Harvard University
Harvard University Biostatistics Working Paper Series

Year 2009 *Paper* 98

The Importance of Scale for Spatial-confounding Bias and Precision of Spatial Regression Estimators

Christopher J. Paciorek*

*Harvard School of Public Health, paciorek@hsph.harvard.edu

The Importance of Scale for Spatial-confounding Bias and Precision of Spatial Regression Estimators

Christopher J. Paciorek

Abstract

Increasingly, regression models are used when residuals are spatially correlated. Prominent examples include studies in environmental epidemiology to understand the chronic health effects of pollutants. I consider the effects of residual spatial structure on the bias and precision of regression coefficients, developing a simple framework in which to understand the key issues and derive informative analytic results. When the spatial residual is induced by an unmeasured confounder, regression models with spatial random effects and closely-related models such as kriging and penalized splines are biased, even when the residual variance components are known. Analytic and simulation results show how the bias depends on the spatial scales of the covariate and the residual; bias is reduced only when there is variation in the covariate at a scale smaller than the scale of the unmeasured confounding. I also discuss how the scales of the residual and the covariate affect efficiency and uncertainty estimation when the residuals can be considered independent of the covariate. In an application on the association between black carbon particulate matter air pollution and birth weight, controlling for large-scale spatial variation appears to reduce bias from unmeasured confounders, while increasing uncertainty in the estimated pollution effect.

The importance of scale for spatial-confounding bias and precision of spatial regression estimators

Christopher J. Paciorek, Department of Biostatistics, Harvard School of Public Health

March 21, 2009

Abstract

Increasingly, regression models are used when residuals are spatially correlated. Prominent examples include studies in environmental epidemiology to understand the chronic health effects of pollutants. I consider the effects of residual spatial structure on the bias and precision of regression coefficients, developing a simple framework in which to understand the key issues and derive informative analytic results. When the spatial residual is induced by an unmeasured confounder, regression models with spatial random effects and closely-related models such as kriging and penalized splines are biased, even when the residual variance components are known. Analytic and simulation results show how the bias depends on the spatial scales of the covariate and the residual; bias is reduced only when there is variation in the covariate at a scale smaller than the scale of the unmeasured confounding. I also discuss how the scales of the residual and the covariate affect efficiency and uncertainty estimation when the residuals can be considered independent of the covariate. In an application on the association between black carbon particulate matter air pollution and birth weight, controlling for large-scale spatial variation appears to reduce bias from unmeasured confounders, while increasing uncertainty in the estimated pollution effect.

Keywords: epidemiology, identifiability, mixed models, penalized likelihood, random effects, spatial correlation

1 Introduction

Spatial confounding is likely present in many of the applied contexts in which spatial residual structure is considered, particularly in public health and social science contexts. Consider the following motivating example from the field of environmental health. The health effects of exposure to (spatially-varying) air pollution are an important public health issue. Many variables that explain variability in the response, including potential confounding variables that may be correlated with exposure, also vary spatially. For example, large-scale regional patterns in air pollution may be correlated with regional patterns in diet, income and other factors that are risk factors for the same health outcomes as the air pollutant. Small-scale patterns in air pollution from local sources may be correlated with risk factors as well, for example if lower-income people live nearer to busy roads or industrial sources. If some of the confounding variables are not measured, it will be difficult to distinguish the effect of air pollution from residual spatial variation in the

health outcome. I use the term spatial confounding to characterize this situation. In the non-spatial setting, unmeasured confounding biases regression coefficient estimates. The hope in the spatial setting is that by accounting for the spatial structure of the residuals one reduces or eliminates bias, but the mechanism of achieving this goal appears not to be well understood nor investigated rigorously in the statistical or applied literature. Note that bias is an important consideration in this context, in addition to mean squared error, in part because multiple studies of similar types of data are conducted and the results reviewed jointly or considered together in formal meta-analysis.

To consider the problem formally, start with simple linear regression with spatial structure:

$$\begin{aligned} \boldsymbol{Y} &= \beta_0 \mathbf{1} + \beta_x \boldsymbol{X} + \boldsymbol{e} \\ \boldsymbol{e} &\sim \mathcal{N}(\mathbf{0}, \Sigma), \end{aligned} \tag{1}$$

in which estimating β_x is of primary interest and where Σ captures any spatial correlation in the residuals, as well as independent variation. Spatial statistics and regression texts note that the OLS estimator under residual spatial correlation is unbiased but inefficient, and the usual OLS variance estimator is incorrect. Assuming known Σ, the GLS estimator is the most efficient estimator. However, there is little information in the literature about how the spatial scale of the residual structure affects inference. Nor is there information about the effects of spatial structure in $\boldsymbol{X}$ on inference; such structure is very common in applications and complicates the problem because $\boldsymbol{X}$ and the residual spatial structure compete to explain the variability in the response (Waller and Gotway, 2004).

One can obtain the basic spatial regression model (1) using a simple mixed model,

$$Y_i = \beta_0 + \beta_x X(s_i) + g(s_i) + \epsilon_i, \tag{2}$$

with spatially-correlated, normally-distributed random effects, $\boldsymbol{g} \sim \mathcal{N}(\mathbf{0}, \sigma_g^2 R(\theta_g))$. Marginalizing over $\boldsymbol{g}$ gives

$$\boldsymbol{Y} \sim \mathcal{N}(\beta_0 \mathbf{1} + \beta_x \boldsymbol{X}, \sigma_g^2 R(\theta_g)) + \tau^2 I). \tag{3}$$

Here Σ is explicitly decomposed into spatial and non-spatial components:$R(\theta_g)$ is a spatial correlation matrix parameterized by θ_g, a spatial range parameter, and $\text{Var}(\epsilon_i) = \tau^2$. A penalized spline formulation of the model would specify $g(\cdot)$ as a penalized spline where a penalty parameter replaces the role of $\{\theta_g, \sigma_g^2\}$ in the marginal likelihood in penalizing complexity of the spatial structure. $\boldsymbol{X}$, the covariate of interest, which I will call the 'exposure', may itself be spatially correlated, e.g., $\boldsymbol{X} \sim \mathcal{N}(\mathbf{0}, \sigma_x^2 R(\theta_x))$ if $\boldsymbol{X}$ is generated by a Gaussian process. To demonstrate processes operating at different spatial scales, Fig. 1 shows simulated spatial surfaces operating at three scales, as one varies θ.

The limited material on spatial regression in the spatial statistics literature assumes that the residual is independent of the covariate(s) (Cressie, 1993; Waller and Gotway, 2004) with little or no discussion of the possibility of confounding by the residual. Consider the possibility that $\boldsymbol{g}$ might play the role of an unmeasured variable, $\boldsymbol{g} \equiv \beta_z \boldsymbol{Z}$, with $\boldsymbol{Z} \sim \mathcal{N}(\mathbf{0}, \sigma_z^2 R(\theta_z))$, such that $\sigma_z^2 = \sigma_g^2/\beta_z^2$. The marginal likelihood

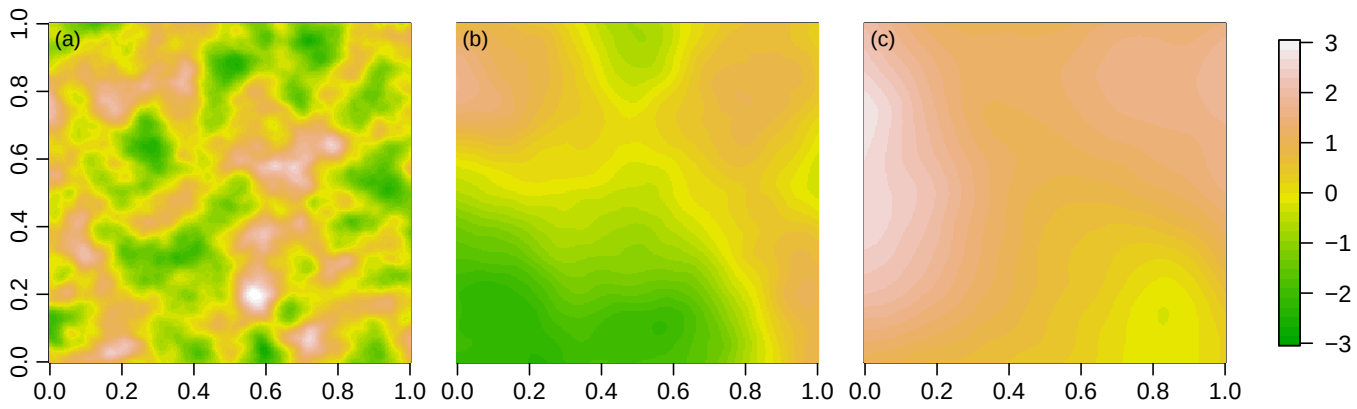

Figure 1: Gaussian process realizations for three values of θ, with high-frequency, small (fine)-scale variability for $\theta = 0.1$ (a), moderate scale variability for $\theta = 0.5$ (b), and low-frequency, large-scale variability for $\theta = 0.9$ (c).

obtained by integrating over the marginal distribution of $\boldsymbol{Z}$ is as in (3) but with $\beta_z^2\sigma_z^2 R(\theta_z)$ in place of $\sigma_g^2 R(\theta_g)$. To consider the effects of unmeasured spatial confounding, consider $\boldsymbol{Z}$ and $\boldsymbol{X}$ to be correlated, such that for any location, s, the correlation of $X(s)$ and $Z(s)$ over repeated sampling at that location is ρ. In this case, derivation of the marginal likelihood should be done by integrating over the conditional distribution of $\boldsymbol{Z}$ given $\boldsymbol{X}$. Note that if $\boldsymbol{X}$ and $\boldsymbol{Z}$ are considered fixed, then association between $\boldsymbol{X}$ and $\boldsymbol{g} \equiv \beta_z \boldsymbol{Z}$ is known as concurvity.

In the applied literature, practitioners often recognize the need to consider spatial residual structure with language of 'control' or 'accounting' for autocorrelation, and fit models (such as kriging or spatial random effects) that implicitly assume independence of the residual and the exposure (Burnett et al., 2001; Cakmak et al., 2003; Cho, 2003; Burden et al., 2005; Augustin et al., 2007; Molitor et al., 2007). Formal statements of the goals and properties of fitting spatial models are generally absent from the applied literature, but much of the interest appears to lie in using the spatial residual structure to try to account for spatial confounding, with the implicit assumption that such models reduce or eliminate bias from unmeasured spatially-varying confounders (e.g., Pope et al., 2002; Cakmak et al., 2003; Richardson, 2003; Biggeri et al., 2005). One potential approach is to explicitly consider the spatial scales involved, hoping that by accounting for variation at a relatively large spatial scale, this allows identification of the parameter of interest based on exposure heterogeneity at a smaller spatial scale (e.g., Burnett et al., 2001; Cakmak et al., 2003; Zeger et al., 2007), ideally heterogeneity that is not spatially correlated at all. This smaller-scale variability may be less prone to confounding in a given application than the larger-scale variation. However, this consideration of spatial scale is often not explicit in the applied literature and effects of scale on potential bias reduction, while sometimes hinted at, have not been developed formally.

In the analogous context of time series modeling of air pollution, Dominici et al. (2004) attempt to attribute all the temporally-correlated variability in the outcome to the residual term so as to identify the effect of exposure based on the temporally-uncorrelated (and presumably unconfounded) heterogeneity in the exposure. Dominici et al. (2004) provide no guidance in the scenario that the exposure cannot be de-

composed into autocorrelated and uncorrelated components. In the spatial setting, in which measurements cannot generally be made at all locations, accurate estimation of the uncorrelated component, if such a component even exists, is rare: consider atmospheric phenomena such as temperature and air pollution. In fact, in estimating environmental exposures, a common scenario involves predicting spatially-varying exposure values using averages of measurements within areal units or spatial smoothing techniques, approaches that do not provide fine-scale heterogeneity. Thus there is need to address the problem when all of the measured components of exposure are spatial.

Here I address estimation in simple regression models with spatial residual structure. In particular I focus on the properties of penalized models, using a simple mixed model fit by generalized least squares, equivalent to universal kriging, to analyze the effects of spatially-correlated residual structure on fixed effect estimators. Section 2 focuses on bias from spatial confounding. I report analytic results when the full covariance structure is known and supporting simulations when the covariance is estimated from the data or when the model is fit as an additive model with penalized splines. I assess the use of sensitivity analysis approaches in which the spatial scale of the residual variation is considered at a number of fixed values in a given application, as a means of explicitly considering the bias-variance tradeoff. Section 3 focuses on precision of estimators when there is no association between exposure and residual (no spatial confounding). Finally, I close with an example of the effects of air pollution on birthweight (Section 4).

2 Spatial confounding and bias

2.1 Identifiability

A key consideration in the basic model (2) is identifiability and the closely-related question of how the estimation procedure attributes variability between the exposure and the spatial residual term (the random effects). In the simple linear model, attribution of variability to $X\boldsymbol{\beta}$ rather than $\boldsymbol{\epsilon}$ is favored because this allows the estimate of the residual variance to decrease, with the normalizing constant of the likelihood favoring smaller residual variance. In the spatial model, if the spatial term $\boldsymbol{g}$ is unconstrained, then $\beta_x X(s)$ and $g(s)$ are not identifiable in the likelihood; one could redefine $g^*(s) = \beta_x X(s) + g(s)$ and then set $\beta_x \equiv 0$, with no change in the likelihood. Identifiability comes through constraints on $g(\cdot)$, either by penalizing lack of smoothness, considering $g(\cdot)$ to be a random effects term, or by having a prior on $g(\cdot)$. In decomposing variability between $\beta_x X(s)$ and $g(s)$, these approaches give higher penalized likelihood, marginal likelihood, or posterior density, respectively, when variability is attributed to $\beta_x X(s)$, the fixed effects term, because it is not penalized. In contrast, attribution to $g(s)$ would increase the penalty term or decrease the marginal likelihood or posterior density. The danger of penalized models in the spatial confounding context is that this dynamic causes bias in estimation of β_x, for example as seen in the simulations of Peng et al. (2006). An alternative constraint on $g(s)$ is to specify the term in a reduced dimension basis, say as a regression spline, in which case the model is identifiable if there is a component of variability in $X(s)$ that cannot be explained by the spline structure, i.e., if $\boldsymbol{X}$ is not perfectly collinear with the columns of the chosen basis matrix.

2.2 Analytic Framework

To consider bias from unmeasured spatially-varying confounders, consider the model,

$$\boldsymbol{Y} \sim \mathcal{N}(\beta_0 \mathbf{1} + \beta_x \boldsymbol{X} + \beta_z \boldsymbol{Z}, \tau^2 I), \tag{4}$$

where $\boldsymbol{Z}$ is a confounder, with the correlation of $X(s)$ and $Z(s)$ over repeated sampling at location s taken to be $\rho \neq 0$. Assume $\boldsymbol{Z}$ is not observed and that one seeks to account for spatial confounding by modeling the residual as spatially-correlated random effects, $\boldsymbol{g} \sim \mathcal{N}(0, \sigma_g^2 R(\theta_g))$ as in (2). Suppose one ignores the correlation between $\boldsymbol{g} \equiv \beta_z \boldsymbol{Z}$ and $\boldsymbol{X}$ and integrates over the marginal distribution for $\boldsymbol{g}$, giving (3). Equivalently, $\boldsymbol{Y} = \beta_0 \mathbf{1} + \beta_x \boldsymbol{X} + \boldsymbol{\epsilon}^*$. The induced correlation between $\boldsymbol{X}$ and $\boldsymbol{\epsilon}^*$ violates the usual regression assumption that the residual is independent of the covariate, leading to bias. From the random effects perspective, we have (incorrectly) assumed that the random effects are independent of the covariate(s) considered to be fixed effects term(s), a key assumption of mixed effects models (Breslow and Clayton, 1993; Diggle et al., 2002, p. 170), but one that is often unstated.

While treating $X(s)$ and $Z(s)$ as random naturally induces spatial structure, in a given dataset $X(s)$ and $Z(s)$ may reflect spatial structure that is not easily posed as arising from a stochastic data generating process, depending on the type of repeated sampling that seems reasonable for a given problem. This is particularly true for large-scale variation in $\boldsymbol{X}$ and $\boldsymbol{Z}$, which mimics the partial spline/partial linear setting. However, even in this situation, the above generative model is useful because realizations give plausible values for $\boldsymbol{X}$ and $\boldsymbol{Z}$ that would arise in real applications for which there is no plausible stochastic mechanism but where there is non-zero association,

$$\hat{\rho} = \frac{\sum (x_i - \bar{x})(z_i - \bar{z})}{s_x s_z} \neq 0, \tag{5}$$

over the collection of locations (e.g., the concurvity in the simulations of Ramsay et al. (2003); Peng et al. (2006); He et al. (2006)). Similarly, the bias seen in the partial spline/partial linear setting (Rice, 1986, eq. 28; Speckman, 1988) is caused by non-zero association between the smooth term and the exposure, calculated over points in the domain of the smooth term. I choose to treat $\boldsymbol{X}$ and $\boldsymbol{Z}$ as stochastic and use ρ as a parameter that explicitly quantifies the strength of association of residual and exposure. In any real dataset, $\hat{\rho} \approx 0$ seems particularly unlikely if both $\boldsymbol{X}$ and $\boldsymbol{Z}$ vary at large scale relative to the size of the domain (though $\hat{\rho} < 0$ may be as much a possibility as $\hat{\rho} > 0$), because it is unlikely that there would be two large-scale processes that are orthogonal. The treatment of $\boldsymbol{Z}$ as random allows for some simple, useful analytic results and is further justified in that the variation that an unmeasured $\boldsymbol{Z}$ induces in $\boldsymbol{Y}$ is necessarily treated stochastically as part of the residual in actual application.

I assume that one fits a regression model based on maximizing the marginal likelihood (3) using GLS, thereby ignoring any correlation between the residual and the exposure, and I assess bias as a function of the spatial scales of $\boldsymbol{X}$ and $\boldsymbol{Z}$. I take $\boldsymbol{X}$ and $\boldsymbol{Z}$ to be generated based on Gaussian processes with Matérn spatial covariance function with $\nu = 2$, which gives continuous and differentiable realizations. The model (3) is equivalent to both a mixed model and a universal kriging model if one knows the variance and spatial dependence parameters.

2.3 Bias with known parameters

I first consider bias when the variance parameters are known and only the regression coefficients, β_0 and β_x, are unknown. To start, assume that $\boldsymbol{X}$ and $\boldsymbol{Z}$ share the same spatial correlation range (θ, where increasing θ corresponds with increasingly large-scale variation, i.e., more smooth spatial processes, as in Fig. 1), but may have different marginal variances, namely, $\text{Cov}(\boldsymbol{X}) = \sigma_x^2 R(\theta)$ and $\text{Cov}(\boldsymbol{Z}) = \sigma_z^2 R(\theta)$ and $\text{Cov}(\boldsymbol{X}, \boldsymbol{Z}) = \rho\sigma_x\sigma_z R(\theta)$. If $\boldsymbol{X}$ and $\boldsymbol{Z}$ are assumed to be normally distributed with means μ_x and μ_z, then straightforward algebra shows

$$\begin{aligned}
\text{E}(\hat{\beta}_x|\boldsymbol{X}) &= \beta_x + \left[(\mathcal{X}^T\Sigma^{-1}\mathcal{X})^{-1}\mathcal{X}^T\Sigma^{-1}\text{E}(\boldsymbol{Z}|\boldsymbol{X})\beta_z\right]_2 \\
&= \beta_x + \left[(\mathcal{X}^T\Sigma^{-1}\mathcal{X})^{-1}\mathcal{X}^T\Sigma^{-1}\mathcal{X}\begin{pmatrix} \mu_z - \rho\frac{\sigma_z}{\sigma_x}\beta_z\mu_x) \\ \rho\frac{\sigma_z}{\sigma_x}\beta_z \end{pmatrix}\right]_2 \\
&= \beta_x + \rho\frac{\sigma_z}{\sigma_x}\beta_z, \qquad (6)
\end{aligned}$$

where $\mathcal{X} = [\mathbf{1}\ \boldsymbol{X}]$, $[\cdot]_2$ indicates the second element of the 2-vector, and $\Sigma = \sigma_g^2 R(\theta) + \tau^2 I$. I use standard conditional normal calculations for $\text{E}(\boldsymbol{Z}|\boldsymbol{X}) = \mu_z\mathbf{1} + \rho\sigma_x\sigma_z R(\theta)\sigma_x^{-2}R(\theta)^{-1}(\boldsymbol{X} - \mu_x\mathbf{1}) = (\mu_z - \rho\frac{\sigma_z}{\sigma_x}\mu_x)\mathbf{1} + \rho\frac{\sigma_z}{\sigma_x}\boldsymbol{X}$. The resulting bias, $\rho\frac{\sigma_z}{\sigma_x}\beta_z$, is the same as if $\boldsymbol{X}$ and $\boldsymbol{Z}$ were not spatially structured and is also equal to the bias under OLS. This demonstrates that we have not accounted for confounding at all by fitting the model that accounts for spatial structure. As with OLS, the model attributes as much of the variability as possible to the exposure, including all of the variability in $\boldsymbol{Z}$ that is related to $\boldsymbol{X}$, indicating that inclusion of a spatial residual does not account for all the variability in the outcome at that scale. If $\rho = 0$, the bias is zero because we average over stochastic variability in $\boldsymbol{Z}$, so any non-orthogonality between $\boldsymbol{X}$ and $\boldsymbol{Z}$ in individual realizations contributes to variance rather than bias. This stands in contrast to the presence of bias terms in Rice (1986) and Dominici et al. (2004) caused by non-orthogonality of fine-scale variation in $\boldsymbol{X}$ and the nonparametric component of the model, since neither is treated stochastically.

When the spatial scales of $\boldsymbol{X}$ and $\boldsymbol{Z}$ are different, but known, one can most easily introduce correlation by assuming $\boldsymbol{X}$ is a multi-scale process and introducing correlation between $\boldsymbol{Z}$ and one of the components of $\boldsymbol{X}$. This multiscale scenario is similar to the temporal scenario of Dominici et al. (2004) and is implicitly the framework underlying Peng et al. (2006) and Zeger et al. (2007), in which one hopes to reduce bias from unmeasured variables varying at large temporal or spatial scales.

Let $\boldsymbol{X} = \boldsymbol{X_c} + \boldsymbol{X_u}$ be decomposed into a component, $\boldsymbol{X_c}$, that is at the same scale as the confounder, $\boldsymbol{Z}$, and a component at a different scale, $\boldsymbol{X_u}$, with $\text{Cov}(\boldsymbol{X}) = \sigma_c^2 R(\theta_c) + \sigma_u^2 R(\theta_u)$, $\text{Cov}(\boldsymbol{Z}) = \sigma_z^2 R(\theta_c)$ and $\text{Cov}(\boldsymbol{X}, \boldsymbol{Z}) = \text{Cov}(\boldsymbol{X_c}, \boldsymbol{Z}) = \rho\sigma_c\sigma_z R(\theta_c)$, and other assumptions as above. After some straightforward algebra and matrix manipulations, we have

$$\begin{aligned}
\text{E}(\hat{\beta}_x|\boldsymbol{X}) &= \beta_x + \left[(\mathcal{X}^T\Sigma^{*-1}\mathcal{X})^{-1}\mathcal{X}^T\Sigma^{*-1}\text{E}(\boldsymbol{Z}|\boldsymbol{X})\beta_z\right]_2 \\
&= \beta_x + c(\boldsymbol{X})\rho\frac{\sigma_z}{\sigma_c}\beta_z \qquad (7)
\end{aligned}$$

where

$$c(\boldsymbol{X}) \equiv \left[(\mathcal{X}^T \Sigma^{*-1} \mathcal{X})^{-1} \mathcal{X}^T \Sigma^{*-1} M(\boldsymbol{X} - \mu_x \mathbf{1})\right]_2 p_c,$$

$$\Sigma^* \equiv \frac{\beta_z^2 \sigma_z^2 R(\theta_c) + \tau^2 I}{\beta_z^2 \sigma_z^2 + \tau^2} = \left((1 - p_z) I + p_z R(\theta_c)\right),$$

$$M \equiv (p_c I + (1 - p_c) R(\theta_u) R(\theta_c)^{-1})^{-1},$$

and $p_z \equiv \beta_z^2 \sigma_z^2 / (\beta_z^2 \sigma_z^2 + \tau^2)$. We see that the bias term is proportional to that in the single scale setting, multiplied by an additional term, $c(\boldsymbol{X})$, that modulates the bias. $c(\boldsymbol{X})$ necessarily includes an extra multiplicative factor, $p_c \equiv \sigma_c^2 / (\sigma_c^2 + \sigma_u^2)$, indicating the magnitude of the confounded component of $\boldsymbol{X}$ relative to the total variation in $\boldsymbol{X}$. While the term $c(\boldsymbol{X})$ is complicated, we can explore its dependence on the spatial scales and the magnitudes of the variance component ratios, in particular, θ_c, θ_u, p_z, and p_c to to see how the bias compares to the same-scale setting.

For a grid of $n = 100$ locations on the unit square, Fig. 2 shows the average of $c(\boldsymbol{X})$ over 1000 simulations as a function of θ_c and θ_u, for combinations of p_c and p_z. There is a simple pattern to the bias modification relative to the same-scale setting. For $\theta_c = \theta_u$, $\mathrm{E}_X c(\boldsymbol{X}) = p_c$, which is equivalent to the same-scale result (6), after accounting for the proportion of variability in $\boldsymbol{X}$ that is confounded. Above the diagonal, for $\theta_u > \theta_c$, $\hat{\mathrm{E}}_X c(\boldsymbol{X}) > p_c$, indicating the potential for even more bias when the scale of confounding is smaller than the scale of unconfounded variability. Only when $\theta_u < \theta_c$, and particularly when $\theta_u \ll \theta_c$, do we see reduced bias compared to the same-scale setting, with the potential for bias reduction from modeling the spatial residual. If one calculates the analogous bias to (7) for OLS applied to spatial data, it is nearly constant ($\hat{\mathrm{E}}_X c(\boldsymbol{X}) \approx p_c$; not shown), again assuming known variance and covariance parameters. For these scales, the results support the intuition that including a spatial residual in the model reduces bias, but highlight that even if the covariance parameters are known, inclusion of the spatial residual does not give unbiased estimates and that bias may be substantial in many scenarios. As in the single scale setting, if $\rho = 0$, the bias is zero because one averages over stochastic variability in $\boldsymbol{Z}$. In contrast, if one were to consider single realizations of $\boldsymbol{X}$ and $\boldsymbol{Z}$, bias would not be zero to the extent there is concurvity between the realization of $\boldsymbol{X}$ and that of $\boldsymbol{g} \equiv \beta_z \boldsymbol{Z}$. Results are very similar when locations are sampled uniformly on the unit square or in a clustered fashion (using a Poisson cluster process).

A complication in the simulations is that the sample variance of spatial process values (calculated over the domain) decreases as θ increases (because the sample variance over the domain in a single spatial replicate underestimates population variability - see Fig. 1b-c for examples of the reduced variability). I want to have fixed ratios of average sample variances $p_z \equiv \beta_z^2 \mathrm{E}_Z s_z^2 / (\beta_z^2 \mathrm{E}_Z s_z^2 + \tau^2) \in \{0.1, 0.5, 0.9\}$ and $p_c \equiv \mathrm{E}_{X_c} s_c^2 / (\mathrm{E}_{X_c} s_c^2 + \mathrm{E}_{X_u} s_u^2) \in \{0.1, 0.5, 0.9\}$ for all values of θ_c and θ_u, thereby avoiding the introduction of artifacts caused solely by having ratios of sample variances change with the spatial ranges. To achieve this I generate $\boldsymbol{X_c} \sim \mathcal{N}(\mathbf{0}, k_c^2 \sigma_c^2 R(\theta_c))$ and $\boldsymbol{X_u} \sim \mathcal{N}(\mathbf{0}, k_u^2 \sigma_u^2 R(\theta_u))$ and modify the calculation of $c(\boldsymbol{X})$ in (7) accordingly. k_c and k_u are functions of θ_c and θ_u, respectively, that are chosen such that

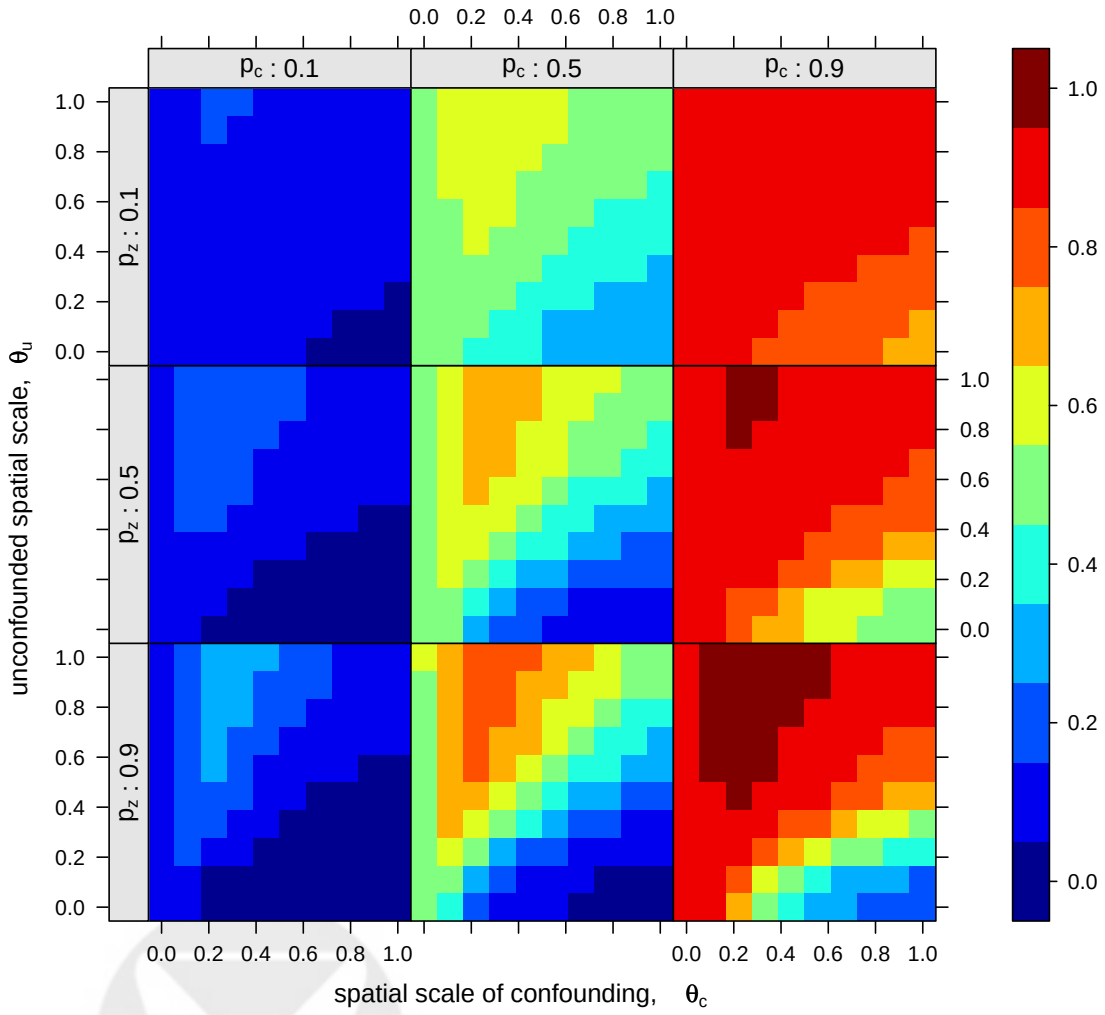

Figure 2: The expected value of the bias modification term, $c(\boldsymbol{X})$, as a function of θ_c and θ_u for a selection of values of p_z and p_c. $c(\boldsymbol{X})$ quantifies the amount of bias relative to the bias in the same-scale setting or with non-spatial confounding ($\rho\sigma_z\beta_z/\sigma_x$). Along the diagonal ($\theta_c = \theta_u$) $\mathrm{E}_X c(\boldsymbol{X}) = p_c$, which is equivalent to no bias reduction. Values near zero indicate substantial bias reduction.

$\mathrm{E}_{X_c} s_c^2(\theta_c) = \sigma_c^2$ and $\mathrm{E}_{X_u} s_u^2(\theta_u) = \sigma_u^2$, where $s_c^2(\theta_c)$ is the sample variance of $\boldsymbol{X_c}$ for a given realization under θ_c and analogously for $s_u^2(\theta_u)$. These manipulations allow me to present bias for scenarios that correspond to specific ratios of average sample variability of $\boldsymbol{X_c}$, $\boldsymbol{X_u}$, $\boldsymbol{Z}$, and $\boldsymbol{\epsilon}$ over the spatial domain.

In an application, there is likely to be small-scale spatial variability in the residual, in addition to any large-scale variability and non-spatial heterogeneity, giving the model

$$\boldsymbol{Y} = \beta_0 \mathbf{1} + \beta_x \boldsymbol{X} + \beta_z \boldsymbol{Z} + \boldsymbol{h} + \boldsymbol{\epsilon}. \tag{8}$$

I considered data generated from this model with $\boldsymbol{h} \sim \mathcal{N}(\mathbf{0}, \sigma_h^2 R(\theta_u))$ independent of $\boldsymbol{X}$, $\boldsymbol{Z}$, and $\boldsymbol{\epsilon}$, and all of the other details as before. Under this data-generating model and again assuming all variance parameters are known, simulations of $\mathrm{E}_X c(\boldsymbol{X})$ indicate that bias is somewhat reduced relative to that seen in Fig. 2 for $\theta_c > \theta_u$ and increased when $\theta_c < \theta_u$ (not shown). Apparently, since the residual covariance accounts for spatial structure at a scale finer than the scale of confounding, bias is reduced because the model attributes more of the variability at the confounded scale to the residual than when the unconfounded fine-scale variability is not included.

2.4 Bias and precision with parameters estimated under penalized models

I set up a simulation study to assess the impact of estimating variance and spatial dependence parameters. In addition to maximum likelihood estimation of a mixed effects/kriging model based on the marginal likelihood (3), I consider the use of penalized likelihood to fit the model (2) with a penalized thin plate spline spatial term for $g(s)$, implemented in gam() in R, which uses generalized cross-validation for data-driven smoothing parameter estimation (Wood, 2006). For the core simulations, I set the following parameter values, $\sigma_u^2 = \sigma_c^2 = \beta_z^2 \sigma_z^2 = 1$, $\tau^2 = 4$, $\beta_x = 0.5$, $\rho = 0.3$, and sampled 100 spatial locations uniformly from the unit square. For a range of values of θ_c and θ_u, I simulated 2000 datasets for each pair $\{\theta_c, \theta_u\}$. For each simulated dataset, new spatial locations are generated, as are new values of $\boldsymbol{X}$ and $\boldsymbol{Z}$; $\boldsymbol{Y}$ is then generated using (4). Again we have to account for the reduced empirical spatial variability as θ increases; for comparison with results in Fig. 2 the simulations have effective values of $p_c = 0.5$ and $p_z = 0.2$.

For bias, the simulation results (Fig. 3) for the mixed/kriging model reasonably match the theoretical value with known variance parameters, albeit with less bias reduction when $\theta_u \ll \theta_c$, because the model fit sometimes finds little or no spatial structure in the residuals, pushing bias results toward the increased bias seen under OLS. Results for the penalized spline model show less bias for $\theta_u < \theta_c$ than the mixed model, presumably caused by the difference between estimating the amount of smoothing by GCV compared to maximum likelihood. Note that the random effects distribution induces a natural penalty on the complexity of the spatial residual, $\boldsymbol{g}$, through the determinant in the marginal likelihood (3), rather than imposing a penalty directly through a smoothing parameter as in gam(). In either case, spatial scales are critical and substantial bias reduction occurs only when the scale of confounding is larger than the scale of the unconfounded variability. Not shown are additional simulations that indicate that as the correlation of confounder and exposure increases, or the magnitude of variation in the confounder increases, or the effect size decreases, relative bias increases. In such scenarios, substantial bias reduction is seen only for

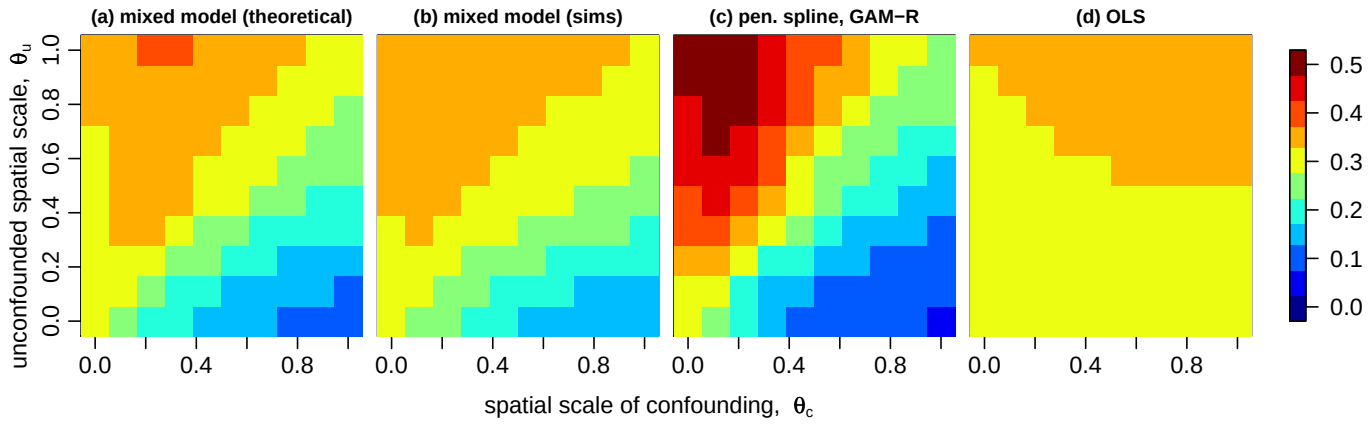

Figure 3: Relative bias, comparing (a) theoretical bias for the mixed model with known variance parameters to simulated bias with fitted variance parameters for (b) the mixed model, (c) a penalized spline model, and (d) OLS.

very small spatial scales in the exposure and large scales of confounding (i.e., the lower right areas of the subplots) (not shown).

Fig. 4 shows that there is a substantial bias-variance tradeoff with the lower bias of the penalized spline model trading off with increased variance and resulting increased MSE (except when θ_u is very small). Both model variance estimates (third column) understate the variance in the coefficient estimates (second column), with particular underestimation of uncertainty and low coverage for the mixed/kriging model, and with coverage decreasing as one moves away from the region of $\theta_c \gg \theta_u$. Of course much of the poor coverage is caused by the bias.

Fitting the mixed/kriging model by REML rather than ML showed moderate improvement in coverage with the average variance estimate more similar to the variance of the estimated coefficients. Using $\nu = 0.5$ (i.e., an exponential spatial correlation function) in the fitting rather than the true $\nu = 2$ had little effect on results, but when the unconfounded variability is generated based on $\nu = 0.5$, bias is reduced substantially relative to the core results (particularly note that there is bias reduction when $\theta_c = \theta_u$), apparently because the non-differentiable sample paths of processes with exponential covariance play the role of very fine-scale, unconfounded variability. There was little change in results when using spatial locations simulated using a Poisson cluster process with an average of seven children per cluster and cluster kernel standard deviation of 0.03. Finally, simulations under $\rho = 0$, i.e., no confounding, support the lack of theoretical bias for the mixed/kriging model estimation and also find no bias for the penalized spline model.

Our bias results when $\rho \neq 0$ are analogous to the bias seen with penalized spline models in He et al. (2006) and Peng et al. (2006) . There, concurvity (i.e., $\hat{\rho} \neq 0$) between the smooth temporal term (analogous to our spatial residual) and the exposure emerged from the fixed basis coefficients chosen based on empirical data examples (R. Peng, personal communication, He et al., 2006). Similar results are seen in the spatial settings of Ramsay et al. (2003). Viewed through the framework developed here, the non-zero $\hat{\rho}$ arising in these various published simulation studies causes the bias.

When there is small-scale independent variation in the residual (8), the presence of this variation causes

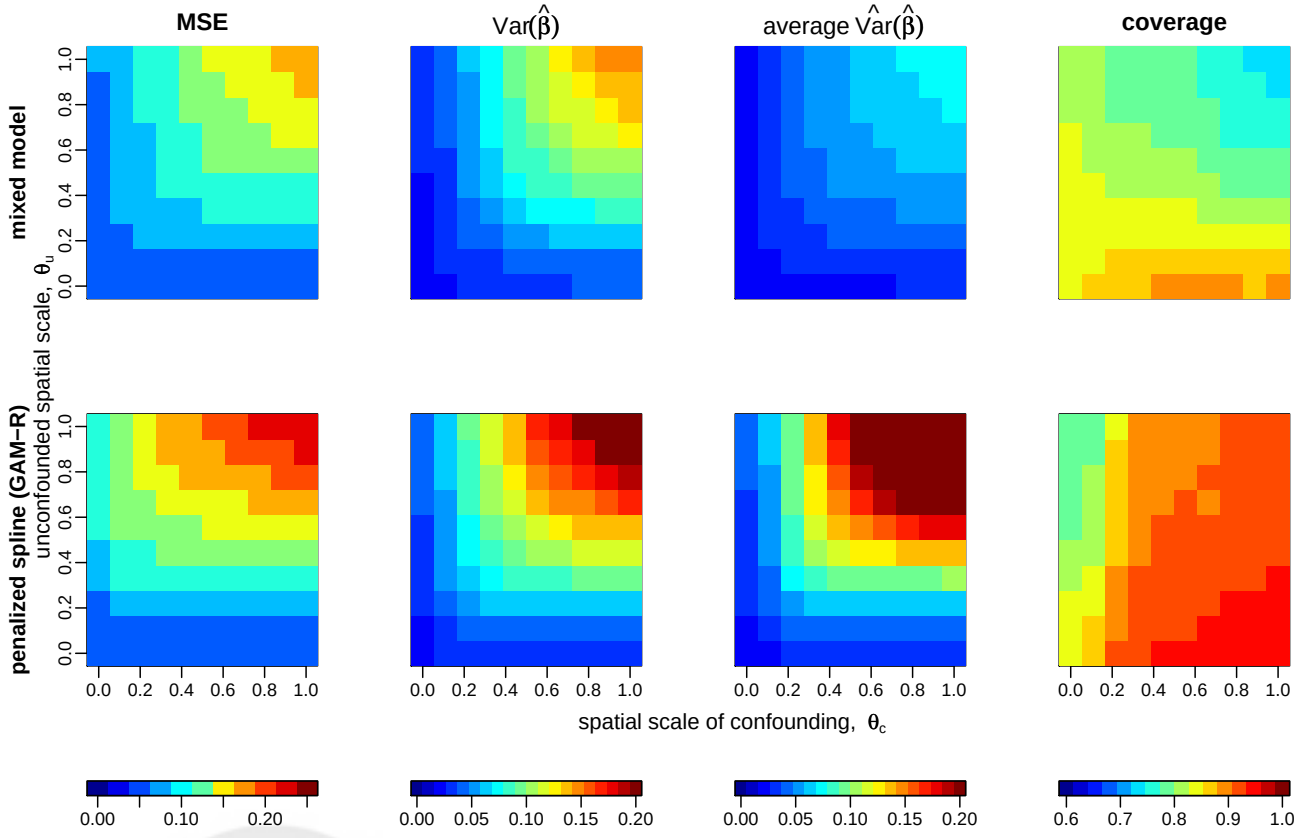

Figure 4: Simulation results for (top row) mixed model/kriging fit and (bottom row) penalized spline model. Each plot shows results as a function of the spatial scales of the confounded, θ_c, and unconfounded variability, θ_u, with MSE (first column), variance of the estimates over the simulations (second column), average squared standard error (third column) and coverage (fourth column).

reduced bias for $\theta_u < \theta_c$ (not shown), relative to the results presented above. This occurs through an increase in the number of degrees of freedom estimated from the data to capture residual variability, i.e., undersmoothing with respect to the variation at the θ_c scale. This scenario seems quite likely in applications: if there is large-scale residual spatial structure, there is likely to be finer-scale structure as well, so analyses that attempt to best fit the data may in the process reduce bias. The bias reduction is analogous to that seen with undersmoothing in the partial spline setting (Rice, 1986; Speckman, 1988). Of course as discussed in Section 2.3, one would not expect a reduction in bias if the residual $\boldsymbol{g}$ were correlated with $\boldsymbol{X_u}$, so we can only expect a reduction in bias from confounding occurring at the larger scales.

2.5 The bias-variance tradeoff

We have seen that even when all covariance parameters are known and the scale of confounding is much larger than the scale of unconfounded variability in $\boldsymbol{X}$, bias remains, albeit at a much reduced level.

In principle, if the structure at the confounded scale could be exactly fit using a set of basis functions, such as a regression spline with a certain number of basis functions, then the exposure effect estimate is unbiased, as in any multiple regression, and as demonstrated by Dominici et al. (2004) in the time series context when $\boldsymbol{X_u}$ is white noise, because one is projecting on the basis functions. The partial residual kernel smoothing approach of Speckman (1988) reduces bias in similar fashion, albeit without using a projection, through the technique of twicing, which reduces bias compared to partial spline models. However, in a real application, one has to choose the basis functions, and if one has not chosen a set of basis functions that fully explain the confounded, large-scale variability, even with a basis of seemingly sufficient dimension, this will induce a bias. One could instead consider a penalized spline approach with penalty parameter chosen in advance to give the desired effective degrees of freedom (edf). For fixed df, since the penalized spline smoother is not a true projection (Speckman, 1988; Peng et al., 2006), one might expect the penalized spline approach to have more bias than the regression spline approach. Heuristically, bias in this approach is caused by the confounded component of the variability in the outcome not being fully explained through the estimated spatial term, with bias analogous to that seen in the partial spline setting (Rice, 1986; Speckman, 1988). However, we would expect the penalized spline to be less sensitive to the exact form of the basis functions and number and placement of knots, as is seen to be the case in the example (Section 4). Furthermore, one can always undersmooth to reduce the bias, following the recommendation in the partial spline literature (Rice, 1986; Speckman, 1988). Thus, using a penalized spline seems reasonable, albeit without the clean interpretation of estimating the exposure effect conditional on the variation accounted for through the regression spline. I show below that simulations comparing regression spline and penalized spline models support these theoretical results from the literature in the spatial context considered here, with regression splines reducing bias more than penalized modeling, but with a price paid in terms of variance.

The primary issue in an application is choosing the amount of smoothing to reduce bias, if this is the goal rather than best fitting the data. Estimating the amount of smoothing based on the data might reduce bias (if there is small-scale residual correlation) or might have little effect on bias (if the data suggest large-scale residual correlation). However any reduction will depend on the scales involved and the actual amount of smoothing done, and the analysis will reveal little about the sensitivity of estimation to scale. Instead, one

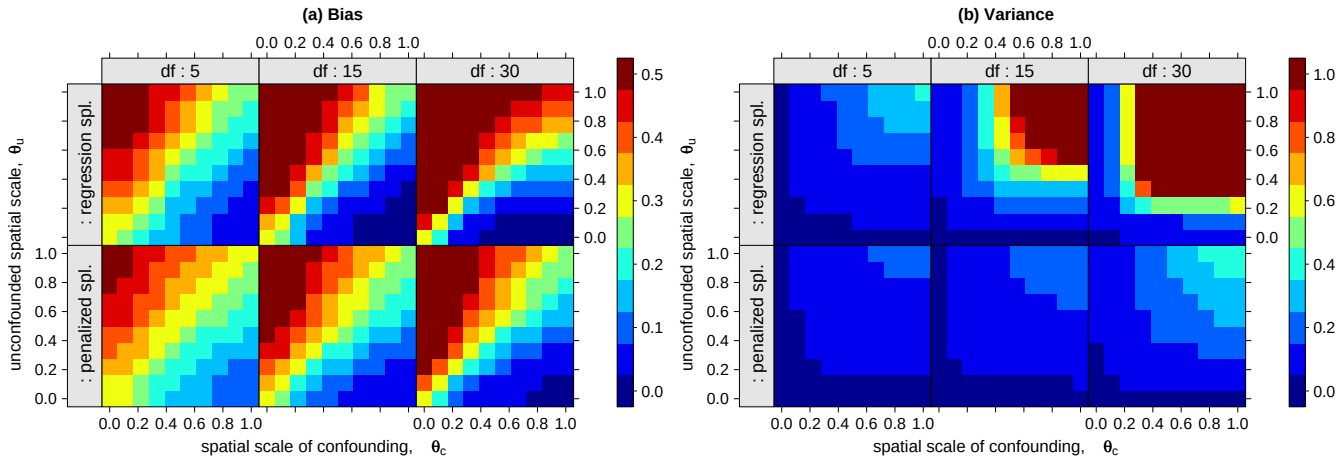

Figure 5: Simulation results for bias (a) and variance (b) for regression splines (top row) and penalized splines (bottom row) with 5, 15 and 30 df, where the df are pre-specified, rather than estimated based on the data.

might consider explicitly assessing the bias-variance tradeoff by considering various amounts of smoothing and assessing sensitivity of the exposure effect estimate to a metric for the amount of smoothing, such as the effective degrees of freedom. One approach is a spatial analogue to the sensitivity analysis approaches of Peng et al. (2006): fit a model with spatial basis functions and vary the edf (e.g., Zeger et al., 2007). Plotting $\hat{\beta}_x$ and uncertainty intervals as a function of df provides an assessment of the robustness of results to potential spatial confounding at various scales, with an explicit focus on the bias-variance tradeoff. If one is concerned about confounding at a particular scale, then one can report the results for a value of df that would undersmooth with respect to that scale to reduce bias, accepting the tradeoff of increased uncertainty. Depending on the smoothing approaches used, other metrics than edf may be appropriate. One might consider estimating the variability of $\hat{\boldsymbol{g}}$ to assess the magnitude of the variation in the outcome that is captured through $\hat{\boldsymbol{g}}$. With linear smoothers, one could also consider the spatial extent of the effective kernels based on the hat matrix.

In Fig. 5a, I show simulation results under the same settings as in Section 2.4, using regression splines (i.e., unpenalized fixed effects) and varying the df, again assessing effects of bias as a function of the spatial scales involved. By choosing a large number of df, one can decrease bias more effectively than when estimating the amount of smoothing from the data (Fig. 3). However, as the unconfounded scale of variability increases to moderate and large scales, the variability in the estimates in the fixed effects model increases dramatically (Fig. 5b), causing a concordant increase in the mean squared error (not shown), highlighting the bias-variance tradeoff. In contrast, using penalized splines with fixed df shows much more stable results; bias is not reduced as much as with regression splines (Fig. 5a), consistent with the results of Peng et al. (2006), but there is much less variability than with regression splines (Fig. 5b).

A diagnostic approach to understanding whether the residual may be confounded with the variable of interest is to assess the correlation of $X(s)$ and the residual. Of course the residual is unknown without knowing β_x. One might use a variety of plausible values of β_x to estimate $g(s)$ and then calculate the

correlation with $X(s)$ (and potentially with filtered versions of $X(s)$ that exclude small-scale variation). This will help to understand the potential for confounding by letting one assess the potential spatial scales that are operating and the implications for bias and precision.

2.6 Accounting for residual spatial correlation

If one accounts for large-scale variation as a means of reducing bias, there may still be small-scale residual variation, such as fine-scale correlation in health outcomes related to residential sorting. As I have shown, any confounding bias from this variation can be reduced only if there is variability in the exposure at an even smaller scale. If there is not, then one is effectively assuming that the small-scale residual is uncorrelated with the exposure. Given this assumption, one may need to consider methods to account for the small-scale residual spatial variation so that uncertainty estimation for $\hat{\beta}_x$ is not compromised. (Of course to the extent that there is small-scale spatial confounding bias, this will not be corrected and the resulting uncertainty estimate will not account for this.) In Section 3.3 I suggest that if the smoothing employed to reduce bias removes variation at large and moderate scales, the remaining variability in the residual may be at sufficiently small scales that it has little effect on uncertainty estimation.

Alternatively, we have seen that the mixed model approach favors attributing variability to the exposure, so one could fit a penalized model with the amount of smoothing determined from the data. This has two effects; first it naturally accounts for the effect of the spatial structure on uncertainty estimation. Second, in the presence of small-scale residual variability, fitting the small-scale variability will naturally undersmooth with respect to large-scale variability that may cause confounding, thereby reducing bias from confounding at the larger scale, as discussed previously.

A different approach to characterizing uncertainty is to use a robust analysis, such as fitting the marginal model using an estimating equation (Liang and Zeger, 1986) with uncertainty based on the sandwich estimator, which is robust to misspecification of the residual variance. Note that this does not account for spatial confounding bias at any scale because the marginal variance is assumed not to be related to the exposure. However, one could fit the marginal model with regression spline terms in the mean to account for large-scale spatial confounding bias.

3 Spatially-correlated residuals and precision

Assuming unbiased estimation, namely that the residual and the covariate are not correlated ($\rho = 0$ in the framework of Section 2), I next consider effects of spatial scale on the following questions about efficiency of estimators and quantification of uncertainty:

1.) Given a fixed amount of residual variation, how is efficiency affected by the proportion of that variation that is spatial?

2.) What is the magnitude of the improvement in efficiency when accounting for residual spatial variation, relative to OLS?

3.) If one uses the naive estimator for the variance of $\hat{\beta}_{\text{OLS}}$, what is the magnitude of the error in uncertainty estimation compared to the true variance estimator for $\hat{\beta}_{\text{OLS}}$?

The first question does not appear to have been raised in the literature. With regard to the second, while we know that GLS is the most efficient estimator when the residuals are correlated, here I investigate the magnitude of this efficiency advantage as a function of the spatial scales involved. Regarding the third, the conventional wisdom in the statistical literature and among applied practitioners appears to be that not accounting for spatial structure leads to underestimation of uncertainty (e.g., Legendre, 1993; Burnett et al., 2001; Schabenberger and Gotway, 2005, p. 324), but see also Ma et al. (2007). However, I have not seen a formal quantification of this underestimation for a regression coefficient. Both Cressie (1993, Sec. 1.3) and Schabenberger and Gotway (2005, Sec. 1.5) comment on the (potentially severe) underestimation of uncertainty for the mean of a spatial process, in an infill asymptotic setting, showing that the sample average (i.e., the OLS estimator) is inconsistent, but do not develop the problem for a regression coefficient.

Note that there are three variance estimators (i.e., estimators for the squared standard error of the regression coefficient) under consideration here: the true GLS variance estimator, and the true and naive OLS variance estimators. When $\rho = 0$, OLS is unbiased, so it makes sense to consider OLS for estimation, provided we adjust the usual OLS variance estimator to account for the residual spatial correlation. While actual applications will likely involve more complicated modeling, consideration of these questions in this simple setting helps to understand the basic issues. Similarly, I conduct the analysis assuming known variance components for simplicity.

3.1 Relationship between spatial scale and GLS efficiency

Given a fixed amount of residual variation, how is efficiency affected by the proportion of variation that is spatial?

Lemma 1: Consider the model (3) and assume all parameters are known except β_0 and β_x. The expectation of the precision of $\hat{\beta}_{\text{GLS}}$, with respect to the sampling distribution of $\boldsymbol{X}$ is

$$\text{E}_X(\text{Var}(\hat{\beta}_{\text{GLS}})^{-1}) = \frac{\sigma_x^2}{\tau^2 + \sigma_g^2} \left(\text{tr}\{\tilde{\Sigma}^{-1} R_x\} - \frac{\mathbf{1}^T \tilde{\Sigma}^{-1} R_x \tilde{\Sigma}^{-1} \mathbf{1}}{\mathbf{1}^T \tilde{\Sigma}^{-1} \mathbf{1}} \right), \tag{9}$$

where $\tilde{\Sigma} \equiv (1 - p_g) I + p_g R_g$ and $p_g \equiv \sigma_g^2/(\sigma_g^2 + \tau^2)$. See the Appendix for the proof.

Note that the term in parentheses is an effective sample size, analogous to $n - 1$ in the nonspatial problem, adjusted here for spatial structure in residual and exposure, with the second component in parentheses analogous to the degree of freedom lost for estimating the mean in the nonspatial problem.

Fig. 6(top) shows the expected precision as a function of θ_x and θ_g, averaging (9) over 500 sets of $n = 100$ locations simulated uniformly on the unit square. I report the expected precision divided by a baseline of $\sigma_x^2(n-1)/(\tau^2 + \sigma_g^2)$, which is the expected precision in the nonspatial setting, assuming that the total residual variation, $\tau^2 + \sigma_g^2$ remains constant. Compared to the nonspatial setting, lower precision occurs unless the exposure varies at small spatial scale. When the exposure varies at small spatial scale and the residual at larger spatial scale, precision can be substantially greater than the nonspatial setting, because the model is able to account for part of the residual variance through the spatial structure, as if the spatial structure were an additional covariate to which variation in the response is attributed. GLS implicitly estimates the process, $\boldsymbol{g}$ in (2), that gives rise to the marginalized model (3), reducing the residual variability

and thereby causing the GLS precision to be larger than with independent residuals but equivalent overall residual variability. In contrast, when $\boldsymbol{X}$ varies only at large spatial scales, then the GLS precision is smaller, which is related to difficulty in distinguishing $\beta X(s)$ from $g(s)$. Note that when fixing θ_x at moderate to large values, as θ_g increases, at first precision decreases because the scales of $\boldsymbol{X}$ and $\boldsymbol{g}$ are similar, making it difficult to identify the effect of $\boldsymbol{X}$, but as θ_g increases further, precision increases, as the model is able to attribute smooth residual variability to $\boldsymbol{g}$. Results are similar using points on a regular grid or clustered based on a Poisson cluster process.

3.2 Efficiency of GLS and OLS estimators

Here I consider how spatial scale affects the relative efficiency of spatial and non-spatial estimators, comparing the precisions of the OLS and GLS estimators. Since the true OLS variance estimator,

$$\left[(\mathcal{X}^T\mathcal{X})^{-1}(\mathcal{X}^T(\tau^2 I + \sigma_g R_g)\mathcal{X})(\mathcal{X}^T\mathcal{X})^{-1}\right]_{2,2},$$

is a complicated function, it is difficult to derive closed form expressions for efficiency relative to the GLS estimator. Instead I conducted a small simulation study. For a regular grid of values of θ_x and θ_g I carried out 500 simulations for each pair of values, with $n = 100$ observations whose spatial locations are drawn uniformly over the unit square domain. Note that because I consider the ratio of the variance estimators, the values of σ_x^2 and $\sigma_g^2 + \tau^2$ cancel out of the ratio and do not affect the results.

Fig. 6(middle) shows the the expected relative precision as a function of the spatial scales, θ_g and θ_x, and the proportion of the residual variability that is spatial. When little of the residual variability is spatial ($p_g = 0.1$), there is little gain in precision, as expected. When more is spatial, the gains in precision are small when $\boldsymbol{g}$ varies only at small scale, but substantial for large values of θ_g, namely for spatially-smooth residual structure. Unfortunately, this is precisely the case in which one would be concerned about spatial confounding. If we assume that the large-scale structure in the residual has been controlled for in an effort to reduce the potential for bias, then with the remaining residual variability being fine-scale, there is limited gain in precision regardless of the spatial scale of the exposure. With locations on a regular grid, the gains in precision are slightly less for small values of θ_g, while with more clustered sampling (I used a Poisson cluster process), the gains increase somewhat for small values of θ_g. See also Dow et al. (1982) for similar simulation results in a setting with correlation induced through a Markov random field structure.

3.3 Underestimation of uncertainty by the naive OLS variance estimator

Applications often ignore residual spatial correlation. Assuming OLS is used, a key question involves the impact of using the incorrect naive OLS variance estimator rather than the true variance of the OLS estimator. One can express the ratio of the true OLS variance to the naive OLS variance as follows. First define $\boldsymbol{W} \equiv (\boldsymbol{X} - \bar{X}\mathbf{1})/s$ where $s^2 \equiv \frac{1}{n}\sum(X_i - \bar{X})^2$. After expressing $\hat{\beta}_x = [(\mathcal{X}^T\mathcal{X})^{-1}\mathcal{X}^T\boldsymbol{Y}]_2 = (\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^T\boldsymbol{Y}$

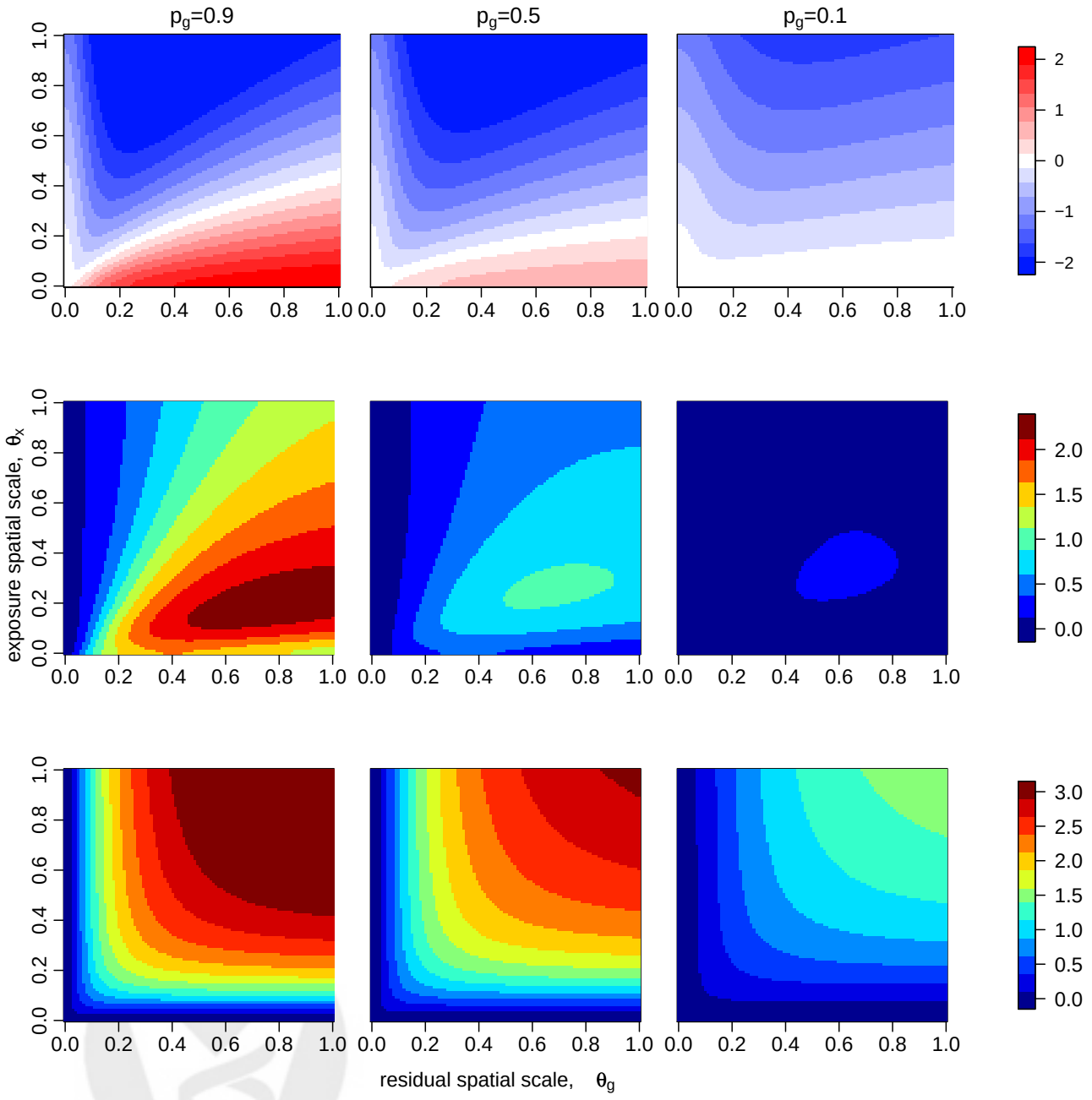

Figure 6: Efficiency and precision results for three values of $p_g = \sigma_g^2/(\sigma_g^2 + \tau^2)$ (columns) as a function of the spatial scales (ranges) of the residual and the exposure. (top row) The log of the expected ratio of the precision of the GLS estimator (9) relative to the precision in the non-spatial setting. (middle row) The log of the expected ratio of GLS to OLS precision. (bottom row) The log of the expected ratio of the correct and naive OLS variance estimators. The results are based on 500 simulations for each set of parameter values, with a Matérn correlation with $\nu = 2$ and 100 locations sampled uniformly over the unit square.

where $\tilde{\boldsymbol{X}} = \boldsymbol{X} - \bar{X}\mathbf{1}$, we have

$$\frac{\text{Var}_{\text{true}}(\hat{\beta}_x)}{\text{Var}_{\text{naive}}(\hat{\beta}_x)} = \frac{(\sigma_g^2+\tau^2)^{-1}(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})}{(\sigma_g^2+\tau^2)^{-1}(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})(\tilde{\boldsymbol{X}}^T\tilde{\Sigma}\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}} = \frac{\tilde{\boldsymbol{X}}^T\tilde{\Sigma}\tilde{\boldsymbol{X}}}{\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}} = \frac{1}{n}\boldsymbol{W}^T\tilde{\Sigma}\boldsymbol{W}.$$

Then we have $\text{E}_X(\frac{1}{n}\boldsymbol{W}^T\tilde{\Sigma}\boldsymbol{W}) = \frac{1}{n}\text{tr}(\tilde{\Sigma}\text{Cov}(\boldsymbol{W}))$. So for $\tilde{\Sigma} \approx I$ or $\text{Cov}(\boldsymbol{W}) \approx I$, i.e. when either θ_g or θ_x are close to zero, we expect the ratio to be near one. Note also that with spatial correlation functions that are non-negative, the only negative contribution to the ratio can be from negative covariances induced by standardizing $\boldsymbol{X}$, which should diminish as the sample size increases, so we expect the ratio to generally be no smaller than one, indicating that the naive variance does underestimate uncertainty. Finally, the largest values of the ratio would occur with large positive correlations in the corresponding elements of $\tilde{\Sigma}$ and $\text{Cov}(\boldsymbol{W})$, which is to be expected when both $\boldsymbol{g}$ and $\boldsymbol{X}$ show large-scale variation.

Fig. 6(bottom) supports these heuristic results, showing the average ratio of variances in simulations, where the simulations were conducted as in Section 3.2. The ratio is close to one when either of the spatial terms has fine-scale variability and far from one when both have large-scale behavior; a result similar to that of Bivand (1980) for inference about a correlation coefficient and to Johnston and DiNardo (1997, p. 178) under serial autocorrelation in a regression setting. As expected, as the proportion of residual variability that is spatial decreases, the expected ratio gets closer to one, indicating that when non-spatial variation dominates the residual and the spatial structure in the residual or exposure is not too large in scale, the naive variance may be reasonable. A lack of large-scale residual structure might be induced by having accounted for large-scale variation in attempting to reduce spatial confounding bias. Results with gridded locations show ratios slightly closer to one, and with clustered locations, ratios further from one.

Simple simulations with spatial ranges and sampling designs specific to an analysis could be easily carried out for further guidance in a given setting, allowing one to assess whether ignoring the spatial structure has substantial impact on uncertainty estimation. Accounting for small-scale spatial correlation requires estimation of the spatial structure and is often computationally burdensome, so being able to use a naive model that assumes independence can have tremendous practical benefit. Of course in some analyses, any underestimation of variability may be cause for concern, in which case use of the naive variance estimator would not be tenable.

It is important to note that these results do not indicate that if one uses a naive analysis and then compares to fitting a spatial model that the uncertainty estimate will increase. Simple simulations (not shown) show that the estimated uncertainty may well decrease because the more sophisticated model both corrects the variance estimate, which increases the estimated uncertainty (but in some cases only by a small amount, as shown above), and uses a more efficient estimator, which decreases uncertainty.

4 Case study: Massachusetts birthweights and air pollution

Chronic health effects of ambient air pollution in developed countries tend to involve small relative risks, but are of considerable public health importance because of widespread exposure. As a result, air pollution epidemiology is an important application in which spatial confounding bias is of critical concern. Studies

need to estimate a small effect from data with high levels of variability and stronger effects from other covariates, including potential confounders such as socioeconomic status.

I reanalyze data on the association between air pollution (ambient estimates of black carbon, a component of particulate matter) and birthweight in eastern Massachusetts (Zeka et al., 2008; Gryparis et al., 2009). These analyses found significant negative effects of traffic proxy variables and black carbon, respectively, on birthweight, with Gryparis et al. (2009) using several methods to try to account for effects of measurement error in the predicted black carbon concentrations, which are based on a regression model that accounts for spatial and temporal structure and key covariates.

I follow these analyses in using an extensive set of covariates to try to account for potential confounding. I use smooth terms for mother's age, gestational age, and mother's cigarette use, to account for nonlinearities, a linear term for census tract income, and categorical variables for the following: presence of a health condition of the mother, previous preterm birth, previous large birth, sex of baby, year of birth, index of prenatal care, and maternal education. The exposure of interest is estimated nine-month average black carbon concentration at the geocoded address of the mother, based on a black carbon prediction model (Gryparis et al., 2007). Following Gryparis et al. (2009), for simplicity I exclude the 13,347 observations with any missing covariate values, giving 205,713 births.

In Gryparis et al. (2009) we found no evidence of residual spatial correlation based on a spatial semivariogram. Further analysis here indicates that there is significant residual spatial variation but that the magnitude of this variation is overwhelmed by non-spatial variation. Fig. 7a shows this lack of pattern in the semivariogram, with the one dot at distance of zero corresponding to correlation of residuals for babies born to mothers living at the same location. If spatial structure were present, the semivariance would be smaller at shorter distances and increase with larger distances. However, this lack of evidence for spatial structure occurs because the spatial variability (Fig. 7b) is swamped by individual non-spatial variability amongst babies. While the spatial variation is small relative to the overall variation, it is large relative to the estimated pollution effect (note the surface values in the range of -40 to 40, for comparison with effect estimates in Fig. 8), so if it is caused by spatially-varying confounders, it could bias estimation of the pollution effect.

I consider both regression splines, an unpenalized approach, and the use of penalized splines, both with df chosen in advance, as in Zeger et al. (2007) and proposed in Section 2.5, as well as a penalized spline with data-driven smoothing parameter estimation based on generalized cross-validation, all as implemented in gam() in R, using the thin plate spline basis (Wood, 2006). For the regression spline, a basis with the desired rank is generated based on a truncated eigendecomposition of the basis (Wood, 2006), so sensitivity to knot placement should be minimal.

I first add a spatial term to the model with the full set of covariates to assess whether some of the estimated effect may be biased by spatial confounding. Fig. 8a shows how the estimated effect of black carbon varies with the df and the spatial smoothing approach. The estimate increases somewhat as more df are used to account for the spatial structure. For the penalized spline, as more than about 10 df are used, the upper confidence limit exceeds zero, and for larger df, the upper limit increases further. GCV chooses 157 df, indicating fairly small-scale spatial structure in the data. For context note that with 129 edf in Fig. 7b

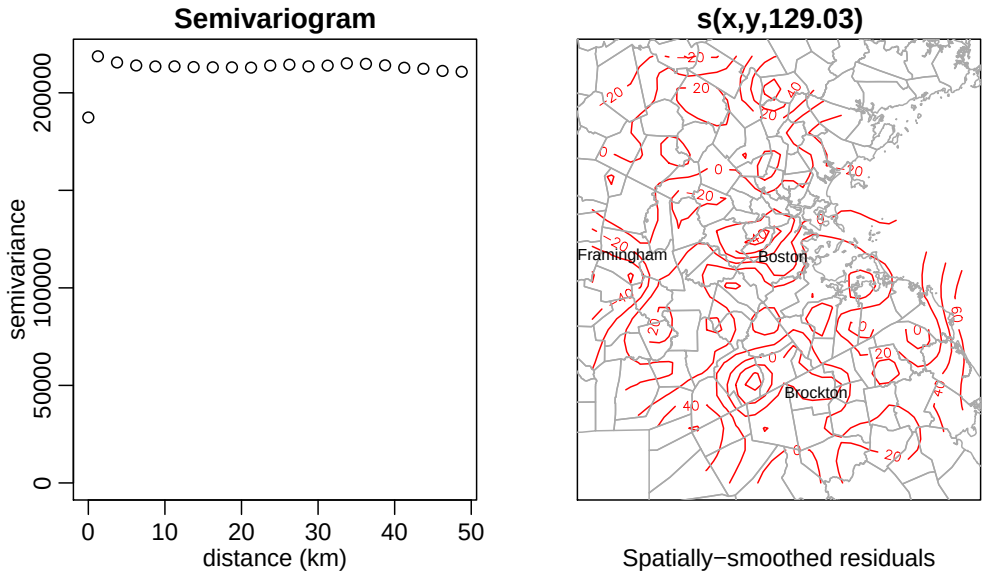

Figure 7: (left) Semivariogram of full model residuals, with the first point representing births at the same location and (right) spatial smooth of residuals. The spatial smooth, with 129 df chosen by GCV, is highly significant.

we see spatial features at the scale of individual towns. While the regression spline approach implemented here avoids having to choose the knots, the empirical results are still very sensitive to df, in contrast to the stability of the penalized spline solution as the edf varies. For both penalized and regression splines, there is a clear bias-variance tradeoff, with increased variance as the number of df increases, though for this problem with a very large sample size, the confidence intervals do not increase drastically, nor is there much difference in the uncertainty between the regression and penalized spline approaches. The spatial confounding assessment suggests that we have somewhat reduced confidence in the black carbon effect. However, the effect estimate is reasonably stable, albeit with some attenuation, even when using a spatial term with a large number of degrees of freedom.

Next I consider what might have happened if most of the covariates (particularly the ones related to socioeconomic status) were not measured, potentially inducing serious confounding. Fig. 8b indicates that without any spatial term in the model, the effect estimate is -23.0 with a confidence interval of (-26.8,-19.2), indicating a much more substantial effect of black carbon than the fully-adjusted model. As soon as one accounts for spatial structure, even with a small number of df, the estimate attenuates, approaching the fully-adjusted estimate, with the upper confidence limit rising above zero. The reduced model estimate appears to suffer from serious confounding, but the spatial analysis is able to account for much of this confounding, substituting for a rich set of covariates. Much of the pollution effect in the reduced model apparently occurs due to large-scale association of pollution and birthweight, which seems likely to be caused by confounding.

Ideally I would fit a model that accounts for fine-scale spatial structure, to ensure that the estimated uncertainty is correct. However, with 205,713 observations, this is a computational challenge that I do not take up for this example. Given the results in Section 3.3 that indicate that most of the variance underestimation occurs from large-scale structure, one can hope that the uncertainty at the larger values of spatial df in Fig.

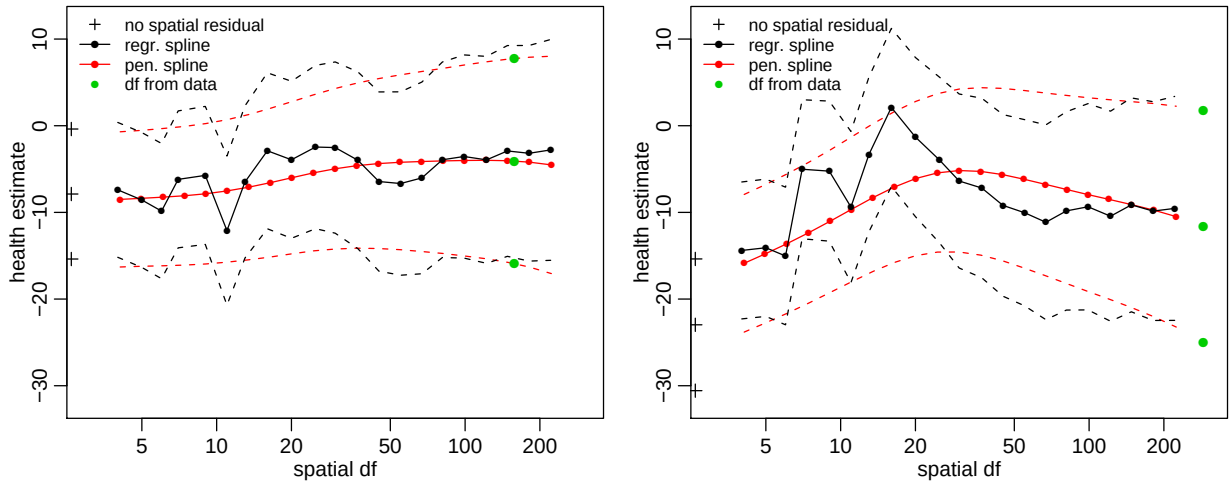

Figure 8: For the model with the full set of covariates (left) and the reduced set of covariates (right), black carbon effect estimates and 95% confidence intervals based on different specifications for the spatial term in an additive model: black pluses indicate the model with no spatial term and green dots with the df chosen by GCV, while black (regression spline) and red (penalized spline) dots indicate results when fixing the degrees of freedom at a set of discrete values. The lines through the points and corresponding dashed lines are taken by connecting the effect estimate and confidence interval bounds for the discrete set.

8 may reasonably approximate the true uncertainty.

5 Discussion

Considerations of scale are critical in spatial regression problems. Standard spatial regression models, based on spatial random effects, kriging specifications, or penalized splines to represent the spatial structure, are penalized models with inherent bias-variance tradeoffs in estimating the smooth function. Under unmeasured spatial confounding, the bias carries over into estimating the coefficient for the exposure of interest, but the degree of bias depends on the spatial scales involved. Inclusion of a spatial residual term accounts for spatial correlation in the sense of reducing bias from unmeasured spatial confounders only when there is unconfounded variability in the covariate of interest at a scale smaller than the scale of the confounding. As in the time series (Dominici et al., 2002; Peng et al., 2006) and partial linear/partial spline (Speckman, 1988) contexts, models that fit better based on model selection criteria have no guarantee of having low bias, since penalized models all trade off bias and variance in search of best explaining the outcome variability.

If the variation in exposure is solely at large scales, there is little opportunity to reduce spatial confounding bias, but with a component of small-scale exposure variability, large-scale spatial confounding bias can be reduced substantially. Accounting for large-scale residual correlation is also important for improving precision of regression estimators and for correctly estimating uncertainty. In contrast, when residual correlation occurs at small scales, there is little opportunity for reducing spatial confounding bias at those scales or improving regression estimator precision. However, under the assumption of no small-scale confounding, fitting such residual structure can reduce bias from larger-scale confounding by causing undersmoothing

with respect to the large-scale structure.

Sensitivity analyses that show the bias-variance tradeoff as a function of the scale at which the spatial residual structure is modeled, such as described in Peng et al. (2006) and Zeger et al. (2007), offer one approach that helps to frame the issue of bias in the context of the spatial scales involved. In choosing a spline formulation to carry out such an analysis, while regression splines have an appealing interpretation and in theory result in less bias in estimating the effect of interest, penalized splines with a fixed effective degrees of freedom may give more stable results. Of course the sensitivity analysis approach does not answer the question of how to get a single estimate of the effect of interest. One might also consider an approach similar to that of Beelen et al. (2007) and explicitly decompose the exposure into multiple scales, including exposure at each scale as a separate covariate and focus causal interpretation on the effect estimates for the smaller scales (e.g., Janes et al., 2007). Lu and Zeger (2007) use matching estimators for each pair of observations and assess how effect estimates vary with spatial lag between the pairs to assess sensitivity, as an alternative to the Peng et al. (2006) approach. Note that estimating equation approaches don't attribute variation to a spatial term and hence are not capable of reducing bias from unmeasured spatial confounding.

Other statistical work has commented on sensitivity of effect estimates to inclusion of a spatial residual term when the covariates vary spatially, pointing out identifiability problems from the difficulty in separating spatial residual from spatial covariate effects (Breslow and Clayton, 1993; Burden et al., 2005; Lawson, 2006, p. 187; Augustin et al., 2007). A different perspective has been taken by Reich et al. (2006) and Houseman et al. (2006), who estimate the effect of the covariate of interest, $\boldsymbol{X}$, by forcing the spatial residual to be orthogonal to $\boldsymbol{X}$, attributing as much variability as possible to $\boldsymbol{X}$, with remaining spatial variability that is uncorrelated with $\boldsymbol{X}$ estimated as the spatial residual. Note that orthogonality of the residual does not occur in GLS estimation (Schabenberger and Gotway, 2005, p. 349). The approaches of Reich et al. (2006) and Houseman et al. (2006) make a very strong assumption of no confounding to avoid bias from accidentally accounting for some of the effect of the covariate in the residual. In contrast the primary question of interest here is whether use of a spatial residual term helps to decrease bias from unmeasured confounding when the confounder varies spatially. Gustafson and Greenland (2006) confront a similar problem of modeling systematic residual confounding in a context with identifiability problems, finding that imposing structure through a prior distribution in a nonidentified model can help account for a portion of the confounding, improving bias and precision of estimators.

While the results here are limited to the simple setting of linear regression/additive models with a single covariate and single unmeasured confounder, I expect that the qualitative results and principles hold in more complicated settings, with no reason to believe that the bias results would improve in more complicated models. Given the need to address the issues in even the most basic setting and provide intuition supported by simple analytic results, I have restricted attention in this way, but work in more complicated, realistic settings to extend the results here is needed.

Note that measurement error in the exposure is of critical concern, because reducing bias relies on estimating variability in exposure at scales smaller than the confounding. In many contexts, measurement error increasingly becomes a concern at small scales because of limitations in measurement resources. In contrast, large-scale exposure estimates may be well-estimated using spatial smoothing and regression mod-

els, thereby inducing Berkson-type error through what is effectively regression calibration (Gryparis et al., 2009). To the extent to which accounting for bias forces one to rely on exposure estimates more likely contaminated by classical measurement error, one may find oneself reducing bias from confounding only to increase it from measurement error. To the extent small-scale variation is affected by Berkson error, one would increase variance but not incur bias by relying on the small-scale variation.

Finally note that in many settings one has aggregated exposure and outcome data, so one has limited ability to identify effects of exposure based on fine-scale variation because the aggregation eliminates the fine-scale variation (e.g., Janes et al., 2007). This suggests that accounting for spatial confounding with areal data, for which researchers often use standard conditional auto-regressive models, is likely to be ineffective when aggregating over large areal units, which is consistent with the bias seen in Richardson (2003).

Acknowledgements

The author thanks Louise Ryan and Francesca Dominici for feedback and encouragement, Andy Houseman, Eric Tchetgen, and Brent Coull for comments, Ben Armstrong and John Rice for thought-provoking discussions, Joel Schwartz for access to the birthweight data, Alexandros Gryparis and Steve Melly for assistance with the birthweight data, and Brent Coull for funding through NIEHS R01 grant ES01244. This work was also funded by NIEHS Center grant ES000002 and NCI P01 grant CA134294-01.

References

Augustin, N., S. Lang, M. Musio, and K. von Wilpert (2007). A spatial model for the needle losses of pine-trees in the forests of Baden-Wurttemberg: an application of Bayesian structured additive regression. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 56*(1), 29–50.

Beelen, R., G. Hoek, P. Fischer, P. Brandt, and B. Brunekreef (2007). Estimated long-term outdoor air pollution concentrations in a cohort study. *Atmospheric Environment 41*(7), 1343–1358.

Biggeri, A., M. Bonannini, D. Catelan, F. Divino, E. Dreassi, and C. Lagazio (2005). Bayesian ecological regression with latent factors: Atmospheric pollutants, emissions, and mortality for lung cancer. *Environmental and Ecological Statistics 12*(4), 397–409.

Bivand, R. (1980). A Monte Carlo study of correlation coefficient estimation with spatially autocorrelated observations. *Quaestiones Geographicae 6*, 5–10.

Breslow, N. E. and D. G. Clayton (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association 88*, 9–25.

Burden, S., S. Guha, G. Morgan, L. Ryan, R. Sparks, and L. Young (2005). Spatio-temporal analysis of acute admissions for ischemic heart disease in NSW, Australia. *Environmental and Ecological Statistics 12*(4), 427–448.

Burnett, R., R. Ma, M. Jerrett, M. Goldberg, S. Cakmak, C. Pope, III, and D. Krewski (2001). The spatial association between community air pollution and mortality: A new method of analyzing correlated geographic cohort data. *Environmental Health Perspectives 109*(3), 375–380.

Cakmak, S., R. Burnett, M. Jerrett, M. Goldberg, C. Pope, III, R. Ma, T. Gultekin, M. Thun, and D. Krewski (2003). Spatial regression models for large-cohort studies linking community air pollution and health. *Journal of Toxicology and Environmental Health, Part A 66*, 1811–1823.

Cho, W. (2003). Contagion effects and ethnic contribution networks. *American Journal of Political Science 47*, 368–387.

Cressie, N. (1993). *Statistics for Spatial Data* (Rev. ed.). New York: Wiley-Interscience.

Diggle, P., P. J. Heagerty, K.-Y. Liang, and S. Zeger (2002). *Analysis of Longitudinal Data* (2 ed.). Oxford: Oxford University Press.

Dominici, F., A. McDermott, and T. Hastie (2004). Improved semiparametric time series models of air pollution and mortality. *Journal of the American Statistical Association 99*(468), 938–949.

Dominici, F., A. McDermott, S. Zeger, and J. Samet (2002). On the use of generalized additive models in time-series studies of air pollution and health. *American Journal of Epidemiology 156*(3), 193–203.

Dow, M., M. Burton, and D. White (1982). Network autocorrelation: A simulation study of a foundational problem in regression and survey research. *Social Networks 4*, 169–200.

Gryparis, A., B. Coull, J. Schwartz, and H. Suh (2007). Latent variable semiparametric regression models for spatio-temporal modeling of mobile source pollution in the greater Boston area. *Journal of the Royal Statistical Society, Series C 56*, 183–209.

Gryparis, A., C. Paciorek, A. Zeka, J. Schwartz, and B. Coull (2009). Measurement error caused by spatial misalignment in environmental epidemiology. *Biostatistics 10*, 258–274.

Gustafson, P. and S. Greenland (2006). The performance of random coefficient regression in accounting for residual confounding. *Biometrics 62*(3), 760–768.

He, S., S. Mazumdar, and V. Arena (2006). A comparative study of the use of GAM and GLM in air pollution research. *Environmetrics 17*(1), 81–93.

Houseman, E., B. Coull, and J. Shine (2006). A nonstationary negative binomial time series with time-dependent covariates: enterococcus counts in Boston harbor. *Journal of the American Statistical Association 101*, 1365–1376.

Janes, H., F. Dominici, and S. Zeger (2007). Trends in air pollution and mortality: An approach to the assessment of unmeasured confounding. *Epidemiology 18*(4), 416–423.

Johnston, J. and J. DiNardo (1997). *Econometric Methods* (4th ed.). New York: McGraw-Hill.

Lawson, A. (2006). *Statistical Methods in Spatial Epidemiology* (2nd ed.). New York: John Wiley & Sons.

Legendre, P. (1993). Spatial autocorrelation: Trouble or new paradigm? *Ecology 74*(6), 1659–1673.

Liang, K.-Y. and S. Zeger (1986). Longitudinal data analysis using generalized linear models. *Biometrika 73*, 13–22.

Lu, Y. and S. Zeger (2007). Decomposition of regression estimators to explore the influence of 'unmeasured' time-varying confounders. Technical Report 159, Johns Hopkins University.

Ma, B., A. Lawson, and Y. Liu (2007). Evaluation of Bayesian models for focused clustering in health data. *Environmetrics 18*, 871–887.

Molitor, J., M. Jerrett, C. Chang, et al. (2007). Assessing uncertainty in spatial exposure models for air pollution health effects assessment. *Environmental Health Perspectives 115*(8), 1147–1153.

Peng, R., F. Dominici, and T. Louis (2006). Model choice in time series studies of air pollution and mortality. *Journal of the Royal Statistical Society Series A 169*(2), 179–203.

Pope, III, C., R. Burnett, M. Thun, E. Calle, D. Krewski, K. Ito, and G. Thurston (2002). Lung cancer, cardiopulmonary mortality and long-term exposure to fine particulate air pollution. *Journal of the American Medical Association 287*, 1132–1141.

Ramsay, T., R. Burnett, and D. Krewski (2003). Exploring bias in a generalized additive model for spatial air pollution data. *Environmental Health Perspectives 111*(10), 1283–1288.

Reich, B., J. Hodges, and V. Zadnik (2006). Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models. *Biometrics 62*(4), 1197–1206.

Rice, J. (1986). Convergence rate for partially linear splined models. *Statistics and Probability Letters 4*, 203–208.

Richardson, S. (2003). Spatial models in epidemiological applications. In P. Green, N. Hjort, and S. Richardson (Eds.), *Highly Structured Stochastic Systems*, pp. 237–259. Oxford University Press.

Schabenberger, O. and C. Gotway (2005). *Statistical Methods for Spatial Data Analysis*. Boca Raton: Chapman & Hall.

Speckman, P. (1988). Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society Series B 50*(3), 413–436.

Waller, L. and C. Gotway (2004). *Applied Spatial Statistics for Public Health Data*. Hoboken, New Jersey: Wiley.

Wood, S. (2006). *Generalized Additive Models: An Introduction with R*. Boca Raton: Chapman & Hall.

Zeger, S., F. Dominici, A. McDermott, and J. Samet (2007). Mortality in the Medicare population and chronic exposure to fine particulate air pollution. Technical Report 133, Johns Hopkins University.

Zeka, A., S. Melly, and J. Schwartz (2008). The effects of socioeconomic status and indices of physical environment on reduced birth weight and preterm births in eastern Massachusetts. *Environmental Health 7*, 60.

Appendix: Proof

Lemma 1: Consider the model (3) and assume all parameters are known except β_0 and β_x. The expectation of the precision of $\hat{\beta}_{\text{GLS}}$ is

$$\text{E}_X\left(\text{Var}(\hat{\beta}_{\text{GLS}})^{-1}\right) = \frac{\sigma_x^2}{\tau^2+\sigma_g^2}\left(\text{tr}\{\tilde{\Sigma}^{-1}R_x\} - \frac{\mathbf{1}^T\tilde{\Sigma}^{-1}R_x\tilde{\Sigma}^{-1}\mathbf{1}}{\mathbf{1}^T\tilde{\Sigma}^{-1}\mathbf{1}}\right),$$

where $\tilde{\Sigma} \equiv (1-p_g)I + p_g R_g$ and $p_g \equiv \sigma_g^2/(\sigma_g^2+\tau^2)$.

Proof: From the definition of the GLS estimator, we have

$$\text{Var}(\hat{\beta}_{\text{GLS}}) = [\mathcal{X}^T\Sigma^{-1}\mathcal{X}]^{-1}_{2,2} = \frac{\mathbf{1}^T\Sigma^{-1}\mathbf{1}}{\mathbf{1}^T\Sigma^{-1}\mathbf{1}\boldsymbol{X}^T\Sigma^{-1}\boldsymbol{X} - \boldsymbol{X}^T\Sigma^{-1}\mathbf{1}\mathbf{1}^T\Sigma^{-1}\boldsymbol{X}}.$$

Using the definitions of $\tilde{\Sigma}$ and p_g, and taking the reciprocal, we have

$$\text{Prec}(\hat{\beta}_{\text{GLS}}) = \frac{1}{\sigma_g^2+\tau^2}\left(\boldsymbol{X}^T\tilde{\Sigma}^{-1}\boldsymbol{X} - \frac{\boldsymbol{X}^T\tilde{\Sigma}^{-1}\mathbf{1}\mathbf{1}^T\tilde{\Sigma}^{-1}\boldsymbol{X}}{\mathbf{1}^T\tilde{\Sigma}^{-1}\mathbf{1}}\right).$$

We then take the expectation with respect to the sampling distribution of $\boldsymbol{X}$, using the expectation of a quadratic form. Rearranging the matrices inside the second trace gives a scalar, so dropping the second trace, this gives

$$\begin{aligned}\text{E}_X\left(\text{Prec}(\hat{\beta}_{\text{GLS}})\right) &= \frac{\sigma_x^2}{\sigma_g^2+\tau^2}\left(\text{tr}(\tilde{\Sigma}^{-1}R_x) + \mu_x^2\mathbf{1}^T\tilde{\Sigma}^{-1}\mathbf{1} - \frac{\text{tr}(\tilde{\Sigma}^{-1}\mathbf{1}\mathbf{1}^T\tilde{\Sigma}^{-1}R_x)}{\mathbf{1}^T\tilde{\Sigma}^{-1}\mathbf{1}} - \frac{\mu_x^2\mathbf{1}^T\tilde{\Sigma}^{-1}\mathbf{1}\mathbf{1}^T\tilde{\Sigma}^{-1}\mathbf{1}}{\mathbf{1}^T\tilde{\Sigma}^{-1}\mathbf{1}}\right) \\ &= \frac{\sigma_x^2}{\sigma_g^2+\tau^2}\left(\text{tr}(\tilde{\Sigma}^{-1}R_x) - \frac{\mathbf{1}^T\tilde{\Sigma}^{-1}R_x\tilde{\Sigma}^{-1}\mathbf{1}}{\mathbf{1}^T\tilde{\Sigma}^{-1}\mathbf{1}}\right).\end{aligned}$$