SMaC: Spatial Matrix Completion method

SMaC: Spatial Matrix Completion method

Giulio Grossi[a], Alessandra Mattei[a], and Georgia Papadogeorgou[b]

[a]Viale Morgagni 59, 50134, Florence, Italy; giulio.grossi@unifi.it,
alessandra.mattei@unifi.it,
[b]102 Griffin-Floyd Hall, 32611, Gainesville, Florida; gpapadogeorgou@ufl.edu,

Abstract

Synthetic control methods are commonly used in panel data settings to evaluate the effect of an intervention. In many of these cases, the treated and control time series correspond to spatial areas such as regions or neighborhoods. Synthetic control methods can be used to evaluate the effect that the treatment had in the treated area, but it is often unclear how far the treatment's effect propagates, as this approach ignores the spatial structure of the data, and can lead to efficiency loss in spatial settings. We propose to deal with these issues by developing a Bayesian spatial matrix completion framework that allows us to predict the missing potential outcomes in the different areas around the intervention point while accounting for the spatial structure of the data. Specifically, the missing time series in the absence of treatment for the treated areas of all sizes are imputed using a weighted average of control time series, where the weights are assumed to vary smoothly over space according to a Gaussian process.

Keywords: Causal inference, Spatial Econometrics, Synthetic Control Method, Gaussian Process, Potential Outcomes

1. Introduction

The Synthetic Control method (SCM hereinafter) is a widespread methodology to estimate causal effects in presence of a single treated unit and many control units, observed over time (2). With this method, the impact of an intervention is evaluated as the difference between the observed value of some primary outcome and its counterfactual value, imputed by using a weighted average of control units.

Evidence of interest in SCM is the flurry of methodological developments. Recently, the exploration of SCM alternatives heads toward Bayesian regression models. (6), (5), (7) and (8) use Bayesian methods for causal effects estimation, illustrating a simple and effective proposal for inference in SCM-like settings. Finally, recent work from (3) investigates the use of multitask Gaussian Processes for weights estimations. In Many fields where SCM is commonly used study outcomes which are measured in spatial areas such as municipalities, states or regions. (1) suggests these as the specific framework of application for SCM-like methods. In such contexts, it is common to see treatment assigned to a single area, and the focus being to estimate the treatment effect on this treated unit. Usually, scholars consider no second-round effects from the treatment, neither in terms of spillovers nor in terms of effect propagation. However, no previous work has addressed spatial treatment effect propagation explicitly within the scope of SCM. In practice, researchers often evaluate the extent to which treatment effects propagate through space by applying SCM to areas of different sizes around the treated location. In this work, we propose a Bayesian estimator for missing potential outcomes in presence of spatial correlation among treated units. We exploit a Gaussian process prior for the vertical regression coefficients that take into account spatial correlation, encouraging regression coefficients across similar areas to be similar. We aim to exploit this spatial information to estimate counterfactual quantities that are still unbiased, but have improved properties in terms of mean bias and mean square error of the point estimate with respect to the separated SCM or vertical regression methods. We refer to this method as *Spatial Matrix Completion* or SMaC. Our motivating application is the impact evaluation arising from the construction of the first line of the Florentine tramway network. In particular, we wish to assess the infrastructural impact on the commercial vitality of the treated neighbourhood, measured as the number of stores located within some distance d from a tramway stop.

2. Causal Framework

Consider a space Ω that can be partitioned into N areas, indexed in $i \in \mathbf{N} = \{1, \ldots, i, \ldots, N\}$, such that $\bigcup_{i=1}^{N} \Omega^i = \Omega$ and $\Omega^i \cap \Omega^j = \emptyset$ for each couple $i, j \in \{1, \ldots, N\}$. In our study, we consider the natural partition of our sample space into the clusters representing the Florentine neighbourhoods. We observe treatment arising from some specific locations $\boldsymbol{\omega}_1 \in \Omega_1$. We can consider treatment locations as a point treatment (e.g.: pollution created by a power plant), a linear treatment or even a polygonal treatment. Let be $\boldsymbol{\omega}^1$ the set of treatment locations, in our application we consider $\boldsymbol{\omega}^1$ as the tramway stops located in the treated area. We also consider sets of locations $\boldsymbol{\omega}^i, i \in \{2, \ldots, N\}$ as sets of locations located in neighbourhoods located far away from the tramway line, in streets similar to the one that receives the treatment. We define our observation units as the areas around the treatment sites $\boldsymbol{\omega}$. Therefore, for each neighbourhood i we construct a set of buffers areas $\boldsymbol{A}_i = \{A_i^1, \ldots, A_i^d, \ldots A_i^H\}$ around the treatment locations $\boldsymbol{\omega}_i$, using the vector of distances $\mathbf{D} = (d_1, \ldots, d_h, \ldots, d_H)$ representing the distance of the $h-$th area from the treatment site. We sort units and distances such that $d_{h+1} \geq d_h \quad \forall h \in (1, 2, \ldots, H-1)$. Let also d denote a generic distance between a treated area and the treatment site. We repeatedly observe units over time, so we consider a panel data setting, with $H \times N$ areas observed for $T^0 = (1, \ldots, t_0 - 1)$ pre-treatment periods, and $T^1 = (t_0, \ldots, T)$ post-treatment periods. Let $Y_{i,t}^d$ be our primary outcome, the number of stores in neighbourhood i within distance d from the tramway stops in each time period $t \in T$. Let be $\boldsymbol{z} = \{z_i^d\}_{i \in N}^{d \in D} \quad z_i^d \in \{0, 1\}$ be a neighbourhood-level treatment for each area A_i^d considered. Thus following, units belonging to the same cluster i can be only treated or not-treated together. We consider two alternative situations for $\boldsymbol{z}$: $\boldsymbol{z}^1$ is the scenario in which each area $A_1^d \in \Omega_1$ receives the treatment, and no one outside. Instead, $\boldsymbol{z}^0$ represents the scenario in which no area results treated, in our scenario the situation in which the tramway was never built in Florence. We consider that areas $\boldsymbol{A}_1 = \{A_1^1, \ldots, A_1^d, \ldots, A_1^H\} \in \Omega_1$ will receive the treatment starting from the period t_0, and remain treated afterwards. In our application, we consider the treated space as the Legnaia neighbourhood in which the tramway stops are located, and t_0 =2006. Units located in other part of Florence will be considered non-treated units with $\boldsymbol{A}^0 = \{A_2^1, \ldots, A_i^d, \ldots, A_N^H\} \notin \Omega_1$. We adopt the potential outcome approach to causal inference (9). Under consistency assumption, for each unit A_i^d in each period t we define the following couple of potential outcomes: $Y_{i,t}^d(1) \equiv Y_{i,t}(\boldsymbol{z}^1)$ as the potential outcome under $\boldsymbol{z}^1$ assignment and $Y_{i,t}^d(0) \equiv Y_{i,t}(\boldsymbol{z}^0)$ as the potential outcome under $\boldsymbol{z}^0$ assignment. In contexts with cluster-level treatment allocation, scholars often invoke a partial interference assumption (10), which rules that interference may occur, but not within groups. Moreover, we exploit the *non-anticipating treatment* assumption to rule out anticipatory effects. We define the causal effect for the treated units as

$$\Delta_{1,t}^d = Y_{1,t}^d(\boldsymbol{z}^1) - Y_{1,t}^d(\boldsymbol{z}^0) \quad \forall t \in T^1, d \in \mathbf{D} \tag{1}$$

For the treated units we observe $Y_{1,t}^d = Y_{1,t}^d(\boldsymbol{z}^1)$ when $t \geq t_0$, so we need to impute the missing quantity $Y_{1,t}^d(\boldsymbol{z}^0)$. From the comparison of effects at different distances from the treatment site, we can get precious insights into the transmission of treatment effects through space. In general, we could expect decaying treatment effects up to some boundary of spatial treatment.

3. Estimation of causal effects

One might be interested in understanding the effect that treating the specific location ω_i had on the area comprised within a specific distance $d \in \mathbf{D}$ versus not treating it. To do this, they can use synthetic control methodology. Specifically, one can find $\beta_{0d} \in \mathbb{R}$ and $\boldsymbol{\beta}_d = (\beta_{2d}, \dots, \beta_{Nd})^T \in \mathbb{R}^{N-1}$ such that:

$$\begin{pmatrix} \beta_{0d} \\ \boldsymbol{\beta}_d^T \end{pmatrix} = \underset{\boldsymbol{\beta}_d \in \mathbb{R}^N}{\operatorname{argmin}} \left\{ \sum_{t=1}^{t_0 - 1} \left(Y_{1,t}^d - (1\ \mathbf{Y}_{i,t}^T)^T \boldsymbol{\beta}_d \right)^2 \right\}. \tag{2}$$

The synthetic control weights and vertical regression coefficients can be calculated separately for different choices of $d \in (d_1, d_H)$. For example, to find the synthetic control weights at distances $d_1 < d_2 < \cdots < d_H$, one could solve the minimization problem in 2 using a constrained optimization procedure, separately for each of these distances. Alternatively, the H different minimization problems could be stacked, and one could solve the combined minimization problem

$$\begin{pmatrix} \beta_{0d_1} \\ \boldsymbol{\beta}_{d_H}^T \\ \beta_{0d_2} \\ \boldsymbol{\beta}_{d_2}^T \\ \vdots \\ \beta_{0d_H} \\ \boldsymbol{\beta}_{d_H}^T \end{pmatrix} = \underset{\boldsymbol{\beta}_0, \boldsymbol{\beta}_2, \dots, \boldsymbol{\beta}_N \in \mathbb{R}^N}{\operatorname{argmin}} \left\{ \sum_{d=1}^{H} \sum_{t=1}^{T_0 - 1} \left(Y_{1,t}^d - (1\ \mathbf{Y}_{i.t}^T)^T \boldsymbol{\beta}_i \right)^2 \right\} \tag{3}$$

which will return the exact same solutions as solving 2 separately for each distance. Thus we can obtain weights that minimise the pre-treatment distance between treated unit and the synthetic control, but ignore the spatial structure of data.

Exploting the spatial structure of data, we introduce the Bayesian framework we will use to impute the missing outcome $Y_{1,t}^d(\boldsymbol{z}^0)$. Building on the vertical regression idea (2, (4)) we will propose a matrix completion algorithm that smooths regression coefficient values through contiguous treated units.

In order to consider the spatial structures of the observed treated units, yet being flexible in the parameter estimation, we follow a Bayesian regression approach to solve the optimization problem in 3, using Gaussian processes as priors for control unit coefficients. In our setting, Gaussian processes can be particularly useful, as we could exploit the spatial information in our data for the specification of regression coefficients. We consider β varying smoothly through space, in particular, the vector of coefficients β^d will be more similar for physically close units. We specify such structure by using a Gaussian process prior for β such that

$$\boldsymbol{\beta}_i(\mathbf{D}) \sim \mathcal{GP}\left(\mathbf{0}, \mathcal{K}_{\alpha_i,\rho_i}(\mathbf{D})\right)$$

with $\mathcal{K}_{\alpha_i,\rho_i}(\mathbf{D})$ as a quadratic exponential smoothing kernel with parameter ρ_i. Thus, the (p, q) entry of $\mathcal{K}_{\alpha_i,\rho_i}(\mathbf{D})$ is

$$[\mathcal{K}_{\alpha_i,\rho_i}(\mathbf{D})]_{pq} = \alpha_i \exp\left\{ -\frac{(d_p - d_q)^2}{2\rho_i^2} \right\}$$

As stated above, other kernel specifications are possible, in order to consider different correlation structures between the treated units. To solve the pooled regression problem in 3, we specify the total vector of coefficients $\boldsymbol{\beta} = (\boldsymbol{\beta}_2, \dots, \boldsymbol{\beta}_i, \dots, \boldsymbol{\beta}_N)$ with $\boldsymbol{\beta} \sim \mathcal{MVN}(0, \Sigma)$, where Σ is an appropriate block covariance matrix for the pooled coefficient estimation for multiple treated units. With $\mathcal{K}_{\alpha_i,\rho_i}(\mathbf{D})$ on the block diagonal. We define the Bayesian regression model as

$$\begin{aligned}
\mathbf{Y} &\sim \mathcal{N}(\beta^T \mathbf{X}, \sigma_y \mathbf{I}) \\
\boldsymbol{\beta_i} &\sim \mathcal{GP}(\mathbf{0}, \mathcal{K}_{\alpha_i,\rho_i}(\mathbf{D})) \\
\alpha_i &\sim \Gamma^{-1}(50, 5) \\
\rho_i &\sim \Gamma^{-1}(5, 5) \\
\sigma_y &\sim \Gamma^{-1}(5, 5)
\end{aligned}$$

This framework has simple yet powerful relapses. In context with spatially correlated units, Gaussian process priors can improve the point estimate quality both in terms of bias and in terms of efficiency. Moreover, from the posterior distribution of $\beta_{\mathbf{i}}$ we can derive the smoothed path of the coefficient for some control unit i across the treated units $d \in \mathbf{D}$. Lastly, we can easily derive credibility intervals for the posterior distribution of the causal effect, retrieving it from the posterior distribution of β_i.

4. Estimating the effect of the Florentine tramway construction

Figure 1 show the results of our computation. Our results show that the tramway has provoked generally an increase in the commercial vitality of the area considered. These results are particularly significant for the areas closer to the tramway stops, as we find significant average treatment effects for the areas within 50 and 100 meters of the treatment sites. The positive, yet non-statistically significant effects are present for the outer areas, from 150 to 400 meters away from the tramway stops. Worksites have not extensively damaged the commercial environment of the treated area. We can note a significant and negative effect for the area within 100 meters during the period 2006-2010. That time span was the construction period of the tramway, and thus we could expect worse outcomes for areas close to the construction site. However, the number of stores steadily recovered in 2010, the inauguration year, and the overall effect, even for this particularly affected area, is still positive. For this purpose, it is worth noting that in the closer area to the treatment site, the positive effect is present since the start of the construction period, some retailers anticipate their competitors by locating the shops in the most served areas even before the start of tramway operations. The effect on the outer bands is similar to the ones found for inner areas. In particular, we notice that worksites has not affected the commercial environment of the outer areas, while the tramway has improved the accessibility of the area, leading to an increase in the number of shops present. The estimated causal effect has a growing tendency, especially for the outer areas, that exhibits statistically significant effects in the last observational periods. Concluding, in this work we propose a framework for matrix completion with spatial data, an open challenge in policy evaluation literature. We provide convincing results for our motivating application, showing the spatial diffusion of the causal effect. Simulation results, not provided here, confirms the good properties of the proposed estimation framework.

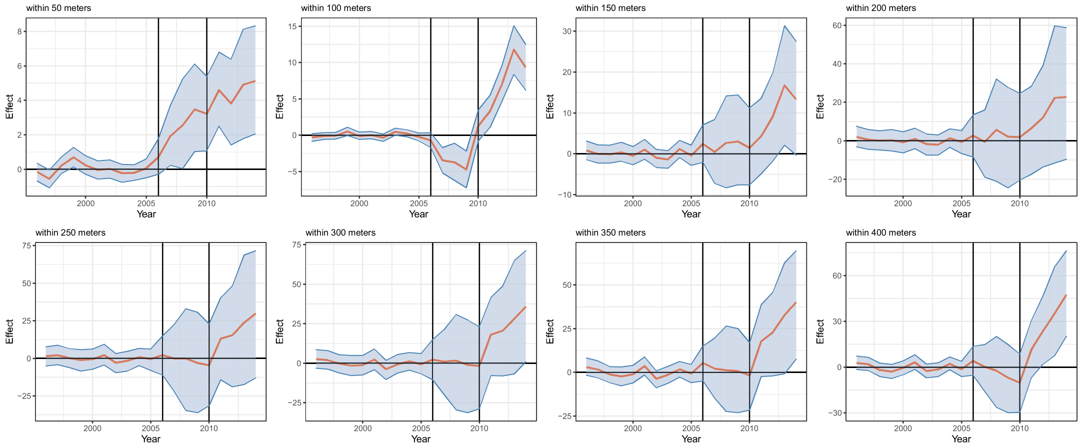

Figure 1: Treatment effect for areas within d meters from a tramway stop, Red line: Treatment effect, Blue area: 90% Credibility interval - First vertical line: tramway worksite starts (2006) - Second vertical line: tramway operational (2010)

References

[1] Abadie, A. (2021). Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of Economic Literature*, 59(2):391–425.

[2] Abadie, A., Diamond, A., and Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of californiaâs tobacco control program. *Journal of the American statistical Association*, 105(490):493–505.

[3] Arbour, D., Ben-Michael, E., Feller, A., Franks, A., and Raphael, S. (2021). Using multitask gaussian processes to estimate the effect of a targeted effort to remove firearms. *arXiv preprint arXiv:2110.07006*.

[4] Doudchenko, N. and Imbens, G. W. (2016). Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. Technical report, National Bureau of Economic Research.

[5] Kim, S., Lee, C., and Gupta, S. (2020). Bayesian synthetic control methods. *Journal of Marketing Research*, 57(5):831–852.

[6] Menchetti, F. and Bojinov, I. (2020). Estimating causal effects in the presence of partial interference using multivariate bayesian structural time series models. *Harvard Business School Technology & Operations Mgt. Unit Working Paper*, (21-048).

[7] Pang, X., Liu, L., and Xu, Y. (2022). A bayesian alternative to synthetic control for comparative case studies. *Political Analysis*, 30(2):269–288.

[8] Pinkney, S. (2021). An improved and extended bayesian synthetic control. *arXiv preprint arXiv:2103.16244*.

[9] Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

[10] Sobel, M. E. (2006). What do randomized studies of housing mobility demonstrate? causal inference in the face of interference. *Journal of the American Statistical Association*, 101(476):1398–1407.