# SCM application to environmental analysis

Jessica Cremonese

March 2023

# Contents

**Abstract**

Lorem ipsum

# 1 Introduction

Lorem ipsum

# 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. 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 $Y^N_{jt}$. For the trated unit $j = 1$, define the potential response under the treatment as $Y^I_{jt}$ in the post treatment period $t > T_0$. The effect of the intervetion for the affected unit $j = 1$ for $t > T_0$ is:

$$\tau_{1t} = Y^I_{1t} - Y^N_{1t}$$

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:

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

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

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 minimizes:

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

Subject to $\sum_{j=2}^{J+1} w_j$ and $w_j \geq 0$. 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:

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

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 $\mathcal{T}_0 = 1, ..., T_0$ :

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

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:

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

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

Abadie and Hainmueller (2010) provide an analysis of the bias properties of synthetic control estimators in the case of a linear factor model and a vector autoregressive (VAR) model.

Complete!!!

### 2.2.2 Predictor selection

Complete!!!

## 2.3 Issues of SCM - Klößner (2018)

An important note must be made about the validation weights, as noted by Klößner et al. (2018). Predictor weights are not 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 (STATA versus R) and on the specific donor units ordering (alphabetical versus custom) when using out-of-sample validation.

- numerical instability - see Kuosmanen 2021:1 and 2021:2 in OSE JDX mafia project

# 3 Literature review and general applications in macro

Lorem ipsum

# 4 Potential applications to Fleurbeay project

Lorem ipsum

# 5 Data sourcing and explanations

Lorem ipsum

# 6 Conclusion

Lorem ipsum

# 7 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.

- 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.

- 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.