Predictor Selection for Synthetic Controls

REDD+ and carbon offsets

Jaume Vives-i-Bastida

Massachusetts Institute of Technology

Motivation

Contribution: propose a new penalized synthetic control method for policy evaluation.

- **Variable Selection**: identify which predictors should not be used in building the synthetic control.
  - Allows researchers to not have to search for predictors.
- **Performance**: achieves lower bias and MSE in sparse settings.
- **Just for this workshop**: REDD+ and carbon offsets!

Outline:

1. Overview of Synthetic Controls.
2. Related Literature.
3. The Sparse Synthetic Control.
4. Variable Selection Result.
5. Simulation Study.
6. Empirical application.

Synthetic Controls are a method to estimate the effects of large scale interventions using aggregate data.

- We observe $J + 1$ units for T periods.
- There is an **aggregate intervention** that affects unit one during periods $T_0 + 1, \dots, T$.
- The other J unaffected units are our **donor** pool.
- **Outcome** variable Y_{it} with potential outcomes N, I.
- **Predictors**: $k \times (J + 1)$ matrix $\mathrm{X} = [\mathrm{X}_1, \mathrm{X}_0]$ of pre-intervention characteristics of the units.

We are interested in a **TET** for $t > T_0$:

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

Synthetic Controls Example I

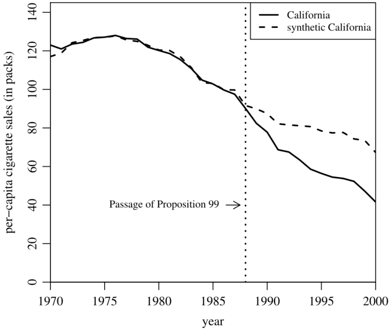
A classic example in Abadie et al. 2010 is the passage of proposition 99 in California.

- The donor units are the other states.
- The predictors are important variables for cigarette consumption.

Table 1. Cigarette sales predictor means

Variables	California Real	California Synthetic	Average of 38 control states
Ln(GDP per capita)	10.08	9.86	9.86
Percent aged 15–24	17.40	17.40	17.29
Retail price	89.42	89.41	87.27
Beer consumption per capita	24.28	24.20	23.75
Cigarette sales per capita 1988	90.10	91.62	114.20
Cigarette sales per capita 1980	120.20	120.43	136.58
Cigarette sales per capita 1975	127.10	126.99	132.81

NOTE: All variables except lagged cigarette sales are averaged for the 1980–1988 period (beer consumption is averaged 1984–1988). GDP per capita is measured in 1997 dollars, retail prices are measured in cents, beer consumption is measured in gallons, and cigarette sales are measured in packs.



Synthetic Controls Example II

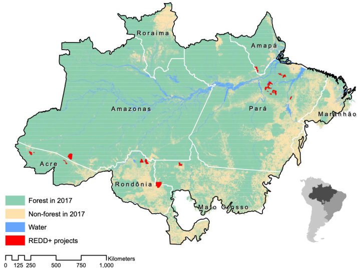
Recent media attention on carbon offsets impact on reducing deforestation using SC (The Guardian).

- Thales et al. 2020 (PNAS) compare regions with **REDD+** (reducing emissions from deforestation and forest degradation) projects with control regions.
- **Outcome**: cumulative deforestation (sq. kms).
- **Predictors**: soil, infrastructure, agriculture, hydrology etc. (up to 18)

Revealed: more than 90% of rainforest carbon offsets by biggest certifier are worthless, analysis shows

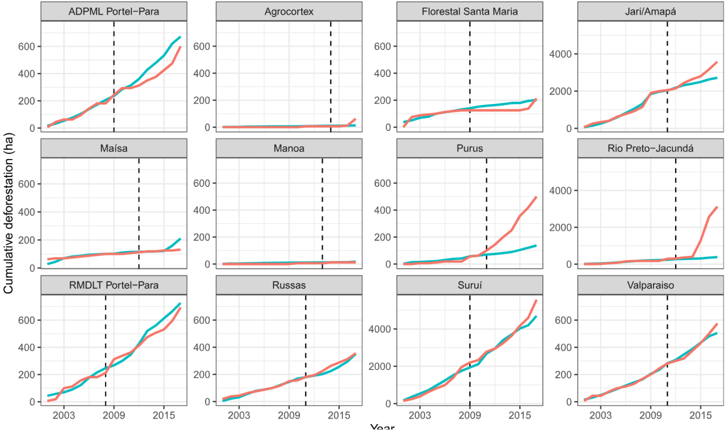
Investigation into Verra carbon standard finds most are 'phantom credits' and may worsen global heating

- 'Nowhere else to go': Alto Mayo, Peru, at centre of conservation row
- Greenwashing or a net zero necessity? Scientists on carbon offsetting
- Carbon offsets flawed but we are in a climate emergency



Synthetic Controls Example II

Thales et al. 2020 find that in general the **REDD+ projects did not decrease deforestation**.



How to build Synthetic Controls?

A **synthetic control** is defined by a weight vector $\mathbf{W} = (W_2, \ldots, W_{J+1})'$ such that $\sum_j W_j = 1$ and $W_j \geq 0$.

- We choose $\mathbf{W}$ to minimize:

$$\|\mathbf{X}_1 - \mathbf{X}_0\mathbf{W}\|_V = \left(\sum_{h=1}^{k} v_h (X_{h1} - W_2 X_{h2} - \cdots - W_{J+1} X_{hJ+1})^2 \right)^{1/2},$$

  subject to the weight constraints.
- Intuitively, the **W weights** recreate the treated unit in the predictor space.
- **Predictor Weights**: The researcher can choose $v_1, \ldots, v_k$ or use a data-driven procedure.

Synthetic control **estimator** for $t > T_0$:

$$\hat{\tau}_1 = Y_{1t} - \sum_{j=2}^{J+1} W_j^* Y_{jt}.$$

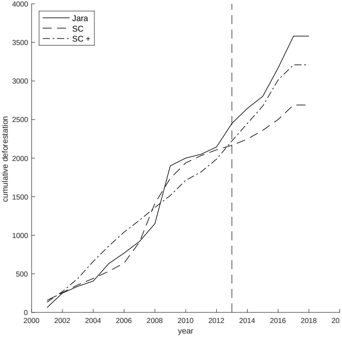
So what is the problem?

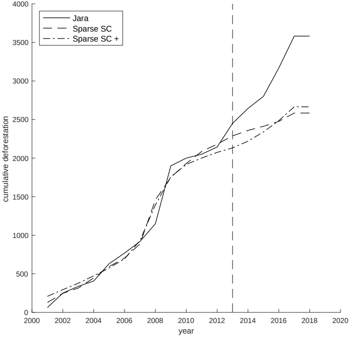
- **The choice of predictor set matters**: like OVB if we don't match relevant predictors the SC is **biased**!
- **The matching problem may be hard**: the more predictors we have to match the worse the **finite sample** properties of SC.
- **Predictor choice** opens the door for **specification search**.

Questions: How do you choose predictors? Can I just put them all in? What about interactions? What about time-varying covariates?

Synthetic Control Example II

- 18 predictors vs. 172 interactions ('+')
- *Sparse* Synthetic Control is **robust** to predictor size





- Training set $(\mathbf{X}_0^{train}, \mathbf{X}_1^{train}, \mathbf{Y}_0^{train}, \mathbf{Y}_1^{train})$ for $t \in \{1, \ldots, T_v\}$.
- Validation set $(\mathbf{X}_0^{val}, \mathbf{X}_1^{val}, \mathbf{Y}_0^{val}, \mathbf{Y}_1^{val})$ for $t \in \{T_v + 1, \ldots, T_0\}$.

The **Sparse Synthetic Control** solves

- Upper level problem:

$$(\mathbf{V}^*, \mathbf{w}^*) \in \text{argmin}_{\mathbf{V},\mathbf{w}} L_V(\mathbf{V}, \mathbf{w}, \lambda) = \frac{1}{T_{val}} \|\mathbf{Y}_1^{val} - \mathbf{Y}_0^{val}\mathbf{w}(\mathbf{V})\|^2 + \lambda \|\mathbf{V}\|_1,$$
$$\text{s.t. } \mathbf{w}(\mathbf{V}) \in \psi(\mathbf{V}), V \in \mathbb{R}_+^K.$$

- Lower level problem:

$$\psi(\mathbf{V}) \equiv \text{argmin}_{\mathbf{w} \in \mathcal{W}} L_W(\mathbf{V}, \mathbf{w}) = \|\mathbf{X}_1^{train} - \mathbf{X}_0^{train}\mathbf{w}\|_V^2,$$

where,

$$\mathbf{w} \in \mathcal{W} \equiv \left\{ \mathbf{w} \in \mathbb{R}^J \mid \mathbf{1}'\mathbf{w} = 1,\ w_j \geq 0,\ j = 2, \ldots, J+1 \right\}$$

The Sparse Synthetic Control II

Algorithm 0: *Sparse* Synthetic Control

Result: $\mathbf{w}^*, \mathbf{V}^*$

Data: $(\mathbf{X}_0^{train}, \mathbf{X}_1^{train}, \mathbf{Y}_0^{train}, \mathbf{Y}_1^{train})$, $(\mathbf{X}_0^{train}, \mathbf{X}_1^{train}, \mathbf{Y}_0^{val}, \mathbf{Y}_1^{val})$

set $v_{k_0} = 1$;

initialize v_k for $k \neq k_0$ to $(\mathbf{X}_0^{train\prime} \mathbf{X}_0^{train})^{-1}$;

for *each* λ *in a grid* **do**

  get $(\mathbf{V}_\lambda, \mathbf{w}_\lambda)$ by jointly minimizing $L_W(\mathbf{V}, \mathbf{w}, \lambda)$ and $L_V(\mathbf{V}, \mathbf{w})$ for the training data;

    s.t. $\mathbf{w} \in \mathcal{W}$, $v_k \geq 0\ \forall k \neq k_0$ and $v_{k_0} = 1$;

  scale $\mathbf{V}_\lambda$ to $[0, 1]$;

  get $\mathbf{w}_\lambda^*$ by minimizing $L_W(\mathbf{V}_\lambda, \mathbf{w}, \lambda)$ for the training data;

  store $\text{MSE}(\mathbf{Y}_1^{val}, \mathbf{Y}_0^{val}\mathbf{w}_\lambda^*)$ and $\mathbf{V}_\lambda$;

end

choose λ^* with minimum $\text{MSE}(\mathbf{Y}_1^{val}, \mathbf{Y}_0^{val}\mathbf{w}_\lambda^*)$;

$\mathbf{V}^* = \mathbf{V}_{\lambda^*}$;

get $\mathbf{w}^*$ by minimizing $L_V(\mathbf{V}_\lambda^*, \mathbf{w})$ for the *shifted* training data.[a]

- **Classic synthetic controls**: Abadie and Gardeazabal (2003), Abadie, Diamond and Hainmueller (2010, 2015).
- **About the donor weights**:
  - **Dis-aggregated synthetic controls**: Abadie and L'Hour (2019), Athey et al. (2018), Gunsilius (2020), Gardeazabal and Vegayo (2017).
  - **Penalized synthetic Controls**: Abadie and L'Hour (2019), Doudchenko and Imbens (2017), Chernozhukov et al. (2019a), Arkhangelsky et al. (2019), Quistorff et al. (2020).
- **About the predictor weights**: Klosner et al. (2018), Abadie (2020), Ben-Michael et al. (2018).
- **Model selection**: Pouliot and Xie (2021).

$\Longrightarrow$ Focus: How to choose the V weights to improve **performance** and do **variable selection**.

Linear factor model

$$Y_{it}^N = \delta_t + \boldsymbol{\theta}_t \mathbf{Z}_i + \boldsymbol{\lambda}_t \boldsymbol{\mu}_i + \epsilon_{it}.$$

- $\mathbf{Z}_i$ is a $(k \times 1)$ vector of observed features.
- $\boldsymbol{\lambda}_t$ is a $(1 \times F)$ vector of unobserved common factors.

Sparse representation:

- $\boldsymbol{\theta}_t$ is partitioned conformably into $(\tilde{\boldsymbol{\theta}}_t, \mathbf{0})'$ where $\tilde{\boldsymbol{\theta}}_t$ is a $(k_1 \times 1)$ vector of non-zero parameters.
- $Z_i = (Z_i^1, Z_i^2)$, where $\boldsymbol{Z}_i^2$ is $k_2 \times 1$ vector such that $k = k_1 + k_2$.

Variable selection is important because:

1. Only using the "useful" predictors improves fit and lowers bias.
2. Researchers need not choose predictors (specification search).

Variable Selection II

Oracle covariate match: For fixed J, let the oracle weights be defined by

$$w^* \in \text{argmin}_{w \in \Delta^J} \mathbb{E}\|Y_1 - Y_0 w\|^2.$$

We consider two assumptions:

1. For all $k \in S = \{k \mid \theta_{tk} = 0 \text{ for all } t\}$, $|Z_{1k} - Z'_{Jk} w^*| > 0$.
2. (1) holds true and for $l \in S^c$, $|Z_{1l} - Z'_{Jl} w^*| = 0$.

Theorem: Variable Selection

Under technical assumptions if ψ is an injective function and $\hat{\lambda} \to 0$ as $T_0 \to \infty$, for a fixed k and J, as $T_0 \to \infty$ the following holds

1. If $k \in S = \{k \mid \theta_{tk} = 0 \text{ for all } t\}$, then $P(v_k = 0) \to 1$.
2. If (2) holds and $l \in S^c$ then $P(v_l = 0) \to 0$.

where v_k is the predictor weight for predictor m assigned by the sparse synthetic control algorithm.

Why predictors matter

$$\hat{\tau}_{1t}^{w} - \tau_{1t} = \boldsymbol{\theta}_t' \left(Z_1 - \sum_{j=2}^{J+1} w_j Z_j \right) + \boldsymbol{\lambda}_t' \left(\boldsymbol{\mu}_1 - \sum_{j=2}^{J+1} w_j \boldsymbol{\mu}_j \right) + \sum_{j=2}^{J+1} w_j (\epsilon_{1t} - \epsilon_{jt}).$$

Under technical assumptions:

$$\mathbb{E}|\hat{\tau}_{1t}^{w} - \tau_{1t}| \leq \frac{\gamma}{T_0} \sum_{m=1}^{T_0} \mathbb{E}|Y_{1m} - \sum_{j=2}^{J+1} w_j Y_{jm}| + \left| \bar{\theta} \left(1 - \frac{\gamma}{T_0} \right) \right| \sum_{k=1}^{k_1} \mathbb{E}|Z_{1k}^1 - \sum_{j=2}^{J+1} w_j Z_{jk}^1| + O\left(T_0^{-1} \right).$$

So, the SC bias is bounded above by:

1. Expected pre-treatment fit (rule of thumb).
2. Expected predictor fit! (like OVB)

MSE Rates

Let $\boldsymbol{Z}_1 = \boldsymbol{Z}_0 w^* + \boldsymbol{u}$ for $u_i \sim_{ind} \text{subG}(\sigma_z^2)$. Then, under technical assumptions as $T_0 \to \infty$, almost surely for the sparse synthetic control $\hat{\boldsymbol{w}}$,

$$MSE(\boldsymbol{Z}_1, \boldsymbol{Z}_0\hat{\boldsymbol{w}}) = \frac{1}{k}\|\boldsymbol{Z}_1 - \boldsymbol{Z}_0\hat{\boldsymbol{w}}\|^2 \lesssim \frac{\sigma_z\sqrt{k_1}}{k}\sqrt{2\log J}.$$

For the standard synthetic control $\tilde{\boldsymbol{w}}$,

$$MSE(\boldsymbol{Z}_1, \boldsymbol{Z}_0\tilde{\boldsymbol{w}}) = \frac{1}{k}\|\boldsymbol{Z}_1 - \boldsymbol{Z}_0\tilde{\boldsymbol{w}}\|^2 \lesssim \sigma_z\sqrt{\frac{2\log J}{k}}.$$

1. In sparse settings, the **MSE rate** for the Sparse SC is faster than the standard SC!
2. More precise estimation, lower s.e. (not easy to compute).

Simulation Study I

We compare three synthetic control estimators:

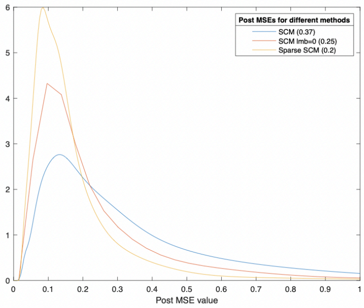
1. The standard synthetic control (**SCM**).
2. The SCM with choosing V to minimize the validation fit (**SCM** $\lambda = 0$).
3. The Sparse synthetic control (**Sparse SCM**).

Under the following setting:

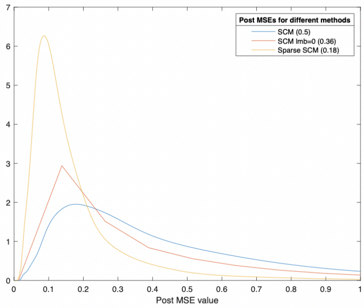
$$
\begin{aligned}
& T = 30,\ T_0 = 20,\ T_v = 10, \\
& \delta_t = 100, \\
& \boldsymbol{Z}_i = [\boldsymbol{Z}_i^1,\ \boldsymbol{Z}_i^2],\ \text{where } \boldsymbol{Z}_i^1, \boldsymbol{Z}_i^2 \sim_{iid} U[0,1], \\
& \boldsymbol{Z}_1^1 = \frac{1}{2}\boldsymbol{Z}_2^1 + \frac{1}{2}\boldsymbol{Z}_3^1, \\
& \lambda_t \text{ follows an } AR(1) \text{ with coefficient } \rho = 0.5, \\
& \epsilon_{it} \sim N(0, \sigma^2) \text{ with } \sigma = 0.25, \\
& F = 7 \text{ in groups of 3 units and } J + 1 = 21, \\
& k_1 = k_2 = 5 \text{ and } k_1 = 1, k_2 = 9, \\
& \boldsymbol{X} \text{ also includes 10 lags.}
\end{aligned}
$$

Simulation Study II - MSEs

- Smaller and more concentrated post-treatment MSEs.
- Improvement larger when k_1 small with respect to k_2.



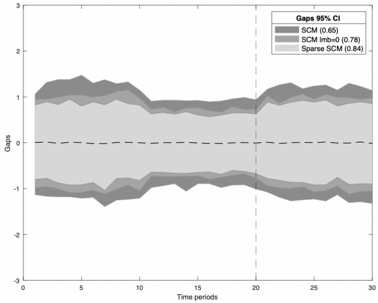
(a) $k_1 = k_2 = 5$.



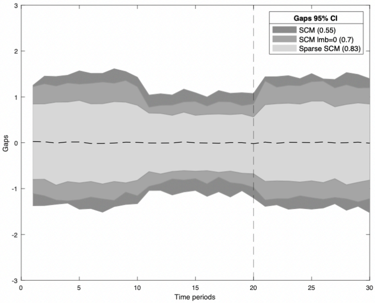
(b) $k_1 = 1,\ k_2 = 9$.

Simulation Study III - Gaps

- Better pre-treatment fit.
- Less over-fitting and closer to optimal.

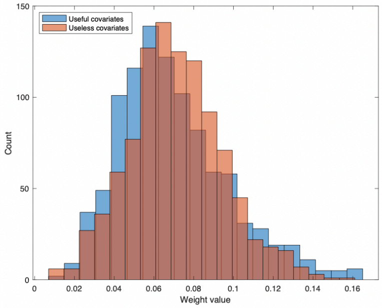


(c) $k_1 = k_2 = 5.$

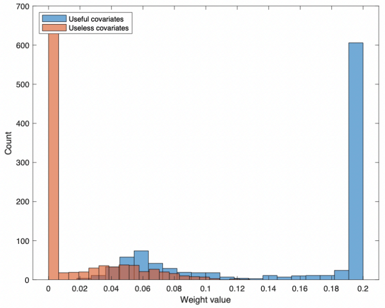


(d) $k_1 = 1,\ k_2 = 9.$

- Plot for $k_1 = k_2 = 5$.
- **Sparse** SCM distinguishes between types of predictors.

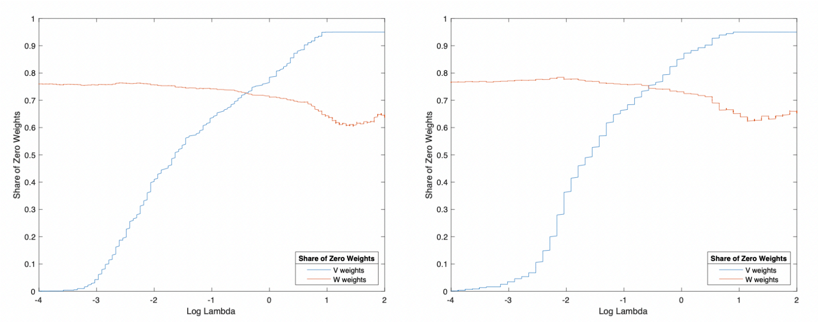


(e) SCM $\lambda^* = 0$ $\boldsymbol{V}^*$



(f) *Sparse* SCM $\boldsymbol{V}^*$

- Evidence that $\psi(V)$ is well behaved.



(a) $k_1 = k_2 = 5$.

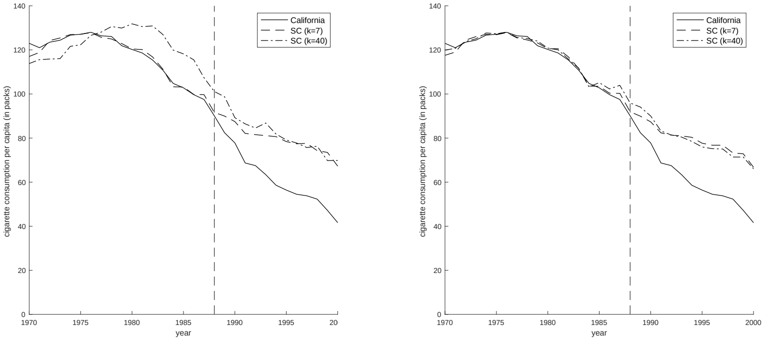
(b) $k_1 = 1$, $k_2 = 9$.

Empirical Application I

California Proposition 99: In 1988 California increased the cigarette excise tax by 25 cents per pack and shifted public policy towards a clean air agenda.

- **Compare** DID, SCM $\lambda = 0$ and Sparse SCM.
- With **augmented predictors**: 50 additional predictors from the IPPSR (MSU) dataset on policy correlates. These include demographic variables, income related variables, political participation measures and government spending statistics.

- 7 vs. 40 predictors (including garbage predictors)
- *Sparse* Synthetic Control is **robust** to predictor size



	DID	SCM	*Sparse* SCM	SCM	*Sparse* SCM
$\hat{\tau}$ estimate	-27.4	-18.9	-18.5	-21.0	-18.2
$\hat{V}_{\tau}^{1/2}$	(16.7)	(13.2)	(12.2)	(12.9)	(11.7)
k	-	7	7	40	40

Notes: variance calculated using the placebo bootstrap.

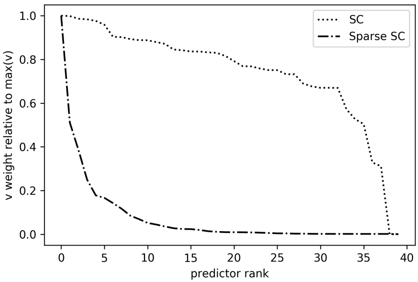
Takeaways:

- DID is badly biased (parallel trends violated).
- In the non-augmented setting SCM and Sparse SCM are similar.
- In the augmented setting the Sparse SCM does not over-fit.
- Sparse SC has lower variance (8% - 10%).

(a) Top 7 predictors

SCM	*Sparse* SCM
smk_80	smk_75
general_revenue_inc	incshare_top
smk_75	smk_88
smk_88	pc_inc_ann
loginc	region
general_expenditure_inc	budget_surpl
pc_inc_ann	taxes_gsp

(b) Predictor weight distribution



Takeaways:

- Sparse SC is more sparse.
- Sparse SC recovers the original predictors of ADH 2010.

Conclusion

Recap:

- **What goes into the synthetic control matters!**
- **Variable selection** can be achieved using a simple penalized procedure.
- Benefits of automatic variable selection:
  1. Avoid predictor search.
  2. Improve performance and interpretability.

Future work:

- Relax theoretical assumptions.
- R package.

Other projects:

- Uniform risk consistency of shrinkage estimators.
- Bayesian and Frequentist Inference for SC as $J, T_0 \to \infty$.
- Bagged polynomial regression as an alternative for neural networks.
- Synthetic controls for experimental design.