Synthetic Control Method: Inference, Sensitivity Analysis and Confidence Sets*

Sergio Firpo† Vitor Possebom‡
Insper Yale University

September 1st, 2017

Abstract

We extend the inference procedure for the synthetic control method in two ways. First, we impose a parametric form to the treatment assignment probabilities that includes, as a particular case, the uniform distribution that is assumed by Abadie et al. (2015). By changing the value of this parameter, we can analyze the sensitiveness of the test's decision to deviations from the usual uniform distribution assumption. 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. (2015). 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 & Gardeazabal (2003) and Abadie et al. (2011).

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

JEL Codes: C21, C23, C33

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

†Corresponding author. Address: Insper, Rua Quata 300, Sao Paulo, SP, 04645-042, Brazil. E-mail: firpo@insper.edu.br.

‡E-mail: vitoraugusto.possebom@yale.edu.

1 Introduction

The Synthetic Control Method (SCM) was proposed by Abadie & Gardeazabal (2003), Abadie et al. (2010) and Abadie et al. (2015) to address counterfactual questions involving only one treated unit and a few control units. Intuitively, this method constructs a weighted average of control units that is as similar as possible to the treated unit regarding the pretreatment 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 inference procedure for small samples using the synthetic control estimator was developed by Abadie et al. (2010) and Abadie et al. (2015). Using placebo tests similar to Fisher's Exact Hypothesis Testing Procedure described by Fisher (1971), Imbens & Rubin (2015) and Rosenbaum (2002), 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 p-value proposed by Abadie et al. (2015) implicitly assumes that the treatment assignment probabilities are the same across all units, i.e., any observed region has the same probability of facing the analyzed intervention. By making this assumption clear, we can address the robustness of the decision based on the hypothesis test to alternative treatment assignment distributions by applying a sensitivity analysis mechanism similar to the one proposed by Rosenbaum (2002) and Cattaneo et al. (2016). We impose a parametric form to the treatment assignment probabilities that allows the empirical researcher to compute p-values for different assumptions regarding these probabilities. As we change the value

[1] This tool was applied to an extremely diverse set of topics, including, for instance, issues related to terrorism, civil wars and political risk (Abadie & Gardeazabal (2003), Bove et al. (2014), Li (2012), Montalvo (2011), Yu & Wang (2013)), natural resources and disasters (Barone & Mocetti (2014), Cavallo et al. (2013), Coffman & Noy (2011), DuPont & Noy (2012), Mideksa (2013), Sills et al. (2015), Smith (2015)), international finance (Jinjarak et al. (2013), Sanso-Navarro (2011)), education and research policy (Belot & Vandenberghe (2014), Chan et al. (2014), Hinrichs (2012)), health policy (Bauhoff (2014), Kreif et al. (2015)), economic and trade liberalization (Billmeier & Nannicini (2013), Gathani et al. (2013), Hosny (2012)), political reforms (Billmeier & Nannicini (2009), Carrasco et al. (2014), Dhungana (2011) Ribeiro et al. (2013)), labor (Bohn et al. (2014), Calderon (2014)), taxation (Kleven et al. (2013), de Souza (2014)), crime (Pinotti (2012*b*), Pinotti (2012*a*), Saunders et al. (2014)), social connections (Acemoglu et al. (2013)), and local development (Ando (2015), Gobillon & Magnac (2016), Kirkpatrick & Bennear (2014), Liu (2015), Possebom (2017) and Severnini (2014)).

of the parameter that affects the decision based on the hypothesis test, we can gauge the sensitiveness of the decision made to the usual assumption of an uniform distribution of treatment assignment probabilities. 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. (2010) and Abadie et al. (2015) 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 & Rubin (2015) and Rosenbaum (2002) to calculate confidence intervals based on Fisher's Exact Hypothesis Testing Procedure and apply it to SCM. 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. Although the assumptions that make them valid are strong, 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. (2010) and Abadie et al. (2015) to use many different test statistics beyond the already traditional Ratio of the Mean Squared Prediction Errors (*RMPSE*) test statistic, discussed in those papers. We run a Monte Carlo experiment and present results on size, power and our robustness analysis of five test statistics applied to 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 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 & Taber (2011). We find that the inference procedure based on the original test statistic proposed by Abadie et al. (2015), *RMPSE*, performs much better than alternative test statistics when we compare their size, power and in our sensitivity analysis.

We also show how to apply our new tools to contexts that differ from the ones analyzed by Abadie & Gardeazabal (2003), Abadie et al. (2010) and Abadie et al. (2015) 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. (2013), 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 (2008) to SCM, is important, for example, to evaluate political reforms (Billmeier & Nannicini (2009), Billmeier & Nannicini (2013), Carrasco et al. (2014), Jinjarak et al. (2013), Sanso-Navarro (2011)) that generally affect multiple outcomes variables, such as income levels and investment 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. (2010) and Abadie et al. (2015), 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 estimated by Abadie & Gardeazabal (2003) and Abadie et al. (2011). 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. Our sensitivity analysis mechanism also shows that Abadie & Gardeazabal (2003) and Abadie et al. (2011) reach very robust conclusions about the economic impact of ETA's terrorism on the Basque Country.

Literature Review

Regarding the inference of the Synthetic Control Method, other authors have surely made important previous contributions. Abadie et al. (2010)[2] are the first authors to propose a inference procedure that consists in estimating p-values through permutation tests and Abadie et al. (2015) suggest a different test statistic for the same procedure. Ando & Sävje (2013) propose two new test statistics that have adequate size and more power when applied to the above mentioned hypothesis test than the ones proposed by Abadie et al. (2010) and Abadie et al. (2015).

Bauhoff (2014), Calderon (2014) and Severnini (2014) propose a way to apply SCM to many treated and control units that is similar to a matching estimator for panel data. Following a similar approach, Wong (2015) extends the synthetic control estimator to a cross-sectional setting where individual-level data is available and derives its asymptotic distribution when the number of observed individuals goes to infinity. Wong (2015) also explores the synthetic control estimator when panel data (or repeated cross-sections) are available in two levels: an aggregate level (regions), where treatment is assigned, and an individual level, where outcomes are observed. In this framework, he derives the asymptotic distribution of the synthetic control estimator when the number of individuals in each region goes to infinity.

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

Finally, Acemoglu et al. (2013), Cavallo et al. (2013) and Dube & Zipperer (2013) 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. (2010) and Abadie et al. (2015).

Gobillon & Magnac (2016), also working in a context with more than one treated unit, propose a way to compute confidence intervals for policy effect function based on the bootstrap. 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. (2010) and Abadie et al. (2015) allows the empirical researcher to test any function of time as the intervention effect.

Moreover, Carvalho et al. (2015) propose the Artificial Counterfactual Estimator (ArCo), that is similar in purpose to 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. (2015)'s method.[3]

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

Finally, the sensitivity analysis literature in a context of observational studies is vast. For

[3] Wong (2015) and Hahn & Shi (2016) also conduct an asymptotic analysis when the pre-intervention period goes to infinity. Ferman & Pinto (2017*b*), Ferman & Pinto (2017*a*) and Ferman et al. (2017) discuss asymptotic biases, size distortions and specification-search possibilities within SCM.

example, Rosenbaum (1987), Rosenbaum (1988), Rosenbaum & Krieger (1990), Rosenbaum (2002), Rosenbaum (2007), and Rosenbaum & Silber (2009) made important contributions to this field, particularly with respect to matching estimators. Cattaneo et al. (2016) 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 synthetic control estimator.

This paper is divided as follows. Section 2 formally presents SCM as proposed by Abadie & Gardeazabal (2003), Abadie et al. (2010) and Abadie et al. (2015). Section 3 proposes a sensitivity analysis related to the assumption of a uniform distribution of treatment assignment probabilities. 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 on the Basque Country used by Abadie et al. (2011) 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 its inference procedure based on permutation tests. The ideas and notation that are used in the next two subsections are mostly based on Abadie et al. (2010) and Abadie et al. (2015). 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[4] during $T \in \mathbb{N}$ time periods. Additionally, assume that there is an intervention (policy)[5] that affects only region 1[6] from period $T_0 + 1$ to period T uninterruptedly[7], where $T_0 \in (1, T) \cap \mathbb{N}$. Let the scalar $Y^N_{j,t}$ be the potential outcome that would be observed for region j in period t if there were no intervention for $j \in \{1, ..., J+1\}$ and $t \in \{1, ..., T\}$. Let the scalar $Y^I_{j,t}$ be the potential outcome that would be observed for region j in period t if region j faced the intervention at period t. Define

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

as the **intervention (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 $\mathbf{Y_j}$.[8] 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.

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

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

[6] In subsection 6.2, we extend this framework to include the case when multiple units face the same or a similar intervention.

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

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

Since we want to make region 1's synthetic control as similar as possible to the actual region 1, SCM produces, for each $t \in \{1, ..., T\}$, $\widehat{Y}_{1,t}^{N} := \sum_{j=2}^{J+1} \widehat{w}_j Y_{j,t}$, which is an estimator of $Y_{1,t}^{N}$. 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:

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

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 SCM.[9]

Finally, we define the synthetic control estimator of $\alpha_{1,t}$ (or the estimated gap) as $\widehat{\alpha}_{1,t} :=$

[9] Abadie & Gardeazabal (2003), Abadie et al. (2010) and Abadie et al. (2015) propose two other ways to choose $\widehat{\mathbf{V}}$. The first and most simple one is to use subjective and previous knowledge about the relative importance of each predictor. Since one of the advantages of 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 of a longer pre-intervention period. Moreover, the Stata command made available by those authors also allows the researcher to use a regression-based method in order to compute matrix $\widehat{\mathbf{V}}$. It basically regress matrix $\mathbf{Y_1}$ on $\mathbf{X_1}$ and imposes $v_k = {}^{|\beta_k|}/\left(\sum_{k'=1}^{K} |\beta_{k'}|\right)$, where v_k is the k-th diagonal element of matrix $\mathbf{V}$ and β_k is the k-th coefficient of the regression of $\mathbf{Y_1}$ on $\mathbf{X_1}$. The choice method that we have chosen to present in the main text is the most used one in the empirical literature.

$Y_{1,t} - \widehat{Y}_{1,t}^N$ for each $t \in \{1, ..., T\}$.

2.2 Hypothesis Testing

Abadie et al. (2010) and Abadie et al. (2015) develop a small sample inference procedure for SCM that is similar to Fisher's Exact Hypothesis Test described by Fisher (1971), Imbens & Rubin (2015) and Rosenbaum (2002). Abadie et al. (2010) 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.*[10]

Abadie et al. (2015) 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, 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, such as the traditional value of $\gamma = 0.1$.

As Abadie et al. (2010) and Abadie et al. (2015) make clear, this inference procedure

[10] As an anonymous referee pointed out, Abadie et al. (2010) and Abadie et al. (2015) exclude the treated unit from the donor pool of the placebo runs in order to avoid bias when there is a treatment effect. Here, we follow Imbens & Rubin (2015) and keep the treated unit in the donor pool of the placebo runs, because, under *sharp null hypotheses*, the treatment effect is known. In our empirical application, we show that our conclusions do not change whether we follow the suggestions by Imbens & Rubin (2015) or by Abadie et al. (2010) and Abadie et al. (2015).

is valid under three assumptions. First, we need to impose the *stable unit treatment value assumption* (SUTVA):

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

Second, we need to impose some form of random assignment to the treatment:

Assumption 2 *The choice of which unit will be treated (i.e., which region is our region 1) is random* conditional on the choice of the donor pool, the observable variables included as predictors and the unobservable variables captured by the path of the outcome variable during the pre-intervention period.[11]

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

Since the differences-in-differences estimator controls only for the unobservable variables that are common among all the observed units (but varies over time) while the synthetic control estimator controls for unobservable variables captured by the path of the outcome variable during the pre-intervention period, assumption 2 is weaker than assumption DID.[12]

Regarding the applicability of assumption 2, it holds true for many empirical applications of the synthetic control estimator. For example, Barone & Mocetti (2014), Cavallo et al. (2013), Coffman & Noy (2011) and DuPont & Noy (2012) evaluate the economic effect of

[11] The unobservable variables captured by the path of the outcome variable during the pre-intervention period are denoted by unobserved common factors, $\boldsymbol{\lambda}_t$, and unknown factor loadings, $\boldsymbol{\mu}_i$ in the factor model discussed by Abadie et al. (2010).

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

large scale natural disasters, such as earthquakes, hurricanes or volcano eruptions. Although the regions in the globe that frequently faces these disasters are not random, the specific region among them that will be hit by a natural disaster and the timing of this phenomenon are fully random.[13] Moreover, Pinotti (2012*b*) and Pinotti (2012*a*) evaluate the economic and political cost of organized crime in Italy exploring the increase in Mafia activities after two large earthquakes. Two other examples of the plausibility of assumption 2 are Smith (2015), who argues that the discovery of large natural resources reserves is *as-if-random*, and Liu (2015), who argues that the location of land-grand universities in the 19th century is *as-if-random* too.[14]

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

Assumption 3 *The potential outcomes* $\mathbf{Y}_j^I := \left[Y_{j,1}^I ... Y_{j,T}^I\right]'$ *and* $\mathbf{Y}_j^N := \left[Y_{j,1}^N ... Y_{j,T}^N\right]'$ *for each region* $j \in \{1, ..., J+1\}$ *are fixed but* a priori *unknown quantities.*[15]

Implicitly, we assume that we observe the *realization* of a random variable for the *entire population of interest* instead of a random sample of a larger superpopulation.[16]

We note that assumptions 2 and 3 implies that the units in the donor pool are *exchangeable*. In reality, *exchangeability* is the weakest assumption that guarantees the validity of this inference procedure, because it is simply based in a permutation test.[17] However, we believe that, although stronger, assumptions 2 and 3 makes interpretation easier, providing a useful framework in order to discuss the validity of the synthetic control method in applied topics.

Finally, Abadie et al. (2010) and Abadie et al. (2015) propose to test the the *exact null*

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

[14] Even in randomized control trials, the synthetic control method may be more interesting than traditional statistical methods when there are only a few treated units — an issue that may emerge due to budget constraints. As we show in section 5, test statistics that use the synthetic control estimator are more powerful than the ones that do not use it.

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

[16] See Imbens & Rubin (2015) for details regarding this interpretation.

[17] Hahn & Shi (2016) discuss permutation tests within the framework of SCM when this assumption is violated. By implementing a Monte Carlo Experiment, they show that the test size is distorted when the data generating process is based on sampling from a larger super-population, which creates a non-negligible stochastic component to the time seires dimension. For this reason, they propose an inference procedure that is based on large T asymptotic analysis.

hypothesis,

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

that is the precise definition of the *null hypothesis of no effect whatsoever*.[18] Note that rejecting the null hypothesis implies that there is some region with a non-zero effect for some time period. As mentioned before, under assumptions 1-3 and the null hypothesis (6), the p-value in equation (5) is valid and known as *Fisher's Exact p-Value*, after Fisher (1971).

3 Sensitivity Analysis

When we propose the rejection rule (5), we implicitly assume that all regions have the same probability to be chosen to face the intervention and only one region can face the intervention. An obvious generalization of this assumption imposes no restriction on the treatment assignment probabilities, keeping the assumption that only one region faces the intervention.[19] 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 probability that region $j \in \{1, ..., J+1\}$ faces the intervention, and reject the *exact null hypothesis* (6) if p is less than some pre-specified significance level, such as the traditional value of $\gamma = 0.1$.

Abadie et al. (2010) and Abadie et al. (2015) implicitly assume the following.

Assumption 4 *The treatment assignment probability distribution is uniform, i.e.,* $\pi_j = {}^1/_{(J+1)}$ *for all* $j \in \{1, ..., J+1\}$.

Under assumption 4, the general p-value (equation (7)) simplifies to the p-value proposed by Abadie et al. (2010) and Abadie et al. (2015) and described in equation (5). However, in

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

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

most of the applications of SCM, Assumption 4 may be considered a strong assumption, as control units may have different treatment probabilities. If the true probabilities $(\pi_1, ..., \pi_{J+1})$ associated to each treatment assignment vector differs from the assumed discrete uniform distribution, we suffer from *hidden bias* according to Rosenbaum (2002). Consequently, a test based on Assumption 4 will present a true size that is different from its nominal size.

Because the vector $(\pi_1, ..., \pi_{J+1})$ is unknown, the general p-value in equation (7) is unfeasible. In the situation that the researcher does not know anything about the true treatment probabilities, Assumption 4 can be interpreted as providing an 'uninformative prior'. In some cases, however, one has access to an estimate of the vector $(\pi_1, ..., \pi_{J+1})$.[20] That is a rare situation and, in most cases, there is no data available or the nature of the analyzed intervention does not allow to conduct such estimation.[21]

The lack of information on the vector $(\pi_1, ..., \pi_{J+1})$ does not mean that we need to assume that the uniform distribution is the correct configuration. For example, one could start distorting the treatment assignment probabilities in the direction of changing the decision of the testing procedure, but preserving the relative lack of knowledge on that distribution. This is exactly what is done in Rosenbaum (2002) and Cattaneo et al. (2016). They consider a sensitivity analysis that allows the empirical researcher to measure the robustness of his or her conclusions (i.e., the test's decision regarding rejecting the *exact null hypothesis*) to Assumption 4, by distorting as little as possible the uniform distribution. We present this sensitivity analysis step-by-step in the framework of the synthetic control estimator:

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 $RMSPE_{(1)} > RMSPE_{(2)} > ... > RMSPE_{(J+1)}$.

[20] For example, when analyzing the economic impact of natural disasters (see Cavallo et al. (2013)), one could use time series data about the frequency that each region faces an intervention.

[21] For example, how can we define and estimate the probability of Switzerland merging with East Germany in the case of the German Reunification analyzed by Abadie et al. (2015)?

3. Define $\overline{j} \in \Omega := \{(1), ..., (J+1)\}$ such that $RMSPE_{\overline{j}} = RMSPE^{obs}$. If there are more than one $j' \in \Omega$ that presents this property, take the largest one.

4. Define the probability of each placebo treatment assignment $(j) \in \Omega$ as

$$\pi_{(j)}(\phi) = \frac{\exp\left(\phi \upsilon_{(j)}\right)}{\sum_{j' \in \Omega} \exp\left(\phi \upsilon_{j'}\right)}, \tag{8}$$

where $\phi \in \mathbb{R}_+$ is the sensitivity parameter and $\upsilon_{j'} \in \{0,1\}$ for each $j' \in \Omega$. Note that, when $\phi = 0$, all placebo treatment assignments present the same probability of being chosen, i.e., $\mathbb{P}(\omega = (j)) = {}^{1}/_{(J+1)}$ for all $(j) \in \Omega$. Consequently, assumption 4 imposes that $\phi = 0$. Moreover, the sensitivity parameter $\phi \in \mathbb{R}_+$ has a very intuitive interpretation: a region $j_1 \in \Omega$ with $\upsilon_{j_1} = 1$ is $\Phi := \exp(\phi)$ times more likely to face the intervention than a region $j_2 \in \Omega$ with $\upsilon_{j_2} = 0$.

5. Under assumption (8), the permutation test's p-value, originally defined in (7), is now given by

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

where $\boldsymbol{\upsilon} := (\upsilon_1, ..., \upsilon_{J+1})$. Observe that, given a sensitivity parameter $\phi \in \mathbb{R}_+$ and a vector $\boldsymbol{\upsilon}$, we reject the *exact null hypothesis* if $p(\phi, \boldsymbol{\upsilon})$ 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 one defined by equation (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[22] is given by

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

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

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

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

where γ is a pre-specified significance level. If $\underline{\phi} \in \mathbb{R}_+$ is close to zero, the permutation test's decision is not robust to small violations of assumption (4), i.e., $\phi = 0$.

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[23] is given by

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

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

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

where γ is a pre-specified significance level. If $\overline{\phi} \in \mathbb{R}_+$ is close to zero, the permutation test's decision is not robust to small violations of assumption (4), i.e., $\phi = 0$.

8. Based on the permutation test's decision, we can fix the vector $\boldsymbol{\upsilon} = (\upsilon_1, ..., \upsilon_{J+1})$ 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 $p(\phi, \boldsymbol{\upsilon})$ in the vertical axis. If $p(\phi, \boldsymbol{\upsilon})$ changes too quickly when we change ϕ, the permutation test is too sensitive to assumptions regarding the treatment assignment probabilities.

We discuss the meaning of large sensitivity parameter values — $\underline{\phi} \in \mathbb{R}_+$ and $\overline{\phi} \in \mathbb{R}_+$ — in subsection 5.2 and illustrate the importance of this sensitivity analysis mechanism in our empirical application (section 7).

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

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. (2010) and Abadie et al. (2015) 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:

$$H_0 : Y^I_{j,t} = Y^N_{j,t} + f_j(t) \text{ for each region } j \in \{1, ..., J+1\} \text{ and time period } t \in \{1, ..., T\}\,, \quad (10)$$

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.

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*[24]:

$$H_0 : Y^I_{j,t} = Y^N_{j,t} + f(t) \text{ for each region } j \in \{1, ..., J+1\} \text{ and time period } t \in \{1, ..., T\}\,, \quad (11)$$

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

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

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

given by equation (4) becomes

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

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

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

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 $p^f \left(\phi, \boldsymbol{\upsilon}\right)$ 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 inference procedure proposed by Abadie et al. (2010) and Abadie et al. (2015) to test any *sharp null hypothesis* (equation (11)) instead of only the *exact null hypothesis* ((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.

Moreover, function f can be chosen in order to test a theory that predicts a specific form for the intervention effect. For example, imagine that a researcher is interested in analyzing the economic impact of natural disasters (Barone & Mocetti (2014), Cavallo et al. (2013), Coffman & Noy (2011), DuPont & Noy (2012)). Theory predicts three different possible intervention effects in this case: (i) GDP initially increases due to the aid effect and, then, decreases back to its potential level; (ii) GDP initially decreases due to the destruction effect and, then, increases back to its potential level; and (iii) GDP decreases permanently due to a reduction in its potential level. The researcher can choose a inverted U-shaped function f_i, a U-shaped function f_{ii} and a decreasing function f_{iii} and apply 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 & Rubin (2015) and Rosenbaum (2002), we can invert the test statistic to estimate confidence sets for the treatment effect function. Formally, under assumptions 1-3 and assumption (8), we can construct a $(1-\gamma)$-confidence set in the space $\mathbb{R}^{\{1,...,T\}}$ as

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

where $p^f(\phi, \boldsymbol{\upsilon})$ is given by equation (13). Note that it is easy to interpret $CS_{(1-\gamma)}(\phi, \boldsymbol{\upsilon})$: 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 $\mathbb{R}^{\{1,\dots,T\}}$ is too large to be informative and estimating such a confidence set would be computationally infeasible. For these reasons, we believe that it is worth focusing in two subsets of $CS_{(1-\gamma)}\left(\phi, \boldsymbol{v}\right)$.

Firstly, we propose to assume the following null hypothesis:

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

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 $RMSPE^{c}$. Under assumptions 1-3 and assumption (8), we can then construct a $(1-\gamma)$-confidence interval for the constant intervention effect as

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

where $c \in \mathbb{R}$ and $\gamma \in (0,1) \subset \mathbb{R}$. It is easy to interpret $CI_{(1-\gamma)}\left(\phi, \boldsymbol{v}\right)$: 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 (15) and (16) to a linear in time intervention effect (with intercept equal to zero). Assume

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

for each region $j \in \{1, ..., J+1\}$ and time period $t \in \{1, ..., T\}$, where $\widetilde{c} \in \mathbb{R}$. 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 $\widetilde{c} \in \mathbb{R}$, estimating the empirical distribution of $RMSPE^{\widetilde{c}}$. Under assumptions 1-3 and assumption (8), we can then construct a $(1-\gamma)$-confidence set for the linear intervention effect as

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

where $\gamma \in (0,1) \subset \mathbb{R}$. It is also easy to interpret $\widetilde{CS}_{(1-\gamma)}\left(\phi,\boldsymbol{\upsilon}\right)$: 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 (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 assumption 4 by comparing the areas of confidence set for different values of $\phi \in \mathbb{R}_+$. As before, we highlight three interesting choices for the sensitivity parameter $\phi \in \mathbb{R}_+$ and the vector $\boldsymbol{\upsilon} = (\upsilon_1, \ldots, \upsilon_{J+1})$. The first one simply assumes $\phi = 0$ and $\boldsymbol{\upsilon} = (1, \ldots, 1)$, i.e., it imposes assumption 4. 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 $\left(\widehat{\alpha}_{1,t'}\right)^2$ 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. (2010) and Abadie et al. (2015), 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 & Rubin (2015), we define a test statistic θ^f as a known $\underline{\text{positive}}$ 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 assumptions 1-3 and 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 $\mathbb{R}^{J+1}$. Adding assumption (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.[25]

Moreover, observe that *RMSPE* and any linear combination of the absolute estimated synthetic control gaps are test statistics according to this definition. Consequently, the hypothesis tests proposed by Abadie et al. (2010) and Abadie et al. (2015) are inserted in this framework.

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 assuming that $\phi = 0$ and $\boldsymbol{\upsilon} = (1, ..., 1)$. 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 $\boldsymbol{\upsilon} = (1, ..., 1)$ 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:

- $\theta^1 := mean\left(\left|\widehat{\alpha}_{\widetilde{j},t}\right| |t \geq T_0 + 1\right)$ is one way to aggregate the information provided by placebo gaps graphs that were introduced by Abadie et al. (2010).

- $\theta^2 := RMSPE_{\widetilde{j}}$ is strongly recommended by Abadie et al. (2015) because it controls

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

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,

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

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

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

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

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

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

where $\widetilde{j}$ is the region that is assumed to face the intervention in each permutation, $mean(\mathbf{B}|\mathbf{A})$ is the mean of variable $\mathbf{B}$ conditional on event $\mathbf{A}$. 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 and framework. For this same reason, we also report rejection rates for the differences-in-differences inference procedure proposed by Conley & Taber (2011) (CT).[26] However, we stress that Conley & Taber (2011) propose an asymptotic inference procedure and that our Monte Carlo Experiment has a small sample size, implying that the CT method may not work properly.

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 Abadie et al. (2010):

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

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[27]. The scalar $u_{j,t+1}$ and each element of the $(K-1) \times 1$-dimension vector $\mathbf{v}_{j,t+1}$ are independent random draws from a standard normal distribution. The scalars δ_t and κ_t and each element of $\boldsymbol{\beta}_{t+1}$ and $\boldsymbol{\rho}_t$ 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 $Y^N_{j,0} = \boldsymbol{\beta}_0 \mathbf{Z}_{j,0} + u_{j,0}$. Finally, the potential outcome when region 1 faces the intervention

[26] We estimate model (20) and test the null hypothesis $H_0 : \eta_2 = 0$ using the confidence intervals recommend by Conley & Taber (2011). Since their inference procedure uses only the control regions in order to estimate the test statistic distribution, the true nominal size of this test is 10.53%.

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

in period $t \in \{1, ..., T\}$ is given by

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

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 alternative hypothesis[28].

Note that, in each one of the 21,000 Monte Carlo repetitions (3,000 repetitions for each different intervention effect λ), we create an entire population of regions. Hence, after realizing the values of the potential outcome variables, we can interpret them as fixed but *a priori* unknown quantities in accordance to assumption 3.[29] Moreover, our data generating process satisfies Assumptions 2 and 4, i.e., the treatment assignment is random and each region is equally likely to face the intervention.[30]

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 & Taber (2011) when the true intervention effect is given by the value mentioned in the column's heading. Consequently, while column (1) presents tests' sizes, the columns (2)-(7) present their power.

Analyzing column (1), we note that the five permutation tests of our Monte Carlo Experiment $\left(\theta^1\text{-}\theta^5\right)$ present the correct nominal size as expected by the decision rule of Fisher's Exact Inference Procedure. Moreover, the asymptotic inference procedure proposed by Conley & Taber (2011) (CT) has a true size close to the correct one (10.53%).

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

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

[30] We also analyzed data generating processes that violate assumptions 2 and 4. The results based on them are very similar to the ones presented here and are available upon request.

Table 1: Monte Carlo Experiment's Rejection Rates

	Intervention Effect						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Test Statistic	$\boldsymbol{\lambda}$**= .0**	$\boldsymbol{\lambda}$**= .05**	$\boldsymbol{\lambda}$**= .1**	$\boldsymbol{\lambda}$**= .25**	$\boldsymbol{\lambda}$**= .5**	$\boldsymbol{\lambda}$**= 1.0**	$\boldsymbol{\lambda}$**= 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 & Taber (2011).

Analyzing the other columns, we note that the test statistic *RMSPE*, proposed by Abadie et al. (2015) $\left(\theta^2\right)$, is uniformly more powerful than the simple test statistics $\left(\theta^4, \theta^5\right)$ that are commonly used in the evaluation literature. This result suggests that, in a context where we observe only one treated unit, we should use the synthetic control estimator even if the treatment were randomly assigned. We also stress that the hypothesis test based on the statistic *RMSPE* $\left(\theta^2\right)$ outperforms the test proposed by Conley & Taber (2011) (CT) in terms of power, suggesting that, in a context with few control regions, we should use the synthetic control estimator instead of a differences-in-differences model that applies a 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 $\left(\theta^4\right)$ or the permuted differences-in-differences test $\left(\theta^5\right)$. 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 (1959) looks to a simple test of mean differences, Ibragimov & Muller (2010) analyzes a two-sample test of mean differences where samples' variances are different from each other, and Young (2015) focus on a linear regression coefficient.

Finally, we note that the simple average of the absolute post-intervention treatment effect $\left(\theta^1\right)$, despite using the synthetic control method, is as powerful as the simple test statistics that are commonly used in the evaluation literature $\left(\theta^4, \theta^5\right)$. Consequently, we do not recommend to use it to conduct inference, because it is as time-consuming to estimate as the more powerful test statistics that uses the synthetic control method, $\left(\theta^2 \text{ and, specially, } \theta^3\right)$. Following Abadie et al. (2010) and Abadie et al. (2015), a possible explanation for the low power of $\left(\theta^1\right)$ is the fact that this test statistic ignores the quality of the pre-intervention fit.

At this point, we avoid making any stronger suggestion about which test statistic the empirical researcher should use, because, as (Eudey et al. 2010, p. 14) makes clear, this choice is data dependent since the empirical researcher's goal is to match the test statistic to the meaning of the data. For example, if outliers are extremely important, θ^2 may be a better option than θ^3 even though the latter is more powerful than the former.

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 DGP described above. We focus on the average values of the sensitivity parameter that change a hypothesis test's decision, 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 $\boldsymbol{v} = (1, ..., 1)$ and using the five test statistics described in subsection 5.1. Based on each test's decision, 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 2 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. Each cell presents the average value of the sensitivity parameter that changes the hypothesis' test decision 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.

Table 2: Sensitivity Analysis

	Intervention Effect						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Test Statistics	**λ= .0**	**λ= .05**	**λ= .1**	**λ= .25**	**λ= .5**	**λ= 1.0**	**λ= 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
$\widehat{\theta}^4$	0.38	0.51	0.56	0.62	0.66	0.70	0.71
$\widehat{\theta}^5$	0.38	0.49	0.55	0.61	0.63	0.71	0.70
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
$\widehat{\theta}^4$	2.50	2.33	2.34	2.22	2.18	2.03	2.14
$\widehat{\theta}^5$	2.50	2.34	2.30	2.20	2.13	2.00	2.07

Source: Authors' own elaboration. *Notes:* Each cell presents the value of the sensitivity parameter that changes the hypothesis' test decision 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. $\widehat{\theta}^4$-$\widehat{\theta}^5$ are associated to permutation tests that are frequently used in the evaluation literature.

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 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 2 shows, the sensitivity analysis parameter

$\underline{\phi} \in \mathbb{R}_+$ for the five 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 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 2 shows, the sensitivity analysis parameter $\overline{\phi} \in \mathbb{R}_+$ for all test statistics but $\widehat{\theta}^3$ decreases when the intervention effect $\lambda \in \mathbb{R}_+$ increases, as desired. Consequently, even if $\widehat{\theta}^3$ is the most powerful test statistic in our Monte Carlo experiment, it does not have good properties when applied to the sensitivity analysis mechanism.

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 $\widehat{\theta}^1$, $\widehat{\theta}^4$ and $\widehat{\theta}^5$ 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 meaning of the data. However, tests statistic $\widehat{\theta}^2$ — the traditional *RMPSE* statistic proposed by Abadie et al. (2015) — performs better than all the other test statistics with respect to either power or robustness to deviations from assumption 4.

To conclude, we stress that similar Monte Carlo experiments can help the empirical researcher to evaluate the robustness of his or her findings. For example, when we redo this Monte Carlo experiment with seventeen regions, we find that the best case scenario $\overline{\phi}$ for the *RMSPE* test statistic $\left(\widehat{\theta}^2\right)$ varies from 2.38 to 1.53. Since, in the empirical application of section 7 that analyzes data from seventeen Spanish provinces, we find a sensitivity parameter $\overline{\phi}$ that ranges from 1.85 to 3.99, we can conclude that the empirical results from section 7 are very robust to deviations from assumption 4.

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 & Rubin (2015) 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 (2008) shows that simultaneously testing hypotheses about a large number of outcomes can be dangerous, leading to an increase in the number of false rejections.[31] 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 (2008) to our framework, we propose one way to test any *sharp null hypothesis* for a large number of outcome variables, preserving the correct test size for each variable of interest.

First, we modify the framework described in 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. In this case, assumptions 1-3 must hold for all outcome variables $\mathbf{Y}^1, ..., \mathbf{Y}^M$. Now, our null hypothesis is also more complex than the one described in equation (11):

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

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. 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.

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

Based on the inference procedure developed by Abadie et al. (2010) and Abadie et al. (2015), we can, for each $m \in \{1, ..., M\}$, calculate an observed test statistic, $\theta^{obs}_{f_m} = \theta^m(e_1, \tau, \mathbf{Y}^m, \mathbf{X}, f_m)$, and their associated observed p-value,

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

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

Since this p-value is itself a test statistic, we can estimate, for each outcome $m \in \{1, ..., M\}$, its empirical distribution by computing

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

for each region $\widetilde{j} \in \{1, ..., J+1\}$, where $\theta^{m,\widetilde{j}} := \theta^m\left(e_{\widetilde{j}}, \tau, \mathbf{Y}^m, \mathbf{X}, f_m\right)$. Our next step is to calculate $p^{\widetilde{j}}_{\theta_{f_m},*} := \min\left\{p^{\widetilde{j}}_{\theta_{f_m}}, p^{\widetilde{j}}_{\theta_{f_{m+1}}}, ..., p^{\widetilde{j}}_{\theta_{f_M}}\right\}$ for each $m \in \{1, ..., M\}$ and each $\widetilde{j} \in \{1, ..., J+1\}$. Then, we estimate, for a given value of the sensitivity parameter $\phi \in \mathbb{R}_+$ and a given vector $\boldsymbol{v} = (v_1, ..., v_{J+1})$ and under assumption (8),

$$p^{fwer*}_{\theta^{obs}_{f_m}}(\phi, \boldsymbol{v}) := \sum_{j=1}^{J+1} \frac{\exp(\phi v_j)}{\sum_{j'=1}^{J+1} \exp(\phi v_{j'})} \times \mathbb{I}\left[p^{j}_{\theta_{f_m},*} \leq p^{obs}_{\theta_{f_m}}\right] \tag{24}$$

for each $m \in \{1, ..., M\}$. We enforce monotonicity one last time by computing $p^{fwer}_{\theta^{obs}_{f_m}}(\phi, \boldsymbol{v}) := \min\left\{p^{fwer*}_{\theta^{obs}_{f_m}}(\phi, \boldsymbol{v}), p^{fwer*}_{\theta^{obs}_{f_{m+1}}}(\phi, \boldsymbol{v}), ..., p^{fwer*}_{\theta^{obs}_{f_M}}(\phi, \boldsymbol{v})\right\}$ for each $m \in \{1, ..., M\}$. Finally, for each outcome variable $m \in \{1, ..., M\}$, we reject the *sharp null hypothesis* (23) if $p^{fwer}_{\theta^{obs}_{f_m}}(\phi, \boldsymbol{v}) \leq \gamma$, where γ is a pre-specified significance level.

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 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 (2008), because we summarized the entire time information in a single test statistic. Since Anderson (2008) argues that the *summary index test*[32] 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, he or she should interpret each post-intervention time period as a different outcome variable and apply the inference procedure described in this subsection. Both approaches deliver valid statistical inference in small samples under assumption (8).

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 inference procedure proposed by Abadie et al. (2010) and Abadie et al. (2015) 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 $p^{fwer}_{\theta^{obs}_{fm}}(\phi, \boldsymbol{\upsilon})$ to define either $\underline{\phi} \in \mathbb{R}_+$ or $\overline{\phi} \in \mathbb{R}_+$.

6.2 Hypothesis Testing and Confidence Sets with Multiple Treated Units

Cavallo et al. (2013) extend the SCM developed by Abadie & Gardeazabal (2003) and Abadie et al. (2010) 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

[32] 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 $t \in \{1, ..., T\}$ for some outcome variable $m \in \{1, ..., M\}$.

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 synthetic control estimator is given by $\overline{\widehat{\alpha}}_{1,t} := \sum_{g=1}^{G} \widehat{\alpha}_{1^g,t}/G$ for each $t \in \{1, ..., T\}$.

In the same way Cavallo et al. (2013) do, we impose assumptions 1-3 to all interventions $g \in \{1, ..., G\}$ and all regions $j^g \in \{1, ..., J^g + 1\}$ in each intervention. We modify their inference procedure by summarizing the entire time information in a single test statistic in order to avoid over-rejecting the null hypothesis as pointed out by Anderson (2008)[33]. We also adapt their discrete uniform probability distribution of treatment assignments to consider an assumption (equation (26)) similar to 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 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.

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

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.*[34]

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

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

1. Estimate the test statistics θ_1^f, θ_2^f,...,θ_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

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

We stress that 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,

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

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 inference procedure can cast new light on empirical studies that use SCM. In particularly, we can analyze the robustness of empirical results to the assumption about assignment probabilities, 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 provinces made available by Abadie & Gardeazabal (2003) and discussed by Abadie et al. (2011) too.

We start by evaluating the statistical significance of the economic impact of ETA's terrorism using the *RMSPE* test statistic and the inference procedure described in section 2.2, repeating the exercise implemented by Abadie et al. (2011). Then, we analyze the robustness of this result to the assumption of uniform assignment probabilities using the procedure explained in section 3. After that, we estimate $88.\overline{9}\%$-Confidence Sets[35] 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 *RMSPE* test statistic. 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 & Gardeazabal (2003) is available for download using the software *R*. We observe, as our outcome variable, annual real GDP per-capita in thousands of 1986 USD from 1955 to 1997 and, as covariates, biannual sector shares as a percentage of total production for agriculture, forestry and fishing, energy and water, industry, construction and engineering, marketable services and nonmarketable services from 1961 to 1969; annual shares of the working age population that was illiterate, that completed at most primary education

[35] The confidence level is only approximately 90% due to the discreteness of the p-value in permutation tests.

and that completed at least secondary education from 1964 to 1969; the population density in 1969; and annual gross total investment as a proportion of GDP from 1964 to 1969. All those variables are observed at the province level and there are seventeen provinces, including the Basque Country ($J + 1 = 17$). For historical details and descriptive statistics about this data set, see Abadie & Gardeazabal (2003) and Abadie et al. (2011).

ETA's terrorism acts gained strength and relevance during the 70s. For this reason, our post intervention period goes from 1970 to 1997 ($T_0 = 1969$). In order to estimate the synthetic control unit, we plug, in equation (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 provinces, we find that the estimated intervention effect does not look abnormally large when compared to the estimated placebo effects as subfigure 1a shows. This intuitive perception is confirmed by the inference procedure proposed by Abadie et al. (2010) and Abadie et al. (2015) (see subsection 2.2) when we use the *RMSPE* test statistic. More specifically, we have that $p = 0.41$, implying that we can not reject the *null hypothesis of no effect whatsoever*.

Abadie et al. (2010) and Abadie et al. (2011) suggest that we should exclude provinces with a poor pre-intervention fit (i.e., provinces whose pre-intervention MSPE is five times greater than the Basque Country's pre-intervention MSPE) because placebo studies for those provinces provinces are not informative about the relative rarity of the post-intervention effect for the Basque Country[36]. By doing so, we exclude the provinces of Madrid, Extremadura and Balearic Islands when computing the p-value of the hypothesis test and find that $p = 0.50$, implying that we can not reject the *null hypothesis of no effect whatsoever*. The remaining placebos effects are plotted in subfigure 1b.

In the last two results, we keep the Basque Country in the donor pool of the placebo tests as recommend by Imbens & Rubin (2015) because, under the null hypothesis, not only we know the intervention effect but we know that it is equal to zero at every period. However,

[36] We thank an anonymous referee for stressing this point.

this decision may reduce the power of our hypothesis test by inducing placebo effects of the opposite sign than the intervention effect on the Basque Country.[37] For this reason, we redo both tests excluding the Basque Country from the donor pool of the placebo tests. Subfigures 1c and 1d show the estimated placebo effects. We find the same p-values of our previous exercise, implying that the conclusion about not rejecting the *null hypothesis of no effect whatsoever* is robust to the exclusion of the Basque Country from the donor pool of placebo tests.

Now, we evaluate the robustness of our findings to the assumption of uniform assignment probabilities using the sensitivity analysis proposed in section 3. We can conclude that not rejecting the *null hypothesis of no effect whatsoever* is a very robust result to deviations from assumption 4 because we must impose a sensitivity parameter $\overline{\phi} = 1.845$ in order to reject it at the 10%-significance level, implying that the treatment assignment probability of the region that is most likely to receive the treatment is more than six times larger than the treatment assignment probability of the Basque Country.[38] Moreover, we note that the permutation test's p-value decreases very slowly as a function of the sensitivity parameter $\phi \in \mathbb{R}_+$ as subfigure 2a. In subfigure 2b, we plot the p-value as a function of the sensitivity parameter when we drop provinces with a poor pre-intervention fit when computing the p-value. Once more, this exercise shows that not rejecting the *null hypothesis of no effect whatsoever* is a very robust result.[39]

We also estimate two $88.\overline{9}$%-Confidence Sets[40]. While subfigure 3a considers a Constant in Time Intervention Effect following equation (16), subfigure 3b considers a Linear in Time Intervention Effect whose intercept is equal to zero following equation (18). Both confidence sets uses the *RMSPE* test statistic.

In gray, we plot confidence sets under assumption 4, i.e., imposing that the treatment

[37] We thank an anonymous referee for stressing this point.

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

[39] We highlight that, according to subsection 2, a sensitivity parameter $\overline{\phi}$ between 1.5 and 2.4 is reasonably large.

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

assignment probability distribution is uniform across Spanish provinces. These gray areas not only quickly show that we can not reject the null hypothesis of no effect whatsoever (because the confidence sets contains the zero function), but also show that the economic impact of ETA's terrorism is not precisely estimated, precluding even conclusions about its true sign.[41]

When we apply our sensitivity analysis mechanism to these confidence sets, we impose a parameter $\overline{\phi} = 1.845$ for the best case scenario described in section 3 and find the confidence sets between the dashed black lines. Observe that we need to impose a very large sensitivity analysis parameter in order to barely exclude the *null hypothesis of no effect whatsoever* from the confidence intervals explained in subsection 4.2. Again, this exercise illustrates the robustness of the empirical conclusions. Moreover, this graphical device quickly show the robustness of this finding by visually comparing the areas of the two confidence sets. Since the area between the dashed lines is much narrower than the gray area, it is easy to see that the test's decision is very robust to deviations from assumption 4.

Moreover, note also that these conclusions are robust to the choice of functional form for the intervention effect (constant or linear) and to the choice of keeping or excluding the Basque Country from the donor pool of the placebo runs (subfigures 3c and 3d) since all gray confidence sets include very positive and very negative null hypothesis and all dashed confidence sets are much narrower than the gray ones. 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 4 shows this fitted quadratic function. Applying the inference procedure described in section 4.1 and using the *RMSPE* test statistic, we do not reject the null hypothesis

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

that the true intervention effect follows this quadratic function because $p_{quadratic} = 0.65$. In this case, we must impose a sensitivity parameter $\overline{\phi} = 2.805$ in order to reject it at the 10%-significance level, implying that the treatment assignment probability of the region that is most likely to receive the treatment is more than sixteen times larger than the treatment assignment probability of the Basque Country. When we exclude the Basque Country from the donor pool of the placebo tests, we find $p_{quadratic} = 0.71$ and $\overline{\phi} = 3.075$. Finally, excluding provinces with a poor pre-intervention fit, we find $p_{quadratic} = 0.79$ and $\overline{\phi} = 3.495$ for the usual hypothesis test and $p_{quadratic} = 0.86$ and $\overline{\phi} = 3.985$ for the hypothesis test that excludes the Basque Country from the donor pool.

Differently from what we do in the last paragraphs, we can treat each year as a different outcome variable and apply the inference procedure described in subsection 6.1. This interpretation allow us to analyze the timing of the economic impact of ETA's terrorism, which may be significant for some time periods even though we have not reject the null hypothesis of no effect whatsoever when we pooled together all the years using the *RMSPE* test statistic. We use the squared value of the estimated intervention effect for each year of the post-intervention period as a test statistic. Using the notation of subsection 6.1, we have that

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

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 between 0.42 and 0.88 for all 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 $\boldsymbol{\upsilon}_m$ 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 $p^{fwer}_{\theta^{obs}_{f_m}}(\phi, \boldsymbol{\upsilon})$ to determine one sensitivity parameter $\overline{\phi}_m$ for each outcome $m \in \{1, ..., M\}$. We find sensitivity parameters that range from 1.845 to 4.895. Clearly, we cannot reject the null hypothesis that ETA's terrorism has no economic impact whatsoever and this conclusion is very robust to deviations from assumption 4. When we exclude the Basque Country from the donor pool of placebo runs, the results are similarly, but slightly less robust since the sensitivity parameter range from 1.015 to 4.965.

Given our choice of test statistic for the exercise about multiple outcomes, the results might be sensitive to the inclusion or exclusion of units with a poor intervention fit. Excluding the three units with a poor intervention fit, we find p-values between 0.29 and 0.86 and sensitivity analysis parameters between 1.285 and 4.675 for all years. When we also exclude the Basque Country from the donor pool of placebo runs, we find p-values between 0.29 and 0.93 and sensitivity analysis parameters between 1.285 and 4.765. We can, then, conclude that the conclusions described above are robust to the exclusion of units with a poor pre-intervention fit.

As a consequence of all our empirical exercises, we conclude that terrorists acts in the Basque Country had no statistically significant economic consequence as discussed by Abadie et al. (2011). Moreover, we can conclude that this conclusion is very robust to deviations from assumption 4. 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. (2010) and Abadie et al. (2015) in two ways. First, we highlight that the p-value proposed by Abadie et al. (2015) assumes that the treatment assignment probability is uniformly distributed across all observed regions. Since most applications using SCM are observational studies, the researcher cannot know whether this assumption is valid or not. To address this issue, we propose a parametric form to the treatment assignment probabilities that allow the researcher to implement a sensitivity analysis mechanism similar to the one suggested by Rosenbaum (2002) and Cattaneo et al. (2016). By analyzing the sensitivity analysis parameter that changes the test's decision, we can gauge the robustness of a conclusion to deviations from the assumption that imposes a uniform distribution of treatment assignment probabilities.

Second, we extend the test proposed by Abadie et al. (2010) and Abadie et al. (2015) 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 last extension and procedures described by Imbens & Rubin (2015) and Rosenbaum (2002), 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. Although the assumptions that guarantee the validity of those confidence sets are strong, 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, 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 SCM. In this simulation, we find that test statistics designed for SCM perform much better than its competitors when there is only one region that faces the intervention. In particular, the traditional *RMSPE* statistic proposed by Abadie et al. (2015) 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 & Gardeazabal (2003), Abadie et al. (2010) and Abadie et al. (2015) in important dimensions: testing null hypothesis about a pooled effect among few treated units and simultaneously testing null hypothesis 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 & Gardeazabal (2003) and Abadie et al. (2011). In the same way as those authors, we do not reject the null hypothesis of no effect whatsoever. Our sensitivity analysis mechanism allows us to conclude that this result is very robust to deviations from the assumption about uniform treatment assignment probabilities. 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.

References

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

Abadie, A., Diamond, A. & Hainmueller, J. (2011), 'Synth: An R Package for Synthetic Control Methods in Comparative Case Studies', *Journal of Statistical Software* **42**(13), 1–17.

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

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

Acemoglu, D., Johnson, S., Kermani, A., Kwak, J. & Mitton, T. (2013), The Value of Connections in Turbulent Times: Evidence from the United States. NBER Working Paper 19701. Available at: http://www.nber.org/papers/w19701.pdf.

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

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

Ando, M. & Sävje, F. (2013), Hypothesis Testing with the Synthetic Control Method. Working Paper, http://www.eea-esem.com/files/papers/eea-esem/2013/2549/scm.pdf.

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

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

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

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

Billmeier, A. & Nannicini, T. (2013), 'Assessing Economic Liberalization Episodes: A Synthetic Control Approach', *The Review of Economics and Statistics* **95**(3), 983–1001.

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

Bove, V., Elia, L. & Smith, R. P. (2014), The Relationship between Panel and Synthetic Control Estimators on the Effect of Civil War. Working Paper, http://www.bbk.ac.uk/ems/research/BirkCAM/working-papers/BCAM1406.pdf.

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

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

Carvalho, C. V., Mansini, R. & Medeiros, M. C. (2015), ArCo: An Artificial Counterfactual Approach for Aggregate Data. Working Paper.

Cattaneo, M., Titiunik, R. & Vazquez-Bare, G. (2016), 'rdlocrand: Inference in Regression Discontinuity Designs under Local Randomization', *The Stata Journal* . Forthcoming. Available at: http://goo.gl/ukyoZi.

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

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

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

Conley, T. G. & Taber, C. R. (2011), 'Inference with Difference-in-Differences with a Small Number of Policy Changes', *The Review of Economics and Statistics* **93**(1), 113–125.

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

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

Dube, A. & Zipperer, B. (2013), Pooled Synthetic Control Estimates for Recurring Treatments: An Application to Minimum Wage Case Studies. Available at: http://www.irle.berkeley.edu/events/spring14/zipperer/dubezipperer_pooledsyntheticcontrol.pdf.

DuPont, W. & Noy, I. (2012), What Happened to Kobe? A Reassessment of the Impact of the 1995 Earthquake in Japan. Available at: http://www.economics.hawaii.edu/research/workingpapers/WP_12-4.pdf.

Eudey, T. L., Kerr, J. & Trumbo, B. (2010), 'Using R to Simulate Permutation Distributions for Some Elementary Experimental Designs', *Journal of Statistics Education* **18**(1).

Ferman, B. & Pinto, C. (2017*a*), Placebo Tests for Synthetic Controls. Available at https://dl.dropboxusercontent.com/u/12654869/Ferman%20and%20Pinto%20-%20placebo%20tests%20for%20SC.pdf.

Ferman, B. & Pinto, C. (2017*b*), Revisiting the Synthetic Control Estimator. Available at https://dl.dropboxusercontent.com/u/12654869/Ferman%20and%20Pinto%20-%20revisiting%20the%20SC.pdf.

Ferman, B., Pinto, C. & Possebom, V. (2017), Cherry Picking with Synthetic Controls. Available at https://dl.dropboxusercontent.com/u/12654869/FPP%20-%20Cherry%20Picking.pdf.

Fisher, R. A. (1971), *The Design of Experiments*, 8th edition edn, Hafner Publishing Company, United States.

Gathani, S., Santini, M. & Stoelinga, D. (2013), Innovative Techniques to Evaluate the Impacts of Private Sector Developments Reforms: An Application to Rwanda and 11 other Countries. Working Paper, https://blogs.worldbank.org/impactevaluations/files/impactevaluations/methods_for_impact_evaluations_feb06-final.pdf.

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

Hahn, J. & Shi, R. (2016), Synthetic Control and Inference. Available at https://ruoyaoshi.github.io/HahnShi_Synthetic.pdf.

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

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

Ibragimov, R. & Muller, U. K. (2010), 't-Statistic Based Correlation and Heterogeneity Robust Inference', *Journal of Business & Economic Statistics* **28**(4), 453–468.

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

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

Kaul, A., Klöbner, S., Pfeifer, G. & Schieler, M. (2015), Synthetic Control Methods: Never Use All Pre-Intervention Outcomes as Economic Predictors. Working Paper. Available at: http://www.oekonometrie.uni-saarland.de/papers/SCM_Predictors.pdf.

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

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

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

Lehmann, E. (1959), *Testing Statistical Hypotheses*, John Wiley & Sons, New York.

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

List, J., Shaikh, A. M. & Xu, Y. (2016), Multiple Hypothesis Testing in Experimental Economics. NBER Working Paper 21875. Available at http://www.nber.org/papers/w21875.

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

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

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

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

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

Possebom, V. (2017), 'Free Trade Zone of Manaus: An Impact Evaluation using the Sythetic Control Method', *Revista Brasileira de Economia* **71**.

Ribeiro, F., Stein, G. & Kang, T. (2013), The Cuban Experiment: Measuring the Role of the 1959 Revolution on Economic Performance using Synthetic Control. Available at: http://economics.ca/2013/papers/SG0030-1.pdf.

Rosenbaum, P. R. (1987), 'Sensitivity Analysis for Certain Permutation Inferences in Matched Observational Studies', *Biometrika* **74**(1), 13–26.

Rosenbaum, P. R. (1988), 'Sensitivity Analysis for Matching with Multple Controls', *Biometrika* **75**(3), 577–581.

Rosenbaum, P. R. (2002), *Observational Studies*, 2$^{\text{nd}}$ edition edn, Springler Science + Business Media, New York.

Rosenbaum, P. R. (2007), 'Sensitivity Analysis for m-Estimates, Tests, and Confidence Intervals in Matched Observational Studies', *Biometrics* **63**, 456–464.

Rosenbaum, P. R. & Krieger, A. M. (1990), 'Sensitivity of Two-Sample Permutation Inferences in Observational Studies', *Journal of the American Statiscal Association* **85**(410), 493–498.

Rosenbaum, P. R. & Silber, J. H. (2009), 'Amplification of Sensitivity Analysis in Matched Observational Studies', *Journal of the* **104**(488), 1398–1405.

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

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

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

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

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

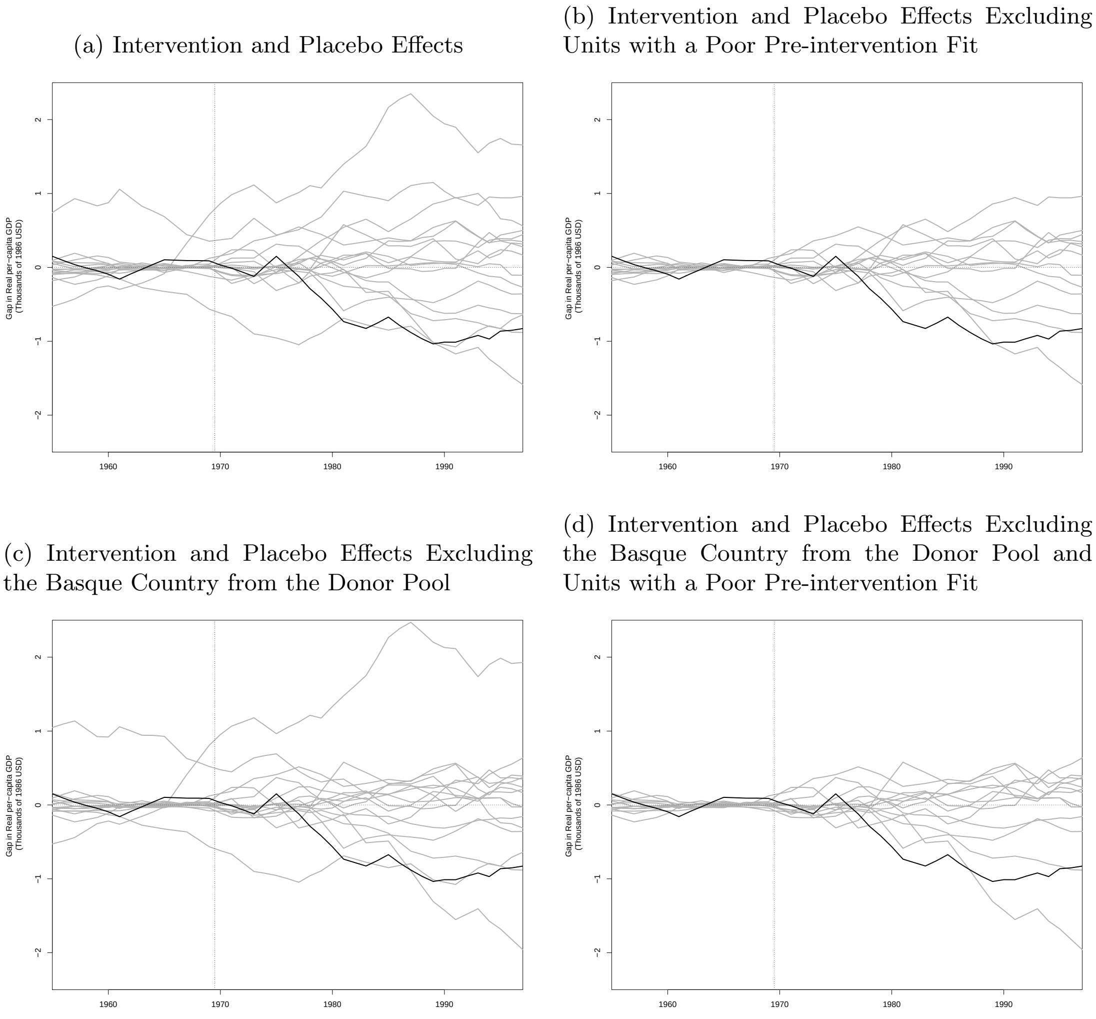
Wong, L. (2015), Three Essays in Causal Inference, PhD thesis, Stanford University.

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

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

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

Figure 1: Estimated Effects using the Synthetic Control Method



(a) Intervention and Placebo Effects

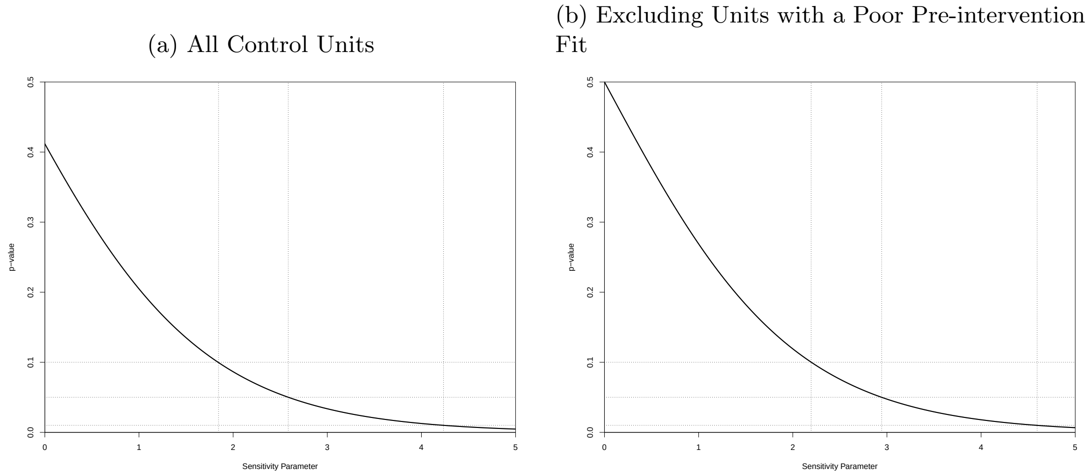
(b) Intervention and Placebo Effects Excluding Units with a Poor Pre-intervention Fit

(c) Intervention and Placebo Effects Excluding the Basque Country from the Donor Pool

(d) Intervention and Placebo Effects Excluding the Basque Country from the Donor Pool and Units with a Poor Pre-intervention Fit

Note: While the gray lines show the estimated placebo effect for each Spanish province, the black lines show the estimated impact of ETA's terrorism on the Basque Country's economy. A poor pre-intervention is defined as a pre-intervention MSPE five times greater than the Basque Country's pre-intervention MSPE.

Figure 2: Sensitivity Analysis

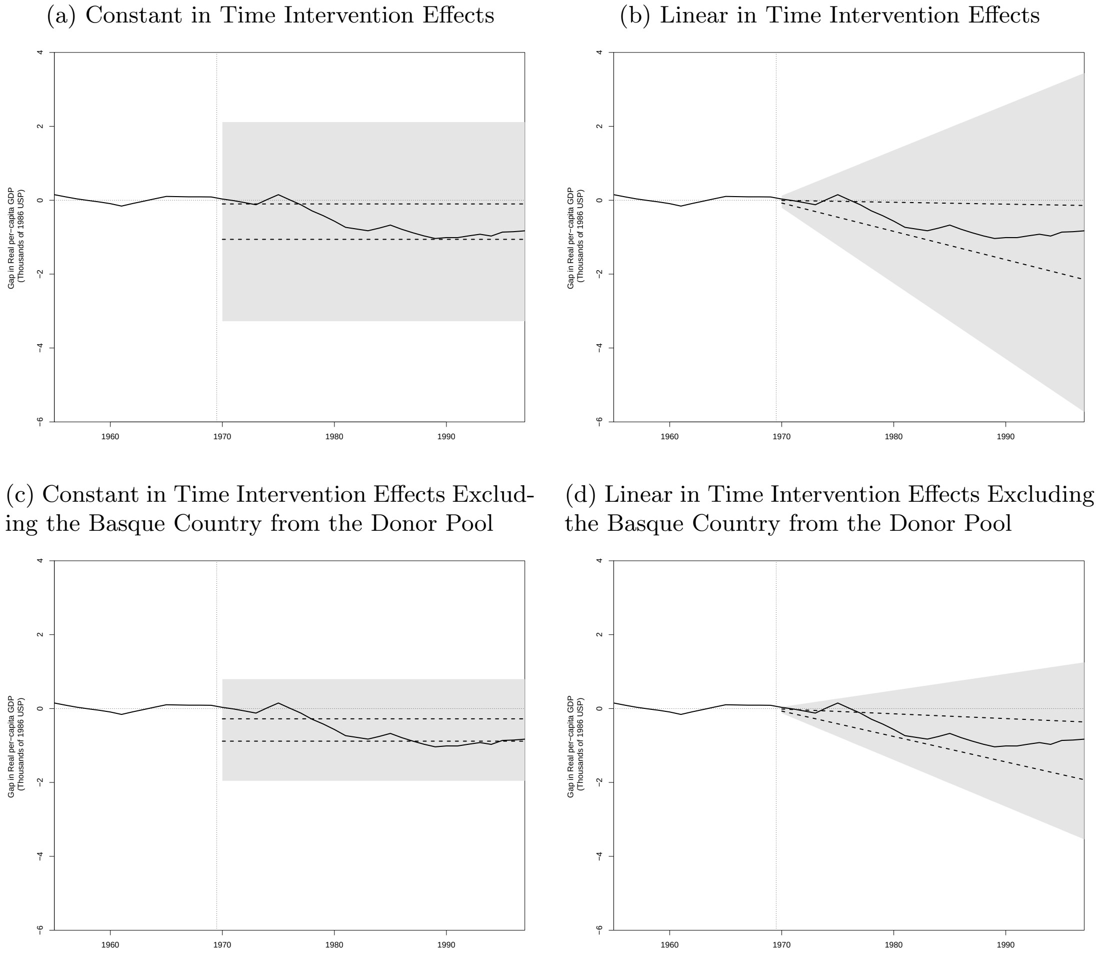


(a) All Control Units

(b) Excluding Units with a Poor Pre-intervention Fit

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

Figure 3: 88.$\overline{9}$%-Confidence Sets for the Intervention Effect



(a) Constant in Time Intervention Effects

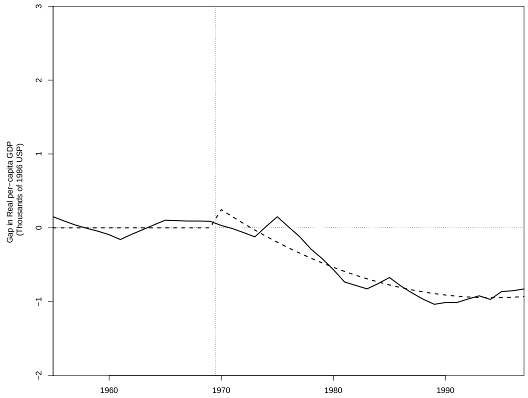
(b) Linear in Time Intervention Effects

(c) Constant in Time Intervention Effects Excluding the Basque Country from the Donor Pool

(d) Linear in Time Intervention Effects Excluding the Basque Country from the Donor Pool

Note: The solid black line shows the estimated impact of ETA's terrorism on the Basque Country's economy while the gray areas show the 88.$\overline{9}$%-Confidence Set for Constant in Time or Linear in Time (with intercept equal to zero) Intervention Effects that were constructed using the *RMSPE* test statistic and imposing that the treatment assignment probabilities are uniformly distributed (Assumption 4). The dashed black lines are the upper and lower bounds of 88.$\overline{9}$%-Confidence Sets that were constructed using the *RMSPE* test statistic and imposing a sensitivity parameter $\overline{\phi} = 1.845$ for the best case scenario described in section 3.

Figure 4: Quadratic Intervention Effect Function



Note: While the black line show the estimated impact of ETA's terrorism on the Basque Country's economy, the dashed line shows the quadratic function that best approximates this effect.