2-9-2024

Discussion on "Spatial+: A novel approach to spatial confounding" by Dupont, Wood, and Augustin

Brian J. Reich

Shu Yang

Yawen Guan

HHS Public Access

Author manuscript

Biometrics. Author manuscript; available in PMC 2024 February 09.

Published in final edited form as:

Biometrics. 2022 December ; 78(4): 1291–1294. doi:10.1111/biom.13651.

Discussion on "Spatial+: A novel approach to spatial confounding" by Dupont, Wood, and Augustin

Brian J. Reich[1], **Shu Yang**[1], **Yawen Guan**[2]

[1]Department of Statistics, North Carolina State University, Raleigh, North Carolina, USA

[2]Department of Statistics, University of Nebraska-Lincoln, Lincoln, Nebraska, USA

1 | INTRODUCTION

Congratulations to the authors for this thoughtful and timely contribution to the spatial confounding literature. The intuitive nature of the method and simplicity of the estimation procedure will surely make `Spatial+` popular with practitioners, and the theoretical developments are a major advance for researchers in this area. There is much to discuss! We have formatted our discussion in two sections: in Section 2 we consider the assumptions and statistical properties of `Spatial+`, and in Section 3 we examine how `Spatial+` fits in the wider literature on spatial causal inference.

2 | ASSUMPTIONS AND STATISTICAL PROPERTIES OF `Spatial+`

Identification

`Spatial+` uses partial linear regression (PLR) to adjust for spatial confounding. It assumed in Equation (4) that x_i, the covariate at spatial location $\mathbf{t}_i$, can be written as $x_i = f^x(\mathbf{t}_i) + \epsilon_i^x$ and $\epsilon_i^x \overset{iid}{\sim} N(0, \sigma_x^2)$ for some smooth process f^x. The estimation procedure utilizes a two-stage smoothing spline regression, where the first stage obtains the residual of the covariate x_i that is uncorrelated with spatial confounding and the second stage replaces x_i by its residual. This trick is well established in the econometrics literature; see, for example, Robinson (1988) and Speckman (1988), where both the covariate x_i and its support $\mathbf{t}_i$ can be vectors. Identification of model parameters is important in the PLR framework. Robinson (1988) shows that $\Phi = E\left[\{x - E(x \mid \mathbf{t})\}\{x - E(x \mid \mathbf{t})\}^{\mathrm{T}}\right]$ being positive definite is a necessary and sufficient condition for β to be identified and $\hat{\beta}$ to be root-n consistent. It will be critical to establish identification conditions in the `Spatial+` framework.

Model assumptions

One of the strengths of the paper is to lay bare the assumptions needed for consistency in the spatial setting. One assumption is that the covariate is Gaussian and can be decomposed into

Correspondence: Brian J. Reich, Department of Statistics, North Carolina State University, Raleigh, NC, USA. bjreich@ncsu.edu.

SUPPORTING INFORMATION

Web Appendices (code) referenced in Section 2 are available with this paper at the Biometrics website on Wiley Online Library.

smooth and independent components as in Equation (4). It seems the methods would still perform well if the errors were slightly non-Gaussian, but we would be curious to learn what the authors recommend for more extreme cases such as binary x_i. In a recent review (Reich et al., 2021), we found that (a slight variation of) the method of Davis et al. (2019) that fits a spatial logistic regression model $\text{logit}\{\text{Prob}(x_i = 1)\} = f^x(\mathbf{t}_i)$ and adjusts for the estimate of $f^x(\mathbf{t}_i)$ in the response model effectively reduced confounding bias. As discussed further in Section 3, this has connections with the propensity-score (PS) adjustment that is common in causal inference. Perhaps a similar approach can be taken for `Spatial+`?

A more challenging scenario is when x is a continuous spatial surface, that is, $\sigma_x = 0$. While this may appear to be a pathological case, it is in fact quite common in the epidemiological literature. For example, Schnell and Papadogeorgou (2020) studied the health effect of supermarket access, and one could envision a study where the exposure of interest is the distance from a subject's residence to the nearest supermarket, which is a continuous spatial surface. Another common setting that gives spatially smooth exposure is the study of neighborhood effects, for example, Giffin et al. (2020) regressed air pollution concentration onto kernel-smoothed measures of wildland fire indicators. There are many other examples such as extreme temperature, some forms of air pollution, distance to a point source, and so forth. To extend `Spatial+` to this case would be seem to require fundamentally different assumptions to avoid the residuals being zero, that is, $r_i^x = 0$, and thus the exposure effect being unidentifiable. For example, Guan et al. (2020) allow for a continuous exposure variable with assumptions about the local correlation between the exposure and confounding variables. Can a similar approach be applied to `Spatial+`?

We examine the performance of `Spatial+` for spatiallysmooth covariates by extending the simulation study to include smoother covariate processes. The data generation and implementation of `Spatial+` are identical to the simulation in Section 4 of the main paper except that we consider a range of σ_x. Coverage of 95% intervals for β is computed using the standard error provided by the `mgcv` package (although the authors do not use these standard errors). Figure 1 shows low bias and nominal coverage for all but the smallest value of σ_x. Modifying the approach to accommodate smooth covariates and/or providing a rule of thumb to caution against this source of bias would be useful in practice.

Another assumption is that the unmeasured spatial confounder $f^x(\mathbf{t})$ in the covariate and the spatial dependence $f(\mathbf{t})$ can be fit with spline regression. To understand the performance of `Spatial+` under model misspecification, we repeat the simulation study in Section 4 with a slightly different data-generation scheme. We simulate $f^x(\mathbf{t})$ from a Gaussian process with the same parameter setting, but instead of taking the fitted value from a thin plate spline (as is done in Section 4) we use the Gaussian process realization as the covariate, and we simulate $f(\mathbf{t})$ similarly. We considered this data-generation scheme as it is the most problematic case for the standard spatial linear model (Paciorek, 2010) and it is often more realistic in data applications. Figure 2 shows an example of the unmeasured confounder $f^x(\mathbf{t})$ used to form the covariate from the different simulation schemes, and the bias and coverage from `Spatial+` when the model is misspecified. While the difference in the unmeasured

confounders is small, the remaining residuals from fitting the smoothing spline may cause collinearity-induced bias in estimating β. The magnitude of the bias is large and coverage is low for small σ_x. Therefore, while splines are generally a robust semiparametric method for function estimation, in this case users should check for sensitivity to their modeling assumptions.

Spline and kernel smoothing

Spline smoothing requires choosing a basis and knot locations. An alternative is kernel smoothing, that is, the mean functions are assumed to be locally well approximated by polynomial functions. For PLR, Robinson (1988) proposed a two-stage kernel smoothing estimator, a counterpart of the proposed estimator in `Spatial+`. Speckman (1988) conducted a theoretical comparison of the asymptotic behaviors of the two types of estimators. It would be interesting to compare the two parallel frameworks in `Spatial+`.

Smoothing selection

As with most semiparametric estimators, tuning parameter selection is a key step. Following Chen and Shiau (1994), the authors suggest minimizing the mean squared error of the estimated spatial effects to select the smoothing parameters; however, in implementation, the authors use generalized cross-validation which in fact targets minimizing the prediction error. Thus, there is a gap between the authors' target and implementation. We are curious if there is an objective function for smoothing selection that directly targets estimating β, and if not, whether the authors could provide intuition for why minimizing these indirect objective functions leads to a good performance of $\widehat{\beta}^{+}$.

Inference

We are disappointed that the paper does not mention how to conduct inference in `Spatial+`. Given the asymptotic results, $\widehat{\beta}^{+}$ is root-n consistent. Will resampling approaches such as the bootstrap work to estimate its variance and conduct inference on β? Also, the authors comment on the equivalence between modeling spatial random effects through the use of a smoothing penalty and Bayesian modeling. When inference under the frequentist framework is a daunting task, will Bayesian modeling offer a remedy?

3 | CONNECTIONS WITH SPATIAL CAUSAL INFERENCE

We would like to take this opportunity to place `Spatial+` in the broader context of spatial casual inference. Causal inference provides a rigorous mathematical foundation to define the causal effect/estimand of interest and clarify the assumptions required to achieve identifiability. Causal effects are typically defined via potential outcomes under different treatments, but defining a potential outcomes framework for spatial problems is nontrivial due to correlation between observations and possible interference between the treatment at one location and the response at another. Reich et al. (2021) reviewed several casual estimands and procedures to estimate the casual effect in spatial problems. `Spatial+` is categorized as a "neighborhood adjustment" method (a discussion of other methods placed

in this category is below), and this is contrasted with other approaches such as matching methods, PS adjustments, and instrumental variables.

In practice of course, one should consider all of these options when conducting a given analysis. For example, if the tuning and inference issues with `Spatial+` discussed above are concerns, matching nearby observations with different treatments and analyzing the difference in their responses is a simple way to adjust for unmeasured spatial confounders, perhaps at the expense of statistical efficiency under a correctly specified model. Also, for binary exposure variables, a PS adjustment might be more appropriate as discussed below. In the remainder of this section, we discuss various connections with `Spatial+` and casual inference methods.

PS methods

To establish a causal effect, the authors require that $f(\mathbf{t}_i)$ captures all confounding effects of treatment (x_i) and outcome (y_i). Under the PLR assumption that $y_i = \beta x_i + f(\mathbf{t}_i) + \epsilon_i,\ \beta$ can be interpreted causally. One strategy of obtaining an unbiased estimator of β with spatial confounding is using (generalized) PS adjustments, where the propensity score is the conditional (density or) probability of x_i given the confounders, $e(\mathbf{t}_i) = \text{Prob}(x_i \mid \mathbf{t}_i)$. Reich et al. (2021) studied a PS-adjusted PLR (Zhou et al., 2019) defined as $E(y_i \mid x_i, \mathbf{t}_i) = \beta x_i + f_1(\mathbf{t}_i) + f_2\{e(\mathbf{t}_i)\}$, where $f_2(\cdot)$ is a flexible nonparametric model such as splines. Zhou et al. (2019) showed that their PS-adjusted PLR estimator is doubly robust in that it is consistent if either $f_1(\cdot)$ or the PS model is correctly specified, but not necessarily both. We would like to solicit opinions from the authors on such PS adjustments in the context of spatial confounding and causal inference.

Comparison with other confounder adjustment methods

Different assumptions on the structure of the missing confounder have led to different adjustment methods. In a very similar approach to `Spatial+`, Keller and Szpiro (2019) partition the covariate into a smooth component and its complement. The smooth component is then removed from the covariate, and the association between the adjusted covariate and response is estimated as a function of the degree of smoothness. Thaden and Kneib (2018) assumes geographic confounding and removes the spatial patterns from both covariate and response, then the causal effect is estimated by regressing the local variation in the covariate and response. `Spatial+` assumes a spatially smooth unmeasured confounder, which is equivalent to confounding only at large spatial scales, while Guan et al. (2020) allows for different degrees of confounding at different spatial resolutions with the assumption that confounding dissipates for smaller scales. Guan et al. (2020) proposed an estimation procedure in the spectral domain, but intuitively it decomposes both the covariate and response into new variables at different spatial scales and estimates their association at each level. The effect is estimated as a function of spatial scale and causal interpretation is drawn at local levels. Schnell and Papadogeorgou (2020) mitigates unmeasured spatial confounding for county-level data by proposing a joint model for the covariate and missing confounder, and put forth assumptions required for identifiability.

In summary, there are now several approaches to reducing the effect of spatial confounding and `Spatial+` will clearly play a central role moving forward. We reiterate our congratulations to the authors for their important contribution to this emerging field and look forward to further developments in this area.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Funding information

National Institutes of Health, Grant/Award Number: R01ES031651-01

REFERENCES

Chen H and Shiau J-JH (1994) Data-driven efficient estimators for a partially linear model. Annals of Statistics, 22, 211–237.

Davis ML, Neelon B, Nietert PJ, Hunt KJ, Burgette LF, Lawson AB et al. (2019) Addressing geographic confounding through spatial propensity scores: a study of racial disparities in diabetes. Statistical Methods in Medical Research, 28, 734–748. [PubMed: 29145767]

Giffin A, Reich BJ, Yang S and Rappold AG (2020) Generalized propensity score approach to causal inference with spatial interference. arXiv preprint arXiv:2007.00106.

Guan Y, Page GL, Reich BJ, Ventrucci M and Yang S (2020) A spectral adjustment for spatial confounding. arXiv preprint arXiv:2012.11767.

Keller JP and Szpiro AA (2019) Selecting a scale for spatial confounding adjustment. arXiv preprint arXiv:1909.11161.

Paciorek CJ (2010) The importance of scale for spatial-confounding bias and precision of spatial regression estimators. Statistical Science, 25, 107–125. [PubMed: 21528104]

Reich BJ, Yang S, Guan Y, Giffin AB, Miller MJ and Rappold AG (2021) A review of spatial causal inference methods for environmental and epidemiological applications. International Statistical Review, 89(4). 10.1111/insr.12452.

Robinson PM (1988) Root-n-consistent semiparametric regression. Econometrica: Journal of the Econometric Society, 56, 931–954.

Schnell P and Papadogeorgou G (2020) Mitigating unobserved spatial confounding when estimating the effect of supermarket access on cardiovascular disease deaths. Annals of Applied Statistics, 14, 2069–2095.

Speckman P. (1988) Kernel smoothing in partial linear models. Journal of the Royal Statistical Society: Series B (Methodological), 50, 413–436.

Thaden H and Kneib T (2018) Structural equation models for dealing with spatial confounding. American Statistician, 72, 239–252.

Zhou T, Elliott MR and Little RJ (2019) Penalized spline of propensity methods for treatment comparison. Journal of the American Statistical Association, 114, 1–19.

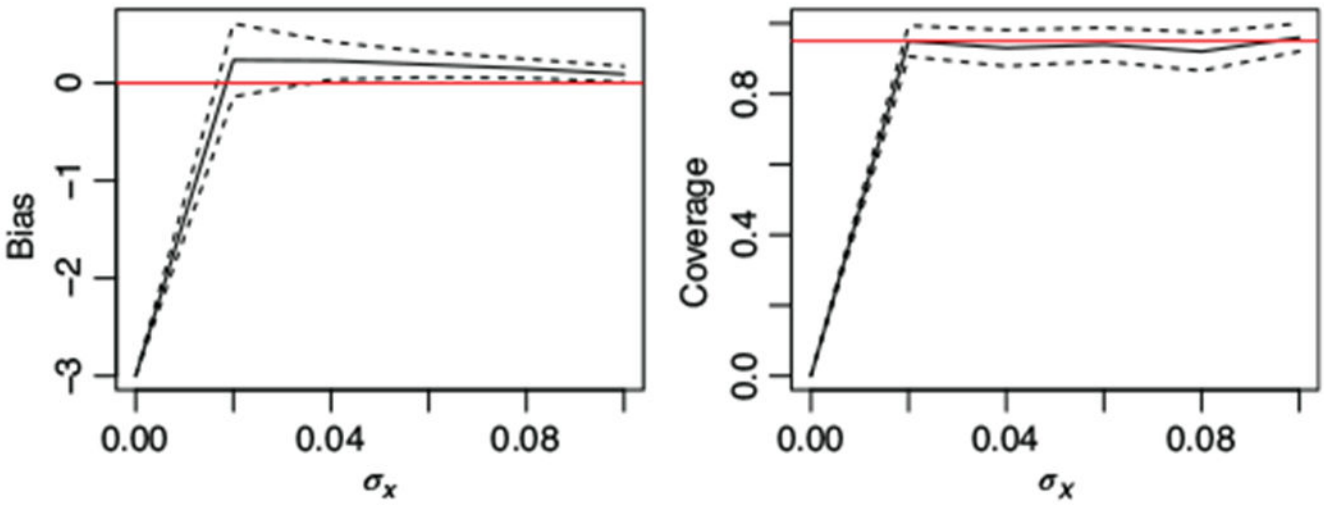

FIGURE 1.

Bias and coverage of `Spatial+` (the dashed lines 95% intervals) as a function of σ_x for the simulation study. This figure appears in color in the electronic version of this article, and any mention of color refers to that version

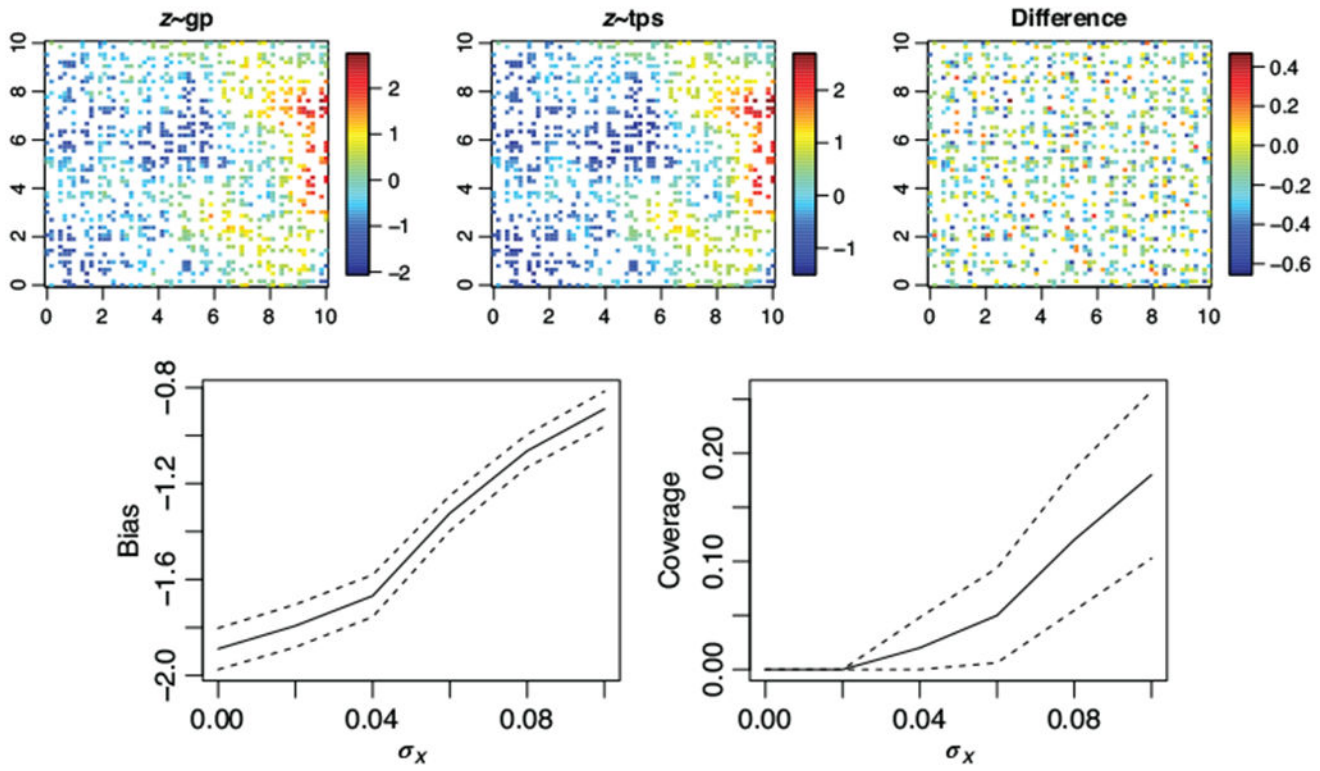

FIGURE 2.

Unmeasured spatial confounder simulated from a Gaussian process (top left), from the fitted thin plate spline (top middle), and their difference (top right). Bias and coverage of `Spatial+` (the dashed lines 95% intervals) as a function of σ_x under misspecified model. The covariates are generated as $x_i = z_i/2 + \epsilon_i$, where z_i is a spatial process and $\epsilon_i \overset{iid}{\sim} \text{Normal}(0, \sigma_x^2)$ and the plots in the top row are the spatial component, z_i. This figure appears in color in the electronic version of this article, and any mention of color refers to that version