticity is a concern. We stress that choosing a test statistic that is robust to this issue — e.g., the t-test or a modified version of the *RMSPE* test statistic — allows us to apply our generalized inference procedure and our confidence sets to empirical problems that present heteroskedasticity.

Moreover, since our inference procedure assumes that the probability of each unit being treated is known, we propose a sensitivity analysis to check the robustness of the permutation test's decision to changes in this assumption. In particular, the basic form of our inference procedure implicitly assumes that all treatment assignment probabilities are equal. In order to verify the robustness of our permutation test's decision to this specific assumption, we propose, based on the work of Rosenbaum [2002] and Cattaneo et al. [2016], a parametric form to the treatment assignment probabilities that allows the empirical researcher to compute p-values for different assumptions regarding the treatment assignment probabilities.

At the end, we apply our generalized inference procedure, its associated new confidence sets, its extension to the case of simultaneous hypothesis testing and its associated sensitivity analysis to evaluate the statistical significance of the economic impact of ETA's terrorism estimated by

Abadie and Gardeazabal [2003]. With this empirical exercise, we illustrate how our proposed confidence set summarizes a large amount of information in a simple graph. Differently from Abadie and Gardeazabal [2003], we find a non-significant impact of ETA's terrorism on Basque Country's GDP per-capita, implying that it is not possible to draw any conclusion about its size or sign.

Literature Review

Regarding the inference of the Synthetic Control Method, other authors have surely made important previous contributions. Abadie et al. [2010][2] are the first authors to propose a inference procedure that consists in estimating p-values through permutation tests and Abadie et al. [2015] suggest a different test statistic for the same procedure. However, they do not make clear the sufficient hypotheses that guarantee that their proposed p-values are valid. One recent advancement in this direction is Ando and Sävje [2013], who discuss the importance of the *Identical and Independent Distribution* hypothesis for the inference procedure[3] and propose two new test statistics that have adequate size and more power when applied to the above mentioned hypothesis test than the ones proposed by Abadie et al. [2010] and Abadie et al. [2015].

Bauhoff [2014], Calderon [2014] and Severnini [2014] propose a way to apply the Synthetic Control Estimator to many treated and control units that is similar to a matching estimator for panel data, but none of them discusses its statistical properties in detail. Following a similar but more formal approach, Wong [2015] extends the synthetic control estimator to a cross-sectional setting where individual-level data is available and derives its asymptotic distribution when the number of observed individuals goes to infinity. Wong [2015] also explores the synthetic control estimator when panel data (or repeated cross-sections) are available in two levels: an aggregate level (regions), where treatment is assigned, and an individual level, where outcomes are observed. In this framework, he derives the asymptotic distribution of the synthetic control estimator when the number of individuals in each region goes to infinity. Finally, Cavallo et al. [2013] and Dube and Zipperer [2013] develop different ways to apply the Synthetic Control Estimator when there are more than one treated unit and propose tests[4] that are similar to the ones proposed by Abadie et al. [2010], although they do not address the statistical properties of their inference procedures either.

Gobillon and Magnac [2016], also working on a context with more than one treated unit, propose a way to compute confidence intervals for their synthetic control estimator based on bootstrapping the point estimate. Although the authors have not clearly stated the assumptions behind their inference procedure either, it requires a large number of treated and control regions

2 They also discuss the asymptotic unbiasedness of their method. Kaul et al. [2015] deepen this topic by arguing that using all pre-intervention outcomes as economic predictors might provoke bias by forcing the synthetic control estimator to ignore all other predictor covariates.

3 Our hypotheses are different than the ones advocated by Ando and Sävje [2013]. In particular, we do not assume that units are independent and identically distributed.

4 Acemoglu et al. [2013] follows a procedure similar to the one proposed by Cavallo et al. [2013]. However, the former is less computationally demanding than the latter.

in order to be valid and focus exclusively on the time average of the post-intervention effect. Our approach differs from theirs in two ways: it is valid in small samples and allow the construction of confidence sets for the post-intervention effect as a function of time. Consequently, while their inference procedure allows the empirical researcher to test only constant in time intervention effects, our generalized inference procedure allows the empirical researcher to test any function of time as the intervention effect.

Moreover, Carvalho et al. [2015] propose the Artificial Counterfactual Estimator (ArCo), that is similar in purpose to the Synthetic Control Estimator, and derive its asymptotic distribution when the time dimension is large (long panel data sets). However, many of the problems to which the Synthetic Control Method is applied present a cross-section dimension larger than their time dimension, making it impossible to apply the ArCo to them. Wong [2015] also conducts an asymptotic analysis when the pre-intervention period goes to infinity.

Finally, our approach is similar to the way Conley and Taber [2011] estimate confidence intervals for the difference-in-differences estimator in the sense that we also construct confidence sets by inverting a test statistic. However, we differ from them in many aspects. Firstly, while they make a contribution to the difference-in-differences framework, our contribution is inserted in the Synthetic Control literature. Secondly, they assume a functional form for the potential outcomes — imposing that the treatment effect is constant in time — and an arbitrarily large number of control units, while we assume a fixed and (possibly) small number of control units and make no assumptions concerning the potential outcome functional form — i.e., treatment effects can vary in time.

This paper is divided as follows: section 2 explains the Synthetic Control Method as proposed by Abadie and Gardeazabal [2003], Abadie et al. [2010] and Abadie et al. [2015], and formalizes and generalizes its inference procedure; section 3 proposes a way to estimate Confidence Sets for the Synthetic Control Estimator; section 4 analyzes size and power of different tests statistics employed in this hypothesis test through a Monte Carlo experiment; section 5 develops possible extensions to our framework; section 6 proposes a sensitivity analysis related to a underlying assumption of our inference procedure; section 7 applies our proposed inference procedure to the data set about the Basque Country made available by Abadie et al. [2011] and section 8 concludes. Finally, in the appendices, we didactically explain how to compute the test statistics described in section 2, expand the results of our Monte Carlo Experiment and offer a guide to empirical researchers who wish to employ the synthetic control method in their studies.

2 SYNTHETIC CONTROL METHOD

This section is organized in three subsections. The first one presents the Synthetic Control Estimator, while the second one explains its inference procedure based on permutation tests.

The ideas and notation that are used in the next two subsections are mostly based on Abadie et al. [2010]. Finally, in the third subsection, we clearly state the hypotheses that guarantees that the current existing inference procedure is valid, generalizing it to test any *sharp null hypothesis* using any test statistic.

2.1 Synthetic Control Estimator

Suppose that we observe data for $(J+1) \in \mathbb{N}$ regions[5] during $T \in \mathbb{N}$ time periods. Additionally, assume that there is an intervention[6] that affects only region 1[7] from period $T_0 + 1$ to period T uninterruptedly[8], where $T_0 \in (1, T) \cap \mathbb{N}$. Let the scalar $Y^N_{j,t}$ be the potential outcome that would be observed for region j in period t if there were no intervention for $j \in \{1, ..., J+1\}$ and $t \in \{1, ..., T\}$. Let the scalar $Y^I_{j,t}$ be the potential outcome that would be observed for region j in period t if region j faced the intervention at period t. Define

$$\alpha_{j,t} := Y^I_{j,t} - Y^N_{j,t} \tag{1}$$

as the intervention effect (or gap) for region j in period t and $D_{j,t}$ as a dummy variable that assumes value 1 if region j faces the intervention 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

$$Y_{j,t} := Y^N_{j,t} + \alpha_{j,t} D_{j,t}.$$

Since only the first region faces the intervention 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 estimate $(\alpha_{1,T_0+1}, ..., \alpha_{1,T})$. Since $Y^I_{1,t}$ is observable for $t > T_0$, equation (1) guarantees that we only need to estimate $Y^N_{1,t}$ to accomplish this goal.

5 We use the word "region" instead of more generic terms, such as "unit", because most synthetic control applications analyze data that are aggregated at the state or country level. We use the term *donor pool* to designated the entire group of $(J+1)$ observed regions.

6 Although the treatment effect literature commonly uses the more generic expression "treated unit", we adopt the expression "the region that faced an intervention" because it is more common in the comparative politics literature, an area where the synthetic control method is largely applied.

7 In subsection 5.2, we extend this framework to include the case when multiple units face the same or a similar intervention.

8 Two famous examples of interventions that affect uninterruptedly a region are Proposition 99 — an Tobacco Control Legislation in California — and the German Reunification, that were studied by Abadie et al. [2010] and Abadie et al. [2015]. If the intervention is interrupted (e.g.: ETA's Terrorism in the Basque Country studied by Abadie and Gardeazabal [2003]), we just have to interpret our treatment differently. Instead of defining the treatment as "region 1 faces an intervention", we define treatment as "region 1 have been exposed to an event that potentially has long term consequences". For example, instead of defining our treatment as "the Basque Country faces constant bombings perpetrated by ETA", we define our treatment as "the Basque Country suffered some bombings perpetrated by ETA".

Let $\mathbf{Y_j} := \left[Y_{j,1}...Y_{j,T_0}\right]'$ be the vector of observed outcomes for region $j \in \{1, ..., J+1\}$ in the pre-intervention period and $\mathbf{X_j}$ a $(K \times 1)$-vector of predictors of $\mathbf{Y_j}$.[9] Let $\mathbf{Y_0} = \left[\mathbf{Y_2}...\mathbf{Y_{J+1}}\right]$ be a $(T_0 \times J)$-matrix and $\mathbf{X_0} = \left[\mathbf{X_2}...\mathbf{X_{J+1}}\right]$ be a $(K \times J)$-matrix.

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

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

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

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

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

Intuitively, $\widehat{\mathbf{W}}$ is a weighting vector that measures the relative importance of each region in the synthetic control of region 1 and $\widehat{\mathbf{V}}$ measures the relative importance of each one of the K predictors. Consequently, this technique makes the synthetic control of region 1 as similar as possible with the actual region 1 considering the K predictors and the pre-intervention values of the outcome variable when we choose the Euclidean metric (or a reweighed version of it) to evaluate the distance between the observed variables for region 1 and the values predicted by the Synthetic Control Method.[10]

9 Some lines of matrix $\mathbf{X_j}$ can be linear combinations of the variables in $\mathbf{Y_j}$.

10 Abadie and Gardeazabal [2003], Abadie et al. [2010] and Abadie et al. [2015] propose two other ways to choose $\widehat{\mathbf{V}}$. The first and most simple one is to use subjective and previous knowledge about the relative importance of each predictor. Since one of the advantages of the Synthetic Control Method is to make the choice of comparison groups in comparative case studies more objective, this method of choosing $\mathbf{V}$ is discouraged by those authors. Another choice method for $\widehat{\mathbf{V}}$ is to divide the pre-intervention period in two sub-periods: one training period and one validation period. While data from the training period are used to solve problem (3), data for the validation period are used to solve problem (4). Intuitively, this technique of cross-validation chooses matrix $\widehat{\mathbf{W}}(\widehat{\mathbf{V}})$ to minimize the out-of-sample prediction errors, an advantage when compared to the method described in the main text. However, the cost of this improvement is the need of a longer pre-intervention period. Moreover, the Stata command made available by those authors also allows the researcher to use a regression-based method in order to compute matrix $\widehat{\mathbf{V}}$. It basically regress matrix $\mathbf{Y_1}$ on $\mathbf{X_1}$ and imposes $v_k = |\beta_k| / \left(\sum_{k'=1}^{K} |\beta_{k'}|\right)$, where v_k is the k-th diagonal element of matrix $\mathbf{V}$ and β_k is the k-th coefficient of the regression of $\mathbf{Y_1}$ on $\mathbf{X_1}$. The choice method that we have chosen to present in the main text is the most used one in the empirical literature.

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

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

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

2.2 Hypothesis Testing

Abadie et al. [2010] propose a inference procedure that

> examines whether or not the estimated effect of the actual intervention is large relative to the distribution of the effects estimated for the regions not exposed to the intervention. This is informative inference if under the hypothesis of no intervention effect the estimated effect of the intervention is not expected to be abnormal relative to the distribution of the placebo effects. (p. 497)

In order to do that, they run a permutation test, i.e., they permute which region 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 in subsection 2.1. Then, they compare the entire vector $\widehat{\boldsymbol{\alpha}}_1 = \left[\widehat{\alpha}_{1,T_0+1} ... \widehat{\alpha}_{1,T}\right]'$ with the empirical distribution of $\widehat{\boldsymbol{\alpha}}_j = \left[\widehat{\alpha}_{j,T_0+1} ... \widehat{\alpha}_{j,T}\right]'$ estimated through the permutation test. If the vector of estimated effects for region 1 is very different (i.e., large in absolute values), they reject the null hypothesis of no effect.

Abadie et al. [2015] note a problem with this approach: $|\widehat{\alpha}_{1,t}|$ can be abnormally large when compared to the empirical distribution of $\left|\widehat{\alpha}_{j,t}\right|$ for some $t \in \{T_0 + 1, ..., T\}$, but not for other time periods. In this case, it is not clear at all whether one should reject the null hypothesis of no effect or not. In order to solve this problem, they recommend to use the empirical distribution of

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

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

$$p := \frac{\sum_{j=1}^{J+1} \mathbb{1}\left[\text{RMSPE}_j \geqslant \text{RMSPE}_1\right]}{J+1}, \tag{6}$$

where $\mathbb{1}[\diamond]$ is the indicator function of event $\diamond$, and reject the null hypothesis of no effect if p is less than some pre-specified significance level, such as the traditional value of 0.1.

Although this RMSPE test statistic solve the problem generated by the time dimension of the Synthetic Control Estimator, Abadie et al. [2015] does not state the sufficient conditions that

guarantee the validity of this procedure[11] nor discuss this test's size and power. We address the former issue in the next subsection and the latter in section 4.

2.3 Formalizing and Generalizing the Inference Procedure

In this section, we follow Imbens and Rubin [2015], adapting their framework to a synthetic control context. We want to formalize and generalize the inference procedure described in subsection 2.2. The first hypothesis that we make is the *stable unit treatment value assumption* (SUTVA):

Assumption 1. The potential outcome vectors $\mathbf{Y}_j^I := \left[Y_{j,1}^I ... Y_{j,T}^I\right]'$ and $\mathbf{Y}_j^N := \left[Y_{j,1}^N ... Y_{j,T}^N\right]'$ for each region $j \in \{1, ..., J+1\}$ do not vary based on whether other regions face the intervention or not (i.e., no spill-over effects in space) and, for each region, there are no different forms or versions of intervention (i.e., single dose treatment), which lead to different potential outcomes [Imbens and Rubin, 2015, p. 19].

The second assumption[12] concerns the treatment assignment:

Assumption 2. The choice of which unit will be treated (i.e., which region is our region 1) is random *conditional on the choice of the donor pool, the observable variables included as predictors and the unobservable variables captured by the path of the outcome variable during the pre-intervention period.*[13]

Assumption 2 is closely related to the literature about *selection on unobservables*. The analogous assumption for the difference-in-differences estimator would be (**Assumption DID**) "The choice of which unit will be treated (i.e., which region is our region 1) is random conditional on the choice of the donor pool, the observable variables included as control variables and *the unobservables variables that are common among all the observed units (but varies over time)*".

Since the differences-in-differences estimator controls only for the unobservables variables that are common among all the observed units (but varies over time) while the synthetic control estimator controls for unobservable variables captured by the path of the outcome variable during the pre-intervention period, assumption 2 is weaker than assumption DID.[14]

Regarding the applicability of assumption 2, it holds true for many empirical applications of the Synthetic Control Estimator. For example, Barone and Mocetti [2014], Cavallo et al.

11 Particularly, it is not clear at all what they mean by "the hypothesis of no intervention effect" [Abadie et al., 2010, p. 497]. Is it a null average effect? Or a null median effect? Or even a null effect for all units in all time periods? Moreover, asking what these authors mean by "the estimated effect of the intervention is not expected to be abnormal" [Abadie et al., 2010, p. 497] is also a valid question.

12 This assumption is our precise definition of "not expected to be abnormal" in footnote 11.

13 The unobservable variables captured by the path of the outcome variable during the pre-intervention period are denoted by unobserved common factors, λ_t, and unknown factor loadings, μ_i in the factor model discussed by Abadie et al. [2010].

14 The differences-in-differences model is actually nested in the factor model discussed by Abadie et al. [2010].

[2013], Coffman and Noy [2011] and DuPont and Noy [2012] evaluate the economic effect of large scale natural disasters, such as earthquakes, hurricanes or volcano eruptions. Although the regions in the globe that frequently faces these disasters are not random, the specific region among them that will be hit by a natural disaster and the timing of this phenomenon is fully random.[15] Moreover, Pinotti [2012a] and Pinotti [2012b] evaluate the economic and political cost of organized crime in Italy exploring the increase in Mafia activities after two large earthquakes. Two other examples of the plausibility of assumption 2 are Smith [2015], who argues that the discovery of large natural resources reserves is *as-if-random*, and Liu [2015], who argues that the location of land-grand universities in the 19$^{\text{th}}$ century is *as-if-random* too.[16]

The third assumption is related to how we interpret the potential outcomes:

Assumption 3. The potential outcomes $\mathbf{Y}_j^I := \left[Y_{j,1}^I ... Y_{j,T}^I\right]'$ and $\mathbf{Y}_j^N := \left[Y_{j,1}^N ... Y_{j,T}^N\right]'$ for each region $j \in \{1, ..., J+1\}$ are fixed but *a priori* unknown quantities.[17]

Implicitly, we assume that we observe the *realization* of a random variable for the *entire population of interest* instead of a random sample of a larger superpopulation.[18]

We note that assumptions 2 and 3 implies that the units in the donor pool are *exchangeable*. In reality, *exchangeability* is the weakest assumption that guarantees the validity of our formal and generalized inference procedure, because it is simply based in a permutation test. However, we believe that, although stronger, assumptions 2 and 3 makes interpretation easier, providing a useful framework in order to discuss the validity of the synthetic control method in applied topics. In particular, assumption 2 justifies one of the robustness checks described in appendix C.

Finally, our null hypothesis is given by a *sharp null hypothesis*:

$$H_0 : Y_{j,t}^I = Y_{j,t}^N + f_j(t) \text{ for each region } j \in \{1, ..., J+1\} \text{ and time period } t \in \{1, ..., T\}, \quad (7)$$

where $f_j : \{1, ..., T\} \to \mathbb{R}$ is a function of time that is specific to each region j.

Observe that a *sharp null hypothesis* allows us to know all potential outcomes for each region regardless of its treatment assignment. Note also that the *exact null hypothesis*

$$H_0 : Y_{j,t}^I = Y_{j,t}^N \text{ for each region } j \in \{1, ..., J+1\} \text{ and time period } t \in \{1, ..., T\}, \quad (8)$$

15 In this example, the donor pool contains all countries that frequently faces natural disasters. Conditional on being in the donor pool, being treated (i.e., being hit by a natural disaster in the analyzed time window) is random.

16 Even in randomized control trials, the synthetic control method may be more interesting than traditional statistical methods when there are only a few treated units — an issue that may emerge due to budget constraints. As we show in section 4, test statistics that use the synthetic control estimator are more powerful than the ones that do not use it.

17 As a consequence of this assumption, all the randomness of our problem come from the treatment assignment.

18 See Imbens and Rubin [2015] for details regarding this interpretation.

is a particular case of the *sharp null hypothesis* (7) and can be interpreted as an hypothesis of no intervention effect whatsoever. We underscore that equation (8) is our precise definition of "no intervention effect" in footnote 11.[19] We also note that, under assumptions 1-3 and the null hypothesis (8), the p-value in equation (6) is valid and known as *Fisher's Exact p-Value*, after Fisher [1971]. In this sense, our inference procedure with the *sharp null hypothesis* is a generalization of the inference procedure proposed by Abadie et al. [2015].

Although the *sharp null hypothesis* (7) is theoretically interesting due to its generality, we almost never have a meaningful null hypothesis that is precise enough to specify individual intervention effects for each observed region. For this reason, we can simply assume a simpler *sharp null hypothesis*[20]:

$$H_0 : Y^I_{j,t} = Y^N_{j,t} + f(t) \text{ for each region } j \in \{1, ..., J+1\} \text{ and time period } t \in \{1, ..., T\}, \quad (9)$$

where $f : \{1, ..., T\} \to \mathbb{R}$.

After formally stating conditions that guarantee the validity of the inference procedure proposed by Abadie et al. [2010] and Abadie et al. [2015], we generalize it to other test statistics and to any *sharp null hypothesis*. We, again, follow Imbens and Rubin [2015].

We define a test statistic θ_f as a known positive real-valued function $\theta_f(\iota, \tau, \mathbf{Y}, \mathbf{X}, f)$ of:

1. the vector $\iota := [\iota_1 ... \iota_{J+1}]' \in \mathbb{R}^{J+1}$ of treatment assignment, where $\iota_j = 1$ if region j faces the intervention at some moment in time and zero otherwise;

2. $\tau := [\tau_1 ... \tau_T]' \in \mathbb{R}^T$, where $\tau_t = 1$ if $t > T_0$ and zero otherwise;

3. the matrix

$$\mathbf{Y} := \begin{bmatrix} Y^I_{1,1}\iota_1\tau_1 + Y^N_{1,1}(1-\iota_1\tau_1) & \dots & Y^I_{1,T}\iota_1\tau_T + Y^N_{1,T}(1-\iota_1\tau_T) \\ & \ddots & \\ Y^I_{J+1,1}\iota_{J+1}\tau_1 + Y^N_{J+1,1}(1-\iota_{J+1}\tau_1) & \dots & Y^I_{J+1,T}\iota_{J+1}\tau_T + Y^N_{J+1,T}(1-\iota_{J+1}\tau_T) \end{bmatrix}$$

   of observed outcomes;

4. the matrix $\mathbf{X} := [\mathbf{X}_1 \ \mathbf{X}_0]$ of predictor variables;

5. the intervention effect function $f : \{1, ..., T\} \to \mathbb{R}$ given by the *sharp null hypothesis* (9).

The observed test statistic is given by $\theta^{obs}_f := \theta(e_1, \tau, \mathbf{Y}, \mathbf{X}, f)$ and, under assumptions 1-3 and the *sharp null hypothesis* (9), we can estimate the entire empirical distribution of θ_f by permuting which region faces the intervention, i.e., by estimating $\theta_f(e_j, \tau, \mathbf{Y}, \mathbf{X}, f)$ for each

19 Observe that the *exact null hypothesis* (8) is stronger than assuming that the *typical* (mean or median) effect across regions is zero.

20 We stress that the *exact null hypothesis* is still a particular case of the simpler *sharp null hypothesis* (9).

$j \in \{1, ..., J+1\}$, where e_j is the j-th canonical vector of $\mathbb{R}^{J+1}$.[21] We reject the *sharp null hypothesis* (9) if

$$p_{\theta_f} := \frac{\sum_{j=1}^{J+1} \mathbb{1}\left[\theta(e_j, \tau, \mathbf{Y}, \mathbf{X}, f) \geqslant \theta_f^{obs}\right]}{J+1} \leqslant \gamma \tag{10}$$

γ is some pre-specified significance level.[22,23] Note that rejecting the null hypothesis implies that there is some region with a non-zero effect for some time period. Moreover, observe that *RMSPE* and any linear combination of the absolute estimated synthetic control gaps are test statistics according to this definition. Consequently, the hypothesis tests proposed by Abadie et al. [2010] and Abadie et al. [2015] are inserted in this framework.[24]

Regarding the choice of function f, there are many interesting options for a empirical researcher. For example, after estimating the intervention effect function $(\widehat{\alpha}_{1,1}, ..., \widehat{\alpha}_{1,T_0+1}, ..., \widehat{\alpha}_{1,T})$, the researcher may want to fit a linear, a quadratic or a exponential function to the estimated points associated with the post-intervention period. He or she can then test whether this fitted function is rejected or not according to our inference procedure. This possibility is useful in order to predict, in a very simple way, the future behavior of the intervention effect function.

Another and possibly the most interesting option for function f is related to cost-benefit analysis. If the intervention cost and its benefit are in the same unit of measurement, function f can be the intervention cost as a function of time and our inference procedure allows the researcher to test whether the intervention effect is different than its costs.

Moreover, function f can be chosen in order to test a theory that predicts a specific form for the intervention effect. For example, imagine that a researcher is interested in analyzing the economic impact of natural disasters (Barone and Mocetti [2014], Cavallo et al. [2013], Coffman and Noy [2011], DuPont and Noy [2012]). Theory predicts three different possible intervention effects in this case: (i) GDP initially increases due to the aid effect and, then, decreases back to its potential level; (ii) GDP initially decreases due to the destruction effect and, then, increases back to its potential level; and (iii) GDP decreases permanently due to a reduction in its potential level. The researcher can choose a inverted U-shaped function f_i, a U-shaped function f_{ii} and a decreasing function f_{iii} and apply our inference procedure to each one of those three *sharp null hypotheses* in order to test which theoretical prediction is not rejected by the data.

21 For a step-by-step guide on how to compute the test statistic and its empirical distribution, see appendix A.

22 In order to compute the p-value p_{θ_f}, we implicitly assume that all regions can be chosen to be the one that faces the intervention with equal probability. In section 6, we propose a sensitivity analysis to measure the robustness of the test's decision to this implicit assumption.

23 Yates [1984] stresses that γ should be chosen carefully and always clearly reported since the discreteness of data (the number of regions is always a finite, usually small, natural number) may preclude the choice of the usual significance levels of 10% or 5%.

24 In section 4, we analyze five different test statistics that were previously proposed in the synthetic control literature in order to select the ones that have power against an alternative hypothesis similar to $H_a : Y_{1,t}^I = Y_{1,t}^N + c_t$ for all time periods $t \in \{T_0 + 1, ..., T\}$, where $c_t \in \mathbb{R}$.

3 CONFIDENCE SETS FOR THE SYNTHETIC CONTROL ESTIMATOR

Following the inference procedure described at the end of subsection 2.3, we can test many different types of *sharp null hypothesis*. Consequently, we can invert the test statistic to estimate confidence sets for the treatment effect function. Formally, under assumptions 1-3, we can construct a γ-confidence set in the space $\mathbb{R}^{\{1,\dots,T\}}$ as

$$CS_{\gamma,\theta} := \left\{ f \in \mathbb{R}^{\{1,\dots,T\}} : p_{\theta_f} > \gamma \right\}. \tag{11}$$

Note that it is easy to interpret $CS_{\gamma,\theta}$: it contains all intervention effect functions whose associated *sharp null hypotheses* are not rejected by the inference procedure described in subsection 2.3.

However, although theoretically possible to define such a general confidence set, null hypothesis (9) might be too general for practical reasons since the space $\mathbb{R}^{\{1,\dots,T\}}$ is too large to be informative and estimating such a confidence set would be computationally infeasible. For these reasons, we believe that it is worth focusing in two subsets of $CS_{\gamma,\theta}$.

Firstly, we propose to assume the following null hypothesis:

$$H_0 : Y^I_{j,t} = Y^N_{j,t} + c \times \mathbb{1}(t \geqslant T_0 + 1) \tag{12}$$

for each region $j \in \{1, \dots, J+1\}$ and time period $t \in \{1, \dots, T\}$, where $c \in \mathbb{R}$. Intuitively, we assume that there is a constant (in space and in time) intervention effect. Note that we can apply the inference procedure described in subsection 2.3 to any $c \in \mathbb{R}$, estimating the empirical distribution of θ_c. Under assumptions 1-3, we can then construct a γ-confidence interval for the constant intervention effect as

$$CI_{\gamma,\theta} := \left\{ f \in \mathbb{R}^{\{1,\dots,T\}} : f(t) = c \text{ and } p_{\theta_c} > \gamma \right\} \subseteq CS_{\gamma,\theta} \tag{13}$$

where $c \in \mathbb{R}$ and $\gamma \in (0,1) \subset \mathbb{R}$. It is easy to interpret $CI_{\gamma,\theta}$: it contains all constant in time intervention effects whose associated *sharp null hypotheses* are not rejected by the inference procedure described in subsection 2.3.

Secondly, we can easily extend (12) and (13) to a linear in time intervention effect (with intercept equal to zero). Assume

$$H_0 : Y^I_{j,t} = Y^N_{j,t} + \tilde{c} \times (t - T_0) \times \mathbb{1}(t \geqslant T_0 + 1) \tag{14}$$

for each region $j \in \{1, \dots, J+1\}$ and time period $t \in \{1, \dots, T\}$, where $\tilde{c} \in \mathbb{R}$. Intuitively, we assume that there is a constant in space, but linear in time intervention effect (with intercept equal to zero). Note that we can apply the inference procedure described in subsection 2.3 to

any $\widetilde{c} \in \mathbb{R}$, estimating the empirical distribution of $\theta_{\widetilde{c}}$. Under assumptions 1-3, we can then construct a γ-confidence set for the linear intervention effect as

$$\widetilde{CS}_{\gamma,\theta} := \left\{ \begin{array}{c} f \in \mathbb{R}^{\{1,\dots,T\}}: \quad f(t) = \widetilde{c} \times (t - T_0) \times \mathbb{1}(t \geqslant T_0 + 1) \\ \text{and } p_{\theta_{\widetilde{c}}} > \gamma \end{array} \right\} \subseteq CS_{\gamma,\theta} \qquad (15)$$

where $\gamma \in (0,1) \subset \mathbb{R}$. It is also easy to interpret $\widetilde{CR}_{\gamma,\theta}$: it contains all linear in time intervention effects (with intercept equal to zero) whose associated *sharp null hypotheses* are not rejected by the inference procedure described in subsection 2.3.

We also note that extending our confidence intervals to two-parameter functions (e.g.: quadratic, exponential and logarithmic functions) is theoretically straightforward as equation (11) makes clear. However, since we believe that computationally estimating such confidence sets would be extremely time consuming for the practitioner, we opted for restricting our main examples to one-parameter functions (equations (13) and (15)).

Moreover, we highlight that confidence sets (13) and (15) summarizes a large amount of relevant information since they not only show the statistical significance of the estimated intervention effect, but also provide a measure of the precision of the point-estimate, indicating the strength of qualitative conclusions. Section 7 exemplifies the communication efficacy of this graphical device.[25]

Finally, we note that our confidence sets are uniform in the sense that they combine information about all time periods in order to describe which *intervention effect functions* are not rejected by the data. If the empirical researcher is interested in only computing point-wise confidence intervals for each period intervention effect, he or she can apply our inference procedure and confidence sets separately for each post-intervention time period $t' \in \{T_0 + 1, ..., T\}$ using the $\left|\widehat{\alpha}_{1,t'}\right|$ as a test statistic.

4 ANALYZING SIZE AND POWER

In this section, we analyze the size and the power of five different test statistics when they are applied to the inference procedure described in subsection 2.3.[26] In order to do that, we assume seven different intervention effects, simulate 5,000 data sets for each intervention effect through a Monte Carlo experiment and, for each data set, we test, at the 10% significance level, the *exact null hypothesis* (equation (8)), following the mentioned inference procedure and using each test statistic. Firstly, we explain how we generated our data sets. Then, we describe our five test statistics. Finally, at the end of this section, we present and discuss the results of our Monte Carlo experiment.

25 In this empirical example, we used a R function that implements confidence sets (13) and (15) using the *RMSPE* as a test statistic. This R function is available at the authors' webpage (https://goo.gl/4Jvd2W).

26 In appendix 2, we discuss the size and the power of other thirteen test statistics.

The first step in our Monte Carlo experiment is to decide the values of the parameters: $J+1$ (number of regions), T (number of time periods), T_0 (number of pre-intervention time periods) and K (number of predictors). In our review of the empirical literature, we found that typical values of these parameters are, approximately, $T = 25$, $T_0 = 15$ and $K = 10$ (nine control variables and the pre-intervention average of the outcome variable). We also set $J+1 = 20$ (one treated region and nineteen control regions). Our data generating process follows equation (5) of Abadie et al. [2010] and is different from the one used by Ando and Sävje [2013]:

$$\begin{aligned} Y^N_{j,t+1} &= \delta_t Y^N_{j,t} + \boldsymbol{\beta}_{t+1}\mathbf{Z}_{j,t+1} + u_{j,t+1} \\ \mathbf{Z}_{j,t+1} &= \kappa_t Y^N_{j,t} + \boldsymbol{\pi}_t \mathbf{Z}_{j,t} + \mathbf{v}_{j,t+1} \end{aligned} \tag{16}$$

for each $j \in \{1, ..., J+1\}$ and $t \in \{0, ..., T-1\}$, where $\mathbf{Z}_{j,t+1}$ is a $(K-1) \times 1$-dimension vector of control variables[27]. The scalar $u_{j,t+1}$ and each element of the $(K-1) \times 1$-dimension vector $\mathbf{v}_{j,t+1}$ are independent random draws from a standard normal distribution. The scalars δ_t and κ_t and each element of $\boldsymbol{\beta}_{t+1}$ and $\boldsymbol{\pi}_t$ are independent random draws from a uniform distribution with lower bound equal to -1 and upper bound equal to +1. We make $\mathbf{Z}_{j,0} = \mathbf{v}_{j,0}$ and $Y^N_{j,0} = \boldsymbol{\beta}_0 \mathbf{Z}_{j,0} + u_{j,0}$. Finally, the potential outcome when region 1 faces the intervention in period $t \in \{1, ..., T\}$ is given by

$$Y^I_{1,t} = Y^N_{1,t} + \lambda \times sd(Y^N_{1,t} | t \leqslant T_0) \times (t - T_0) \times \mathbb{1}[t \geqslant T_0 + 1], \tag{17}$$

where $\lambda \in \{0, 0.05, 0.1, 0.25, 0.5, 1.0, 2.0\}$ is the intervention effect and $sd(\clubsuit|\diamond)$ is the standard deviation of variable $\clubsuit$ conditional on event $\diamond$. Hence, our alternative hypothesis is that there is a linear intervention effect only for region 1, implying that our Monte Carlo experiment investigates what are the most powerful test statistics against this alternative hypothesis[28].

Note that, in each one of the 35,000 Monte Carlo repetitions, we create an entire population of regions. Hence, after realizing the values of the potential outcome variables, we can interpret them as fixed but *a priori* unknown quantities in accordance to assumption 3.[29]

Now that we have explained our data generating process for our 35,000 Monte Carlo repetitions (5,000 repetitions for each different intervention effect λ), we describe the five different test statistics that we use to analyze the size and the power of the inference procedure described in subsection 2.3:

- $\theta^1 := mean\left(\left|\widehat{\alpha}_{\tilde{j},t}\right| \, | t \geqslant T_0 + 1\right)$ is implicitly suggested by Abadie et al. [2010].
- $\theta^2 := RMSPE_{\tilde{j}}$ is used by Abadie et al. [2015].

27 $\mathbf{X}_j$ is a vector that contains the pre-intervention averages of the control variables and the outcome variable.

28 In a previous version of this text, that circulated under the title *Synthetic Control Estimator: A Walkthrough with Confidence Intervals*, we used a constant in time intervention effect. The results of that smaller Monte Carlo experiment were similar to the ones presented below.

29 If we treat our hypothesis test as conditional on the realized outcome variable, assumption 3 holds automatically.

- θ^3 is the absolute value of the statistic of a t-test that compares the estimated average post-intervention effect against zero. More precisely,

$$\theta^3 := \left| \frac{\overline{\alpha}_{post} / (T - T_0)}{\widehat{\sigma} / \sqrt{T - T_0}} \right|$$

where $\overline{\alpha}_{post} := \frac{\left(\sum_{t=T_0+1}^{T} \widehat{\alpha}_{\widetilde{j},t} \right)}{(T - T_0)} =: \theta^1$ and $\widehat{\sigma} := \frac{\left(\sum_{t=T_0+1}^{T} \left(\widehat{\alpha}_{\widetilde{j},t} - \overline{\alpha}_{post} \right)^2 \right)}{(T - T_0)}$. This test statistic is used by Mideksa [2013].

- $\theta^4 := \left| mean\left(Y_{\widetilde{j},t} | t \geqslant T_0 + 1 \right) - \frac{\sum_{t=T_0+1}^{T} \sum_{j \neq \widetilde{j}} Y_{j,t}}{(T - T_0) \times J} \right|$ is a simple difference in means between the treated region and the control regions for the realized outcome variable during the post-intervention period. This test statistic is suggested by Imbens and Rubin [2015].

- θ^5 is the coefficient of the interaction term in a differences-in-differences model. More precisely, we estimate the model

$$Y_{j,t} = \eta_1 \times \mathbb{1}\left[j = \widetilde{j} \right] + \eta_2 \times \mathbb{1}\left[j = \widetilde{j} \right] \times \mathbb{1}\left[t \geqslant T_0 + 1 \right] + Z_{j,t} \times \zeta + \xi_j + \mu_t + \varepsilon_{j,t}, \quad (18)$$

where ξ_j and μ_t are, respectively, region and time fixed effects, and we make $\widehat{\theta}^5 = |\widehat{\eta}_2|$.

where $\widetilde{j}$ is the region that is assumed to face the intervention in each permutation, $mean(\clubsuit|\diamond)$ is the mean of variable $\clubsuit$ conditional on event $\diamond$. We construct the empirical distribution of each test statistic for each Monte Carlo repetition and test the null hypothesis at the 10% significance level. In practice, we reject the null hypothesis if the observed test statistic is one of the two largest values of the empirical distribution of the test statistic.

Note that, although test statistic θ^4 and θ^5 do not use the synthetic control method, they are included in our Monte Carlo Experiment for being commonly used in the literature about permutation tests. Since the synthetic control estimator is a time-consuming and computer-demanding methodology, it is important to analyze whether it outperforms much simpler methods that are commonly used in the evaluation literature and that are also adequate in our framework. For this same reason, we also report rejection rates for the differences-in-differences inference procedure proposed by Conley and Taber [2011] (CT)[30].

Table 1 shows the results of our Monte Carlo Experiment. Each cell presents the rejection rate of the permutation test described in subsection 2.3 that uses the test statistic in each row or the rejection rate of the test proposed by Conley and Taber [2011] when the true intervention

30 We estimate model (18) and test the null hypothesis $H_0 : \eta_2 = 0$ using the confidence intervals recommend by Conley and Taber [2011]. Since their inference procedure uses only the control regions in order to estimate the test statistic distribution, the true nominal size of this test is 10.53%.

effect is given by the value mentioned in the column's heading. Consequently, while column (1) presents tests' sizes[31], the columns (2)-(7) present their power.

Table 1: Monte Carlo Experiment's Rejection Rates

	Intervention Effect						
Test Statistic	(1) **λ= .0**	(2) **λ= .05**	(3) **λ= .1**	(4) **λ= .25**	(5) **λ= .5**	(6) **λ= 1.0**	(7) **λ= 2.0**
$\widehat{\theta}^1$	0.10	0.19	0.23	0.35	0.45	0.59	0.69
$\widehat{\theta}^2$	0.10	0.30	0.37	0.48	0.56	0.70	0.77
$\widehat{\theta}^3$	0.10	0.62	0.71	0.79	0.88	0.93	0.95
$\widehat{\theta}^4$	0.10	0.20	0.27	0.37	0.46	0.57	0.65
$\widehat{\theta}^5$	0.10	0.19	0.23	0.37	0.45	0.60	0.70
CT	0.06	0.15	0.24	0.36	0.38	0.60	0.64

Source: Authors' own elaboration. *Notes:* Each cell presents the rejection rate of the test associated to each row when the true intervention effect is given by the value λ in the columns' headings. Consequently, while column (1) presents tests' sizes, the columns (2)-(7) present their power. $\widehat{\theta}^1$-$\widehat{\theta}^3$ are associated to permutation tests that uses the Synthetic Control Estimator. $\widehat{\theta}^4$-$\widehat{\theta}^5$ are associated to permutation tests that are frequently used in the evaluation literature. CT is associated with the asymptotic inference procedure proposed by Conley and Taber [2011].

Analyzing column (1), we note that the five permutation tests of our Monte Carlo Experiment $\left(\theta^1\text{-}\theta^5\right)$ present the correct nominal size as expected by the decision rule of Fisher's Exact Inference Procedure. The most interesting result in this column is the conservativeness of the inference procedure proposed by Conley and Taber [2011] (CT), that under-rejects the null hypothesis. This finding can be explained by the fact that, while our sample size is small $(J+1=20)$, their inference procedure is an asymptotic test based on the number of control regions going to infinity.

Analyzing the other columns, we note that the test statistic *RMSPE*, proposed by Abadie et al. [2015] $\left(\theta^2\right)$, is uniformly more powerful than the simple test statistics $\left(\theta^4, \theta^5\right)$ that are commonly used in the evaluation literature. This result suggests that, in a context where we observe only one treated unit, we should use the synthetic control estimator even if the treatment were randomly assigned. We also stress that the hypothesis test based on the statistic *RMSPE* $\left(\theta^2\right)$ outperforms the test proposed by Conley and Taber [2011] (CT) in terms of power, suggesting that, in a context with few control regions, we should use the synthetic control estimator instead of a differences-in-differences model.

We also underscore that the most powerful test statistic is the t-test, θ^3. This result makes clear the gains of power when the researcher chooses to use the synthetic control estimator instead of a simpler method, such as the difference in means $\left(\theta^4\right)$ or the permuted differences-in-differences test $\left(\theta^5\right)$. We also note that the large power of the t-test have been previously observed in contexts that are different from ours: Lehmann [1959] looks to a simple test of

31 Note that one possible measure of the coverage rate of our confidence set is one minus the rejection rates presented in column (1).

mean differences, Ibragimov and Muller [2010] analyzes a two-sample test of mean differences where samples' variances are different from each other, and Young [2015] focus on a linear regression coefficient.

Finally, we note that the simple average of the absolute post-intervention treatment effect (θ^1), despite using the synthetic control method, is as powerful as the simple test statistics that are commonly used in the evaluation literature (θ^4, θ^5). Consequently, we do not recommend to use it to conduct inference, because it is as time-consuming to estimate as the more powerful test statistics that uses the synthetic control method, $(\theta^2$ and, specially, $\theta^3)$. We avoid making any stronger suggestion about which test statistic the empirical researcher should use, because, as [Eudey et al., 2010, p. 14] makes clear, this choice is data dependent since the empirical researcher's goal is to match the test statistic to the meaning of the data. For example, if outliers are extremely important, θ^2 may be a better option than θ^3 even though the latter is more powerful than the former.

In appendix B.1, we expand the results of this section to other test statistics.

5 EXTENSIONS TO THE INFERENCE PROCEDURE

5.1 Simultaneously Testing Hypotheses about Multiple Outcomes

Imbens and Rubin [2015] states that the validity of the procedure described in subsection 2.3 depends on a prior (i.e., before seeing the data) commitment to a test statistic. Moreover, Anderson [2008] shows that simultaneously testing hypotheses about a large number of outcomes can be dangerous, leading to an increase in the number of false rejections.[32] Consequently, applying the inference procedure described in subsection 2.3 to simultaneously test hypotheses about multiple outcomes can be misleading, because there is no clear way to choose a test statistic when there are many outcome variables and because our test's true size may be smaller than its nominal value in this context. After adapting the *familywise error rate control methodology* suggested by Anderson [2008] to our framework, we propose one way to test any *sharp null hypothesis* for a large number of outcome variables, preserving the correct test size for each variable of interest.

First, we modify the framework described in section 2, assuming that there are $M \in \mathbb{N}$ observed outcome variables — $\mathbf{Y}^1, ..., \mathbf{Y}^M$ — with their associated potential outcomes. We change assumptions 1-3 to:

Assumption 4. The potential outcome vectors $\mathbf{Y}_j^{m,I} := \left[Y_{j,1}^{m,I} ... Y_{j,T}^{m,I}\right]'$ and $\mathbf{Y}_j^{m,N} := \left[Y_{j,1}^{m,N} ... Y_{j,T}^{m,N}\right]'$ for each region $j \in \{1, ..., J+1\}$ and each outcome variable $m \in \{1, ..., M\}$ do not vary based on

32 List et al. [2016] argues that false rejections can harm the economy since vast public and private resources can be misguided if agents base decisions on false discoveries. They also point that multiple hypothesis testing is a especially pernicious influence on false positives.

whether other regions face the intervention or not (i.e., no spill-over effects in space) and, for each region, there are no different forms or versions of intervention (i.e., single dose treatment), which lead to different potential outcomes.

Assumption 5. The choice of which unit will be treated (i.e., which region is our region 1) is random *conditional on the choice of the donor pool, the observable variables included as predictors for each outcome variable* $m \in \{1, ..., M\}$ *and the unobservable variables captured by the path of the outcome variables during the pre-intervention period.*

Assumption 6. The potential outcomes $\mathbf{Y}_j^{m,I} := \left[Y_{j,1}^{m,I} ... Y_{j,T}^{m,I}\right]'$ and $\mathbf{Y}_j^{m,N} := \left[Y_{j,1}^{m,N} ... Y_{j,T}^{m,N}\right]'$ for each region $j \in \{1, ..., J+1\}$ and each outcome variable $m \in \{1, ..., M\}$ are fixed but *a priori* unknown quantities.

Now, our null hypothesis is slightly more complex than the one described in equation (9):

$$H_0 : Y_{j,t}^{m,I} = Y_{j,t}^{m,N} + f_m(t) \tag{19}$$

for each region $j \in \{1, ..., J+1\}$, each time period $t \in \{1, ..., T\}$ and each outcome variable $m \in \{1, ..., M\}$, where $f_m : \{1, ..., T\} \to \mathbb{R}$ is a function of time that is specific to each outcome m. Note that we could index each function f_m by region j, but we opt not to do so because we almost never have a meaningful null hypothesis that is precise enough to specify individual intervention effects. Observe also that it is important to allow for different functions for each outcome variable because the outcome variables may have different units of measurement or different scales.

Under assumptions 4-6 and the null hypothesis (19), we can, for each $m \in \{1, ..., M\}$, calculate an observed test statistic, $\theta_{f_m}^{obs} = \theta^m(e_1, \tau, \mathbf{Y}^m, \mathbf{X}, f_m)$, and their associated observed p-value,

$$p_{\theta_{f_m}^{obs}} := \frac{\sum_{j=1}^{J+1} \mathbb{1}\left[\theta^m(e_j, \tau, \mathbf{Y}, \mathbf{X}, f_m) \geqslant \theta_{f_m}^{obs}\right]}{J+1}$$

where we choose the order of the index m to guarantee that $p_{\theta_{f_1}^{obs}} < p_{\theta_{f_2}^{obs}} < ... < p_{\theta_{f_M}^{obs}}$.

Since this p-value is itself a test statistic, we can estimate, for each outcome $m \in \{1, ..., M\}$, its empirical distribution by computing $p_{\theta_{f_m}}^{\tilde{j}} := \frac{\sum_{j=1}^{J+1} \mathbb{1}\left[\theta^m(e_j, \tau, \mathbf{Y}, \mathbf{X}, f_m) \geqslant \theta^{m,\tilde{j}}\right]}{J+1}$ for each region $\tilde{j} \in \{1, ..., J+1\}$, where $\theta^{m,\tilde{j}} := \theta^m\left(e_{\tilde{j}}, \tau, \mathbf{Y}^m, \mathbf{X}, f_m\right)$. Our next step is to calculate $p_{\theta_{f_m,*}}^{\tilde{j}} := \min\left\{p_{\theta_{f_m}}^{\tilde{j}}, p_{\theta_{f_{m+1}}}^{\tilde{j}}, ..., p_{\theta_{f_M}}^{\tilde{j}}\right\}$ for each $m \in \{1, ..., M\}$ and each $\tilde{j} \in \{1, ..., J+1\}$. Then, we estimate $p_{\theta_{f_m}^{obs}}^{fwer*} := \frac{\sum_{j=1}^{J+1} \mathbb{1}\left[p_{\theta_{f_m},*}^{j} \leqslant p_{\theta_{f_m}^{obs}}^{obs}\right]}{J+1}$ for each $m \in \{1, ..., M\}$. We enforce monotonicity one last time by computing $p_{\theta_{f_m}^{obs}}^{fwer} := \min\left\{p_{\theta_{f_m}^{obs}}^{fwer*}, p_{\theta_{f_{m+1}}^{obs}}^{fwer*}, ..., p_{\theta_{f_M}^{obs}}^{fwer*}\right\}$ for each

$m \in \{1, ..., M\}$. Finally, for each outcome variable $m \in \{1, ..., M\}$, we reject the *sharp null hypothesis* (19) if $p^{fwer}_{\theta^{obs}_{f_m}} \leqslant \gamma$, where γ is a pre-specified significance level.

It is important to observe that rejecting it for some outcome variable $m \in \{1, ..., M\}$ implies that there is some region whose intervention effect differs from $f_m(t)$ for some time period $t \in \{1, ..., T\}$ for that specific outcome variable.

We also note that, when we observe only one outcome variable of interest as in section 2, we can reinterpret it as case with multiple outcome variables where each post-intervention time period is seem as a different outcome variable. With this interpretation, the inference procedure described in subsection 2.3 is still valid and is similar in flavor with the *summary index test* proposed by Anderson [2008], because we summarized the entire time information in a single test statistic. Since Anderson [2008] argues that the *summary index test*[33] has more power than the *familywise error rate control* approach, we recommend that the empirical researcher uses the inference procedure described in subsection 2.3 if he or she is interested in knowing whether there is an intervention effect or not, but is not interested in the timing of this effect. If the empirical researcher is interested in the timing of this effect, he or she should interpret each post-intervention time period as a different outcome variable and apply the inference procedure described in this subsection. Both approaches deliver valid statistical inference in small samples.

5.2 Hypothesis Testing and Confidence Sets with Multiple Treated Units

Cavallo et al. [2013] extend the Synthetic Control Method developed by Abadie and Gardeazabal [2003] and Abadie et al. [2010] to the case when we observe multiple treated units. Firstly, we explain their innovation and, then, we clearly state the hypotheses that guarantee the validity of their method since they have not done it either. We also generalize their inference procedure in order to test any kind of *sharp null hypothesis* and, then, propose a way to estimate confidence sets for the pooled intervention effect.

Assume that there are $G \in \mathbb{N}$ similar interventions that we are interested in analyzing simultaneously. For each intervention $g \in \{1, ..., G\}$, there are $J^g + 1$ observed regions and we denote the region that faces the intervention as the first one, 1^g. Following the procedure described in subsection 2.1, we define the Synthetic Control Estimator of $\alpha_{1^g,t}$ as

$$\widehat{\alpha}_{1^g,t} := Y_{1^g,t} - \widehat{Y}^N_{1^g,t} \tag{20}$$

33 The *summary index test* can also be adapted to our framework of multiple outcomes and be applied in place of the procedure described in this subsection. In order to do that, the researcher must aggregate all the information contained in test statistics $\theta^1, ..., \theta^M$ in a single index test statistic $\widetilde{\theta}$ and use $\widetilde{\theta}$ as the test statistic for the inference procedure described in subsection 2.3. In this case, a rejection of the null hypothesis implies that there is some region whose intervention effect differs from $f_m(t)$ for some time period $t \in \{1, ..., T\}$ for some specific outcome variable $m \in \{1, ..., M\}$.

for each $t \in \{1, ..., T\}$ and each intervention $g \in \{1, ..., G\}$. The pooled intervention effect according to the Synthetic Control Estimator is given by:

$$\overline{\widehat{\alpha}}_{1,t} := \frac{\sum_{g=1}^{G} \widehat{\alpha}_{1^g,t}}{G} \tag{21}$$

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

In order to run hypothesis testing for each time period $t \in \{T_0 + 1, ..., T\}$, Cavallo et al. [2013] suggests the following procedure:

1. For each intervention $g \in \{1, ..., G\}$, permute which region is assumed to be treated and estimate, for each control region $j^g \in \{2, ..., J^g + 1\}$, $\widehat{\alpha}_{j^g,t}$ as described in subsection 2.1.
2. Estimate a placebo pooled intervention effect as $\overline{\widehat{\alpha}}_{q,t} := \frac{\sum_{g=1}^{G} \widehat{\alpha}_{\widetilde{j^g},t}}{G}$, where $q \in \mathbb{N}$ indexes placebo estimations and $\widetilde{j^g} \in \{1, ..., J^g + 1\}$ for each intervention $g \in \{1, ..., G\}$. Note that there are $Q := \prod_{g=1}^{G} (J^g + 1)$ possible placebo pooled intervention effects.
3. Compute $p_{CGNP,t} := \frac{\sum_{q=1}^{Q} \mathbb{1}\left[|\overline{\widehat{\alpha}}_{q,t}| \geqslant |\overline{\widehat{\alpha}}_{1,t}|\right]}{Q}$ for each $t \in \{T_0 + 1, ..., T\}$.
4. Reject the null hypothesis if $p_{CGNP,t}$ is less than some pre-specified significance level.

We want to formalize and generalize this inference procedure. Moreover, differently from Cavallo et al. [2013], we summarize the entire time information in a single test statistic in order to avoid over-rejecting the null hypothesis as pointed out by Anderson [2008][34]. We need to assume:

Assumption 7. The potential outcome vectors $\mathbf{Y}^I_{j^g} := \left[Y^I_{j^g,1} ... Y^I_{j^g,T}\right]'$ and $\mathbf{Y}^N_{j^g} := \left[Y^N_{j^g,1} ... Y^N_{j^g,T}\right]'$ for each intervention $g \in \{1, ..., G\}$ and each region $j^g \in \{1, ..., J^g + 1\}$ do not vary based on whether other regions face the intervention or not (i.e., no spill-over effects in space) and, for each region, there are no different forms or versions of intervention (i.e., single dose treatment), which lead to different potential outcomes.

Assumption 8. The choice of which unit will be treated in each intervention (i.e., which region is our region 1^g for each $g \in \{1, ..., G\}$) is random *conditional on the choice of the donor pool of each intervention* $g \in \{1, ..., G\}$*, the observable variables included as predictors and the unobservable variables captured by the path of the outcome variable during the pre-intervention period of each intervention* $g \in \{1, ..., G\}$.

Assumption 9. The potential outcomes $\mathbf{Y}^I_{j^g} := \left[Y^I_{j^g,1} ... Y^I_{j^g,T}\right]'$ and $\mathbf{Y}^N_{j^g} := \left[Y^N_{j^g,1} ... Y^N_{j^g,T}\right]'$ for each intervention $g \in \{1, ..., G\}$ and each region $j^g \in \{1, ..., J^g + 1\}$ are fixed but *a priori* unknown quantities.

34 For more information about over-rejecting the null hypothesis, see the articles mentioned in subsection 5.1.

Finally, our *sharp null hypothesis* is now given by:

$$H_0 : Y^I_{j^g,t} = Y^N_{j^g,t} + f(t) \tag{22}$$

for each intervention $g \in \{1, ..., G\}$, each region $j^g \in \{1, ..., J^g + 1\}$ and time period $t \in \{1, ..., T\}$, where $f : \{1, ..., T\} \to \mathbb{R}$. Note that we could index the function f by intervention g and region j^g, but we opt not to do so because we almost never have a meaningful null hypothesis that is precise enough to specify individual intervention effects for each observed region.

Now, we define a test statistic $\theta_{pld,f}$ as a known positive real-valued function $\theta_{pld}((\iota^g, \tau^g, \mathbf{Y}^g, \mathbf{X}^g)_{g=1}^G, f)$ that summarizes the entire information of the post-intervention period.

The observed test statistic is given by $\theta^{obs}_{pld,f} := \theta_{pld}((e_{1^g}, \tau^g, \mathbf{Y}^g, \mathbf{X}^g)_{g=1}^G, f)$, where e_{j^g} is the j^g-th vector of the canonical base of $\mathbb{R}^{J^g+1}$. Under assumptions 7-9 and the *sharp null hypothesis* (22), we can estimate the entire empirical distribution of $\theta_{pld,f}$ by estimating the test statistics $\theta_q := \theta_{pld}((e_{j^g}, \tau^g, \mathbf{Y}^g, \mathbf{X}^g)_{g=1}^G, f)$ associated to each possible placebo treatment assignment[35] $q \in \{1, ..., Q\}$. We, then, reject the null hypothesis (equation (22)) if

$$p_{\theta_{pld,f}} := \frac{\sum_{q=1}^{Q} \mathbb{1}\left[\theta_q \geqslant \theta^{obs}_{pld,f}\right]}{Q} \leqslant \gamma \tag{23}$$

where γ is some pre-specified significance level.[36] Note that rejecting the null hypothesis implies that there is some intervention with some region whose intervention effect differs from $f(t)$ for some time period $t \in \{1, ..., T\}$.

Now, we extend our confidence sets to the pooled intervention effect. Under assumptions 7-9, we can then construct a γ-confidence set for the pooled intervention effect as

$$CS_{\gamma,\theta_{pld}} := \left\{ f \in \mathbb{R}^{\{1,...,T\}} : p_{\theta_{pld,f}} > \gamma \right\}. \tag{24}$$

Note that it is easy to interpret $CS_{\gamma,\theta_{pld}}$: it contains all pooled intervention effect functions whose associated *sharp null hypotheses* are not rejected by the inference procedure described in this subsection.

However, although theoretically possible to define such a general confidence set, null hypothesis (22) might be too general for practical reasons since the space $\mathbb{R}^{\{1,...,T\}}$ is too large to be informative and estimating such a confidence set would be computationally infeasible. For these reasons,we believe that it is worth focusing in two subsets of $CS_{\gamma,\theta_{pld}}$.

Firstly, we propose to assume the following null hypothesis:

$$H_0 : Y^I_{j^g,t} = Y^N_{j^g,t} + c \times \mathbb{1}(t \geqslant T_0 + 1) \tag{25}$$

35 Note that each possible placebo treatment assignment q is just a possible combination of $(e_{j^1}, ..., e_{j^G})$.

36 In order to compute the p-value $p_{\theta_{pld,f}}$, we implicitly assume that all placebo treatment assignments $q \in \{1, ..., Q\}$ are equally likely. In section 6, we propose a sensitivity analysis to measure the robustness of the test's decision to this implicit assumption.

for each intervention $g \in \{1, ..., G\}$, each region $j^g \in \{1, ..., J^g + 1\}$ and time period $t \in \{1, ..., T\}$, where $c \in \mathbb{R}$, and estimate the empirical distribution of $\theta_{pld,c}$ following the procedure described in this subsection. Under assumptions 7-9, we can then construct a γ-confidence interval for the constant pooled intervention effect as

$$CI_{\gamma,\theta_{pld}} := \left\{ f \in \mathbb{R}^{\{1,...,T\}} : f(t) = c \text{ and } p_{\theta_{pld,c}} > \gamma \right\} \subseteq CS_{\gamma,\theta_{pld}} \tag{26}$$

where $c \in \mathbb{R}$ and $\gamma \in (0, 1) \subset \mathbb{R}$. It is easy to interpret $CI_{\gamma,\theta_{pld}}$: it contains all constant in time pooled intervention effects whose associated *sharp null hypotheses* are not rejected by the inference procedure described in this subsection.

Secondly, We can easily extend (25) and (26) to a linear in time pooled intervention effect. Assume

$$H_0 : Y^I_{j^g,t} = Y^N_{j^g,t} + \widetilde{c} \times (t - T_0) \times \mathbb{1}(t \geqslant T_0 + 1) \tag{27}$$

for each intervention $g \in \{1, ..., G\}$, each region $j^g \in \{1, ..., J^g + 1\}$ and time period $t \in \{1, ..., T\}$, where $\widetilde{c} \in \mathbb{R}$. Note that we can apply the inference procedure described above to any $\widetilde{c} \in \mathbb{R}$, estimating the empirical distribution of $\theta_{pld,\widetilde{c}}$. Under assumptions 7-9, we can then construct a γ-confidence set for the linear pooled intervention effect as

$$\begin{aligned} \widetilde{CS}_{\gamma,\theta_{pld}} &:= \left\{ \begin{array}{ll} f \in \mathbb{R}^{\{1,...,T\}} : & f(t) = \widetilde{c} \times (t - T_0) \times \mathbb{1}(t \geqslant T_0 + 1) \\ & \text{and } p_{\theta_{pld,\widetilde{c}}} > \gamma \end{array} \right\} \\ &\subseteq CS_{\gamma,\theta_{pld}} \end{aligned} \tag{28}$$

where $\gamma \in (0, 1) \subset \mathbb{R}$. It is also easy to interpret $\widetilde{CR}_{\gamma,\theta_{pld}}$: it contains all linear in time pooled intervention effects whose associated *sharp null hypotheses* are not rejected by the inference procedure described in this subsection.

Finally, if the researcher wants to analyze each intervention $g \in \{1, ..., G\}$ separately in order to investigate heterogeneous effects, he or she can apply our framework for multiple outcomes (see subsection 5.1) instead of implementing the pooled analysis describe in this subsection. The more detailed analysis based on the multiple outcomes framework has the cost of losing statistical power since the framework described in this subsection is based on the *summary index test* while the procedure explained in subsection 5.1 is based on the *familywise error rate*.[37]

5.3 Hypothesis Testing and Confidence Sets under Heteroskedasticity

Assumption 3 and its equivalent assumptions 6 and 9 under different contexts may be considered too strong for empirical applications since they implicitly impose homoskedasticity among different regions. In particular, Ferman and Pinto [2016] stress that a common source of het-

37 Anderson [2008] offers a detailed discussion about the differences between inference procedures based on the *summary index test* or on the *familywise error rate*.

eroskedasticity in empirical contexts is that regions with larger population sizes present smaller variances in their potential outcome values.

Fortunately, assumption 3 is stronger than needed. In order to have a valid inference procedure, we only need exchangeability between the control units and the treated unit. If we want to allow for heteroskedasticity, we can use test statistics that are robust to heterogeneity in the variance of the potential outcome values. For example, Hahn et al. [2013] and Pauly et al. [2015] conclude that permutation tests using a t-test statistic are robust to heteroskedasticity. Moreover, Ferman and Pinto [2016] propose a modified RMSPE test statistic that is also robust to heteroskedasticity when not all pre-intervention outcome values are used to construct the synthetic control region since, in this case, its denominator is a measure of each region's variance. More generally, Canay et al. [2015] argues that permutation tests are valid under heteroskedasticity if the approximate symmetry assumption holds. They also show that this assumption implies that the test statistic distribution does not depend on which unit is treated or not and that it holds for the t-statistic.[38]

Finally, we stress that this section is another example of the importance of a careful choice of test statistic as pointed out by Eudey et al. [2010]. If the data present heteroskedasticity, the empirical researcher should use a test statistic that is robust to this issue.

6 SENSITIVITY ANALYSIS

When we propose the rejection rules (10) and (23), we implicitly assume that all regions have the same probability to be chosen to face the intervention. In observational studies such as the most common applications of the Synthetic Control Method, this assumption may be considered too strong since the researcher does not know the true data generating process, i.e., the underlying experimental design. If the true probabilities associated to each treatment assignment vector differs from the assumed discrete uniform distribution, we suffer from *hidden bias* according to Rosenbaum [2002]. Consequently, our test will present a true size that is different from its nominal size. Fortunately, there are, in the literature[39], a common sensitivity analysis that allows the empirical researcher to measure the robustness of his or her conclusions (i.e., the test's decision regarding rejecting the *sharp null hypothesis*) to this implicit assumption. We present this sensitivity analysis step-by-step in the framework of subsection 5.2 (Multiple Treated Units) because it encompasses the simpler framework of section 2.3 (Single Treated Unit).

1. Estimate the test statistics $\theta_1, \theta_2, ..., \theta_Q$ for all possible placebo treatment assignment $q \in \{1, ..., Q\}$, where $\theta_1 = \theta^{obs}_{pld,f}$.

38 We stress that, in a panel data framework, the t-test is valid if the post-intervention period is long enough and if there is no serial correlation.

39 For more information, see Rosenbaum [2002] and Cattaneo et al. [2016].

2. Rename them as θ_{1^*}, θ_{2^*},...,θ_{Q^*} such that $\theta_{1^*} > \theta_{2^*} > ... > \theta_{Q^*}$.

3. Define $\overline{q} \in \Omega := \{1^*, ..., Q^*\}$ such that $\theta_{\overline{q}} = \theta^{obs}_{pld,f}$. If there are more than one $q' \in \Omega$ that presents this property, take the largest one.

4. Define the probability of each placebo treatment assignment $q' \in \Omega$ as

$$\mathbb{P}(\omega = q^*) = \frac{\exp(\phi \upsilon_{q^*})}{\sum_{q' \in \Omega} \exp(\phi \upsilon_{q'})}, \tag{29}$$

where ω is the random variable that denotes the placebo treatment assignment, $\phi \in \mathbb{R}_+$ is the sensitivity parameter and $\upsilon_{q'} \in \{0, 1\}$ for each $q' \in \Omega$. Note that, when $\phi = 0$, all placebo treatment assignments present the same probability of being chosen, i.e., $\mathbb{P}(\omega = q^*) = 1/Q$ for all $q^* \in \Omega$. Consequently, in subsection 5.2, we implicitly assume that $\phi = 0$. Moreover, the sensitivity parameter $\phi \in \mathbb{R}_+$ presents a very intuitive interpretation: a region $q_1 \in \Omega$ with $\upsilon_{q_1} = 1$ is $\Phi := \exp(\phi)$ times more likely to face the intervention than a region $q_2 \in \Omega$ with $\upsilon_{q_2} = 0$.

5. Under assumption (29), the permutation test's p-value, originally defined in (23), is now given by

$$p_{\theta_{pld,f}} := \sum_{q^* \in \Omega} \mathbb{1}[\theta_{q^*} \geqslant \theta_{\overline{q}}] \frac{\exp(\phi \upsilon_{q^*})}{\sum_{q' \in \Omega} \exp(\phi \upsilon_{q'})}. \tag{30}$$

6. If the *sharp null hypothesis* is rejected, we want to measure the robustness of this conclusion to changes in the parameter $\phi \in \mathbb{R}_+$. The worst case scenario[40] is given by

$$\begin{cases} \upsilon_{q^*} = 1 \text{ if } q^* \leqslant \overline{q} \\ \upsilon_{q^*} = 0 \text{ if } q^* > \overline{q}. \end{cases}$$

where $q^* \in \Omega$. Define $\underline{\phi} \in \mathbb{R}_+$ such that

$$p_{\theta_{pld,f}} := \sum_{q^* \in \Omega} \mathbb{1}[\theta_{q^*} \geqslant \theta_{\overline{q}}] \frac{\exp(\phi \upsilon_{q^*})}{\sum_{q' \in \Omega} \exp(\phi \upsilon_{q'})} = \gamma,$$

where γ is a pre-specified significance level. If $\underline{\phi} \in \mathbb{R}_+$ is close to zero, the permutation test's decision is not robust to small violations of the assumption $\phi = 0$.

7. If the *sharp null hypothesis* is not rejected, we want to measure the robustness of this conclusion to changes in the parameter $\phi \in \mathbb{R}_+$. The best case scenario[41] is given by

40 In this case, we pick values for $\upsilon_{q'}$ in order to make as hard as possible the rejection of the *sharp null hypothesis* given a value for $\phi \in \mathbb{R}_+$,

41 In this case, we pick values for $\upsilon_{q'}$ in order to make as easy as possible the rejection of the *sharp null hypothesis* given a value for $\phi \in \mathbb{R}_+$.

$$\begin{cases} \upsilon_{q^*} = 0 \text{ if } q^* \leqslant \overline{q} \\ \upsilon_{q^*} = 1 \text{ if } q^* > \overline{q}. \end{cases}$$

where $q^* \in \Omega$. Define $\overline{\phi} \in \mathbb{R}_+$ such that

$$p_{\theta_{pld,f}} := \sum_{q^* \in \Omega} \mathbb{1}\left[\theta_{q^*} \geqslant \theta_{\overline{q}}\right] \frac{\exp\left(\phi \upsilon_{q^*}\right)}{\sum_{q' \in \Omega} \exp\left(\phi \upsilon_{q'}\right)} = \gamma,$$

where γ is a pre-specified significance level. If $\overline{\phi} \in \mathbb{R}_+$ is close to zero, the permutation test's decision is not robust to small violations of the assumption $\phi = 0$.

8. Regardless of the permutation test's decision, we can always evaluate the impact of $\phi \in \mathbb{R}_+$ in the p-value $p_{\theta_{pld,f}}$ by plotting a graph with ϕ in the horizontal axis and $p_{\theta_{pld,f}}$ in the vertical axis. If $p_{\theta_{pld,f}}$ changes too quickly when we change ϕ, the permutation test is too sensitive to assumptions regarding the treatment assignment probabilities.

7 EVALUATING THE STATISTICAL SIGNIFICANCE OF THE ECONOMIC IMPACT OF ETA'S TERRORISM

In this section, we aim to illustrate that our inference procedure can cast new light on empirical studies that use the Synthetic Control Method. Not only we can test more flexible null hypotheses, but also we can summarize important information in a simple and effective graph. In order to achieve this goal, we use economic data for Spanish provinces made available by Abadie and Gardeazabal [2003]. When they estimated the economic impact of ETA's terrorism on the Basque Country's economy, they did not discuss the statistical significance of their results in detail because the first inference procedure for the synthetic control method would only be proposed by Abadie et al. [2010]. In order to fill this lacuna, we implement four empirical exercises:

1. We evaluate the statistical significance of the economic impact of ETA's terrorism and whether this effect can be reasonably approximated by a quadratic function using the *RMSPE* test statistic and the inference procedure described in section 2.3.

2. We analyze the robustness of the permutation test's decision regarding the significance of the economic impact of ETA's terrorism using the sensitivity analysis described in section 6.

3. We estimate the $88.\overline{9}\%$-confidence set that contains all linear in time intervention effects (with intercept equal to zero) whose associated *sharp null hypotheses* are not rejected by our inference procedure (see equation (15)) when we use the *RMSPE* test statistic.

4. We analyze the timing of the economic impact of ETA's terrorism using the procedure described in subsection 5.1.

The data set used by Abadie and Gardeazabal [2003] is available for download using the software *R*. We observe, as our outcome variable, annual real GDP per-capita in thousands of 1986 USD from 1955 to 1997 and, as covariates, biannual sector shares as a percentage of total production for agriculture, forestry and fishing, energy and water, industry, construction and engineering, marketable services and nonmarketable services from 1961 to 1969; annual shares of the working age population that was illiterate, that completed at most primary education and that completed at least secondary education from 1964 to 1969; the population density in 1969; and annual gross total investment as a proportion of GDP from 1964 to 1969. All those variables are observed at the province level and there are seventeen provinces, including the Basque Country $(J+1=17)$. For historical details and descriptive statistics about this data set, see Abadie and Gardeazabal [2003] and Abadie et al. [2011].

ETA's terrorism acts gained strength and relevance during the 70s. For this reason, our post intervention period goes from 1970 to 1997 $(T_0 = 1969)$. In order to estimate the Synthetic Control Unit, we plug, in equation (3), the averages of our covariates and the average of our outcome variable from 1960 to 1969. Moreover, we use data from 1960 to 1969 in equation (4).

When we estimate the intervention effect for the Basque Country and the placebo effect for all the other Spanish provinces, we find that the estimated intervention effect does not look abnormally large when compared to the estimated placebo effects as subfigure 1a shows.

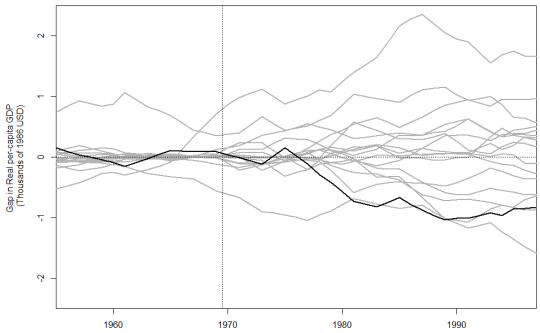
This intuitive perception is confirmed by our formal inference procedure (see subsection 2.3) when we use the *RMSPE* test statistic. More specifically, we have that $p_{RMSPE} = 0.41$, implying that we can not reject the *null hypothesis of no effect whatsoever*. When we conduct the sensitivity analysis proposed in section 6, we can verify that this conclusion is very robust to deviations of the assumption of equal probabilities of treatment for all Spanish provinces. According to figure 2, we must impose a sensitivity parameter $\overline{\phi} = 1.845$ in order to reject the *null hypothesis of no effect whatsoever* at the 10%-significance level, implying that the treatment assignment probability of the region that is most likely to receive the treatment is more than six times larger than the treatment assignment probability of the Basque Country.[42] Moreover, we note that the permutation test's p-value decreases very slowly as a function of the sensitivity parameter $\phi \in \mathbb{R}_+$.

We also test whether the estimated intervention effect can be reasonably approximated by a quadratic function. In order to do that, we fit a second order polynomial to the estimated intervention effect by applying a ordinary least square estimator only in the post-intervention period. Subfigure 1b shows this fitted quadratic function. Applying our formal inference procedure and using the *RMSPE* test statistic, we do not reject the null hypothesis that the true intervention effect follows this quadratic function because $p_{RMSPE_{quadratic}} = 0.65$.

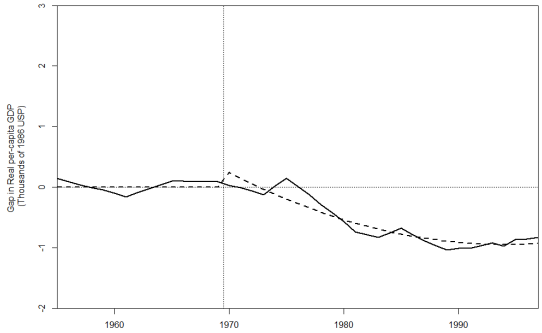
42 In order to reject the *null hypothesis of no whatsoever* at the 5%-significance level or at the 1%-significance level, we must impose $\overline{\phi} = 2.585$ or $\overline{\phi} = 4.235$, respectively.

Figure 1: Estimated Effects using the Synthetic Control Method

(a) Intervention and Placebo Effects

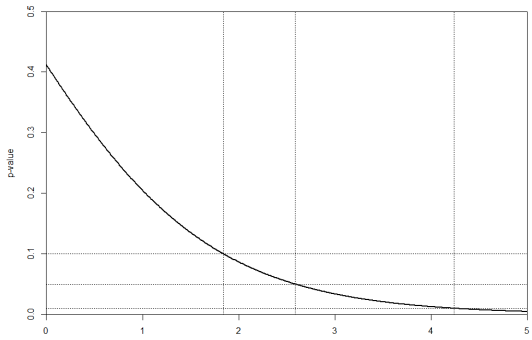


(b) Quadratic Intervention Effect Function



Note: While the gray lines show the estimated placebo effect for each Spanish province, the black lines show the estimated impact of ETA's terrorism on the Basque Country's economy and the dashed line shows the quadratic function that best approximates this effect.

Figure 2: Sensitivity Analysis



Note: The black line denotes the estimated p-value for each value of the sensitivity parameter $\phi \in \mathbb{R}_+$. The horizontal dotted lines denotes the usual p-values of .1, .05 and .01.

We also estimate a $88.\overline{9}\%$-Confidence Set[43] for a Linear in Time Intervention Effect whose intercept is equal to zero following equation (15) and using the *RMSPE* test statistic. This Confidence Set is represented in figure 3. This graph not only quickly shows that we can not reject the null hypothesis of no effect whatsoever (because the confidence set contains the linear function whose slope is equal to zero), but also shows that the economic impact of ETA's terrorism is not precisely estimated, precluding even conclusions about its true sign.[44] Due to its ability to summarize a large amount of information, our preferred confidence set (equation

43 Since we need at least 20 regions in order to estimate a 90%-Confidence Set, we use the possible confidence level that it is closest to 90%. Intuitively, we only reject the null hypothesis that generates one of the two largest values of the empirical distribution of the test statistic.

44 Note that, if our estimated confidence set intersected only a small part of the positive quadrant, we could argue that the analyzed intervention effect is likely to be negative.

(15)) is useful to the empirical researcher even being only a subset of the general confidence set (equation (11)).

Figure 3: $88.\overline{9}\%$-Confidence Set for Linear in Time Intervention Effects

Note: The black line shows the estimated impact of ETA's terrorism on the Basque Country's economy while the gray area shows the $88.\overline{9}\%$-Confidence Set for Linear in Time Intervention Effects (with intercept equal to zero) that were constructed using the *RMSPE* test statistic.

Differently from what we do in the last paragraphs, we can treat each year as a different outcome variable and apply the inference procedure described in subsection 5.1. This interpretation allow us to analyze the timing of the economic impact of ETA's terrorism, which may be significant for some time periods even though we have not reject the null hypothesis of no effect whatsoever when we pooled together all the years using the *RMSPE* test statistic. We use the squared value of the estimated intervention effect for each year of the post-intervention period as a test statistic. Using the notation of subsection 5.1, we have that

$$\theta_{f_m}^{obs} = \theta^m(e_1, \tau, \mathbf{Y}^m, \mathbf{X}, f_m) = (\widehat{\alpha}_{1,m})^2,$$

where $m \in \{1970, ..., 1997\}$ is a year of the post-intervention period.

Applying the procedure described in subsection 5.1, we find p-values between 0.42 and 0.88 for all years. Clearly, we can not reject the null hypothesis that ETA's terrorism has no economic impact whatsoever.

As a consequence of all our empirical exercises, we conclude that terrorists acts in the Basque Country had no statistically significant economic consequence even though the point estimate found by Abadie and Gardeazabal [2003] suggests a strong negative effect. We stress that we analyzed only the impact on GDP per-capita, ignoring possible other macroeconomic and microeconomic costs and, most importantly, social and human costs incurred by the Basque people.

8 CONCLUSION

In this article, we contribute to the theoretical literature about the Synthetic Control Method. First, we clearly state the assumptions that guarantee the validity of the inference procedure proposed by Abadie et al. [2010] and Abadie et al. [2015] and vastly used in the empirical literature. As our main contribution, we, then, generalize this inference procedure to test any kind of *sharp null hypothesis*, allowing us to propose a new way to estimate confidence sets for the Synthetic Control Estimator by inverting a test statistic. Basically, our confidence sets contain any function of time — particularly, the constant and linear ones — whose associated *sharp null hypothesis* is not rejected by the mentioned inference procedure. To the best of our knowledge, that is the first way to estimate confidence sets for the Synthetic Control Estimator in a context with only aggregate level data whose cross-section dimension may be larger than its time dimension. We also extend our framework to the cases when there are more than one observed outcome variable or more than one treated unit. Furthermore, we make some brief comments about applying our generalized inference procedure when heteroskedasticity is a concern. When this issue is present, the empirical researcher should use a test statistic that is robust to this problem, such as the t-test statistic or the *RMSPE*. Finally, we propose a sensitivity analysis to evaluate the robustness of the permutation test's decision to changes in the assumption that all possible treatment assignments probabilities are equal.

Moreover, since our inference procedure assumes that the probability of each region being chosen to face the intervention is known, we propose a sensitivity analysis to check the robustness of the permutation test's decision to changes in this assumption. In particular, the basic form of our inference procedure implicitly assumes that all treatment assignment probabilities are equal. In order to verify the robustness of our permutation test's decision to this specific assumption, we propose, based on the work of Rosenbaum [2002] and Cattaneo et al. [2016], a parametric form to the treatment assignment probabilities that allow the researcher to compute p-values for different assumptions regarding the treatment assignment probabilities.

The possibility to test any *sharp null hypothesis* is important to predict the future behavior of the intervention effect, to compare the costs and the benefits of a policy and to test theories that predict some specific kind of intervention effect. Constructing confidence sets is useful to summarize a large amount of information in a single graph, illustrating the statistical significance of the intervention effect and the precision of a point-estimate. Moreover, simultaneous hypothesis testing can be used to analyze the timing of an intervention effect. Consequently, those tools not only allows the empirical researcher to be more flexible about his or her null hypothesis, but also help him or her to convey a message in a more effective way.

Since our inference procedure works for any test statistic, we analyze, using a Monte Carlo experiment, the size and power of five different test statistics that are applied to hypothesis testing in the empirical literature about the Synthetic Control Method. In this simulation, we compare test statistics that use the Synthetic Control Method to simpler test statistics that are

commonly used in the evaluation literature (e.g.: difference in means and the coefficient associated with the interaction term in a differences-in-differences model) and to an asymptotic inference procedure proposed by Conley and Taber [2011]. We find that test statistics that use the Synthetic Control Method perform much better than its competitors when there is only one region that faces the intervention.

As an application of our generalized inference procedure, its associated new confidence sets, its extension to the case of simultaneous hypothesis testing and its associated sensitivity analysis, we evaluate the statistical significance of the economic impact of ETA's terrorism estimated by Abadie and Gardeazabal [2003]. This application clearly demonstrates the amount of information summarized by our proposed confidence sets, whose graphs quickly show not only the significance of the estimated intervention effect, but also the precision of this estimate. We stress that knowing the precision of a point-estimate is an important measure of the strength of qualitative conclusions. Finally, our empirical exercises clearly do not reject the null hypothesis of no effect whatsoever, a conclusion that is very robust according to our sensitivity analysis.

REFERENCES

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

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

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

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

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

Michael L. Anderson. Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool and Early Training Projects. *Journal of the American Statiscal Association*, 103(484):1481–1495, December 2008.

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

Michihito Ando and Fredrik Sävje. Hypothesis Testing with the Synthetic Control Method. Working Paper, http://www.eea-esem.com/files/papers/eea-esem/2013/2549/scm.pdf, 2013.

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

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

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

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

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

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

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

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

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

A. Colin Cameron, Jonah B. Gelbach, and Douglas L. Miller. Bootstrap-based Improvements for Inference with Clustered Errors. *The Review of Economics and Statistics*, 90(3):414–427, 2008.

Ivan A. Canay, Joseph P. Romano, and Azeem M. Shaikh. Randomization Tests under an Approximate Symmetry Assumption. Working Paper, http://goo.gl/zCvu08., December 2015.

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

Carlos V. Carvalho, Ricardo Mansini, and Marcelo C. Medeiros. ArCo: An Artificial Counterfactual Approach for Aggregate Data. Working Paper, February 2015.

Matias Cattaneo, Rocio Titiunik, and Gonzalo Vazquez-Bare. rdlocrand: Inference in Regression Discontinuity Designs under Local Randomization. *The Stata Journal*, 2016. Forthcoming. Available at: http://goo.gl/ukyoZi.

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

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

Andrew Chang and Phillip Li. Is Economic Research Replicable? Sixty Published Papers from Thirteen Journal Say "Usually Not". Finance and Economics Discussion Series 2015-083. Available at http://dx.doi.org/10.17016/FEDS.2015.083., September 2015.

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

Timothy G. Conley and Christopher R. Taber. Inference with Difference-in-Differences with a Small Number of Policy Changes. *The Review of Economics and Statistics*, 93(1):113–125, 2011.

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

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

Arindrajit Dube and Ben Zipperer. Pooled Synthetic Control Estimates for Recurring Treatments: An Application to Minimum Wage Case Studies. Available

at: http://www.irle.berkeley.edu/events/spring14/zipperer/dubezipperer_pooledsyntheticcontrol.pdf., November 2013.

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

T. Lynn Eudey, Joshua Kerr, and Bruce Trumbo. Using R to Simulate Permutation Distributions for Some Elementary Experimental Designs. *Journal of Statistics Education*, 18(1), 2010.

Bruno Ferman and Cristine Pinto. Inference in Differences-in-Differences with Few Treated Groups and Heteroskedasticity. Working Paper. Available at https://dl.dropboxusercontent.com/u/12654869/Ferman%20and%20Pinto%20-%20Inference%20in%20DID.pdf., February 2016.

Bruno Ferman, Cristine Pinto, and Vitor Possebom. Specification-Search Possibilities with Synthetic Controls. February 2016.

Ronald A. Fisher. *The Design of Experiments*. Hafner Publishing Company, United States, 8th edition edition, 1971.

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

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

Sonja Hahn, Frank Konietschke, and Luigi Salmaso. A Comparison of Efficient Permutation Tests for Unbalanced ANOVA in two by two Designs — and their Behavior under Heterocedasticity. Avaiable at: arXiv:1309.7781., October 2013.

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

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

Rustam Ibragimov and Ulrich K. Muller. t-Statistic Based Correlation and Heterogeneity Robust Inference. *Journal of Business & Economic Statistics*, 28(4):453–468, 2010.

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

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

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

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

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

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

Erich Lehmann. *Testing Statistical Hypotheses*. John Wiley & Sons, New York, 1959.

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

John List, Azeem M. Shaikh, and Yang Xu. Multiple Hypothesis Testing in Experimental Economics. NBER Working Paper 21875. Available at http://www.nber.org/papers/w21875., January 2016.

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

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

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

Markus Pauly, Edgar Brunner, and Frank Konietschke. Asymptotic Permutation Tests in General Factorial Designs. *Journal of the Royal Statistical Society. Series B*, 77(2):461–473, 2015.

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

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

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

Paul R. Rosenbaum. *Observational Studies*. Springler Science + Business Media, New York, 2nd edition edition, 2002.

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

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

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

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

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

Laurence Wong. *Three Essays in Causal Inference*. PhD thesis, Stanford University, March 2015.

Frank Yates. Tests of Significance for 2 x 2 Contingency Tables. *Journal of the Royal Statistical Society. Series A (General)*, 147(3):p. 426–463, 1984.

Alwyn Young. Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Eperimental Results. Available at: http://goo.gl/zR07Bn., October 2015.

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

A COMPUTING THE OBSERVED TEST STATISTIC AND ITS EMPIRICAL DISTRIBUTION

In this appendix, we didactically explain how to compute the observed test statistic $\theta_f^{obs} := \theta(e_1, \tau, \mathbf{Y}, \mathbf{X}, f)$ and the empirical distribution of the test statistic.

First of all, we compute the observed test statistic θ_f^{obs}:

1. Simply solve the nested minimization problem given by equations (3) and (4) using the observed matrices $\mathbf{Y_1}$, $\mathbf{Y_0}$, $\mathbf{X_1}$ and $\mathbf{X_0}$.

2. Save the estimated intervention effect vector $\widehat{\boldsymbol{\alpha}}_{\mathbf{1}} := \left[\widehat{\alpha}_{1,1}, ..., \widehat{\alpha}_{1,T_0+1}, ..., \widehat{\alpha}_{1,T}\right]'$.

3. Save a vector containing the differences between the estimated intervention effect and the hypothesized intervention effect function given by the *sharp null hypothesis*, i.e., $\widetilde{\boldsymbol{\alpha}}_{\mathbf{1}} := \left[\widehat{\alpha}_{1,1} - f(1), ..., \widehat{\alpha}_{1,T_0+1} - f(T_0+1), ..., \widehat{\alpha}_{1,T} - f(T)\right]'$.

4. Aggregate the information contained in vector $\widetilde{\boldsymbol{\alpha}}_{\mathbf{1}}$ using some positive function. This function is your test statistic. For example, if we chose the test statistic *RMSPE*, we compute
$$\theta_f^{obs} = RMSPE_f^1 = \frac{\sum_{t=T_0+1}^{T} (\widehat{\alpha}_{1,t} - f(t))^2/(T - T_0)}{\sum_{t=1}^{T_0} (\widehat{\alpha}_{1,t} - f(t))^2/T_0}$$
or, if you choose the first test statistic of our Monte Carlo Experiment (see section 4), we compute
$$\theta_f^{obs} = \frac{\sum_{t=T_0+1}^{T} |\widehat{\alpha}_{1,t} - f(t)|}{T - T_0}.$$

Now, we start our permutation test by assuming that region $\widetilde{j}$, where $\widetilde{j} \in \{2, ..., J+1\}$ is treated and estimate $\theta_f^{\widetilde{j}} := \theta(e_{\widetilde{j}}, \tau, \mathbf{Y}, \mathbf{X}, f)$:

1. Compute the counterfactual outcomes for regions 1 and $\widetilde{j}$ using the *sharp null hypothesis* (9).

2. Substitute those hypothesized counterfactual outcomes for the realized outcomes in matrices $\mathbf{Y_1}$ and $\mathbf{Y_0}$, saving the new matrices $\widetilde{\mathbf{Y}}_{\mathbf{1}}$ and $\widetilde{\mathbf{Y}}_{\mathbf{0}}$.

3. Change matrices $\mathbf{X_1}$ and $\mathbf{X_0}$ accordingly and save them as $\widetilde{\mathbf{X}}_{\mathbf{1}}$ and $\widetilde{\mathbf{X}}_{\mathbf{0}}$.

4. Solve the nested minimization problem given by equations (3) and (4) using the hypothesized matrices $\widetilde{\mathbf{Y}}_{\mathbf{1}}$, $\widetilde{\mathbf{Y}}_{\mathbf{0}}$, $\widetilde{\mathbf{X}}_{\mathbf{1}}$ and $\widetilde{\mathbf{X}}_{\mathbf{0}}$.

5. Save the estimated intervention effect vector $\widehat{\boldsymbol{\alpha}}_{\widetilde{\mathbf{j}}} := \left[\widehat{\alpha}_{\widetilde{j},1}, ..., \widehat{\alpha}_{\widetilde{j},T_0+1}, ..., \widehat{\alpha}_{\widetilde{j},T}\right]'$.

6. Save a vector containing the differences between the estimated intervention effect and the hypothesized intervention effect function given by the *sharp null hypothesis*, i.e., $\widetilde{\alpha}_{\widetilde{j}} := \left[\widehat{\alpha}_{\widetilde{j},1} - f(1), ..., \widehat{\alpha}_{\widetilde{j},T_0+1} - f(T_0+1), ..., \widehat{\alpha}_{\widetilde{j},T} - f(T)\right]'$.

7. Aggregate the information contained in vector $\widetilde{\alpha}_{\widetilde{j}}$ using some positive function. This function is your test statistic. For example, if we chose the test statistic *RMSPE*, we compute

$$\theta_f^{\widetilde{j}} = RMSPE_f^{\widetilde{j}} = \frac{\sum_{t=T_0+1}^{T}\left(\widehat{\alpha}_{\widetilde{j},t} - f(t)\right)^2/(T-T_0)}{\sum_{t=1}^{T_0}\left(\widehat{\alpha}_{\widetilde{j},t} - f(t)\right)^2/T_0}$$

or, if you choose the first test statistic of our Monte Carlo Experiment (see section 4), we compute

$$\theta_f^{\widetilde{j}} = \frac{\sum_{t=T_0+1}^{T}\left|\widehat{\alpha}_{\widetilde{j},t} - f(t)\right|}{T-T_0}.$$

Repeating this process for each $\widetilde{j} \in \{2, ..., J+1\}$, we compute the entire empirical distribution of the test statistic θ_f, given by $\left(\theta_f^{obs}, \theta_f^2, ..., \theta_f^{J+1}\right)$, and can estimate Fisher's Exact p-Value as in equation (10).

B MONTE CARLO EXPERIMENT'S COMPLETE SET OF RESULTS

B.1 More Test Statistics

In this subsection, we report our Monte Carlo Experiment's rejection rates for thirteen test statistics in addition to the ones shown in section 4. The new test statistics are:

- $\theta^6 := \left|mean\left(\widehat{\alpha}_{\widetilde{j},t} | t \geqslant T_0+1\right)\right|$ is suggested by Mideksa [2013] and used by Ando [2015].
- $\theta^7 := mean\left(\widehat{\alpha}^2_{\widetilde{j},t} | t \geqslant T_0+1\right)$.
- $\theta^8 := \left|median\left(\widehat{\alpha}_{\widetilde{j},t} | t \geqslant T_0+1\right)\right|$ is suggested by Sanso-Navarro [2011].
- $\theta^9 := median\left(\left|\widehat{\alpha}_{\widetilde{j},t}\right| | t \geqslant T_0+1\right)$.
- $\theta^{10} := median\left(\widehat{\alpha}^2_{\widetilde{j},t} | t \geqslant T_0+1\right)$.
- $\theta^{11} := min\left(\left|\widehat{\alpha}_{\widetilde{j},t}\right| | t \geqslant T_0+1\right)$.

- θ^{12} is the absolute value of the first (AS 1) test statistic proposed by Ando and Sävje [2013]. Intuitively, it is a rescaled post-intervention time average of the estimated intervention effects.[45]

- θ^{13} is the statistic of the Kolmogorov-Smirnov Test that compares the vector of estimated post-intervention effects against a vector of zeros. This test statistic is suggested by Imbens and Rubin [2015].

- θ^{14} is the Rank statistic for region $\widetilde{j}$ using the post-intervention time average of the observed outcome. This test statistic is suggested by Imbens and Rubin [2015].

- θ^{15} is the Rank statistic for region $\widetilde{j}$ using the post-intervention time median of the observed outcome.

- θ^{16} is the Rank statistic for region $\widetilde{j}$ using the post-intervention time minimum of the observed outcome.

- θ^{17} is the Rank statistic for region $\widetilde{j}$ using the post-intervention time maximum of the observed outcome.

where $\widetilde{j}$ is the region that is assumed to face the intervention in each permutation, $\mathtt{mean}(\clubsuit|\diamond)$ and $\mathtt{median}(\clubsuit|\diamond)$ are, respectively, the mean and the median of variable $\clubsuit$ conditional on event $\diamond$.

Note that test statistics θ^{14}, θ^{15}, θ^{16} and θ^{17} do not use the synthetic control method and are commonly used in the literature about permutation tests. We also report rejection rates for the differences-in-differences estimator (DID)[46].

Table 2 shows the results of our Monte Carlo Experiment for all our analyzed tests. Each cell presents the rejection rate of the permutation test described in subsection 2.3 that uses the test statistic in each row, or the rejection rates of the tests recommend by Bertrand et al. [2004] or proposed by Conley and Taber [2011] when the true intervention effect is given by the value mentioned in the column's heading. Consequently, while column (1) presents tests' sizes, columns (2)-(7) present their power.

Looking at the size of the tests associated with the new test statistics, there are only two interesting findings that, although not surprising according to the previous literature, are worth noting. On one hand, the Kolmogorov-Smirnov Test Statistic $\left(\theta^{13}\right)$ is associated with a conservative test due to ties in its empirical distribution. On the other hand, the rejection rate associated to a differences-in-differences model that uses standard errors clusterized at the region level is

45 For details on how to calculate θ^{12}, we recommend reading Ando and Sävje [2013]. These authors also propose a second test statistic (AS 2). Since it is very computationally demanding and presents similar size and power to AS 1 in Ando and Sävje [2013]'s Monte Carlo Experiment, we decided to test only AS 1 in our simulation.

46 We estimate the model $Y_{j,t} = \eta_1 \times \mathbb{1}\left[j = \widetilde{j}\right] + \eta_2 \times \mathbb{1}\left[j = \widetilde{j}\right] \times \mathbb{1}\left[t \geqslant T_0 + 1\right] + Z_{j,t} \times \zeta + \xi_j + \mu_t + \varepsilon_{j,t}$, where ξ_j and μ_t are, respectively, region and time fixed effects, and test the null hypothesis $H_0 : \eta_2 = 0$ using standard errors clusterized at the region level as recommend by Bertrand et al. [2004].

Table 2: Monte Carlo Experiment's Rejection Rates

	Intervention Effect						
Test Statistic	(1) $\lambda = .0$	(2) $\lambda = .05$	(3) $\lambda = .1$	(4) $\lambda = .25$	(5) $\lambda = .5$	(6) $\lambda = 1.0$	(7) $\lambda = 2.0$
Test Statistics in the Main Text							
$\widehat{\theta}^1$	0.10	0.19	0.23	0.35	0.45	0.59	0.69
$\widehat{\theta}^2$	0.10	0.30	0.37	0.48	0.56	0.70	0.77
$\widehat{\theta}^3$	0.10	0.62	0.71	0.79	0.88	0.93	0.95
$\widehat{\theta}^4$	0.10	0.20	0.27	0.37	0.46	0.57	0.65
$\widehat{\theta}^5$	0.10	0.19	0.23	0.37	0.45	0.60	0.70
CT	0.06	0.15	0.24	0.36	0.38	0.60	0.64
Additional Test Statistics							
$\widehat{\theta}^6$	0.10	0.31	0.40	0.52	0.63	0.75	0.80
$\widehat{\theta}^7$	0.10	0.17	0.21	0.32	0.41	0.55	0.64
$\widehat{\theta}^8$	0.10	0.24	0.30	0.45	0.55	0.68	0.80
$\widehat{\theta}^9$	0.10	0.24	0.30	0.45	0.55	0.68	0.80
$\widehat{\theta}^{10}$	0.10	0.24	0.29	0.45	0.55	0.68	0.79
$\widehat{\theta}^{11}$	0.10	0.29	0.36	0.50	0.62	0.75	0.86
$\widehat{\theta}^{12}$	0.10	0.32	0.39	0.46	0.51	0.59	0.61
$\widehat{\theta}^{13}$	0.05	0.51	0.61	0.75	0.85	0.91	0.95
$\widehat{\theta}^{14}$	0.10	0.20	0.27	0.37	0.46	0.57	0.65
$\widehat{\theta}^{15}$	0.10	0.26	0.31	0.48	0.53	0.66	0.77
$\widehat{\theta}^{16}$	0.10	0.14	0.15	0.23	0.31	0.40	0.54
$\widehat{\theta}^{17}$	0.10	0.13	0.13	0.19	0.25	0.33	0.44
DID	0.61	0.65	0.67	0.73	0.77	0.82	0.86

Source: Authors' own elaboration. *Notes:* Each cell presents the rejection rate of the test associated to each row when the true intervention effect is given by the value λ in the columns' headings. Consequently, while column (1) presents tests' sizes, the columns (2)-(7) present their power. $\widehat{\theta}^1$-$\widehat{\theta}^3$ and $\widehat{\theta}^6$-$\widehat{\theta}^{13}$ are associated to permutation tests that uses the Synthetic Control Estimator. $\widehat{\theta}^4$-$\widehat{\theta}^5$ and $\widehat{\theta}^{14}$-$\widehat{\theta}^{17}$ are associated to permutation tests that are frequently used in the evaluation literature. CT is associated with the asymptotic inference procedure proposed by Conley and Taber [2011]. DID is associated with the differences-in-differences model that uses standard errors clusterized at the region level.

much higher than the nominal size of 10%. This last result is explained by the small number of cluster $(J + 1 = 20)$, as already pointed out by Bertrand et al. [2004], Cameron et al. [2008] and Conley and Taber [2011].

Analyzing the power of the tests associated with the new test statistics, we observe that all the test statistics that are frequently used in the evaluation literature $\left(\theta^4\text{-}\theta^5 \text{ and } \theta^{14}\text{-}\theta^{17}\right)$ are dominated by the test statistics in the main text that uses the synthetic control method $\left(\theta^2\text{-}\theta^3\right)$. We also stress that the permuted t-test that uses the synthetic control method and the numerator of the t-test $\left(\theta^3 \text{ and } \theta^6, \text{ respectively}\right)$ are the most powerful tests. Those results suggests, again, that, in a context where we observe only one treated unit, we should use the synthetic control estimator even if the treatment were randomly assigned.

Comparing the test statistics *AS 1* proposed by Ando and Sävje [2013] $\left(\theta^{12}\right)$ with the *RMSPE* recommended by Abadie et al. [2015], we find ambiguous results. While *AS 1* is more powerful to detect small intervention effects ($\lambda \in \{0.05, 0.1\}$), *RMSPE* is more powerful to detect intermediate and large intervention effects ($\lambda \in \{0.25, 0.5, 1.0, 2.0\}$). Consequently, choosing between the two test statistics may be influenced by the magnitude of the expected intervention effect, providing another example of the importance of a careful choice of test statistic as pointed out by Eudey et al. [2010].

Finally, we note that the simple average of the squared post-intervention treatment effect $\left(\theta^{7}\right)$, despite using the synthetic control method, is as powerful as the simple test statistics that are commonly used in the evaluation literature $\left(\theta^{4}, \theta^{5}\right)$. We also observe that test statistics θ^{8}-θ^{11}, that uses the synthetic control method, are as powerful as $\left(\theta^{15}\right)$, the most powerful test statistic that does not use the synthetic control estimator. Consequently, we do not recommend to use them θ^{7}-θ^{11} to conduct inference.

B.2 Synthetic Control Method with Bad Controls

In the Monte Carlo Experiment presented in section 4 and subsection B.1, all observed regions follow the same data-generating process, implying that assumption 2 is valid. Since this assumption is non-testable, the applied researcher will never be sure about its validity. For this reason, it is important to analyze the behavior of our generalized inference procedure when there are *bad controls* in the donor pool, invalidating assumption 2.

In our Monte Carlo Experiment, a bad control region $j' \in \{2, ..., J+1\}$ presents the following data generating process:

$$\begin{aligned} \widetilde{Y}^{N}_{j',t+1} &= \delta_t \widetilde{Y}^{N}_{j',t} + \boldsymbol{\beta}_{t+1}\mathbf{Z}_{j',t+1} + u_{j',t+1} \\ \mathbf{Z}_{j',t+1} &= \kappa_t \widetilde{Y}^{N}_{j',t} + \boldsymbol{\pi}_t \mathbf{Z}_{j',t} + \mathbf{v}_{j',t+1} \\ Y^{N}_{j',t+1} &= \widetilde{Y}^{N}_{j',t+1} + 0.25 \times sd(\widetilde{Y}^{N}_{j',t}) \times t \end{aligned} \tag{31}$$

for each $t \in \{0, ..., T-1\}$. Note that regions that follows the data generating process given by (31) differ from regions that follows the data generating process given by (16) because the former regions present a increasing time trend.

We also vary the number of bad control regions (nbc), analyzing the size and power ($\lambda \in \{0, 0.05, 0.5\}$) of the tests discussed in table 2 when there are one bad control region (5% of the total sample) or five bad control regions (25% of the total sample) in the donor pool. Since bad control regions are never treated in our Monte Carlo Experiment, assumption (2) does not hold. In all the other aspects, the Monte Carlo of this subsection is identical to the one described in subsection B.1.

Table 3 shows the results of our Monte Carlo Experiment with 3,000 repetitions for all our analyzed tests when there are bad control regions. Each cell presents the rejection rate of the test associated to each row when the true intervention effect is given by the value λ in the columns'

headings and there are nbc bad control regions in the donor pool. Consequently, while columns (1) and (2) presents tests' sizes, the columns (3)-(6) present their power.

Table 3: Monte Carlo Experiment's Rejection Rates

	Intervention Effect					
	$\lambda = .0$		$\lambda = .05$		$\lambda = .5$	
Test Statistic	(1) nbc = 1	(2) nbc = 5	(3) nbc = 1	(4) nbc = 5	(5) nbc = 1	(6) nbc = 5
Test Statistics in the Main Text						
$\widehat{\theta}^1$	0.06	0.04	0.10	0.06	0.29	0.28
$\widehat{\theta}^2$	0.10	0.09	0.26	0.22	0.55	0.50
$\widehat{\theta}^3$	0.06	0.07	0.26	0.11	0.67	0.48
$\widehat{\theta}^4$	0.06	0.00	0.16	0.00	0.38	0.05
$\widehat{\theta}^5$	0.06	0.00	0.14	0.00	0.41	0.10
CT	0.05	0.08	0.07	0.08	0.15	0.05
Additional Test Statistics						
$\widehat{\theta}^6$	0.06	0.04	0.17	0.08	0.47	0.44
$\widehat{\theta}^7$	0.06	0.04	0.09	0.06	0.26	0.26
$\widehat{\theta}^8$	0.06	0.04	0.13	0.06	0.36	0.34
$\widehat{\theta}^9$	0.06	0.04	0.13	0.06	0.36	0.34
$\widehat{\theta}^{10}$	0.06	0.04	0.13	0.06	0.36	0.33
$\widehat{\theta}^{11}$	0.05	0.04	0.13	0.06	0.43	0.34
$\widehat{\theta}^{12}$	0.10	0.11	0.32	0.29	0.50	0.46
$\widehat{\theta}^{13}$	0.04	0.03	0.21	0.09	0.60	0.45
$\widehat{\theta}^{14}$	0.06	0.00	0.16	0.00	0.38	0.05
$\widehat{\theta}^{15}$	0.06	0.00	0.21	0.00	0.47	0.04
$\widehat{\theta}^{16}$	0.05	0.00	0.09	0.00	0.21	0.00
$\widehat{\theta}^{17}$	0.08	0.02	0.10	0.02	0.20	0.10
DID	0.48	0.72	0.44	0.66	0.61	0.57

Source: Authors' own elaboration. *Notes:* Each cell presents the rejection rate of the test associated to each row when the true intervention effect is given by the value λ in the columns' headings and there are nbc bad control regions in the donor pool. Consequently, while columns (1) and (2) presents tests' sizes, the columns (3)-(6) present their power. $\widehat{\theta}^1$-$\widehat{\theta}^3$ and $\widehat{\theta}^6$-$\widehat{\theta}^{13}$ are associated to permutation tests that uses the Synthetic Control Estimator. $\widehat{\theta}^4$-$\widehat{\theta}^5$ and $\widehat{\theta}^{14}$-$\widehat{\theta}^{17}$ are associated to permutation tests that are frequently used in the evaluation literature. CT is associated with the asymptotic inference procedure proposed by Conley and Taber [2011]. DID is associated with the differences-in-differences model that uses standard errors clusterized at the region level.

Analyzing columns (1)-(2), we note that all tests, with the exception of the one that uses the *RMSPE* as a test statistic ($\widehat{\theta}^2$) and the one proposed by Ando and Sävje [2013] ($\widehat{\theta}^{12}$), are conservative, presenting a true size that is lower than the nominal size.[47] In particular, the tests associated to statistics that are commonly used in the evaluation literature ($\widehat{\theta}^4$-$\widehat{\theta}^5$ and $\widehat{\theta}^{14}$-$\widehat{\theta}^{17}$)

47 Another exception is the behavior of the DID model with standard errors clusterized at the region level. This test, as in subsection B.1, over-rejects the null hypothesis.

are severely under-powered when $nbc = 5$. In this sense, although still under-powered, the tests that are based in statistics that use the Synthetic Control Method ($\widehat{\theta}^1$-$\widehat{\theta}^3$ and $\widehat{\theta}^6$-$\widehat{\theta}^{13}$) present a better behavior than its most common competitors. However, the inference procedure proposed by Conley and Taber [2011] (CT) presents a similar or better behavior than most of tests that uses the Synthetic Control Estimator with the exception of the one that uses the *RMSPE* as a test statistic ($\widehat{\theta}^2$) and the one proposed by Ando and Sävje [2013] ($\widehat{\theta}^{12}$). These two tests, by scaling the test statistics using the estimated pre-intervention gaps, effectively corrects for the large post-interventions gaps found for the bad control regions.

Another advantage of the test that uses the *RMSPE* as a test statistic ($\widehat{\theta}^2$) and the one proposed by Ando and Sävje [2013] ($\widehat{\theta}^{12}$) is that they are the most powerful options. (See columns (3)-(6)). In this sense, those two tests clearly outperforms their main competitor, the inference procedure proposed by Conley and Taber [2011]. Consequently, we find, as in subsection B.1, that test statistics $\widehat{\theta}^2$ and $\widehat{\theta}^{12}$ are equally good options.

C SYNTHETIC CONTROL METHOD: A WALKTHROUGH

In our review of the empirical literature that applied the Synthetic Control Method, we have found, on one hand, some innovative articles that proposed new, intuitive and relevant robustness checks and, on the other hand, confusing articles that were impossible to replicate due to missing information about their data. In this short walkthrough guide, we aim to not only summarize the best practices found in the literature about the Synthetic Control Method, but also point out all the information that an author must provide in order to make his or her research replicable[48].

The first thing that must be extremely clear in a article that applies the Synthetic Control Method is the choice of the donor pool. The author must state not only how many units are in the donor pool $(J + 1)$, but how they have been chosen. In view of assumption 2, understanding the method of choice behind the donor pool is fundamental in order to evaluate whether the treatment assignment is indeed random conditional on *the choice of the donor pool*, the observable variables included as predictors and the unobservable variables captured by the path of the outcome variable during the pre-intervention period. This recommendation is particularly important for articles that analyze more than one intervention with different donor pools since this information must be available for all the analyzed cases.

Related to the choice of the donor pool, Abadie et al. [2015] propose two robustness checks. The first one, the leave-one-out test, consists in dropping, from the donor pool, one of the comparison regions that received a positive weight in the synthetic control unit and reestimating the treatment effect. It aims to verify the influence of a particular unit in the estimated result.

48 Chang and Li [2015] offers a deeper discuss about replicability in Economics.

The second robustness check consists in trying different donor pool sizes in order to verify the sensibility of the estimated results to the choice of the donor pool. This last test is also applied by other authors, e.g.: Ando [2015], Barone and Mocetti [2014], Kreif et al. [2015] and Mideksa [2013].

The size and duration of the pre-intervention period (T_0) and of the entire studied period (T) must also be clearly stated. Since the main identification result about the Synthetic Control Method [Abadie et al., 2010] relies on the pre-intervention period going to infinity $(T_0 \to \infty)$, the value of T_0 is important to know in order to evaluate whether the Synthetic Control Estimates are close to the true counterfactual. Moreover, the author must inform the real date associated with T_0 and T in order to access causality. In a similar way to the differences-in-differences estimator, the Synthetic Control Method relies on a time difference in order to identify the intervention effect. If there are unobservable variables that affect the outcome of interest and change at T_0, it is impossible to know whether the estimated effect is due to the analyzed treatment or due to those unobservable variables. An example will clarify this problem. Imagine that a researcher wants to investigate the effect of winning FIFA World Cup on a country's GDP. Brazil won the FIFA World Cup in 1994 and went through a fast and effective stabilization plan (*Plano Real*). If we apply the synthetic control method using other Latin American Countries as comparison units, GDP per capita as the outcome of interest and 1994 as the beginning of the treatment period $(T_0 + 1 = 1994)$, we have no way to disentangle which part of the estimated effect is due to the World Cup or to the stabilization plan.

One indirect way to test for the presence of unobservable variables that harm the interpretation of the estimated effect is to run the in-time placebo test recommend by Abadie et al. [2015]. This procedure consists in assigning the beginning of the post-treatment period to a earlier period $t^* < T_0$ and look for a treatment effect before T_0. If there is one, there is evidence that the estimated effect is not caused by the investigated treatment.

The components of matrix $\mathbf{X}$ must also be very clear. Not only the author must state which variables are included in this matrix (e.g.: investment rate, population size, pre-intervention outcome values), but also which time periods or linear combinations of time periods are included if the predictor variables are observed in more than one time period (e.g.: investment in T_0, $T_0 - 1$, $T_0 - 2$, $T_0 - 3$ and $T_0 - 4$; population size only in the last census before T_0 and the pre-intervention time average of the outcome variable). In particular, the choice of which linear combinations of pre-intervention outcome values to include in matrix $\mathbf{X}$ is important because it may affect the significance of the estimated intervention effect as pointed out by Ferman et al. [2016]. Consequently, an author must report results for different specifications of matrix $\mathbf{X}$ in order to address the robustness of his or her estimates.

The researcher must also be precise about which procedure he or she have used to estimate matrix $\widehat{\mathbf{V}}$. Since there are four different procedures (ad-hoc choice, nested minimization, cross-validated nested minimization, regression based method), the author could report results for some of them in order to access the robustness of his or her estimates. Since the cross-validated

nested minimization requires a larger pre-intervention period, it may not be reasonable to report results for all the four methods. However, it is always possible to report results for the nested minimization, the regression based method and for the ad-hoc choice that imposes $\widehat{\mathbf{V}} = \mathbf{I}$.

Finally, Abadie et al. [2015] proposes one last robustness check: the restricted synthetic control unit. In order to avoid over-fitting, this test force the synthetic control method to allocate positive weights only to a fixed number of control units, mimicking the $\mathfrak{n}$ nearest neighbors matching estimator.