Sergio Firpo* and Vitor Possebom

Synthetic Control Method: Inference, Sensitivity Analysis and Confidence Sets

https://doi.org/10.1515/jci-2016-0026
Received November 15, 2016; revised August 6, 2018; accepted August 11, 2018

Abstract: We extend the inference procedure for the synthetic control method in two ways. First, we propose parametric weights for the p-value that includes the equal weights benchmark of Abadie et al. [1]. By changing the value of this parameter, we can analyze the sensitivity of the test's result to deviations from the equal weights benchmark. Second, we modify the *RMSPE* statistic to test any *sharp null hypothesis*, including, as a specific case, the *null hypothesis of no effect whatsoever* analyzed by Abadie et al. [1]. Based on this last extension, we invert the test statistic to estimate confidence sets that quickly show the point-estimates' precision, and the test's significance and robustness. We also extend these two tools to other test statistics and to problems with multiple outcome variables or multiple treated units. Furthermore, in a Monte Carlo experiment, we find that the *RMSPE* statistic has good properties with respect to size, power and robustness. Finally, we illustrate the usefulness of our proposed tools by reanalyzing the economic impact of ETA's terrorism in the Basque Country, studied first by Abadie and Gardeazabal [2] and Abadie et al. [3].

Keywords: Synthetic Control Estimator, Hypothesis Testing, Sensitivity Analysis, Confidence Sets

JEL Codes: C21, C23, C33

1 Introduction

The Synthetic Control Method (SCM) was proposed by Abadie and Gardeazabal [2], Abadie et al. [4] and [1] 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. Since the empirical literature applying SCM is vast,[1] developing and expanding this tool's theoretical foundation is an important task. The relevance of this goal is stressed by Athey and Imbens [44], who describe the SCM as arguably the most important innovation in the policy evaluation literature in the last fifteen years.

The inference procedure for small samples using the synthetic control estimator was developed by Abadie et al. [4] and [1]. Using, as a benchmark, placebo tests similar to Fisher's Exact Hypothesis Testing Procedure described by Fisher [45], Imbens and Rubin [46] and Rosenbaum [47], they compare an observed test statistic to its empirical distribution in order to verify whether there is enough evidence to reject the *null hypothesis of no effect whatsoever*. We extend this inference procedure in two ways.

First, we stress that the benchmark for hypothesis testing proposed by Abadie et al. [1] computes the p-value of the test by weighting all units equally. According to Abadie et al. [1, p. 499],

1 This tool was applied to an extremely diverse set of topics, including, for instance, issues related to terrorism, civil wars and political risk [2, 5–8], natural resources and disasters [9–15], international finance [16, 17], education and research policy [18–20], health policy [21, 22], economic and trade liberalization [23–25], political reforms [26–29], labor [30, 31], taxation [32, 33], crime [34–36], social connections [37], and local development [38–43].

***Corresponding author: Sergio Firpo,** Insper, Rua Quata 300, Sao Paulo, SP, 04645-042, Brazil, e-mail: firpo@insper.edu.br
Vitor Possebom, Yale University, New Haven, Connecticut, United States, e-mail: vitoraugusto.possebom@yale.edu

> this alternative model of inference is based on the premise that our confidence that a particular synthetic control estimate reflects the impact of the intervention under scrutiny would be severely undermined if we obtained estimated effects of similar or even greater magnitudes in cases where the intervention did not take place.

We propose a way to boost our confidence that a "particular synthetic control estimate reflects the impact of the intervention under scrutiny" by applying a sensitivity analysis mechanism similar to the one proposed by Rosenbaum [47] and Cattaneo et al. [48]. Intuitively, we impose parametric weights that allows the empirical researcher to compute p-values for weights that continuously differ from the equal weights benchmark of Abadie et al. [1]. As we change the value of the parameter that affects the decision based on the hypothesis test, we can gauge the sensitivity of the decision made to the equal weights benchmark. We also highlight that the results of this sensitivity analysis mechanism can easily be displayed in a graph that quickly shows the robustness of the decision based on the test.

Second, Abadie et al. [4] and [1] only test the *null hypothesis of no effect whatsoever*, which is the most common null hypothesis of interest in the empirical literature, albeit restrictive. We extend their inference procedure to test any kind of *sharp null hypothesis*. This possibility 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 hypotheses 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 from 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 this extension of the current existing inference procedure, we propose a method to compute confidence sets by inverting a test statistic. We modify the method described by Imbens and Rubin [46] and Rosenbaum [47] to calculate confidence intervals based on Fisher's Exact Hypothesis Testing Procedure and apply it to the benchmark inference procedure of the SCM proposed by Abadie et al. [1]. To the best of our knowledge, this is the first work to propose confidence sets for the intervention effect function using SCM when its typical setup is prevailing. That is, when we observe aggregate level data for only one treated unit and few control units (i. e., small finite samples) in a context whose cross-section dimension may be larger than its time dimension. Using the benchmark inference procedure, our confidence sets allow the researcher to quickly and visually show, not only the significance of the estimated intervention effect in a given point in time, but also the precision of the point estimates. This plot summarizes a large amount of information that is important to measure the strength of qualitative conclusions achieved after an econometric analysis. Furthermore, these confidence set plots can easily be combined with the aforementioned sensitivity analysis mechanism to quickly display the robustness of the empirical findings.

We then extend the inference method developed by Abadie et al. [4] and [1] to use many different test statistics beyond the already traditional Ratio of the Mean Squared Prediction Errors (*RMSPE*) test statistic, discussed in those papers. We run a Monte Carlo experiment and present results on size, power and robustness of five test statistics applied to the SCM. We choose these test statistics based on our review of the empirical literature that applies the method. More specifically, we compare test statistics that use the SCM to test statistics typically used in other 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 [49]. We find that the inference procedure based on the original test statistic proposed by Abadie et al. [1], *RMSPE*, performs much better than alternative test statistics in terms of size, power, and robustness.

We also show how to apply our new tools to contexts that differ from the ones analyzed by Abadie and Gardeazabal [2], Abadie et al. [4] and [1] in important dimensions. A researcher that wants to test null hypotheses about a pooled effect among few treated units, as studied by Cavallo et al. [10], can apply our sensitivity analysis mechanism, test any *sharp null hypothesis* and compute our confidence sets too. Moreover, a researcher who wants to simultaneously test null hypotheses for different outcome variables can apply our sensitivity analysis mechanism and test any *sharp null hypothesis*. This last extension, that expands the multiple hypotheses framework described by Anderson [50] to the SCM, is important, for example, to evaluate political reforms [26, 23, 27, 16, 17] that generally affect multiple outcomes variables, such as income levels

and investment rates. 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.

At the end, we apply the inference procedure proposed by Abadie et al. [4] and [1], its associated sensitivity analysis, its extension to other *sharp null hypotheses*, its associated confidence sets and its extension to the case of simultaneous hypothesis testing to reevaluate the economic impact of ETA's terrorism on the Basque Country estimated by Abadie and Gardeazabal [2] and Abadie et al. [3]. With this empirical exercise, we illustrate how our sensitivity analysis mechanism and our proposed confidence sets summarize a large amount of information in simple graphs. Furthermore, we show how testing more flexible null hypotheses and analyzing multiple outcomes can enrich a empirical analysis by finding suggestive evidence of a negative economic impact in the middle run that attenuates in the long run.

Literature review

Regarding the inference of the Synthetic Control Method, other authors have surely made important previous contributions. Abadie et al. [4][2] are the first authors to propose a inference procedure that consists in estimating p-values through permutation tests and Abadie et al. [1] suggest a different test statistic for the same procedure. Ando and Sävje [52] 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. [4] and [1].

Bauhoff [21], Calderon [31] and Severnini [43] propose a way to apply SCM to many treated and control units that is similar to a matching estimator for panel data. Following a similar approach, Wong [53] 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 [53] 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, under some assumptions (i. e., exact matching with population weights) beyond those typically considered in the synthetic control literature, the asymptotic distribution of the synthetic control estimator when the number of individuals in each region goes to infinity. Finally, Acemoglu et al. [37], Cavallo et al. [10] and Dube and Zipperer [54] develop different ways to apply SCM when there are more than one treated unit and propose tests that are similar to the ones proposed by Abadie et al. [4] and [1].

Gobillon and Magnac [39], also working in a context with more than one treated unit, propose a way to compute bootstrap confidence intervals for the policy effect function. Their procedure requires a large number of treated and control regions in order to be valid and focuses 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 allows the construction of confidence sets for the post-intervention effect as a function of time. Consequently, while their inference procedure allows testing a constant (over time) policy effect only, our extension of the inference procedure developed by Abadie et al. [4] and [1] allows the empirical researcher to test any function of time as the intervention effect.

Moreover, Carvalho et al. [55] propose the Artificial Counterfactual Estimator (ArCo), that is similar in purpose to SCM, and derive its asymptotic distribution when the time dimension is large (long panel data sets). However, many of the problems to which the SCM is applied present a cross-section dimension larger than their time dimension, making it impossible to apply Carvalho et al. [55]'s method.[3]

2 They also discuss the asymptotic unbiasedness of their method. Kaul et al. [51] 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 Wong [53] and Hahn and Shi [56] also conduct an asymptotic analysis when the pre-intervention period goes to infinity. Ferman and Pinto [57, 58] and Ferman et al. [59] discuss asymptotic biases, size distortions and specification-search possibilities within the SCM framework.

Our approach is similar to the way Conley and Taber [49] 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.

Finally, the sensitivity analysis literature in a context of observational studies is vast. For example, Rosenbaum et al. [60–62, 47, 63], and [64] made important contributions to this field, particularly with respect to matching estimators. Cattaneo et al. [48] exemplify one way to apply similar tools to a regression discontinuity design. We contribute to this literature by applying a standard sensitivity analysis mechanism to the benchmark inference procedure of the SCM proposed by Abadie et al. [1].

This paper is divided as follows. Section 2 presents the SCM as proposed by Abadie and Gardeazabal [2], Abadie et al. [4] and [1]. Section 3 proposes a sensitivity analysis mechanism for the benchmark p-value formula of Abadie et al. [1] by parametrically reweighing the observed units. In Section 4, we extend the inference procedure to test any *sharp null hypothesis* and propose a way to construct confidence sets for the policy effect function. In Section 5, we run a Monte Carlo experiment to analyze the size, the power and the robustness of different tests statistics. We then extend the sensitivity analysis mechanism and the confidence sets to the cases when we observe multiple treated units or multiple outcomes in Section 6. We revisit, using the methods here developed, the empirical application about the Basque Country [2, 3] in Section 7. Finally, section 8 concludes.

2 Synthetic control method

This section is organized in two subsections. The first one presents the Synthetic Control Method, while the second one explains the benchmark for its inference procedure based on permutation tests. The ideas and notation that are used in the next two subsections are mostly based on [4] and [1]. We present these two topics in a way that will help us explain our sensitivity analysis mechanism, our extension to test any *sharp null hypothesis* using any test statistic and our confidence sets.

2.1 SCM: Policy effect and estimation

Suppose that we observe data for $(J+1) \in \mathbb{N}$ regions during $T \in \mathbb{N}$ time periods.[4] Additionally, assume that there is an intervention (policy) that affects only region 1 from period $T_0 + 1$ to period T uninterruptedly,[5] where $T_0 \in (1, T) \cap \mathbb{N}$. Let the scalar $Y_{j,t}^N$ 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 $Y_{j,t}^I$ be the potential outcome

4 We use the expression "region" and "region that faced an intervention" instead of more generic terms such as "unit" and "treated unit" because the former are typically used in SCM applications. Although, in this section, we assume that only one region faces an intervention, we extend this framework to include the case when multiple units face the same or a similar intervention in subsection 6.2.

5 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. [4] and [1], respectively. If the intervention is interrupted (e. g.: ETA's Terrorism in the Basque Country studied by Abadie and Gardeazabal [2]), 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".

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 (policy) effect** (sometimes simply called the *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 $D_{j,t} := 1$ if $j = 1$ and $t > T_0$, and $D_{j,t} := 0$ otherwise.

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.

Let $\mathbf{Y_j} := [Y_{j,1}...Y_{j,T_0}]'$ 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 $\mathbf{Y_j}$.[6] Let $\mathbf{Y_0} = [\mathbf{Y_2}...\mathbf{Y_{J+1}}]$ be a $(T_0 \times J)$-matrix and $\mathbf{X_0} = [\mathbf{X_2}...\mathbf{X_{J+1}}]$ be a $(K \times J)$-matrix.

Since we want to make region 1's synthetic control as similar as possible to the actual region 1, the SCM produces, for each $t \in \{1, ..., T\}$, $\widehat{Y}^N_{1,t} := \sum_{j=2}^{J+1} \widehat{w}_j Y_{j,t}$, which is an estimator of $Y^N_{1,t}$. The weights are given by $\widehat{\mathbf{W}} = [\widehat{w}_2...\widehat{w}_{J+1}]' := \widehat{\mathbf{W}}(\widehat{\mathbf{V}}) \in \mathbb{R}^J$, which are the solution to a nested minimization problem:

$$\widehat{\mathbf{W}}(\mathbf{V}) := \arg\min_{\mathbf{W} \in \mathcal{W}} (\mathbf{X_1} - \mathbf{X_0}\mathbf{W})'\mathbf{V}(\mathbf{X_1} - \mathbf{X_0}\mathbf{W}) \tag{2}$$

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

$$\widehat{\mathbf{V}} := \arg\min_{\mathbf{V} \in \mathcal{V}} (\mathbf{Y_1} - \mathbf{Y_0}\widehat{\mathbf{W}}(\mathbf{V}))'(\mathbf{Y_1} - \mathbf{Y_0}\widehat{\mathbf{W}}(\mathbf{V})) \tag{3}$$

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 to 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 SCM.[7]

Finally, we define the synthetic control estimator of $\alpha_{1,t}$ (or the estimated gap) as $\widehat{\alpha}_{1,t} := Y_{1,t} - \widehat{Y}^N_{1,t}$ for each $t \in \{1, ..., T\}$.

2.2 Hypothesis testing

Abadie et al. [4] and [1] develop a benchmark for a small sample inference procedure for the SCM that is similar to Fisher's Exact Hypothesis Test described by Fisher [45], Imbens and Rubin [46] and Rosenbaum

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

7 Abadie and Gardeazabal [2], Abadie et al. [4] and [1] 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 SCM 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 (2), data for the validation period are used to solve problem (3). 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 for 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 $v_k = |\beta_k| / (\sum_{k'=1}^{K} |\beta_{k'}|)$, 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.

[47]. Abadie et al. [4] permute regions to the treatment 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 $\widehat{\boldsymbol{\alpha}}_1 = [\widehat{\alpha}_{1,T_0+1} ... \widehat{\alpha}_{1,T}]'$ with the empirical distribution of $\widehat{\boldsymbol{\alpha}}_j = [\widehat{\alpha}_{j,T_0+1} ... \widehat{\alpha}_{j,T}]'$ estimated through the permutation procedure. 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 whatsoever*.

Abadie et al. [1] highlight that $|\widehat{\alpha}_{1,t}|$ can be abnormally large when compared to the empirical distribution of $|\widehat{\alpha}_{j,t}|$ 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 and to account for imperfect fit in the pre-intervention period, they recommend to use the empirical distribution of a summary statistic:

$$RMSPE_j := \frac{\sum_{t=T_0+1}^{T} \left(Y_{j,t} - \widehat{Y_{j,t}^N}\right)^2 / (T - T_0)}{\sum_{t=1}^{T_0} \left(Y_{j,t} - \widehat{Y_{j,t}^N}\right)^2 / T_0}, \tag{4}$$

Moreover, they propose to calculate a p-value

$$p := \frac{\sum_{j=1}^{J+1} \mathbb{I}\left[RMSPE_j \geq RMSPE_1\right]}{J+1}, \tag{5}$$

where $\mathbb{I}[\mathbf{A}]$ is the indicator function of event $\mathbf{A}$, and reject the *null hypothesis of no effect whatsoever* if p is less than some pre-specified significance level,[8] such as the traditional value of $\gamma = 0.1$. We stress that the *null hypothesis of no effect whatsoever* is also known as the *exact null hypothesis*[9] [45] and is given by

$$H_0 : Y_{j,t}^I = Y_{j,t}^N \text{ for each region } j \in \{1, ..., J+1\} \text{ and time period } t \in \{1, ..., T\}. \tag{6}$$

Note that rejecting the null hypothesis implies that there is some region with a non-zero effect for some time period.

3 Sensitivity analysis

The benchmark rejection rule based on the p-value formula (5) weights all units equally. Although natural, such a choice of weights is restrictive and the test result may depend heavily on it. An obvious extension of this benchmark imposes no restriction on the units' weights, while keeping the assumption that only one region faces the intervention.[10] In this case, we compute the p-value as

$$p := \sum_{j=1}^{J+1} \pi_j \times \mathbb{I}\left[RMSPE_j \geq RMSPE_1\right], \tag{7}$$

where π_j denotes the weight for region $j \in \{1, ..., J+1\}$, and reject the *exact null hypothesis* (6) if p is less than some pre-specified significance level, such as the traditional value of $\gamma = 0.1$.

While the benchmark p-value formula (5) is an important, but restrictive, starting point, the general p-value formula (7) provides no guidance on the choice of weights, $\{\pi_j\}_{j \in \{1,...,J+1\}}$, for the hypothesis test and, consequently, it is of no use to evaluate the robustness of the test result to the benchmark choice of equal weights. To analyze the sensitivity of the test result to the benchmark formula, one could start distorting the weights in the direction of changing the decision of the testing procedure by using continuous parametric

8 Yates [65] 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 %.

9 Observe that the *exact null hypothesis* (6) is stronger than assuming that the *typical* (mean or median) effect across regions is zero.

10 The case when more than one unit receives the treatment is explained in subsection 6.2.

weights. This is exactly what is done in [47] and [48]. They consider a sensitivity analysis that allows the empirical researcher to measure the robustness of his or her conclusions (i. e., the test's result regarding the rejection of the *exact null hypothesis*) to the choice of weights of the p-value formula (5), by distorting as little as possible the uniform distribution of weights. We present this sensitivity analysis step-by-step in the framework of the SCM:

1. Estimate the test statistics $RMSPE_1$, $RMSPE_2$,...,$RMSPE_{J+1}$ for all possible placebo treatment assignments $j \in \{1, ..., J+1\}$, where $RMSPE_1 := RMSPE^{obs}$ is the observed test statistic.
2. Rename them as $RMSPE_{(1)}$, $RMSPE_{(2)}$,...,$RMSPE_{(J+1)}$ such that $RMSPE_{(1)} > RMSPE_{(2)} > ... > RMSPE_{(J+1)}$.
3. Define $\bar{j} \in \Omega := \{(1), ..., (J+1)\}$ such that $RMSPE_{\bar{j}} = RMSPE^{obs}$. If there are more than one $j' \in \Omega$ that presents this property, take the largest one.
4. Define the weight of each placebo treatment assignment $(j) \in \Omega$ as

$$\pi_{(j)}(\phi, \boldsymbol{v}) = \frac{\exp(\phi v_{(j)})}{\sum_{j' \in \Omega} \exp(\phi v_{j'})}, \tag{8}$$

   where $\phi \in \mathbb{R}_+$ is the sensitivity parameter, $v_{j'} \in \{0, 1\}$ for each $j' \in \Omega$, and $\boldsymbol{v} := (v_1, ..., v_{J+1})$. Note that, when $\phi = 0$, all placebo treatment assignments have the same weight. Consequently, the benchmark p-value formula (5) imposes that $\phi = 0$. Moreover, the sensitivity parameter $\phi \in \mathbb{R}_+$ has a very intuitive interpretation: a region $j_1 \in \Omega$ with $v_{j_1} = 1$ has a weight $\Phi := \exp(\phi) - 1$ times larger than a region $j_2 \in \Omega$ with $v_{j_2} = 0$.
5. Using the weights given by equation (8), the permutation test's p-value is now given by

$$p(\phi, \boldsymbol{v}) := \sum_{(j) \in \Omega} \frac{\exp(\phi v_{(j)})}{\sum_{j' \in \Omega} \exp(\phi v_{j'})} \times \mathbb{I}\left[RMSPE_{(j)} \geq RMSPE_{\bar{j}}\right]. \tag{9}$$

   Observe that, given a sensitivity parameter $\phi \in \mathbb{R}_+$ and a vector $\boldsymbol{v}$, we reject the *exact null hypothesis* if $p(\phi, \boldsymbol{v})$ is less than some pre-specified significance level, such as the traditional value of $\gamma = 0.1$. Note also that, when $\phi = 0$, the p-value described in equation (9) simplifies to the benchmark p-value formula (5).
6. If the *exact 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[11] is given by

$$\begin{cases} v_{(j)} = 1 \text{ if } (j) \leq \bar{j} \\ v_{(j)} = 0 \text{ if } (j) > \bar{j}. \end{cases}$$

   where $(j) \in \Omega$. Define $\underline{\phi} \in \mathbb{R}_+$ such that

$$p\left(\underline{\phi}, \boldsymbol{v}\right) = \sum_{(j) \in \Omega} \frac{\exp\left(\underline{\phi} v_{(j)}\right)}{\sum_{j' \in \Omega} \exp\left(\underline{\phi} v_{j'}\right)} \times \mathbb{I}\left[RMSPE_{(j)} \geq RMSPE_{\bar{j}}\right] = \gamma,$$

   where γ is a pre-specified significance level. If $\underline{\phi} \in \mathbb{R}_+$ is close to zero, the permutation test's result is not robust to small deviations from the benchmark p-value formula (5), i. e., $\phi = 0$. Observe that, here, the weights of units at least as extreme as the treated unit are equally increased by equally reducing the weight of other units.
7. If the *exact 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[12] is given by

$$\begin{cases} v_{(j)} = 0 \text{ if } (j) \leq \bar{j} \\ v_{(j)} = 1 \text{ if } (j) > \bar{j}. \end{cases}$$

11 In this case, we pick values for $v_{j'}$ in order to make as hard as possible the rejection of the *exact null hypothesis* given a value for $\phi \in \mathbb{R}_+$.

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

where $(j) \in \Omega$. Define $\overline{\phi} \in \mathbb{R}_+$ such that

$$p\left(\overline{\phi}, \boldsymbol{v}\right) = \sum_{(j)\in\Omega} \frac{\exp\left(\overline{\phi} v_{(j)}\right)}{\sum_{j'\in\Omega}\exp\left(\overline{\phi} v_{j'}\right)} \times \mathbb{I}\left[RMSPE_{(j)} \geq RMSPE_{\bar{j}}\right] = \gamma,$$

where γ is a pre-specified significance level. If $\overline{\phi} \in \mathbb{R}_+$ is close to zero, the permutation test's result is not robust to small deviations from the benchmark p-value formula (5), i. e., $\phi = 0$. Observe that, here, the weights of units at least as extreme as the treated unit are equally reduced by equally increasing the weight of other units.

8. Based on the permutation test's result, we can fix the vector $\boldsymbol{v} = (v_1, ..., v_{J+1})$ and evaluate the impact of $\phi \in \mathbb{R}_+$ in the p-value given by equation (9) by plotting a graph with ϕ in the horizontal axis and $p(\phi, \boldsymbol{v})$ in the vertical axis. If $p(\phi, \boldsymbol{v})$ changes too quickly when we change ϕ, the permutation test is too sensitive to the choice of weights.

Consequently, large sensitivity parameter values — $\underline{\phi} \in \mathbb{R}_+$ and $\overline{\phi} \in \mathbb{R}_+$ — boost "our confidence that a particular synthetic control estimate reflects the impact of the intervention under scrutiny" in the same way that the benchmark inference procedure proposed by Abadie et al. [1, p. 499] does, since large sensitivity parameter values implies that the test result is robust to the choice of weights in the p-value formula (5). We discuss the meaning of large sensitivity parameter values in subsection 5.2 and illustrate the importance of this sensitivity analysis mechanism in our empirical application (section 7).

4 Sharp null hypotheses and confidence sets

4.1 Testing sharp null hypotheses

A researcher may be interested in testing not only the *exact null hypothesis*, but also a more general treatment effect function. We extend the inference procedure proposed by Abadie et al. [4] and [1] and the sensitivity analysis mechanism developed in section 3 to test any *sharp null hypothesis*. Now, instead of testing the *exact null hypothesis* given by equation (6), we want to test:

$$H_0^f: \; Y_{j,t}^I = Y_{j,t}^N + f_j(t) \text{ for each region } j \in \{1, ..., J+1\} \text{ and time period } t \in \{1, ..., T\}, \tag{10}$$

where $f_j : \{1, ..., T\} \to \mathbb{R}$ is a function of time that is specific to each region j and describes the treatment effect for each region, and $f := \{f_j\}_{j\in\{1,...,J+1\}}$.

Observe that a *sharp null hypothesis*, such as the one described by equation (10), allows us to know all potential outcomes for each region regardless of its treatment assignment. Note also that the *exact null hypothesis* (equation (6)) is a particular case of the *sharp null hypothesis* (10).

Although the *sharp null hypothesis* (10) 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 assume a simpler *sharp null hypothesis*[13]:

$$H_0^f: \; Y_{j,t}^I = Y_{j,t}^N + f(t) \text{ for each region } j \in \{1, ..., J+1\} \text{ and time period } t \in \{1, ..., T\}, \tag{11}$$

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

Now, for a given intervention effect function $f : \{1, ..., T\} \to \mathbb{R}$, the test statistic RMSPE given by equation (4) becomes

$$RMSPE_j^f := \frac{\sum_{t=T_0+1}^{T}\left(Y_{j,t} - \widehat{Y_{j,t}^N} - f(t)\right)^2/(T-T_0)}{\sum_{t=1}^{T_0}\left(Y_{j,t} - \widehat{Y_{j,t}^N} - f(t)\right)^2/T_0}, \tag{12}$$

13 We stress that the *exact null hypothesis* is still a particular case of the simpler *sharp null hypothesis* (11).

for all $j \in \{1, ..., J+1\}$, while the p-value given by equation (9) becomes

$$p^f(\phi, \boldsymbol{\upsilon}) := \sum_{j=1}^{J+1} \frac{\exp(\phi \upsilon_j)}{\sum_{j'=1}^{J+1} \exp(\phi \upsilon_{j'})} \times \mathbb{I}\left[RMSPE_j^f \geq RMSPE_1^f\right]. \tag{13}$$

For a given value of the sensitivity parameter $\phi \in \mathbb{R}_+$ and a given vector $\boldsymbol{\upsilon} = (\upsilon_1, ..., \upsilon_{J+1})$, we reject the *sharp null hypothesis* (11) if $p^f(\phi, \boldsymbol{\upsilon})$ is less than some pre-specified significance level, such as the traditional value of $\gamma = 0.1$. Note that, now, rejecting the null hypothesis implies that there is some region whose intervention effect is different from $f(t)$ for some time period $t \in \{1, ..., T\}$.

We highlight three interesting choices for the sensitivity parameter $\phi \in \mathbb{R}_+$ and the vector $\boldsymbol{\upsilon} = (\upsilon_1, ..., \upsilon_{J+1})$. The first one simply assumes $\phi = 0$ and $\boldsymbol{\upsilon} = (1, ..., 1)$, extending the benchmark inference procedure proposed by Abadie et al. [4] and [1] to test any *sharp null hypothesis* (equation (11)) instead of only the *exact null hypothesis* (equation (6)). The other two choices are related to the sensitivity parameter for the average worst case scenario $\underline{\phi} \in \mathbb{R}_+$ if the *sharp null hypothesis* (equation (11)) is rejected and for the best case scenario $\overline{\phi} \in \mathbb{R}_+$ if it is not rejected. In order to apply the sensitivity analysis mechanism proposed in section 3 to any *sharp null hypothesis* we follow the same steps described above, but using the test statistic and the p-value described in equations (12) and (13).

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 decision rule (13) allows the researcher to test whether the intervention effect is different than its costs.[14]

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 [9–12]. 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 decision rule (13) to each one of those three *sharp null hypotheses* in order to test which theoretical prediction is not rejected by the data.

4.2 Confidence sets

As described in subsection 4.1, we can, for a given value of the sensitivity parameter $\phi \in \mathbb{R}_+$, a given vector $\boldsymbol{\upsilon} = (\upsilon_1, ..., \upsilon_{J+1})$ and a given significance level $\gamma \in (0, 1)$, test many different types of *sharp null hypotheses*. Consequently, as explained by Imbens and Rubin [46] and Rosenbaum [47], we can invert the test statistic to estimate confidence sets for the treatment effect function. More clearly, using the equal weights benchmark p-value formula (5) or parametric deviations from it (equation (9)), we can construct a $(1-\gamma)$-confidence set in the space $\mathbb{R}^{\{1,...,T\}}$ as

$$CS_{(1-\gamma)}(\phi, \boldsymbol{\upsilon}) := \left\{f \in \mathbb{R}^{\{1,...,T\}} : p^f(\phi, \boldsymbol{\upsilon}) > \gamma\right\}, \tag{14}$$

14 In the empirical example (section 7), we show how to implement a one-sided test that can be used to test whether the intervention effect is greater than its costs.

where $p^f(\phi, \boldsymbol{v})$ is given by equation (13). Note that it is easy to interpret $CS_{(1-\gamma)}(\phi, \boldsymbol{v})$: it contains all intervention effect functions whose associated *sharp null hypotheses* are not rejected by the inference procedure described in subsection 4.1.

However, although theoretically possible to define such a general confidence set, null hypothesis (11) 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 simple subsets of $CS_{(1-\gamma)}(\phi, \boldsymbol{v})$.

Firstly, we propose to assume the following null hypothesis:

$$H_0^c: \; Y_{j,t}^I = Y_{j,t}^N + c \times \mathbb{I}(t \geq T_0 + 1) \tag{15}$$

for each region $j \in \{1, ..., J+1\}$ and time period $t \in \{1, ..., T\}$, 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 4.1 to any $c \in \mathbb{R}$, estimating the empirical distribution of $RMSPE^c$. Using the weights given by equation (8), we can then construct a $(1-\gamma)$-confidence interval for the constant intervention effect as

$$CI_{(1-\gamma)}(\phi, \boldsymbol{v}) := \left\{ f \in \mathbb{R}^{\{1,...,T\}} : f(t) = c \text{ and } p^c(\phi) > \gamma \right\} \subseteq CS_{(1-\gamma)}(\phi, \boldsymbol{v}) \tag{16}$$

where $c \in \mathbb{R}$ and $\gamma \in (0,1) \subset \mathbb{R}$. It is easy to interpret $CI_{(1-\gamma)}(\phi, \boldsymbol{v})$: it contains all constant in time intervention effects whose associated *sharp null hypotheses* are not rejected by the inference procedure described in subsection 4.1.

Secondly, we can easily modify equations (15) and (16) to a linear in time intervention effect (with intercept equal to zero). Assume

$$H_0^{\tilde{c}}: \; Y_{j,t}^I = Y_{j,t}^N + \tilde{c} \times (t - T_0) \times \mathbb{I}(t \geq T_0 + 1) \tag{17}$$

for each region $j \in \{1, ..., J+1\}$ and time period $t \in \{1, ..., T\}$, 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 4.1 to any $\tilde{c} \in \mathbb{R}$, estimating the empirical distribution of $RMSPE^{\tilde{c}}$. Using the weights given by equation (8), we can then construct a $(1-\gamma)$-confidence set for the linear intervention effect as

$$\widetilde{CS}_{(1-\gamma)}(\phi, \boldsymbol{v}) := \left\{ f \in \mathbb{R}^{\{1,...,T\}} : \begin{array}{c} f(t) = \tilde{c} \times (t - T_0) \times \mathbb{I}(t \geq T_0 + 1) \\ \text{and } p^{\tilde{c}}(\phi) > \gamma \end{array} \right\} \subseteq CS_{(1-\gamma)}(\phi, \boldsymbol{v}) \tag{18}$$

where $\gamma \in (0,1) \subset \mathbb{R}$. It is also easy to interpret $\widetilde{CS}_{(1-\gamma)}(\phi, \boldsymbol{v})$: 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 4.1.

We also note that extending our confidence intervals to two-parameter functions (e. g.: quadratic, exponential and logarithmic functions) is theoretically straightforward as equation (14) makes clear. However, since we believe that computationally estimating such confidence sets would be time consuming for the practitioner, we opted for restricting our main examples to one-parameter functions[15] (equations (16) and (18)).

Moreover, we highlight that confidence sets (16) and (18) summarize 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. For example, narrower confidence sets suggest stronger conclusions. Furthermore, by plotting confidence sets for different values of the sensitivity parameter $\phi \in \mathbb{R}_+$, the empirical researcher can access how robust his or her qualitative conclusions are to deviations from the equal weights benchmark p-value formula (5) by comparing the areas of confidence sets for different values of $\phi \in \mathbb{R}_+$. As before, we highlight three interesting

15 On our website https://goo.gl/RBYomh, we provide R and Stata codes to compute the confidence sets in equations (16) and (18).

choices for the sensitivity parameter $\phi \in \mathbb{R}_+$ and the vector $\boldsymbol{\upsilon} = (\upsilon_1, ..., \upsilon_{J+1})$. The first one simply assumes $\phi = 0$ and $\boldsymbol{\upsilon} = (1, ..., 1)$, i. e., it weights all observed units equally as proposed by Abadie et al. [1] as a benchmark. The other two choices are related to the sensitivity parameter for the worst case scenario $\underline{\phi} \in \mathbb{R}_+$ if the *exact null hypothesis* (equation (6)) is rejected and for the best case scenario $\overline{\phi} \in \mathbb{R}_+$ if it is not rejected. Our empirical application (section 7) exemplifies the communication efficacy of those graphical devices.

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 the inference procedure of the SCM and our confidence sets separately for each post-intervention time period $t' \in \{T_0 + 1, ..., T\}$ using $(\widehat{\alpha}_{1,t'})^2$ as a test statistic. In subsection 6.1, we explain why a point-wise confidence interval may not be adequate and propose an alternative inference procedure for multiple outcome variables.

5 Other test statistics and a Monte Carlo experiment

Although we presented the inference procedure proposed by Abadie et al. [4] and [1], our sensitivity analysis mechanism and our confidence sets using the RMSPE as a test statistic, all of them can use any test statistic. Following Imbens and Rubin [46], 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

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

   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* (11).

The observed test statistic is given by $\theta^{f,obs} := \theta(e_1, \tau, \mathbf{Y}, \mathbf{X}, f)$ and, under the *sharp null hypothesis* (11), 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 $j \in \{1, ..., J + 1\}$, where e_j is the j-th canonical vector of $\mathbb{R}^{J+1}$. Using weights (8) and fixing a value of the sensitivity parameter $\phi \in \mathbb{R}_+$ and a vector $\boldsymbol{\upsilon} = (\upsilon_1, ..., \upsilon_{J+1})$, we reject the *sharp null hypothesis* (11) if

$$p_{\theta^f}(\phi, \boldsymbol{\upsilon}) := \sum_{j=1}^{J+1} \frac{\exp(\phi\upsilon_j)}{\sum_{j'=1}^{J+1} \exp(\phi\upsilon_{j'})} \times \mathbb{I}\left[\theta(e_j, \tau, \mathbf{Y}, \mathbf{X}, f) \geq \theta^{f,obs}\right] \leq \gamma, \tag{19}$$

where γ is some pre-specified significance level.

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. [4] and [1] are inserted in this framework.

5.1 Monte Carlo experiment: Rejection rates

In this subsection, we analyze the size and the power of five different test statistics when they are applied to the inference procedure described above imposing that $\phi = 0$ and $\boldsymbol{\upsilon} = (1, ..., 1)$, i. e., we use the benchmark

weights proposed by Abadie et al. [1]. In order to do that, we assume seven different intervention effects, simulate 3,000 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 (6)), following the mentioned inference procedure assuming that $\phi = 0$ and $\boldsymbol{\nu} = (1, ..., 1)$ and using each test statistic.

Firstly, we describe our five test statistics. Then, we explain our data generating process and discuss the results.

We analyze the following test statistics:

- $\theta^1 := mean\left(\left|\widehat{\alpha}_{\tilde{j},t}\right| | t \geq T_0 + 1\right)$ is one way to aggregate the information provided by placebo gaps graphs that were introduced by Abadie et al. [4].
- $\theta^2 := RMSPE_{\tilde{j}}$ is recommended by Abadie et al. [1] because it controls for the quality of the pre-intervention fit.
- 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 $\overline{\alpha}_{post} := \dfrac{\left(\sum_{t=T_0+1}^{T} \widehat{\alpha}_{\tilde{j},t}\right)}{(T - T_0)}$ and $\widehat{\sigma} := \sqrt{\dfrac{\left(\sum_{t=T_0+1}^{T} \left(\widehat{\alpha}_{\tilde{j},t} - \overline{\alpha}_{post}\right)^2\right)}{(T - T_0)}}$. This test statistic is used by Mideksa [13].

- $\theta^4 := \left| mean\left(Y_{\tilde{j},t} | t \geq T_0 + 1\right) - \dfrac{\sum_{t=T_0+1}^{T} \sum_{j \neq \tilde{j}} Y_{j,t}}{(T - T_0) \times J} \right|$ 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 [46].
- θ^5 is the coefficient of the interaction term in a differences-in-differences model. More precisely, we estimate the model

$$Y_{j,t} = \eta_1 \times \mathbb{I}\left[j = \tilde{j}\right] + \eta_2 \times \mathbb{I}\left[j = \tilde{j}\right] \times \mathbb{I}\left[t \geq T_0 + 1\right] + Z_{j,t} \times \boldsymbol{\zeta} + \xi_j + \mu_t + \varepsilon_{j,t}, \tag{20}$$

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

Observe that, in this notation, $\tilde{j}$ is the region that is assumed to face the intervention in each permutation, and $mean(\mathbf{B}|\mathbf{A})$ is the mean of variable **B** conditional on event **A**. 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. For each Monte Carlo repetition and each test statistic, we also compute the worst case scenario $\underline{\phi} \in \mathbb{R}_+$ if the *exact null hypothesis* is rejected and the best case scenario $\overline{\phi} \in \mathbb{R}_+$ if it is not rejected. We discuss the results for the sensitivity analysis parameters in the next subsection.

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 given our data generating process. For this same reason, we also report rejection rates for the differences-in-differences inference procedure proposed by Conley and Taber [49] (CT).[16] However, we stress that Conley and Taber [49] propose an asymptotic inference procedure and that our Monte Carlo Experiment has a small sample size, implying that the CT method is not the most suitable tool in this context.

16 We estimate model (20) and test the null hypothesis $H_0 : \eta_2 = 0$ using the confidence intervals recommend by Conley and Taber [49]. 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 %.

The first step in the data generating process of 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 [4]:

$$\begin{aligned} Y_{j,t+1}^N &= \delta_t Y_{j,t}^N + \boldsymbol{\beta}_{t+1}\mathbf{Z}_{j,t+1} + u_{j,t+1} \\ \mathbf{Z}_{j,t+1} &= \kappa_t Y_{j,t}^N + \boldsymbol{\rho}_t \mathbf{Z}_{j,t} + \mathbf{v}_{j,t+1} \end{aligned} \tag{21}$$

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.[17] 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 $\boldsymbol{\rho}_t$ 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 $Y_{j,0}^N = \boldsymbol{\beta}_0 \mathbf{Z}_{j,0} + u_{j,0}$. Finally, the potential outcome when region 1 faces the intervention in period $t \in \{1, ..., T\}$ is given by

$$Y_{1,t}^I = Y_{1,t}^N + \lambda \times sd(Y_{1,\tilde{\tau}}^N | \tilde{\tau} \leq T_0) \times (t - T_0) \times \mathbb{I}\left[t \geq T_0 + 1\right], \tag{22}$$

where $\lambda \in \{0, 0.05, 0.1, 0.25, 0.5, 1.0, 2.0\}$ is the intervention effect and $sd(\mathbf{B}|\mathbf{A})$ is the standard deviation of variable $\mathbf{B}$ conditional on event $\mathbf{A}$. 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 one-direction alternative hypothesis.[18]

Now that we have explained our data generating process with 21,000 Monte Carlo repetitions, we discuss our findings. Table 1 shows the results of our Monte Carlo Experiment about the size and power of the analyzed tests when we assume $\phi = 0$ and $\boldsymbol{\upsilon} = (1, ..., 1)$. Each cell presents the rejection rate of the permutation test described above that uses the test statistic in each row or the rejection rate of the test proposed by Conley and Taber [49] when the true intervention effect is given by the value mentioned in the column's heading. Consequently, while column (1) presents tests' sizes, the columns (2)–(7) present their power.

Table 1: Monte Carlo Experiment's Rejection Rates.

	Intervention Effect						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Test Statistic	λ= .0	λ= .05	λ= .1	λ= .25	λ= .5	λ= 1.0	λ= 2.0
$\widehat{\theta}^1$	0.10	0.19	0.22	0.36	0.43	0.63	0.68
$\widehat{\theta}^2$	0.10	0.32	0.36	0.49	0.53	0.72	0.77
$\widehat{\theta}^3$	0.10	0.63	0.69	0.80	0.87	0.94	0.95
$\widehat{\theta}^4$	0.10	0.20	0.24	0.36	0.44	0.60	0.65
$\widehat{\theta}^5$	0.10	0.18	0.24	0.36	0.42	0.63	0.70
CT	0.10	0.15	0.21	0.32	0.37	0.61	0.66

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 $\boldsymbol{\lambda}$ 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 [49].

17 $\mathbf{X}_j$ is a vector that contains the pre-intervention averages of the control variables and of the outcome variable.

18 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 and are available upon request.

Analyzing column (1), we note that the five permutation tests of our Monte Carlo Experiment (θ^1-θ^5) present the correct nominal size as expected by the decision rule of Fisher's Exact Inference Procedure [45]. Moreover, the asymptotic inference procedure proposed by Conley and Taber [49] (CT) has a true size close to the correct one (10.53 %).

Analyzing the other columns, we note that the test statistic *RMSPE*, proposed by Abadie et al. [1] (θ^2), is uniformly more powerful than the simple test statistics (θ^4, θ^5) 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 as in our data generating process. We also stress that the hypothesis test based on the statistic *RMSPE* (θ^2) outperforms the test proposed by Conley and Taber [49] (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 that applies an asymptotic inference procedure. This last result can be explained by the fact that, while our sample size is small ($J + 1 = 20$), the CT inference procedure is an asymptotic test based on the number of control regions going to infinity and, therefore, inadequate for this data generating process.

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 (θ^4) or the permuted differences-in-differences test (θ^5). As pointed out by an anonymous referee, we stress that this gain of power is present even though our treatment effect also increases the standard deviation of the potential outcome, i. e., it also increases the denominator of the observed test statistic. We also note that the large power of the t-test have been previously observed in contexts that are different from ours: Lehmann [66] looks to a simple test of mean differences, Ibragimov and Muller [67] analyzes a two-sample test of mean differences where samples' variances are different from each other, and Young [68] focus on a linear regression coefficient. However, it is easy to think about contexts in which the simple t-test (θ^3) would be weak. As pointed out by an anonymous referee, if positive and negative treatment effects for different time periods cancel out in $\overline{\alpha}_{post}$, the simple t-test (θ^3) will not be able to detect deviations with respect to the *null hypothesis of no effect whatsoever*. This example illustrates two important points. First, when the researcher believes that the treatment effect varies a lot over time, he or she should explicitly acknowledge that and treat different time periods as different outcome variables, applying the multiple hypothesis testing framework described in section 6.1. Second, as Eudey et al. [69] stress, the test statistic should be carefully chosen in order to have power against the alternative hypothesis of interest.

Table 2: Rejection Rates Using Only Units with a Good Pre-intervention Fit.

	Intervention Effect						
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**
Test Statistic	**λ= .0**	**λ= .05**	**λ= .1**	**λ= .25**	**λ= .5**	**λ= 1.0**	**λ= 2.0**
$\widehat{\theta}^1$	0.13	0.38	0.43	0.56	0.59	0.76	0.81
$\widehat{\theta}^2$	0.06	0.26	0.31	0.44	0.47	0.65	0.71
$\widehat{\theta}^3$	0.06	0.57	0.62	0.76	0.82	0.9	0.92

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 $\boldsymbol{\lambda}$ 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. Good pre-intervention fit is defined as a pre-intervention MSPE at most five times larger than the MSPE of the treated unit.

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). Following Abadie et al. [1], a possible explanation for the low power of (θ^1) is the fact that this test statistic ignores the quality of the pre-intervention fit. To analyze this possibility, table 2

presents the rejection rates of the test statistics that uses the SCM ($\theta^1 - \theta^3$) when only units with a good pre-intervention fit are used to compute the p-value (equation (5)).[19] Here, we follow Abadie et al. [1] and define units with a good pre-intervention fit as units whose mean squared prediction error (MSPE) is at most five times larger than the MSPE of the treated unit.

We note that test statistic θ^1 becomes, as expected, more powerful when we control for the quality of the pre-intervention fit. However, the cost of this increase in power is a small over-rejection of the null hypothesis when $\lambda = 0$. We also highlight that test statistics θ^2 and θ^3 are, now, slightly conservative, because, when we exclude units with a poor pre-intervention fit, the null hypothesis is only rejected when the observed unit is the most extreme, effectively reducing the level of the permutation test. Most importantly, since the RMSPE test statistic (θ^2) already controls for the pre-intervention fit, excluding units with a poor pre-intervention fit reduces its power.

At this point, we avoid making any stronger suggestion about which test statistic the empirical researcher should use, because, as Eudey et al. [69, p. 14] make clear, this choice is data dependent since the empirical researcher's goal is to match the test statistic to the research question or population object of interest.

5.2 Monte Carlo experiment: Sensitivity analysis

In this subsection, we analyze the behavior of the sensitivity analysis mechanism proposed in section 3 when we generate datasets based on the data generating process described above. We focus on the average values of the sensitivity parameter that change a hypothesis test's result, i. e., the values of $\overline{\phi} \in \mathbb{R}_+$ if the *exact null hypothesis* is not rejected and $\underline{\phi} \in \mathbb{R}_+$ if the *exact null hypothesis* is rejected. As before, we assume seven different intervention effects, simulate 3,000 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 (6)), following the mentioned inference procedure assuming that $\phi = 0$ and $\boldsymbol{\nu} = (1, ..., 1)$ and using the five test statistics described in subsection 5.1. Based on each test's result, we compute either the worst case scenario $\underline{\phi} \in \mathbb{R}_+$ if the *exact null hypothesis* is rejected or the best case scenario $\overline{\phi} \in \mathbb{R}_+$ if the *exact null hypothesis* is not rejected.

Tables 3 shows the sensitivity parameter for the average worst case scenario $\underline{\phi} \in \mathbb{R}_+$ if the *exact null hypothesis* is rejected and for the best case scenario $\overline{\phi} \in \mathbb{R}_+$ if it is not rejected.[20] Each cell presents the average value of the sensitivity parameter that changes the hypothesis' test result associated to the scenario in the panel and to the test statistic in each row when the true intervention effect is given by the value mentioned in the column's heading.

On the one hand, when the *exact null hypothesis* is true (column (1)) and we reject the null hypothesis (Panel A), we want the sensitivity parameter $\underline{\phi} \in \mathbb{R}_+$ to be small, because a less robust result would help us avoid making a Type I error. On the other hand, when the *exact null hypothesis* is false (columns (2)–(7)) and we reject the null hypothesis (Panel A), we want the sensitivity parameter $\underline{\phi} \in \mathbb{R}_+$ to be large, because a more robust result would help us avoid making a Type II error. As table 3 shows, the sensitivity analysis parameter $\underline{\phi} \in \mathbb{R}_+$ for the three test statistics increases when the intervention effect $\lambda \in \mathbb{R}_+$ increases, as desired.

Moreover, when the *exact null hypothesis* is true (column (1)) and we do not reject the null hypothesis (Panel B), we want the sensitivity parameter $\overline{\phi} \in \mathbb{R}_+$ to be large, because a more robust result would help us avoid making a Type I error. Similarly, when the *exact null hypothesis* is false (columns (2)–(7)) and we do not reject the null hypothesis (Panel B), we want the sensitivity parameter $\overline{\phi} \in \mathbb{R}_+$ to be small, because a more robust result would help to avoid making a Type II error. As table 3 shows, the sensitivity analysis parameter $\overline{\phi} \in \mathbb{R}_+$ for test statistics θ^1 and θ^2 decreases when the intervention effect $\lambda \in \mathbb{R}_+$ increases, as desired. However, even if θ^3 is the most powerful test statistic in our Monte Carlo experiment, it does not have good properties when applied to the sensitivity analysis mechanism.

19 To save space, we do not report the results for the test statistics θ^4 and θ^5. They are available upon request.

20 To save space, we do not report the results for the test statistics θ^4 and θ^5. They are available upon request.

Table 3: Sensitivity Analysis.

	Intervention Effect						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Test Statistics	$\lambda = .0$	$\lambda = .05$	$\lambda = .1$	$\lambda = .25$	$\lambda = .5$	$\lambda = 1.0$	$\lambda = 2.0$
Panel A: Worst Case Scenario $\underline{\phi} \in \mathbb{R}_+ - H_0$ is rejected							
$\widehat{\theta}^1$	0.38	0.40	0.34	0.48	0.52	0.57	0.62
$\widehat{\theta}^2$	0.38	0.60	0.62	0.66	0.66	0.70	0.70
$\widehat{\theta}^3$	0.38	0.68	0.68	0.71	0.72	0.74	0.75
Panel B: Best Case Scenario $\overline{\phi} \in \mathbb{R}_+ - H_0$ is not rejected							
$\widehat{\theta}^1$	2.50	1.83	1.77	1.54	1.54	1.33	1.13
$\widehat{\theta}^2$	2.50	2.26	2.22	2.04	1.96	1.86	1.72
$\widehat{\theta}^3$	2.50	2.94	3.00	3.44	3.29	3.05	3.91

Source: Authors' own elaboration. *Notes:* Each cell presents the value of the sensitivity parameter that changes the hypothesis' test result associated to the scenario in the panel and to the test statistic in each row when the true intervention effect is given by the value mentioned in the column's heading. H_0 is the *exact null hypothesis* given by equation (6). $\widehat{\theta}^1$–$\widehat{\theta}^3$ are associated to permutation tests that uses the Synthetic Control Estimator.

Furthermore, when we compare the sensitivity parameters $\underline{\phi} \in \mathbb{R}_+$ and $\overline{\phi} \in \mathbb{R}_+$ across the different test statistics, we find that $\widehat{\theta}^2$ is more robust than θ^1 and θ^3 for some values of the intervention effect parameter $\lambda \in \mathbb{R}_+$. As discussed in subsection 5.1, the best test statistic depends on the population object of interest. However, test statistic θ^2 — the traditional *RMSPE* statistic proposed by Abadie et al. [1] — performs, in our specific data generating process, satisfactorily with respect to power and robustness to deviations from the equal weights benchmark p-value formula (5).

To conclude, we stress that similar Monte Carlo experiments might help the empirical researcher to gauge the robustness of his or her findings. For example, in section 7, we observe a dataset with $J + 1 = 14$, $T_0 = 15$ and $T = 43$, and use a one-sided test statistic. We can implement a Monte Carlo experiment using the data generating process of subsection 5.1 with the parameters just mentioned and with the test statistic of section 7. For each Monte Carlo repetition that rejects the null hypothesis at the $^2/_{14}$-significance level, we save the sensitivity parameter for the worst case scenario $\underline{\phi}$. Averaging across Monte Carlo repetitions, we find that the average worst case scenario $\underline{\phi}$ is 0.995. Since in our empirical exercise, we find a sensitivity parameter $\underline{\phi} = 0.495$, we may conclude that the empirical results from section 7 are not very robust to deviations from the equal weight benchmark p-value formula (5), a problem that is connected to the small sample size of the exercise.

6 Extensions to the inference procedure

In this section, we discuss the inference procedure for SCM when we observe Multiple Outcomes or Multiple Treated Units. By doing so, we also extend the sensitivity analysis mechanism to both cases and the confidence sets to the second case.

6.1 Simultaneously testing hypotheses about multiple outcomes

Imbens and Rubin [46] states that the validity of the procedure described in subsection 4.1 depends on a prior (i. e., before seeing the data) commitment to a test statistic. Moreover, Anderson [50] shows that simultaneously testing hypotheses about a large number of outcomes can be dangerous, leading to an increase in the

number of false rejections.[21] Consequently, applying the inference procedure described in subsection 4.1 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 [50] to our framework, we propose one way to test any *sharp null hypothesis* for a large number of outcome variables, preserving the correct test level for each variable of interest.

First, we modify the framework described in subsection 4.1, assuming that there are $M \in \mathbb{N}$ observed outcome variables — $\mathbf{Y}^1, ..., \mathbf{Y}^M$ — with their associated potential outcomes. Now, our null hypothesis is also more complex than the one described in equation (11):

$$H_0^f : Y_{j,t}^{m,I} = Y_{j,t}^{m,N} + f_m(t) \tag{23}$$

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 and $f := \{f_m\}_{m \in \{1,...,M\}}$. 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.

Based on the benchmark inference procedure developed by Abadie et al. [4] and [1], 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,

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

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

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

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 $p_{\theta_{fm},*}^{\tilde{j}} := \min\left\{p_{\theta_{f_m}}^{\tilde{j}}, p_{\theta_{f_{m+1}}}^{\tilde{j}}, ..., p_{\theta_{f_M}}^{\tilde{j}}\right\}$ for each $m \in \{1, ..., M\}$ and each $\tilde{j} \in \{1, ..., J+1\}$. Then, we estimate, for a given value of the sensitivity parameter $\phi \in \mathbb{R}_+$ and a given vector $\boldsymbol{\upsilon} = (\upsilon_1, ..., \upsilon_{J+1})$ and using the weights given by equation (8),

$$p_{\theta_{f_m}^{obs}}^{fwer*}(\phi, \boldsymbol{\upsilon}) := \sum_{j=1}^{J+1} \frac{\exp(\phi \upsilon_j)}{\sum_{j'=1}^{J+1} \exp(\phi \upsilon_{j'})} \times \mathbb{I}\left[p_{\theta_{f_m},*}^{j} \leq p_{\theta_{f_m}}^{obs}\right] \tag{24}$$

for each $m \in \{1, ..., M\}$. We enforce monotonicity one last time by computing

$$p_{\theta_{f_m}^{obs}}^{fwer}(\phi, \boldsymbol{\upsilon}) := \min\left\{p_{\theta_{f_m}^{obs}}^{fwer*}(\phi, \boldsymbol{\upsilon}), p_{\theta_{f_{m+1}}^{obs}}^{fwer*}(\phi, \boldsymbol{\upsilon}), ..., p_{\theta_{f_M}^{obs}}^{fwer*}(\phi, \boldsymbol{\upsilon})\right\}$$

for each $m \in \{1, ..., M\}$. Finally, for each outcome variable $m \in \{1, ..., M\}$, we reject the *sharp null hypothesis* (23) if $p_{\theta_{f_m}^{obs}}^{fwer}(\phi, \boldsymbol{\upsilon}) \leq \gamma$, where γ is a pre-specified significance level.

21 List et al. [70] 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.

It is important to observe that rejecting the null hypothesis 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 a 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 4.1 is still valid and is similar in flavor with the *summary index test* proposed by Anderson [50], because we summarized the entire time information in a single test statistic. Since Anderson [50] argues that the *summary index test*[22] has more power than the *familywise error rate control* approach, we recommend that the empirical researcher uses the inference procedure described in subsection 4.1 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 (as we are in section 7), he or she should interpret each post-intervention time period as a different outcome variable and apply the inference procedure described in this subsection.

As before, we highlight three interesting choices for the sensitivity parameter $\phi \in \mathbb{R}_+$ and the vector $\boldsymbol{\upsilon} = (\upsilon_1, ..., \upsilon_{J+1})$. The first one simply assumes $\phi = 0$ and $\boldsymbol{\upsilon} = (1, ..., 1)$, extending the benchmark inference procedure proposed by Abadie et al. [4] and [1] to test *sharp null hypotheses* about multiple outcome variables (equation (23)). The other two choices are related to the sensitivity parameter for the average worst case scenario $\underline{\phi} \in \mathbb{R}_+$ if the *sharp null hypothesis* (equation (11)) is rejected and for the best case scenario $\overline{\phi} \in \mathbb{R}_+$ if it is not rejected. We can easily apply the sensitivity analysis mechanism proposed in section 3 to any outcome variable $m \in \{1, ..., M\}$ using $p^{fwer}_{\theta^{obs}_{f_m}}(\phi, \boldsymbol{\upsilon})$ to define either $\underline{\phi}_m \in \mathbb{R}_+$ or $\overline{\phi}_m \in \mathbb{R}_+$.

6.2 Hypothesis testing and confidence sets with multiple treated units

Cavallo et al. [10] extend the SCM developed by Abadie and Gardeazabal [2] and Abadie et al. [4] to the case when we observe multiple treated units. We briefly extend their contribution to allow our sensitivity analysis mechanism and to test any kind of *sharp null hypothesis*. By doing so, we can also 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 $\widehat{\alpha}_{1^g,t} := Y_{1^g,t} - \widehat{Y}^N_{1^g,t}$ for each $t \in \{1, ..., T\}$ and each intervention $g \in \{1, ..., G\}$. The estimated pooled intervention effect according to the SCM is given by $\overline{\widehat{\alpha}}_{1,t} := \sum_{g=1}^{G} \widehat{\alpha}_{1^g,t}/G$ for each $t \in \{1, ..., T\}$.

Differently from [10], 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 [50].[23] We also adapt their benchmark p-value formula to consider deviations from equally weighted units by using parametric weights (equation (26)) that are similar to the ones in equation (8).

Now, our *sharp null hypothesis* is given by:

$$H_0 : Y^I_{j^g,t} = Y^N_{j^g,t} + f(t) \tag{25}$$

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

22 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 4.1. 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 $t \in \{1, ..., T\}$ and some outcome variable $m \in \{1, ..., M\}$.

23 For more information about over-rejecting the null hypothesis, see the articles mentioned in subsection 6.1.

do so because we almost never have a meaningful null hypothesis that is precise enough to specify individual intervention effects for each observed region. Moreover, since most empirical applications with multiple treated units are concerned with interventions that are similar across regions, imposing that the treatment effect does not vary across interventions is a reasonable assumption.

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 6.1) 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 6.1 is based on the *familywise error rate*.[24]

Furthermore, we define a test statistic $\theta_{pld,f}$ for the pooled intervention effect as a known positive real-valued function $\theta_{pld}((t^g, \tau^g, \mathbf{Y}^g, \mathbf{X}^g)_{g=1}^G, f)$ that summarizes the entire information of all interventions.

Now, to apply the inference procedure to the pooled intervention effect allowing for the sensitivity analysis mechanism described in section 3, we recommend the following steps:

1. Estimate the test statistics $\theta_1^f, \theta_2^f, ..., \theta_Q^f$ for each possible placebo treatment assignment $q \in \{1, ..., Q\}$, where $\theta_1^f = \theta_{pld,f}^{obs} := \theta_{pld}((e_{1^g}, \tau^g, \mathbf{Y}^g, \mathbf{X}^g)_{g=1}^G, f)$ is the observed test statistic and e_{j^g} is the j^g-th vector of the canonical base of $\mathbb{R}^{J^g+1}$. A possible placebo treatment assignment simply permutes which region is assumed to be treated in each intervention $g \in \{1, ..., G\}$, i. e., it uses different combinations of canonical vectors $(e_{j^1}, ..., e_{j^G})$. Note that there are $Q := \prod_{g=1}^G (J^g + 1)$ possible placebo pooled intervention effects.
2. Follow the mechanism described in section 3 where the word *region* and the indexes j associated to it are now interpreted as *placebo treatment assignments* and indexes q. In particular, the p-value of equation (9) is now given by

$$p_{\theta_{pld,f}}(\phi, \boldsymbol{\upsilon}) := \sum_{(q)\in\{1,...,Q\}} \frac{\exp(\phi \upsilon_{(q)})}{\sum_{q'\in\{1,...,Q\}} \exp(\phi \upsilon_{q'})} \times \mathbb{I}\left[\theta_{(q)} \geq \theta_{\overline{q}}\right]. \tag{26}$$

We stress that rejecting null hypothesis (25) implies that there is some intervention with some region whose intervention effect differs from $f(t)$ for some time period $t \in \{1, ..., T\}$.

Finally, to extend the confidence sets of subsection 4.2 to the pooled intervention effect, simply follow the definitions of the aforementioned subsection using the p-value given by equation (26).

7 Empirical application

In this section, we aim to illustrate that our extensions of the inference procedure proposed by Abadie et al. [4] and [1] can cast new light on empirical studies that use SCM. In particularly, we can analyze the robustness of empirical results to the equal weights benchmark p-value formula (5), test more flexible null hypotheses, and summarize important information in simple and effective graphs. In order to achieve this goal, we use economic data for Spanish regions made available by Abadie and Gardeazabal [2] and discussed by Abadie et al. [3] too.

We start by evaluating the statistical significance of the economic impact of ETA's terrorism. Since power is a concern due to the small sample size, we implement, as suggested by an anonymous referee, a one-sided test because only negative effects of terrorism on GDP are of interest. To do so, our test statistic,[25]

24 Anderson [50] offers a detailed discussion about the differences between inference procedures based on the *summary index test* and on the *familywise error rate*.

25 This test statistic is an one-sided version of the t-test in section 5, which is powerful against alternative hypothesis that have only one direction of impact such as the one in subsection 5.1 and the one that is of interest in the present empirical exercise. The results for the two-sided test that uses the *RMSPE* test statistic are available upon request.

$\theta = -\frac{\bar{\alpha}_{post}/(T-T_0)}{\hat{\sigma}/\sqrt{T-T_0}}$, can assume positive and negative values (differently from section 5) and we apply the inference procedure described in subsection 2.2, an exercise similar to the one implemented by Abadie et al. [3]. Then, we analyze the robustness of this result to the equal weights benchmark using the procedure explained in section 3. After that, we estimate the upper bound of one-sided Confidence Sets that contains all constant in time intervention effects and all linear in time intervention effects (with intercept equal to zero) whose associated *sharp null hypotheses* are not rejected by our inference procedure (see equations (16) and (18)) when we use the test statistic θ.[26] Furthermore, we test whether the intervention effect can be reasonably approximated by a quadratic function. Finally, we analyze the timing of the economic impact of ETA's terrorism using the procedure described in subsection 6.1.

The data set used by Abadie and Gardeazabal [2] 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 regional level and there are seventeen regions, including the Basque Country ($J + 1 = 17$). For historical details and descriptive statistics about this data set, see [2] and [3].

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 (2), 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 (3).

When we estimate the intervention effect for the Basque Country and the placebo effect for all the other Spanish regions, it is visually unclear (subfigure 1(a)) whether the estimated intervention effect looks abnormally negative when compared to the estimated placebo effects. When we apply the inference procedure proposed by Abadie et al. [1] (see subsection 2.2) using θ as a test statistic, we find $p = 3/17$, a marginal rejection of the *null hypothesis of no effect whatsoever* given the small sample size.

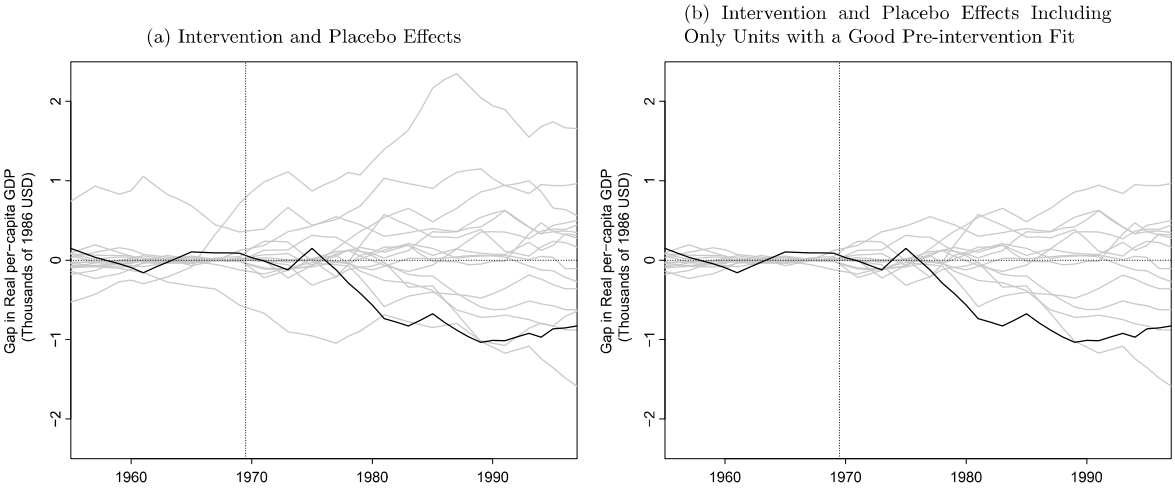

Figure 1: Estimated Effects using the Synthetic Control Method. *Note:* While the gray lines show the estimated placebo effect for each Spanish region, the black lines show the estimated impact of ETA's terrorism on the Basque Country's economy. A good pre-intervention fit is defined as a pre-intervention MSPE at most five times greater than the Basque Country's pre-intervention MSPE.

26 Since only negative effects of ETA's terrorism on the Basque economy are of interest, we only need to report the largest treatment effect constant (c for equation (16) or $\tilde{c}$ for equation (18)) whose associated *sharp null hypothesis* is not rejected.

When using test statistics that do not control for the pre-intervention fit such as the test statistic θ, Abadie et al. [4, 3] and our results in subsection 5.1 suggest that we should include only regions with a good pre-intervention fit (i. e., regions whose pre-intervention MSPE is at most five times greater than the Basque Country's pre-intervention MSPE) because placebo studies for those regions are not informative about the relative rarity of the post-intervention effect for the Basque Country.[27] By doing so, we exclude the regions of Madrid, Extremadura and Balearic Islands when computing the p-value of the hypothesis test (i. e., our sample size is now equal to 14) and find that $p = 2/14$, a reasonable value to reject the *null hypothesis of no effect whatsoever* given the small sample size.[28] The remaining placebos effects are plotted in subfigure 1(b), that visually suggest a negative impact of ETA's terrorism on the Basque Country's GDP.

Now, we evaluate the robustness of our findings to deviations from the equal weights benchmark using the sensitivity analysis mechanism proposed in section 3. We may conclude that, due to the small sample size of the exercise, rejecting the *null hypothesis of no effect whatsoever* is not a very robust conclusion because we must impose a sensitivity parameter of only $\underline{\phi} = 0.495$ in order to stop rejecting it at the $3/14$-significance level, implying the Basque Country has a weight only 64 % times larger than the units with $v_j = 0$. Moreover, we note that the permutation test's p-value increases fast as a function of the sensitivity parameter $\phi \in \mathbb{R}_+$ as subfigure 2(a) shows.[29]

We, now, estimate two one-sided $12/14$-Confidence Sets.[30] While subfigure 3(a) considers a Constant in Time Intervention Effect following equation (16), subfigure 3(b) considers a Linear in Time Intervention Effect whose intercept is equal to zero following equation (18). Both one-sided confidence sets are based on the test statistic $\theta = -\frac{\bar{a}_{post}/(T-T_0)}{\hat{\sigma}/\sqrt{T-T_0}}$ and consider only units with a good pre-intervention fit.

The dashed lines are the upper bounds of the one-sided confidence sets based on the equal weights benchmark given by equation (5). These upper bounds not only quickly show that we reject the null hypoth-

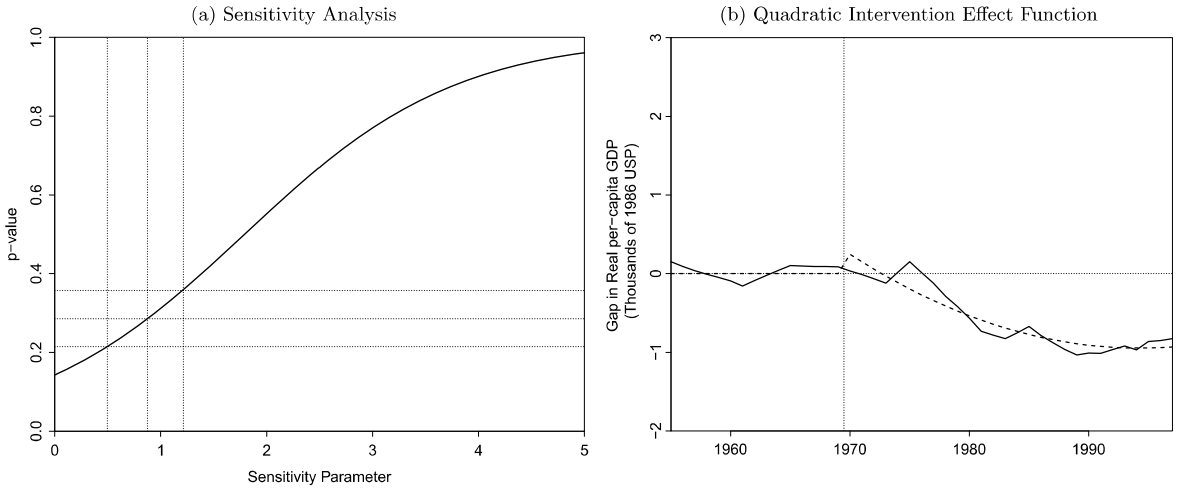

Figure 2: Sensitivity Analysis and Quadratic Intervention Effect. *Note:* In the left panel, the black line denotes the estimated p-value for each value of the sensitivity parameter $\phi \in \mathbb{R}_+$ using only units with a good pre-intervention fit, while the horizontal dotted lines denote the p-values of $3/14$, $4/14$ and $5/14$. In the right panel, the black line show the estimated impact of ETA's terrorism on the Basque Country's economy, while the dashed line shows the quadratic function that best approximates this effect.

27 We thank an anonymous referee for stressing this point.

28 If we impose the usual significance level of 10 %, we would only reject the null hypothesis when the observed test statistic is the most extreme one, a criterion that seems to be too conservative for the problem at hand.

29 We highlight that, according to subsection 5.2, a sensitivity parameter $\underline{\phi}$ smaller than 0.995 may be considered small.

30 Since we need at least 20 regions in order to estimate a 90 %-Confidence Set, we use a confidence level that is close 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.

esis of no effect whatsoever (because the upper bounds of the confidence sets are below the zero function), but also show that the economic impact of ETA's terrorism is far away from zero, suggesting economically relevant negative effects.[31]

When we apply our sensitivity analysis mechanism to these one-sided confidence sets, we impose a parameter $\underline{\phi} = 0.495$ for the worst case scenario described in section 3 and find the upper bounds denoted by the dotted lines. Observe that we need to impose a small sensitivity analysis parameter in order to include not only the *null hypothesis of no effect whatsoever*, but also many positive treatment effect functions in the linear confidence set (equation (18)) of subfigure (3)(b). Again, this exercise illustrates that, due to the small sample size, the test result is not very robust to small deviations of the equal weight benchmark given by (5).

Moreover, note also that these conclusions are robust to the choice of functional form for the intervention effect (constant or linear). Finally, observe that, due to their ability to summarize a large amount of information, our preferred confidence sets (equations (16) and (18)) are useful to the empirical researcher even being only subsets of the general confidence set (equation (14)), particularly because they can also be combined with the sensitivity analysis mechanism proposed in section 3.

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. Figure 2(b) shows this fitted quadratic function. Applying the inference procedure described in section 4.1 and using the test statistic $\theta = -\frac{\bar{\alpha}_{post}/(T-T_0)}{\hat{\sigma}/\sqrt{T-T_0}}$ in a one-sided test that includes only units with a good pre-intervention fit, we do not reject the null hypothesis that the true intervention effect follows this quadratic function because $p_{quadratic} = {}^{6}/_{14}$. In this case, we must impose a sensitivity parameter of $\overline{\phi} = 1.905$ in order to reject it at the 10 %-significance level, implying that the units with $\upsilon_j = 1$ have a weight that is more than five times larger than the weight for the Basque Country. Consequently, not rejection this quadratic treatment effect function is a robust result, suggesting that the economic impact of ETA's terrorism on the Basque Country is initially negative, but attenuates toward zero in the long run.

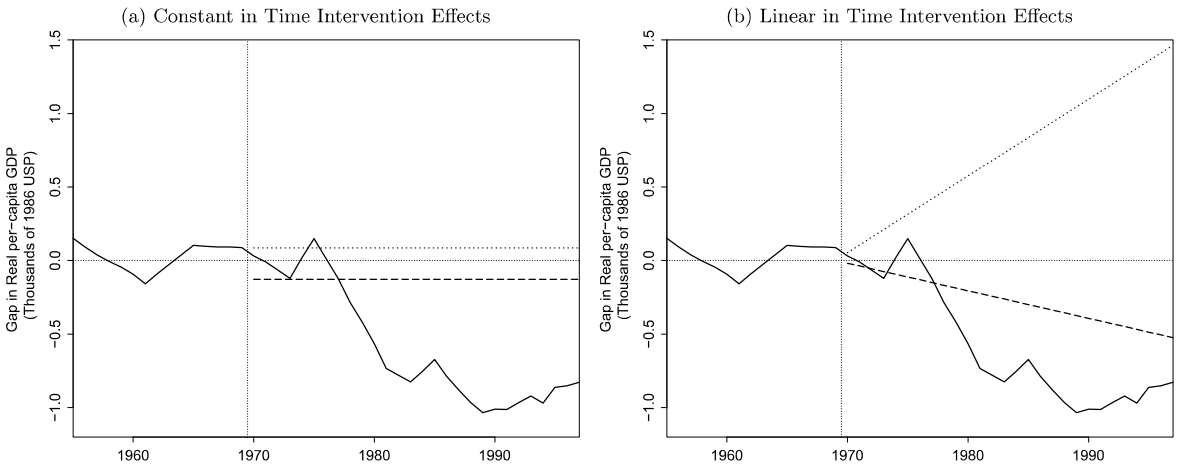

Figure 3: One-sided ${}^{12}/_{14}$-Confidence Sets for the Intervention Effect. *Note:* The solid black lines show the estimated impact of ETA's terrorism on the Basque Country's economy while the dashed lines show the upper-bounds of the one-sided ${}^{12}/_{14}$-Confidence Sets for Constant in Time or Linear in Time (with intercept equal to zero) Intervention Effects that were constructed using the test statistic $\theta = -\frac{\bar{\alpha}_{post}/(T-T_0)}{\hat{\sigma}/\sqrt{T-T_0}}$ and using the equal weights benchmark (equation (5)). The dotted lines are the upper bounds of the one-sided ${}^{12}/_{14}$-Confidence Sets that were constructed using the test statistic θ and imposing a sensitivity parameter $\underline{\phi} = 0.495$ for the worst case scenario described in section 3.

31 Note that, if the estimated upper bounds were close to zero, we could informally argue that the analyzed intervention effect is economically irrelevant.

Differently from what we did in the last paragraphs, we can treat each year as a different outcome variable and apply the inference procedure described in subsection 6.1. This interpretation allow us to analyze the timing of the economic impact of ETA's terrorism, which may be more significant for some time periods. We use the negative of the value of the estimated intervention effect for each year of the post-intervention period as a one-sided test statistic. Using the notation of subsection 6.1, we have that $\theta_{f_m}^{obs} = \theta^m(e_1, \tau, \mathbf{Y}^m, \mathbf{X}, f_m) = -\widehat{\alpha}_{1,m}$, where $m \in \{1970, ..., 1997\}$ is a year of the post-intervention period. Applying the procedure described in subsection 6.1, we find p-values that are equal to $^4/_{14}$ in the 80's, equal to $^6/_{14}$ in the late 70's and early 90's, and greater than $^{10}/_{14}$ in the other years. In order to apply the sensitivity analysis mechanism proposed in subsection 6.1 for the case with Multiple Outcome Variables, we choose one vector $\boldsymbol{v}_m$ for each outcome $m \in \{1, ..., M\}$ based on the test statistic $p^j_{\theta_{f_m},*}$ that is used to compute the p-value described in equation (24). We, then, use the p-value $p^{fwer}_{\theta^{obs}_{f_m}}(\phi, \boldsymbol{v})$ to determine one sensitivity parameter $\overline{\phi}_m$ for each outcome $m \in \{1, ..., M\}$. We find sensitivity parameters that are equal to 1.285 in the 80's, equal to 1.905 in the late 70's and early 90's, and greater than 3.115 in the other years. Unfortunately, the loss of power — due to the use of the *familywise error rate control* approach instead of the *summary index text* as pointed by Anderson [50] — prevents any strong conclusion. However, the *familywise error rate control* approach provides a straightforward numerical interpretation of subfigure 1(b) by suggesting that not rejecting the *null hypothesis of no effect whatsoever* in the 80's is not an extremely robust result in line with a visually extreme (on the negative side) gap during this decade. Moreover, it suggests that, even in the presence of possible negative impacts on the 80's, the Basque economy recovered in the late 90's since, for these years, not rejecting the *null hypothesis of no effect whatsoever* is a robust result. Therefore, the results using the multiple outcome framework are in line with the results for the quadratic trend.

As a consequence of all our empirical exercises and keeping in mind the underpowered context due to the small sample size, we conclude that ETA's terrorists acts had negative and marginally significant, although not robust nor permanent in the long run, impacts on the Basque economy, in line with the conclusion by Abadie et al. [3, p. 15] that "there is a very low probability of obtaining a gap as large as the one obtained for the Basque region". 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 and Spanish peoples.

8 Conclusion

In this article, we contribute to the theoretical literature on SCM by extending the inference procedure proposed by Abadie et al. [4] and [1] in two ways. First, we make the equal weights benchmark p-value proposed by Abadie et al. [1] more flexible by using parametric weights that allow the researcher to implement a sensitivity analysis mechanism similar to the one suggested by Rosenbaum [47] and Cattaneo et al. [48]. By analyzing the sensitivity analysis parameter that changes the test's result, we can gauge the robustness of a conclusion to continuous deviations from the equal weights benchmark.

Second, we extend the test proposed by Abadie et al. [4] and [1] to test any *sharp null hypothesis*, including, as a particular case, the usual *null hypothesis of no effect whatsoever* studied by these authors. 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. Moreover, based on this extension and on procedures described by Imbens and Rubin [46] and Rosenbaum [47], we invert the test statistic to estimate confidence sets. 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. As a benchmark, they are useful to the applied researcher because they represent a graphical device that summarizes a large amount of information, illustrating the statistical significance of the intervention effect, the precision of a point-estimate and the robustness of a test. 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.

We also stress that these two tools can use not only the *RMSPE* test statistic proposed by Abadie et al. [1], but any test statistic. For this reason, we analyze, using a Monte Carlo experiment, the size, power and robustness of five different test statistics that are applied to hypothesis testing in the empirical literature about the SCM. In this simulation, we find that test statistics designed for the SCM perform much better than its competitors when there is only one region that faces the intervention. In particular, the traditional *RMSPE* statistic has good properties with respect to power and the sensitivity analysis mechanism.

Furthermore, we extend our new tools to contexts that differ from the ones analyzed by Abadie and Gardeazabal [2], Abadie et al. [4] and [1] in important dimensions: testing a null hypothesis about a pooled effect among few treated units and simultaneously testing null hypotheses for different outcome variables. These extensions allows researchers to investigate more complex questions such as interventions that have impact in more than one country or in more than one variable, such as policy reforms. In particularly, we can also interpret each post-intervention time period as a different outcome variable, allowing us to analyze short and long term effects.

Finally, in order to show the usefulness of our new tools, we reevaluate the economic impact of ETA's terrorism in the Basque Country, analyzed by Abadie and Gardeazabal [2] and Abadie et al. [3]. By testing a quadratic treatment effect function and combining our sensitivity analysis mechanism and a multiple outcomes framework, we find a negative and marginally significant effect in the 80's, that attenuates in long run, since, in the late 90's, not rejecting the *null hypothesis of no effect whatsoever* is a robust result. Furthermore, 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 and the robustness of the test's conclusion. We stress that this knowledge is an important measure of the strength of qualitative conclusions.

Acknowledgment: We are in debt to the editor and two anonymous referees for useful suggestions. We also thank 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, EPGE-FGV, Yale, the California Econometrics Conference 2015, the 37th Brazilian Meeting of Econometrics, the 2016 Latin American Workshop in Econometrics, the 2017 North American Winter Meeting of the Econometric Society, and the International Workshop on 'Causal Inference, Program Evaluation, and External Validity' for comments. All errors are our own.

Funding: We are grateful to FAPESP that provided financial aid through grant number 2014/23731-3.

Code and Datasets: The author(s) published code and data associated with this article is on Code Ocean, a computational reproducibility platform. We recommend Code Ocean to JCI contributors who wish share, discover, and run code in published research articles. (See: https://doi.org/10.24433/CO.23bd238f-38c5-4b3e-82f4-3a1624fd8a33).

References

1. Abadie A, Diamond A, Hainmueller J. Comparative politics and the synthetic control method. Am J Polit Sci. 2015;59(2):495–510.
2. Abadie A, Gardeazabal J. The economic costs of conflict: A case study of the basque country. Am Econ Rev. 2003;93(1):113–32.
3. Abadie A, Diamond A, Hainmueller J. Synth: An R package for synthetic control methods in comparative case studies. J Stat Softw. 2011;42(13):1–17.
4. Abadie A, Diamond A, Hainmueller J. Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. J Am Stat Assoc. 2010;105(490):493–505.
5. Bove V, Elia L, Smith RP. The relationship between panel and synthetic control estimators on the effect of civil war. Working paper. 2014. Available at http://www.bbk.ac.uk/ems/research/BirkCAM/working-papers/BCAM1406.pdf.
6. Li Q. Economics consequences of civil wars in the post-world war II period. Macrotheme Rev. 2012;1(1):50–60.

7. Montalvo JG. Voting after the bombings: A natural experiment on the effect of terrorist attacks on democratic elections. Rev Econ Stat. 2011;93(4):1146–54.
8. Yu J, Wang C. Political risk and economic development: A case study of China. Econ Res [Eknomska Istrazianja]. 2013;26(2):35–50.
9. Barone G, Mocetti S. Natural disasters, growth and institutions: A tale of two earthquakes. J Urban Econ. 2014;52–66.
10. Cavallo E, Galiani S, Noy I, Pantano J. Catastrophic natural disasters and economic growth. Rev Econ Stat. 2013;95(5):1549–61.
11. Coffman M, Noy I. Hurricane iniki: Measuring the long-term economic impact of natural disaster using synthetic control. Environ Dev Econ. 2011;17:187–205.
12. DuPont W, Noy I. What happened to Kobe? A reassessment of the impact of the 1995 earthquake in Japan. Econ Dev Cult Change. 2015;63(4):777–812. Available at http://www.economics.hawaii.edu/research/workingpapers/WP_12-4.pdf.
13. Mideksa TK. The economic impact of natural resources. J Environ Econ Manag. 2013;65:277–89.
14. Sills EO, Herrera D, Kirkpatrick AJ, Brandao A, Dickson R, Hall S, Pattanayak S, Shoch D, Vedoveto M, Young L, Pfaff A. Estimating the impact of a local policy innovation: The synthetic control method applied to tropica desforestation. PLoS One. 2015;10(7):e0132590.
15. Smith B. The resource curse exorcised: Evidence from a panel of countries. J Dev Econ. 2015;116:57–73.
16. Jinjarak Y, Noy I, Zheng H. Capital controls in Brazil — Stemming a tide with a signal? J Bank Finance. 2013;37:2938–52.
17. Sanso-Navarro M. The effects on American foreign direct investment in the United Kingdom from not adopting the euro. J Common Mark Stud. 2011;49(2):463–83.
18. Belot M, Vandenberghe V. Evaluating the threat effects of grade repetition: Exploiting the 2001 reform by the French-speaking community of Belgium. Educ Econ. 2014;22(1):73–89.
19. Chan HF, Frey BS, Gallus J, Torgler B. Academic honors and performance. Labour Econ. 2014;31:188–204.
20. Hinrichs P. The effects of affirmative action bans on college enrollment, educational attainment, and the demographic composition of universities. Rev Econ Stat. 2012;94(3):712–22.
21. Bauhoff S. The effect of school nutrition policies on dietary intake and overweight: A synthetic control approach. Econ Human Biol. 2014;45–55.
22. Kreif N, Grieve R, Hangartner D, Turner AJ, Nikolova S, Sutton M. Examination of the synthetic control method for evaluating health policies with multiple treated units. Health Econ. 2016;25(12):1514–28.
23. Billmeier A, Nannicini T. Assessing economic liberalization episodes: A synthetic control approach. Rev Econ Stat. 2013;95(3):983–1001.
24. Gathani S, Santini M, Stoelinga D. Innovative techniques to evaluate the impacts of private sector developments reforms: An application to Rwanda and 11 other countries. Working paper. 2013. Available at https://blogs.worldbank.org/impactevaluations/files/impactevaluations/methods_for_impact_evaluations_feb06-final.pdf.
25. Hosny AS. Algeria's trade with GAFTA countries: A synthetic control approach. Transit Stud Rev. 2012;19:35–42.
26. Billmeier A, Nannicini T. Trade openness and growth: Pursuing empirical glasnost. IMF Staff Pap. 2009;56(3):447–75.
27. Carrasco V, de Mello JMP, Duarte I. A Década Perdida: 2003–2012. Texto para Discussão. 2014. Available at http://www.econ.puc-rio.br/uploads/adm/trabalhos/files/td626.pdf.
28. Dhungana S. Identifying and evaluating large scale policy interventions: What questions can we answer? 2011. Available at https://openknowledge.worldbank.org/bitstream/handle/10986/3688/WPS5918.pdf?sequence=1.
29. Jales H, Ribeiro F, Stein G, Kang T. Measuring the role of the 1959 revolution on Cuba economic performance. 2013. Available at https://drive.google.com/file/d/0B-Z_Rf2gRVJzRXozekhMNGpUVUU/view.
30. Bohn S, Lofstrom M, Raphael S. Did the 2007 legal Arizona workers act reduce the state's unauthorized immigrant population? Rev Econ Stat. 2014;96(2):258–69.
31. Calderon G. The effects of child care provision in Mexico. Working paper. 2014. Available at http://goo.gl/YSEs9B.
32. Kleven HJ, Landais C, Saez E. Taxation and international migration of superstars: Evidence from European football market. Am Econ Rev. 2013;103(5):1892–924.
33. de Souza FFA. Tax evasion and inflation: Evidence from the nota fiscal paulista program. Master's thesis. Pontifícia Universidade Católica. 2014. Available at http://www.dbd.puc-rio.br/pergamum/tesesabertas/1212327_2014_completo.pdf.
34. Pinotti P. The economic costs of organized crime: Evidence from Southern Italy. Econ J. 2015;125:203–32.
35. Pinotti P. Lessons from the economics of crime: What reduces offending? The MIT Press. Chapter: Organized crime, violence and the quality of politicians: Evidence from Southern Italy. 2013. Available at http://dx.doi.org/10.2139/ssrn.2144121.
36. Saunders J, Lundberg R, Braga AA, Ridgeway G, Miles J. A synthetic control approach to evaluating place-based crime interventions. J Quant Criminol. 2015;31(3):413–34.
37. Acemoglu D, Johnson S, Kermani A, Kwak J, Mitton T. The value of connections in turbulent times: Evidence from the United States. J Financ Econ. 2013;121(2):368–91.
38. Ando M. Dreams of urbanization: Quantitative case studies on the local impacts of nuclear power facilities using the synthetic control method. J Urban Econ. 2015;85:68–85.

39. Gobillon L, Magnac T. Regional policy evaluation: Interative fixed effects and synthetic controls. Rev Econ Stat. 2016;98(3):535–51.
40. Kirkpatrick AJ, Bennear LS. Promoting clean enery investment: An empirical analysis of property assessed clean energy. J Environ Econ Manag. 2014;68:357–75.
41. Liu S. Spillovers from universities: Evidence from the land-grant program. J Urban Econ. 2015;87:25–41.
42. Possebom V. Free trade zone of manaus: An impact evaluation using the sythetic control method. Rev Brasil Econ. 2017;71(2):217–31.
43. Severnini ER. The power of hydroelectric dams: Agglomeration spillovers. IZA discussion paper, No 8082. Available at http://ftp.iza.org/dp8082.pdf.
44. Athey S, Imbens GW. The state of applied econometrics: Causality and policy evaluation. J Econ Perspect. 2017;31(2):3–32.
45. Fisher RA. The design of experiments. 8th ed. Hafner Publishing Company, United States; 1971.
46. Imbens GW, Rubin DB. Causal inference for statistics, social and biomedical sciences: An introduction, 1st edn. United Kingdom: Cambridge University Press; 2015.
47. Rosenbaum PR. Observational studies. 2nd ed. New York: Springler Science + Business Media; 2002.
48. Cattaneo M, Titiunik R, Vazquez-Bare G. Inference in regression discontinuity designs under local randomization. Stata J. 2016;16(2):331–67.
49. Conley TG, Taber CR. Inference with difference-in-differences with a small number of policy changes. Rev Econ Stat. 2011;93(1):113–25.
50. Anderson ML. Multiple inference and gender differences in the effects of early intervention: A reevaluation of the Abecedarian, Perry preschool and early training projects. J Am Stat Assoc. 2008;103(484):1481–95.
51. Kaul A, Klöbner S, Pfeifer G, Schieler M. Synthetic control methods: Never use all pre-intervention outcomes as economic predictors. Working paper. 2015. Available at http://www.oekonometrie.uni-saarland.de/papers/SCM_Predictors.pdf.
52. Ando M, Sävje F. Hypothesis testing with the synthetic control method. 2013. Working Paper. Available at http://www.eea-esem.com/files/papers/eea-esem/2013/2549/scm.pdf.
53. Wong L. Three essays in causal inference. PhD thesis. Stanford University; 2015.
54. Dube A, Zipperer B. Pooling multiple case studies using synthetic controls: An application to minimum wage policies. Working paper. 2015. Available at http://ftp.iza.org/dp8944.pdf.
55. Carvalho CV, Mansini R, Medeiros MC. ArCo: An artificial counterfactual approach for aggregate data. Working paper. 2017. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2823687.
56. Hahn J, Shi R. Synthetic control and inference. Econometrics. 2017;5(4):52.
57. Ferman B, Pinto C. Revisiting the synthetic control estimator. 2017. Available at https://dl.dropboxusercontent.com/u/12654869/Ferman%20and%20Pinto%20-%20revisiting%20the%20SC.pdf.
58. Ferman B, Pinto C. Placebo tests for synthetic controls. 2017. Available at https://dl.dropboxusercontent.com/u/12654869/Ferman%20and%20Pinto%20-%20placebo%20tests%20for%20SC.pdf.
59. Ferman B, Pinto C, Possebom V. Cherry picking with synthetic controls. 2017. Available at https://dl.dropboxusercontent.com/u/12654869/FPP%20-%20Cherry%20Picking.pdf.
60. Rosenbaum PR. Sensitivity analysis for certain permutation inferences in matched observational studies. Biometrika. 1987;74(1):13–26.
61. Rosenbaum PR. Sensitivity analysis for matching with multple controls. Biometrika. 1988;75(3):577–81.
62. Rosenbaum PR, Krieger AM. Sensitivity of two-sample permutation inferences in observational studies. J Am Stat Assoc. 1990;85(410):493–8.
63. Rosenbaum PR. Sensitivity analysis for m-estimates, tests, and confidence intervals in matched observational studies. Biometrics. 2007;63:456–64.
64. Rosenbaum PR, Silber JH. Amplification of sensitivity analysis in matched observational studies. J Am Stat Assoc. 2009;104(488):1398–405.
65. Yates F. Tests of significance for 2 x 2 contingency tables. J R Stat Soc A. 1984;147(3):426–63.
66. Lehmann E. Testing statistical hypotheses. New York: John Wiley & Sons; 1959.
67. Ibragimov R, Muller UK. T-statistic based correlation and heterogeneity robust inference. J Bus Econ Stat. 2010;28(4):453–68.
68. Young A. Channeling Fisher: Randomization tests and the statistical insignificance of seemingly significant eperimental results. 2016. Available at http://economics.mit.edu/files/11362.
69. Eudey TL, Kerr J, Trumbo B. Using R to simulate permutation distributions for some elementary experimental designs. J Stat Educ. 2010;18(1).
70. List J, Shaikh AM, Xu Y. Multiple hypothesis testing in experimental economics. NBER working paper 21875. 2016. Available at http://www.nber.org/papers/w21875.

Supplemental Material: The online version of this article offers supplementary material (https://doi.org/10.1515/jci-2016-0026).