# An introduction to Synthetic Control Methods and their applications to climate change analysis

Jessica Cremonese

March 2023

École Normale Supérieure

Scuola Galileiana di Studi Superiori

Professor Marc Fleurbaey.

# Contents

**Abstract**

Lorem ipsum

# 1 Introduction

Lorem ipsum SCM termed by Athey and Imbens (2017) as "arguably the most important innovation in the policy evaluation literature in the last 15 years".

# 2 Synthetic Control Methods

Synthetic Control Methods (SCM) have been originally proposed in Abadie and Gardeazabal (2003) and by Abadie et al. (2010) to estimate the effects of aggregate interventions. The key idea behind the method is that, when units are a few aggregate entities, a better counterfactual than using any single unit can be derived by computing a combination of the untreated units that closely resembles the treated one, i.e. a "synthetic control". The selection of the "donor units" is formalized with a data driven procedure. Although the method was originally intended for samples with few units, it has been successfully applied in contexts with large samples, for instance in Acemoglu et al. (2016). Such a synthetic control unit is computed as a weighted average of all potential comparison units that best resemble the treated units. A good synthetic unit will resemble treated one not only in terms of the outcomes variable (outer optimization), but also in terms of the pre-treatment predictors value (inner optimization).

In this section, I will introduce the method and explore feasibility, data requirements and methodological issues. The main references are Abadie and Gardeazabal (2003) and Abadie, Diamond and Hainmueller (2010), which introduced the method in the literature, and Abadie (2021), which provides a useful guide to the application of SCM.

## 2.1 Setting the method

Suppose to have data for $j = 1, .., J+1$ units, and suppose that unit $j = 1$ is the treated unit. The "donor pool" of untreated units which will contribute to the construction of a synthetic control for unit $j = 1$ is then constituted by the remaining $j = 2, ..., J + 1$ units. Assume that data covers $T$ periods, with periods up to $T_0$ being the pre-intervention observations.

For each unit $j$ at time $t$ data is available for the outcome of interest $Y_{jt}$, and for a number $k$ of predictors $X_{1j}, ..., X_{kj}$. Define the $k \times 1$ vectors $\mathbf{X}_1, ..., \mathbf{X}_{J+1}$ which contain values of the predictors for units $j = 1, ..., J + 1$. Define the $k \times J$ matrix $\mathbf{X}_0 = \mathbf{X}_2, ..., \mathbf{X}_{J+1}$ which collects values of the predictors or the untreated units. For each unit $j$, define the potential outcome

without treatment as $Y_{jt}^N$. For the trated unit $j = 1$, define the potential response under the treatment as $Y_{jt}^I$ in the post treatment period $t > T_0$. The effect of the intervetion for the affected unit $j = 1$ for $t > T_0$ is:

$$\tau_{1t} = Y_{1t}^I - Y_{1t}^N \tag{1}$$

For the treated unit, $Y_{1t}^I$ is observed so that $Y_{1t} = Y_{1t}^I$, but $Y_{jt}^N$ is not. SCM provides a way to estimate $Y_{jt}^N$ for $t > T_0$, that is, how the outcome of interest would have been in the absence of treatment. Notice that $\tau_{1t}$ is allowed to change over time.

## 2.2 Estimation

A downside of comparative case studies lies in the attempt to select the control units by informally arguing for an affinity between the treated and the untreated before the intervention. However, when using data from aggregate units such as countries or regions, it can be difficult to find a proper counterfactual. SCM offers a formal procedure to select and combine the comparison units in order to create a scenario where unit $j = 1$ was unaffected by treatment.

Define $\mathbf{W} = w_2, ..., w'_{J+1}$ as a $J \times 1$ vector of nonnegative weights that sum to one. The $\mathbf{W}$ vector attributes a weight to each unit in the donor pool $j = 2, ..., J + 1$ and characterizes its contribution to the synthetic unit.

For a set of weights, $\mathbf{W}$, the estimators of $Y_{1t}^N$ and $\tau_{1t}$ are:

$$\hat{Y}_{1t}^N = \sum_{j=2}^{J+1} w_j Y_{jt} \tag{2}$$

$$\hat{\tau}_{1t} = Y_{1t} - \hat{Y}_{1t}^N \tag{3}$$

Nonnegative weights ensure a convex combination of the donor units, so that the resulting control can be interpreted as a weighted average of the control units with typically sparse weights. Furthermore, it ensures comparability of the outcome variable by giving the synthetic control outcome the same scale of the intervention unit. Abadie (2021) notes that, when using weights that sum to one, variables in the data should be rescaled to correct for differences in size between the units, for instance by using per capita GDP instead of level GDP. This correction is not needed if variables are already comparable, for instance in the case of prices. Allowing for weights outside of the $[0, 1]$ interval may provide a more accurate synthetic control by placing negative emphasis on some donor units that are dissimilar to the

treated one. However, negative unbounded weights may introduce extrapolation, where the assigned weights are used to extrapolate beyond the observed rage of data to estimate the effect of treatment. This can lead to biased estimates and reduced precision, and makes the interpretation of weights less straightforward.

The core of the SCM estimation lies in the definition of the weights. The approach in Abadie et al. (2021) is to optimize the weights with the aim of minimizing the distance between the treatment and control group in the pre-treatment period. Given the nonnegative constant vector $\mathbf{V} = (v_1, ..., v_k)$ that represent the relative importance of the $k$ predictors, the optimal weight vector $\mathbf{W}^*$ is the one that, for some positive constants $v_h$, minimizes:

$$||\mathbf{X}_1 - \mathbf{X_0}\mathbf{W}|| = \left( \sum_{h=1}^{k} v_h \left( X_{h1} - w_2 X_{h2} - \ldots - w_{J+1} X_{hJ+1} \right)^2 \right)^{1/2} \tag{4}$$

Subject to $\sum_{j=2}^{J+1} w_j$ and $w_j \geq 0$. That is, minimizing the distance between the treated unit $\mathbf{X}_1$ and the weighted combination of the control units $\mathbf{X_0}\mathbf{W}$. The output $\mathbf{W}^* = \left(w_2^*, ..., w_{J+1}^*\right)'$ is used in the estimation of the treatment effect for the treated unit across $t = T_0 + 1, ..., T$ as:

$$\hat{\tau}_{1t} = Y_{1t} - \sum_{j=2}^{J+1} w_j^* Y_{jt} \tag{5}$$

Any possible choice of weights in $\mathbf{V}$ produces a different set of optimal weights, which is effectively a different synthetic control $\mathbf{W}(\mathbf{V}) = (w_2(\mathbf{V}), ..., w_{J+1}(\mathbf{V}))'$. For this reason the choice of $\mathbf{V}$ is a key issue. An initial approach could be to divide the predictor weights equally across the $k$ predictors included in the model. However, two more elegant solutions are proposed by the authors.

A first proposed solution is to choose $\mathbf{V}$ such that the synthetic control $\mathbf{W}(\mathbf{V})$ minimizes the mean squared prediction error (MSPE) of the synthetic control with respect to $Y_{1t}^N$ over the pre treatment period $\mathcal{T}_0 = 1, ..., T_0$ :

$$\sum_{t \in \mathcal{T}} \left( Y_{1t} - w_2(V) Y_{2t} - ... - w_{J+1}(\mathbf{V}) Y_{J+1t} \right)^2 \tag{6}$$

Another solution is out-of-sample validation, which requires substantial pre treatment observations. This path exploits the observed pre treatment $Y_{1t}^N$ to gauge the predictive power of the variables $X_{1j}, ..., X_{kj}$ and assign coherent weights $\mathbf{V}$. To use out-of-sample validatio start by dividing the $\mathcal{T}_0$ period in a *training* period and a *validation* period. The lenghts of the

two periods may depend on data availability and frequency of measurement of outputs and variables. Then, for every value $\mathbf{V}$, compute the synthetic control weights on the training period and call them $\tilde{w}_2(V), ..., \tilde{w}_{J+1}(V)$. The MSPE over the validation period will be:

$$\sum_{t=t_1+1}^{T_0} (Y_{1t} - \tilde{w}_2(V)Y_{2t} - ... - \tilde{w}_{J+1}(V)Y_{J+1t})^2 \tag{7}$$

Then, compute $\mathbf{V}^* \in \mathcal{V}$ such that MSPE is minimized, with $\mathcal{V}$ being the set of all potential $\mathbf{V}$. Then, check the synthetic control's ability to emulate the treated unit behavior on the remaining validation period observations by using $\mathcal{V}^*$ to compute $\mathcal{W}^* = \mathcal{W}(V)^*$.
For an analysis of the shortcomings of cross-validation as defined in this section, see subsection 2.3.

### 2.2.1 Bias properties of SCM

Appendix B of Abadie et al. (2010) provides an analysis of the bias properties of synthetic control estimators in the case of a linear factor model and a vector autoregressive (VAR) model. According to their findings, the estimator's bias in a factor model scenario can be constrained by a function that goes to zero as the amount of pre-treatment periods increases. In an VAR scenario, the synthetic control estimator is unbiased.

maybe add something more detailed if you have time

### 2.2.2 Inference in SCM

Complete!!!

## 2.3 A detour on methodological advantages

SCM come with advantages relative to other competing methods. In this subsection, these features are emphasized with respect to linear regression estimators.

*Transparency and goodness of fit.* While both synthetic control estimation and regression estimators are applied to panel data, SCM are transparent relative to the discrepancy between the treated unit and the synthetic unit. Furthermore, the goodness of fit of the synthetic control unit can be easily evaluated through analysis of the pre tratment period. SCM should not be used if the fit in the pre treatment period is not satisfactory.

*Extrapolation prevention.* Another advantage of synthetic controls comes from the restrictions placed on the weights, which prevent extrapolation.

Instead, regression weights may lie outstide of the $[0, 1]$ interval (for a comparison of the outcomes of SCM and regression see Abadie, Diamond and Hainmueller (2015)).

*No need for post-treatment outcomes.* Synthetic control weights can be computed before the observation of post-treatment outcomes, so that a researcher may decide on the design of the study without knowing how it would affect the conclusions.

*Weight sparsity.* When estimating weights for the control group, regression estimators typically provide non zero weights to all control units. Instead, sythetic control weights are sparse. The contribution of each donor unit is straightforward and allows for a geometric interpretation of the sparse weights: the sythetic unit represents a point that lies in the convex hull generated by the donor units with non zero weights. Note that with many treated units the weights are not necessarily unique nor that sparse. Abadie and L'Hour (2019) offers a penalized version of SCM that provides unique and sparse weights under some conditions.

## 2.4 Contextual and data requirements for a credible application

Most of the data and contextual requirements for successful application of SCM are applicable to any other comparative case study. Following Abadie (2021), I will start from contextual requirements, and eventually move on to data requirements.

### 2.4.1 Contextual requirements

The first observation is on the volatility of outcomes. Excessive random noise in the outcome variable increases the likelihood of over-fitting, therefore it is advised to filter out such noise before implementation, otherwise the effect of treatment may be drowned out. This concern does not come from common volatility between the units, but from unit specific volatility.

The second observation pertains to the donor pool. For a unit to be a suitable candidate of the donor units, it should not be affected by or subject to the treatment provided to the unit of interest, and it should have common features with the treated unit (typically, units in a similar region and/or context). This entails the elimination from the donor pool of any unit which may have suffered a shock to the outcome that would not have happened in the absence of treatment in unit of interest.

add reflection on climate change shocks usually being common shocks

Third, a typical concern is anticipatory behavior by the units which would introduce bias in the SCM estimates. Depending on data avilability, a solution would be to backdate the period that marks the start of treatment. Since SCM allow for differential exposure to treatment across time, the initial periods that are barely affected by treatment may show very small effects, whereas the subsequent periods will show larger effects.

Spillover effects are another point of concern that stems from the selecion of an ideal donor unit. Usually, a donor unit is valid if it has common features with the treated one, while also being unaffected by the shock. As a consequence, donor units tend to come from the same regions or be exposed to the same context. Spillover effects are especially an issue if units come from the same geographical area. In order to understand the bias introduced by interference, the researcher may carry out and compare the estimates with and without the affected donor units. Moreover, due to the transparency of the synthetic counterfactual and the sparsity of weights, the researcher may reason as to the potential direction of the bias and account for that in their analysis.

### 2.4.2 Data requirements

For the method to accurately and credibly track the treated unit there must be availability of a window of pre-treatment observations from all units. Using large periods of time may present the issue of structural breaks in the pre-treatment window, affecting the structural stability of the model. Accuracy of prediction may suffer if we add enough observations, thus the predictive ability of the synthetic control fails.

Extensive post-treatment information is crucial for the evaluation of the effects, especially if they are expected to intensify or dissipate over time.

## 2.5 Issues and solutions for successful application

Similar to any other technique, the application of synthetic controls demands careful consideration of the proper data and contextual prerequisites that must be fulfilled, along with fulfilling the required conditions for its use to claim a causal explanation of outcomes. Various features are currently being debated and enhanced in the literature, while some areas remain unexplored. For instance, such an area is the investigation of computational aspects, which is still relatively under-researched and holds significant potential for improvement.

***Kaul et al. (2015) on the inclusion of lagged dependent variables***

The predictors in a synthetic control application can be lagged outcome variables and other economically relevant variables with explanatory power for the dependent. Generally, the lagged outcomes tend to have greater predictive power than the covariates included in the model. Although the intuitive approach would be to include all pre-treatment lagged outcome variables and all the covariates, doing so could nullify the effect of the covariates for both nested and regression optimization approaches. The potental bias arising from this mechanic is something to be mindful of when considering what to include in the model.

To illustrate this, write predictor matrices as:

$$X_1 = \begin{pmatrix} C_1 \\ Z_1 \end{pmatrix}, \; X_0 = \begin{pmatrix} C_0 \\ Z_0 \end{pmatrix} \tag{8}$$

Where $C_i$ represent covariates, and $Z_i$ represent the dependent variable elements. In a similar manner, rewrite matrix $V$ as:

$$V = \begin{pmatrix} V_C & 0 \\ 0 & V_Z \end{pmatrix} \tag{9}$$

So that $V_C$ contains the covariate weights, and $V_Z$ contains the lagged dependent variable weights.

For given predictor weights $V$, we can write the inner optimization to find $W(V)^*$ as:

$$\min_W \sqrt{(X_1 - X_0 W)' V (X_1 - X_0 W)} \tag{10}$$

And the outer optimization to find the optimal $V$ that minimizes the MSPE as:

$$\min_V (Z_1 - Z_0 W^*(V))' (Z_1 - Z_0 W^*(V)) \tag{11}$$

If we denote $W^{**}$ as:

$$W^{**} := \arg\min_W (Z_1 - Z_0 W)' (Z_1 - Z_0 W) \tag{12}$$

Then, for any $V$ we would have:

$$(Z_1 - Z_0 W^{**})' (Z_1 - Z_0 W^{**}) \leq (Z_1 - Z_0 W^*(V))' (Z_1 - Z_0 W^*(V)) \tag{13}$$

However, choosing $V$ such that $V_C$ is zero and $V_Z$ is the identity matrix yields:

$$W^*(V^*) = W^{**} \tag{14}$$

The implication is that the outer optimization would want null predictor weights even if the inner optimization would assign them positive weights. Thus, the synthetic control would end up ignoring the covariates. Consequently, even if positive $V$ elements are reported, they are effectively ignored. Bias emerges from this dynamic if the covariates are relevant explanatory elements of the dependent. To solve for this, they advise to exclude outcomes lags in the inner optimization so that positive weights emerge for other predictors.

The authors illustrate this by replicating a paper by Billmeier and Nannicini (2013), and highlight a stark difference in estimated treatment effects when using all pre-treatment dependent varibles observation as opposed to not doing so.

This dynamic may be exploited to understand the predictive power of predictors via the comparison of a full lagged dependent variable model and a reduced lagged dependent variable one. The discrepancy between results is indicative of whether to include such covariates.

### *Klößner et al. (2018) on the uniqueness of weights when utilizing cross-validation techniques*

Some authors noticed numerical instability in the assigned predictor weights when replicating results from Abadie and Gardeazabal (2003). Klößner et al. (2018) investigate this ambiguity by looking at validation weights. When using the out-of-sample validation technique, predictor weights are not necessarily uniquely defined, so that when replicating Abadie et al. (2015) the authors find different yet equivalent solutions for the weights depending on the software package used (specifically, R with the SYNTH package versus STATA) and on the specific donor units ordering (alphabetical versus custom).

Cross-validation requires the sectioning of the pre-treatment observations in a training period and validation period, and a two-step estimation procedure that computes predictor and donor weights. This is opposed to the standard one step estimation procedure of SCM and generates ambiguity on weights. The authors develop a rule of thumb to gauge the extent of this ambiguity in the estimates of the treatment effect that is centered on the difference $k - \alpha$, where $k$ is the number of predictors and $\alpha$ is the number of donor units that obtains positive weights in the training period. If $k - \alpha > 0$, the predictor weights are not unique and there will be proportionally increasing ambiguity.

On the other hand, the one-step procedure to estimate donor weights $W^*(V^*)$ provides a well-defined estimator that outputs unique donor weights.

The authors stress that the ill-defined cross-validation synthetic control estimator is no failure of the method as such, and that the ambiguity may be negligible, although a positive $k - \alpha$ is the typical scenario.

***Kuosmanen et al. (2021) on the true nature of numerical instability***

In contrast with the findings from Klößner et al. (2018) that point to the cross-validation method as culprit, Kuosmanent et al. (2021) ascribe numerical instability to the commonly used *Synth* package algorithm available for R, STATA and Matlab and the MSCMT algorithm which produce unstable and suboptimal weights that to converge to the existing optimum unique solution. Therefore, different ordering of the donors and predictors affect the results.

The authors stress that numerical instability is not the underlying flaw, but more a symptom. They show that the model has a tendency for corner solutions that is the true design flaw and is caused by joint optimization of donor and predictors weights. The computational complexity of what is an NP-hard bilevel optimization problem causes other packages to fail. Their proposed solution is to determine predictor and donor weights by applying a two-step algorithm that optimizes the donor weights when the predicor weights are given (note that the *Synth* package fails to derive optimal weights even when predictor weights are given), an approach that builds on previous work by Malo et al. (2020). The authors provide the updated code and technical documentation for this approach at their GitHub page (see p.4 of Kuosmanen et al. (2021) for the link).

An explicit restatement of the SCM problem provided by Malo et al. (2020) reveals the computational difficulty of problem. Following the paper's notation:

$$\min_{\mathbf{v},\mathbf{w}} L_V = \frac{1}{T^{pre}} \left(y_1^{pre} - Y_0^{pre} w\right)' \left(y_1^{pre} - Y_0^{pre} w\right) \tag{15}$$

subject to

$$\mathbf{w} = argmin L_W = (x_1 - X_0 w)\, \mathbf{V}\, (x_1 - X_0 w) \tag{16}$$

$$\mathbf{1}'\mathbf{w} = 1$$

$$\mathbf{1}'\mathbf{v} = 1$$

$$\mathbf{w} \geq 0, \mathbf{v} \geq 0$$

Where (15) is the outer optimization problem, and (16) is the inner optimization problem. $T^{pre}$ is the pre intervention period. This could be interpreted

as a social planner playing a Stackelberg game where the leader chooses **v** and the follower chooses **w** cooperatively.
The proposed solution is detailed next.

*Step 1*: Solve the quadratic optimization problem.

$$\min_{w} L_Q = (x_q - X_0 w)' V^* (x_1 - X_0 w)$$

subject to

$$\mathbf{1}'\mathbf{w} = \mathbf{1}$$

$$\mathbf{w} \geq \mathbf{0}$$

*Step 2*: Given the step 1 optimal solution for $L_W^*$, solve the convex optimization problem.

$$\min_{w} L_V = (y_1^{pre} - Y_0^{pre} w)' (y_1^{pre} - Y_0^{pre} w)$$

subject to

$$\mathbf{1}'\mathbf{w} = \mathbf{1}$$

$$\mathbf{w} \geq \mathbf{0}$$

This procedure allows for multiple optima in Step 1, unlike *Synth*.

***Ferman et al. (2020) on specification searching opportunities***

The authors pose the issue of specification searching behavior in synthetic control applications stemming from a lack of consensus about which specifications should be used in SCM applications which has lead to a wide variety of choices in the literature. If different model specifications lead to different synthetic control units, the researcher may be tempted to discretionarily select the ones that lead to stastically significant results. Via Monte Carlo and placebo simulations they find that the probability of detecting a false positive is decreasing in the number of available pre-treatment periods, but still large when the time window is much greater than what is usually used in the literature.

They explore opportunities for specification searching when selecting the number of pre-treatment periods to include in the model. In accordance with Kaul et al. (2015), they suggest to not include all pre-treatment periods if the model includes relevant covariates. If there are no covariates, the inclusion of all pre-treatment periods is the MSPE minimizing choice. The recommendation is to always include multiple sepcifications in synthetic control applications, and use the full pre-treatment period model as a benchmark.

A commonly used criterion to choose between different specifications is to select the ones that minimise the MSPE over the validation period. Some examples of this in practice are Dube and Zipperer (2015) and Donohue, Aneja and Weber (2019).

# 3 Applying SCM to climate change analysis

inizia con riflessione sulle implicazioni della parte DATA E CONTEXT REQUIREMENTS per le applicazioni allo studio del cambiamento climatico

# 4 Literature review and general applications in macro

Lorem ipsum

# 5 Potential applications to Fleurbeay project

Lorem ipsum

# 6 Data sourcing and explanations

Lorem ipsum

# 7 Conclusion

Lorem ipsum

# 8 References


- Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. "Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program." Journal of the American statistical Association 105.490 (2010): 493-505.

- Abadie, Alberto, and Javier Gardeazabal. "The economic costs of conflict: A case study of the Basque Country." American economic review 93.1 (2003): 113-132.

- Abadie, Alberto. "Using synthetic controls: Feasibility, data requirements, and methodological aspects." Journal of Economic Literature 59.2 (2021): 391-425.

- Abadie, Alberto, and Jérémy L'hour. "A penalized synthetic control estimator for disaggregated data." Journal of the American Statistical Association 116.536 (2021): 1817-1834.

- Acemoglu, Daron, et al. "The value of connections in turbulent times: Evidence from the United States." Journal of Financial Economics 121.2 (2016): 368-391.

- Athey, Susan, and Guido W. Imbens. "The state of applied econometrics: Causality and policy evaluation." Journal of Economic perspectives 31.2 (2017): 3-32.

- Doudchenko, Nikolay, and Guido W. Imbens. "Balancing, regression, difference-in-differences and synthetic control methods: A synthesis." No. w22791. National Bureau of Economic Research, 2016.

- Dube, Arindrajit, and Ben Zipperer. "Pooling multiple case studies using synthetic controls: An application to minimum wage policies." (2015).

- Donohue, John J., Abhay Aneja, and Kyle D. Weber. "Right-to-carry laws and violent crime: A comprehensive assessment using panel data and a state-level synthetic control analysis." Journal of Empirical Legal Studies 16.2 (2019): 198-247.

- Malo, Pekka, et al. "Computing Synthetic Controls Using Bilevel Optimization." (2020).

- Kaul, Ashok, et al. "Synthetic control methods: Never use all pre-intervention outcomes together with covariates." (2015).

- Klößner, Stefan, et al. "Comparative politics and the synthetic control method revisited: A note on Abadie et al.(2015)." Swiss journal of economics and statistics 154 (2018): 1-11.