Proceeding Series of the Brazilian Society of Computational and Applied Mathematics, v. 11, n. 1, 2025.

Trabalho apresentado no XLIII CNMAC, Centro de Convenções do Armação Resort - Porto de Galinhas - PE, 2024

Unraveling Spatial Confounding: A Comprehensive Review and Simulation-based Evaluation of Contemporary Methods

Isaque V. M. Pim[1] Luiz M. F. Carvalho[2]
FGV EMAp, Rio de Janeiro, RJ

Health, environmental, demographic, and other public data sets are typically aggregated (to administrative or geopolitical regions) to facilitate analysis and protect privacy. The use of standard regression models for spatially referenced data can result in spatial dependence in the residuals. For decades, the solution for this problem was to use spatial regression models [4]. The usual setup for a spatial regression defines a set of areal-units $I = \{1, 2, \ldots, n\}$ with respective observations y_i, $i \in I$ modelled as:

$$Y_i = \beta_0 + \boldsymbol{X}_i\beta_X + \boldsymbol{Z}_i\beta_Z + \varepsilon_i. \tag{1}$$

Here $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ represent measure and unmeasured covariates for location $i \in I$ respectively. Also, ε_i represents an error term for the observation at location i, having mean 0 and variance σ^2. Assuming that unobserved variables exhibit spatial patterns, the typical modeling approach involves incorporating a spatially structured latent variable, often represented as a correlated Gaussian Markov random field (GMRF), to address the absence of these variables in the analysis [6]. If measured variables do not suffice for confounding adjustment, but the missing confounders exhibit a spatial structure, we face a *spatial confounding* situation. The same terminology is also used in the context of spatial statistics to address correlation between the latent spatial factor and fixed effects [6]. If the unmeasured confounders are spatially varying in that nearby observations have similar values, recent developments utilize this structure to mitigate bias from these unmeasured spatial confounders [5], [2], [3]. In this work, we review, in the lines of [7], current methods in causal inference literature to deal with confounding.

For this purpose, consider assigning treatments $A_i \in \{0, 1\}$ for each region $i \in I$. We simulate treatments having a spatial distribution putting $X_i \sim \text{Bernoulli}[\text{expit}(V_i + \phi U_i)]$. Here U and V represent spatial terms drawn from models $U \sim \text{CAR}(\rho_U, 2)$ and $U \sim \text{CAR}(\rho_V, 2)$. Then we set $Y_i|X_i \sim \text{Normal}(X_i\beta + U_i, 1)$. Note that ρ controls spatial dependence and ϕ the strength of confounding. For the parameter values we choose $\rho_U, \rho_V \in \{0.9, 0.99\}$ to simulate a scenario with moderate and strong spatial correlation. To address misspecification we change $V_i + \phi U_i$ to $V_i + \phi U_i^2$. We generate the data sets on the map of São Paulo municipalities with $\beta = \phi = 0.6$. For each data sets we fit the models: non-spatial OLS (NS); non-spatial OLS + propensity score splines (NS + P); Spatial CAR (S); Spatial CAR + propensity score spline (S + P). Bayesian methods are used to best deal with the latent variables and inherently account for uncertainty. All of these methods are fit in R-Nimble. All of the variance components received $\text{InvGamma}(0.5, 0.005)$ priors. Means had $\text{Normal}(0, 10)$ prior. Latent spatial factors received ICAR [1] priors.

The results show that the spatial model (S) provides little improvement over the null model (NS). The best performers were propensity score splines (NS + P) and (S + P). The propensity score spline, differently than [7] was found to be not robust to misspecification. All parameter estimates are biased on the scenarios with low spatial association. In this case, the unmeasured confounder cannot be explained by known covariates or spatial patterns.

[1] isaque.pim@fgv.br
[2] luiz.fagundes@fgv.br

© 2025 SBMAC

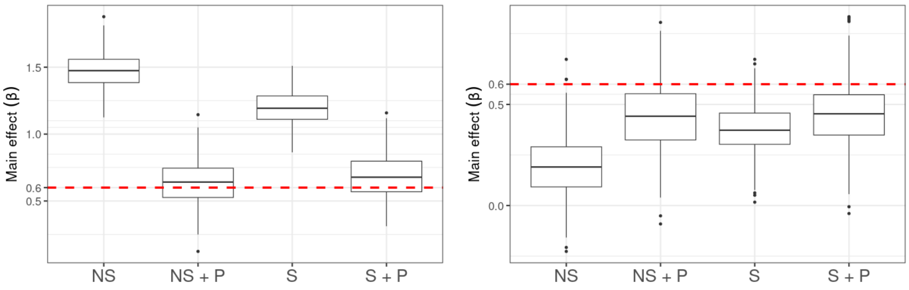

Figura 1: Simulation study results. The boxplots summarize the sampling distribution of the causal estimates across data sets and the solid line at 0.6 is the true value. Both scenarios displayed are with high spatial association ($\rho = 0.99$). The scenario on the right is misspecified. Source: created by the authors

In conclusion, our simulations demonstrate promising results in mitigating unmeasured spatial confounding, paving the way for further exploration and application in real-world scenarios. This methodology holds potential for extension to longitudinal and time series data analysis, offering approaches to investigate confounders across temporal dimensions. Robust non-parametric methods for the propensity scores are also of interest to mitigate misspecification. Moreover, future research should delve into addressing the issue of interference to enhance the robustness and applicability of these approaches.

Referências

[1] S. Banerjee, B. P. Carlin e A. E. Gelfand. **Hierarchical Modeling and Analysis for Spatial Data**. 2nd ed. CRC Press, 2015.

[2] E. Dupont, S. N. Wood e N. H. Augustin. “Spatial+: A novel approach to spatial confounding”. Em: **Biometrics** 78 (2020), pp. 1279–1290. URL: https://api.semanticscholar.org/CorpusID:221819113.

[3] B. A. Gilbert, A. Datta e E. L. Ogburn. “Approaches to spatial confounding in geostatistics”. Em: 2022. URL: https://api.semanticscholar.org/CorpusID:245634816.

[4] K. Khan e C. Berrett. **Re-thinking Spatial Confounding in Spatial Linear Mixed Models**. 2023. arXiv: 2301.05743 [stat.ME].

[5] G. Papadogeorgou, C. Choirat e C. M. Zigler. “Adjusting for unmeasured spatial confounding with distance adjusted propensity score matching”. Em: **Biostatistics** 20.2 (jan. de 2018), pp. 256–272. ISSN: 1465-4644. DOI: 10.1093/biostatistics/kxx074.

[6] B. J. Reich, J. S. Hodges e V. Zadnik. “Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models”. en. Em: **Biometrics** 62.4 (dez. de 2006), pp. 1197–1206.

[7] B. J. Reich, S. Yang, Y. Guan, A. B. Giffin, M. J. Miller e A. Rappold. “A Review of Spatial Causal Inference Methods for Environmental and Epidemiological Applications”. Em: **International Statistical Review** 89.3 (2021), pp. 605–634. DOI: https://doi.org/10.1111/insr.12452. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/insr.12452.

© 2025 SBMAC