A BAYESIAN SPATIO-TEMPORAL REGRESSION APPROACH FOR CONFOUNDING ADJUSTMENT

Carlo Zaccardi [1], Pasquale Valentini[1] and Luigi Ippoliti[1]

[1] University G. d'Annunzio, Chieti-Pescara, Department of Economics, Viale Pindaro 42, 65127 Pescara, Italy (e-mail: carlo.zaccardi@unich.it, pvalent@unich.it, luigi.ippoliti@unich.it)

ABSTRACT: For an accurate evaluation of the harmful impacts of pollution on human health, confounding variables must always be taken into account. Unfortunately, it oftentimes happens that some confounders might result unmeasured, hence, within a regression framework, the parameter that represents the exposure's effect might no longer be recoverable. In this paper, the unmeasured confounder is represented by a linear combination of basis functions, a technique that has been used in the spatial confounding literature, and that we expand to spatio-temporal designs. To reduce dimensionality and confounding bias, spike-and-slab priors are assumed on basis coefficients.

KEYWORDS: confounding, spatio-temporal, pollution, health, Bayesian.

1 Introduction

The principal objective in environmental epidemiology is to evaluate whether exposure to a pollutant has adverse health consequences. To this end, the relationship between exposure and outcome variables can be expressed in regression terms. An accurate evaluation of the relationship of interest requires that all variables correlated with both exposure and outcome (known as *confounders*), such as meteorological variables, should be included in the model as additional regressors (Dominici & Peng, 2008). However, data about some confounders could result not available because of, for example, budget constraints. If the model fails to account for confounding, it would be impossible to recover the parameter of interest. The estimator for the exposure's effect would then become biased, and its bias is known as *confounding bias* in the epidemiological literature (Dominici & Peng, 2008).

While smooth functions of calendar time are usually included in models for time-series data (e.g., see Dominici & Peng, 2008), in purely spatial settings, the simplest and more appealing remedy to the *spatial confounding* problem

is to add into the model a spatial random effect. However, Reich *et al.*, 2006 show that doing so distorts inference on the effect of interest and leads the practitioner to draw incorrect conclusions. Different other solutions are reviewed by Reich *et al.*, 2021 and Urdangarin *et al.*, 2022. To our knowledge, relatively few authors consider confounding adjustment in spatio-temporal designs. Reich *et al.*, 2021 reviews spatio-temporal methods as well to account for unmeasured confounding under causal inference hypotheses. More recently, two approaches in the spatial confounding literature are extended to account for temporal dependence as well (Adin *et al.*, 2023; Prates *et al.*, 2022). In the next Section, we discuss a different approach wherein, extending the work by Valentini *et al.*, 2022, unmeasured confounding is accounted for by including spatio-temporal basis functions into the regression model. We also impose a prior structure on the basis coefficients that encourages sparsity.

2 The Proposed Model

Consider a spatio-temporal process $\{Y(\mathbf{s},t) : \mathbf{s} \in \mathcal{D},\ t = 1,2,\ldots,T\}$, defined for every location, $\mathbf{s}$, over a continuous spatial domain $\mathcal{D} \subseteq \mathbb{R}^2$, and for discrete time periods $t = 1,2,\ldots,T$. Assume that it represents a health outcome observed at a finite set of locations, $\{\mathbf{s}_1,\ldots,\mathbf{s}_N\}$, for the entire study period. Moreover, suppose that $X(\mathbf{s},t)$ and $Z(\mathbf{s},t)$ are two correlated Gaussian spatio-temporal processes representing the exposure (observed at the same spatial locations and time instants as the outcome) and the unmeasured confounder, respectively. Assuming that the distribution F is a member of the exponential family, and that realizations are conditionally independent, it is possible to specify the following hierarchy, for $i = 1,\ldots,N$ and $t = 1,\ldots,T$:

$$Y(\mathbf{s}_i,t) \overset{ind}{\sim} F(\mu(\mathbf{s}_i,t),\phi) \tag{1}$$

$$g\left(\mu(\mathbf{s}_i,t)\right) = \beta_0 + \beta_x X(\mathbf{s}_i,t) + Z(\mathbf{s}_i,t) + \varepsilon(\mathbf{s}_i,t)\,, \tag{2}$$

where $\mu(\mathbf{s}_i,t) = E[Y(\mathbf{s}_i,t)]$, ϕ is a scale parameter, $g(\cdot)$ is an opportune link function, and $\varepsilon(\mathbf{s}_i,t)$ represents a zero-mean stationary Gaussian process with realizations mutually independent in time but correlated in space such that the spatial covariance structure is defined by a parametric function with parameter vector $\boldsymbol{\theta}$, that is $Cov(\varepsilon(\mathbf{s}_i,t),\varepsilon(\mathbf{s}_j,t)) = C(|\mathbf{s}_i - \mathbf{s}_j|;\boldsymbol{\theta})$ for $i,j = 1,\ldots,N$. The primary aim of the analysis is to correctly recover the regression coefficient of the exposure, β_x, while controlling for confounding at the same time.

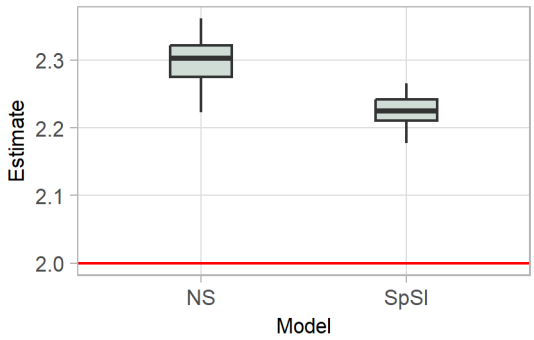

Figure 1. *Boxplots representing the estimated exposure effect in the simulation study. The red line represents the real value,* $\beta_x = 2$.

Thanks to the Karhunen-Loéve theorem (KLT, Banerjee *et al.*, 2014), the process $Z(\mathbf{s},t)$ can be represented as an infinite linear combination of pairwise orthogonal basis functions, but, operationally, a reduced-rank representation is given to it:

$$Z(\mathbf{s},t) \approx \sum_{m=1}^{M} \alpha_m \psi_m(\mathbf{s},t)\,, \tag{3}$$

where $\psi_m(\cdot,\cdot)$ are spatio-temporal basis functions, and α_m are expansion coefficients, for $m = 1,\ldots,M$. These bases are then introduced in Equation (2) in place of the unmeasured confounder. A necessary condition is that they must be correlated to both $X(\mathbf{s}_i,t)$ and $Z(\mathbf{s}_i,t)$, so the aforementioned drawbacks discussed by Reich *et al.*, 2006 could be overcome.

To select the most promising bases and hence obtain a parsimonious model, we assume spike-and-slab priors (Ishwaran & Rao, 2005) on the expansion coefficients. The Bayesian hierarchical specification is completed by assigning prior distributions to all the other parameters, and a Markov chain Monte Carlo (MCMC) algorithm is constructed for inferential purposes. To show whether our model is able to mitigate confounding issues, we set up a simulation study wherein $X(\mathbf{s},t)$ and $Z(\mathbf{s},t)$ are drawn from their joint distribution, under the assumptions that $Cor(X(\mathbf{s},t),Z(\mathbf{s},t)) = 0.5$, and that the second process varies at spatial and temporal scales coarser than those of the first process. The outcome is then generated using Eqs. 1–2, where F is the Gaussian distribution. We then fit a non-spatial (NS) model that does not account for confounding, and our proposal (denoted as SpSl). Figure 1 synthesize the main results: for each model, it depicts a boxplot of the posterior means for β_x obtained from fit-

ting 100 replicates. The red line represents its true value, $\beta_x = 2$. The proposed model can potentially reduce the confounding bias so it should be preferred to the non-spatial one.

Finally, a more extensive simulation study and real-data applications will be discussed in an extended version of this paper, wherein several types of basis functions will be examined as well.

References

ADIN, ARITZ, GOICOA, TOMÁS, HODGES, JAMES S, SCHNELL, PATRICK M, & UGARTE, MARÍA D. 2023. Alleviating confounding in spatio-temporal areal models with an application on crimes against women in India. *Statistical Modelling*, **23**(1), 9–30.

BANERJEE, SUDIPTO, CARLIN, BRADLEY P, & GELFAND, ALAN E. 2014. *Hierarchical modeling and analysis for spatial data*. CRC press.

DOMINICI, FRANCESCA, & PENG, ROGER D. 2008. *Statistical methods for environmental epidemiology with R: a case study in air pollution and health*. Springer.

ISHWARAN, HEMANT, & RAO, J. SUNIL. 2005. Spike and slab variable selection: Frequentist and Bayesian strategies. *The Annals of Statistics*, **33**(2), 730–773.

PRATES, MARCOS O, AZEVEDO, DOUGLAS RM, MACNAB, YING C, & WILLIG, MICHAEL R. 2022. Non-separable spatio-temporal models via transformed multivariate Gaussian Markov random fields. *Journal of the Royal Statistical Society Series C: Applied Statistics*, **71**(5), 1116–1136.

REICH, BRIAN J, HODGES, JAMES S, & ZADNIK, VESNA. 2006. Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models. *Biometrics*, **62**(4), 1197–1206.

REICH, BRIAN J, YANG, SHU, GUAN, YAWEN, GIFFIN, ANDREW B, MILLER, MATTHEW J, & RAPPOLD, ANA. 2021. A review of spatial causal inference methods for environmental and epidemiological applications. *International Statistical Review*, **89**(3), 605–634.

URDANGARIN, ARANTXA, GOICOA, TOMÁS, & UGARTE, MARÍA DOLORES. 2022. Evaluating recent methods to overcome spatial confounding. *Revista Matemática Complutense*, 1–28.

VALENTINI, PASQUALE, SCHMIDT, ALEXANDRA M., ZACCARDI, CARLO, & IPPOLITI, LUIGI. 2022. Adjusting for Unmeasured Spatial Confounding Through Shrinkage Methods. *In: Book of Short Papers of the 51st Scientific Meeting of the Italian Statistical Society*. Pearson.