The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: https://amstat.tandfonline.com/loi/utas20

Adding Spatially-Correlated Errors Can Mess Up the Fixed Effect You Love

James S. Hodges & Brian J. Reich

To cite this article: James S. Hodges & Brian J. Reich (2010) Adding Spatially-Correlated Errors Can Mess Up the Fixed Effect You Love, The American Statistician, 64:4, 325-334, DOI: 10.1198/tast.2010.10052

To link to this article: https://doi.org/10.1198/tast.2010.10052

Supplementary materials for this article are available online. Please click the TAS link at http://pubs.amstat.org.

General

Adding Spatially-Correlated Errors Can Mess Up the Fixed Effect You Love

James S. HODGES and Brian J. REICH

Many statisticians have had the experience of fitting a linear model with uncorrelated errors, then adding a spatially-correlated error term (random effect) and finding that the estimates of the fixed-effect coefficients have changed substantially. We show that adding a spatially-correlated error term to a linear model is equivalent to adding a saturated collection of canonical regressors, the coefficients of which are shrunk toward zero, where the spatial map determines both the canonical regressors and the relative extent of the coefficients' shrinkage. Adding a spatially-correlated error term can also be seen as inflating the error variances associated with specific contrasts of the data, where the spatial map determines the contrasts and the extent of error-variance inflation. We show how to avoid this spatial confounding by restricting the spatial random effect to the orthogonal complement (residual space) of the fixed effects, which we call restricted spatial regression. We consider five proposed interpretations of spatial confounding and draw implications about what, if anything, one should do about it. In doing so, we debunk the common belief that adding a spatially-correlated random effect adjusts fixed-effect estimates for spatially-structured missing covariates. This article has supplementary material online.

KEY WORDS: Confounding; Missing covariate; Random effect; Spatial correlation; Spatial regression.

James S. Hodges is Associate Professor, Division of Biostatistics, University of Minnesota, Minneapolis, MN 55414 (E-mail: *hodges@ccbr.umn.edu*). Brian J. Reich is Assistant Professor, Department of Statistics, North Carolina State University, Raleigh, NC 27695. The authors thank Vesna Zadnik for allowing us to use the Slovenian stomach-cancer data. We benefited from the comments of Sudipto Banerjee, Brad Carlin, Wei Pan, and Melanie Wall of the University of Minnesota Division of Biostatistics, Dave Nelson of the Minneapolis Veterans Affairs Medical Center's Center for Chronic Disease Outcomes Research, Chris Paciorek of the Harvard University Department of Biostatistics, Babette Brumback of the University of Florida Division of Biostatistics, and Montserrat Fuentes of North Carolina State University's Department of Statistics. These helpful people cannot be blamed for the results.

1. STOMACH CANCER IN SLOVENIA: WHERE DOES THE FIXED EFFECT GO?

Dr. Vesna Zadnik, a Slovenian epidemiologist, collected data describing counts of stomach cancers in the 194 municipalities that partition Slovenia, for the years 1995 to 2001 inclusive. She was studying the possible association of stomach cancer with socioeconomic status, as measured by a composite score calculated from 1999 data by Slovenia's Institute of Macroeconomic Analysis and Development. (Her findings were published in Zadnik and Reich 2006.) Figure 1(a) shows the standardized incidence ratio (SIR) of stomach cancer for the 194 municipalities; for municipality $i = 1, \ldots, 194$, $\text{SIR}_i = O_i/E_i$, where O_i is the observed count of stomach cancer cases and E_i is the expected count using indirect standardization, that is, $E_i = P_i \sum_j O_j / \sum_j P_j$, P_i being municipality i's population. Figure 1(b) shows the socioeconomic scores for the municipalities, $\text{SE}c_i$, after centering and scaling, so the $\text{SE}c_i$ have average 0 and finite-sample variance 1. In both panels of Figure 1, dark colors indicate larger values. SIR and SEc have a negative association: western municipalities generally have low SIR and high SEc while eastern municipalities generally have high SIR and low SEc.

Following advice received in a spatial-statistics short course, Dr. Zadnik first did a nonspatial analysis assuming the O_i were independent Poisson observations with $\log\{E(O_i)\} = \log(E_i) + \alpha + \beta \text{SE}c$, with flat priors on α and β. This analysis gave the obvious result: β had posterior median -0.14 and 95% posterior interval $(-0.17, -0.10)$, capturing Figure 1's negative association.

Dr. Zadnik continued following the short course's guidance by doing a spatial analysis using the improper (or implicit) conditionally autoregressive (ICAR) model of Besag, York, and Mollié (1991). Dr. Zadnik's understanding was that ignoring spatial correlation would make β's posterior standard deviation (standard error) too small, while spatial analysis in effect discounts the sample size with little effect on the estimate of β, just as generalized estimating equations (GEE) adjusts standard errors for clustering but (in the authors' experience) has little effect on point estimates unless the working correlations are very large. As we will see, other people have different reasons for introducing spatial correlation.

In the model of Besag, York, and Mollié (1991), the O_i are conditionally independent Poisson random variables with mean

$$\log\{E(O_i)\} = \log(E_i) + \beta \text{SE}c_i + S_i + H_i. \qquad (1)$$

 DOI: 10.1198/tast.2010.10052

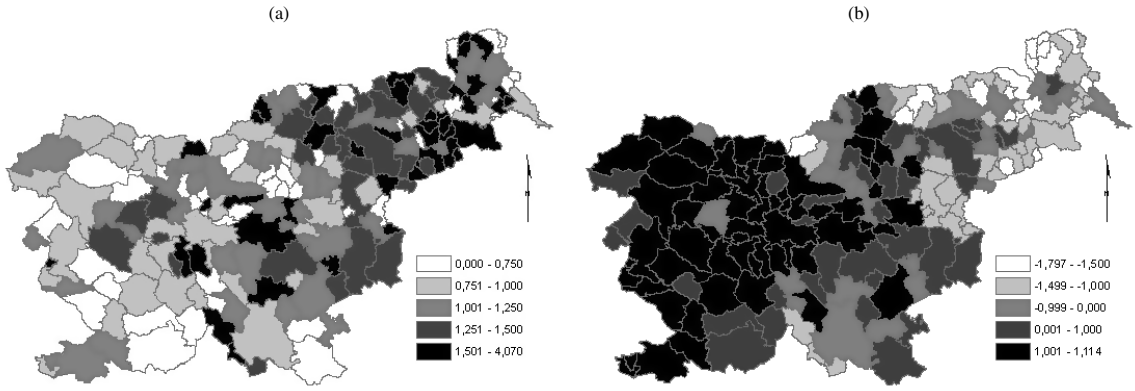

Figure 1. For the Slovenian municipalities, panel (a): observed standardized incidence ratio SIR $= O_i/E_i$; panel (b): centered and scaled socioeconomic status SEc.

The intercept is now the sum of two random effects, $\mathbf{S} = (S_1, \ldots, S_{194})'$ capturing spatial clustering and $\mathbf{H} = (H_1, \ldots, H_{194})'$ capturing heterogeneity. The H_i are modeled as independent draws from a normal distribution with mean zero and precision (reciprocal of variance) τ_h. The S_i are modeled using an L_2-norm ICAR, also called a Gaussian Markov random field, which is discussed in detail below. The ICAR represents the intuition that neighboring municipalities tend to be more similar to each other than municipalities that are far apart, where similarity of neighbors is controlled by an unknown parameter τ_s that is like a precision. This Bayesian analysis used independent gamma priors for τ_h and τ_s with mean 1 and variance 100, and a flat prior for β.

In the spatial analysis, β had posterior mean -0.02 and 95% posterior interval $(-0.10, 0.06)$. Compared to the nonspatial analysis, the 95% interval was wider and the spatial model fit better, with deviance information criterion (DIC; Spiegelhalter et al. 2002) decreasing from 1153 to 1082 even though the effective number of parameters (p_D) increased sharply, from 2.0 to 62.3. These changes were expected. The surprise was that the negative association, which is quite plain in the maps and the nonspatial analysis, had disappeared. What happened?

This spatial confounding effect, which we reported in a previous article (Reich, Hodges, and Zadnik 2006), has been reported elsewhere but is not widely known. The earliest report we have found is the article by Clayton, Bernardinelli, and Montomoli (1993), who used the term "confounding due to location" for a less dramatic but still striking effect in analyses of lung-cancer incidence in Sardinia; they and Wakefield (2007) reported a similar-sized effect in the long-suffering Scottish lip-cancer data. This effect is not yet understood; we have seen five proposed interpretations. The interpretations depend on whether the random effect **S** meets the traditional definition of "random effect" given by, for example, Scheffé (1959, p. 238): the levels of a random effect are draws from a population, and the draws are not of interest in themselves but only as samples from the larger population, which *is* of interest. For reasons to be discussed later, we describe random effects *not* meeting this definition as formal devices to implement a smoother. With this distinction, which Section 3 discusses in more detail, the five interpretations are:

- The random effect **S** is a formal device to implement a smoother.
  - (i) Spatially-correlated errors remove bias in estimating β and are generally conservative (Clayton et al. 1993).
  - (ii) Spatially-correlated errors can introduce or remove bias in estimating β and are not necessarily conservative (Wakefield 2007; implicit in Reich, Hodges, and Zadnik 2006).
- **S** is a Scheffé-style random effect.
  - (iii) The spatial effect **S** is collinear with the fixed effect, but neither estimate of β is biased (David B. Nelson, personal communication).
  - (iv) Adding the spatial effect **S** creates information loss, but neither estimate of β is biased (David B. Nelson, personal communication).
  - (v) Because error is correlated with the regressor SEc in the sense commonly used in econometrics, *both* estimates of β are biased (Paciorek 2011).

Except for (v), these interpretations treat SEc as measured without error and not drawn from a probability distribution.

Our purpose is to examine these interpretations and determine which is appropriate under what circumstances. Section 2 describes the mechanics of spatial confounding. We give derivations for the normal-errors analog to (1); Reich, Hodges, and Zadnik (2006) gave some of this material and an extension to generalized linear mixed models. These derivations suggest a method to restrict spatial smoothing to the orthogonal complement of the fixed effects (i.e., the residual space), which we call restricted spatial regression. We argue briefly that spatial confounding is also present, and for the same reasons, for other models which, like ICAR, represent the intuition that observations taken near each other in space tend to be more similar than observations taken far apart in space. (This argument refers to the online supplementary materials.) Section 3 considers the five interpretations in light of Section 2's theory and draws implications about what, if anything, one should do about spatial confounding. In making this argument, we debunk the common belief that introducing a spatially-correlated random effect adjusts fixed-effect estimates for spatially-structured missing covariates (e.g., Clayton, Bernardinelli, and Montomoli 1993).

2. THE MECHANICS OF SPATIAL CONFOUNDING

This section relies heavily on the theory of linear models. Although our approach is Bayesian, most of our results apply immediately to analyses based on maximizing the restricted likelihood.

2.1 The Model With Spatial Correlation Written as a Linear Model

For an n-dimensional observation vector $\mathbf{y}$, write the normal-errors analog to model (1) as

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{I}_n\mathbf{S} + \boldsymbol{\epsilon}, \tag{2}$$

where $\mathbf{y}$, $\mathbf{X}$, $\mathbf{S}$, and $\boldsymbol{\epsilon}$ are $n \times 1$, β is scalar, $\mathbf{I}_n$ is the n-dimensional identity matrix, $\mathbf{y}$ and $\mathbf{X}$ are known, and β, $\mathbf{S}$, and $\boldsymbol{\epsilon}$ are unknown. In the Slovenian data, $n = 194$ and $\mathbf{X} = \mathrm{SE}c$. $\mathbf{X}$ is centered and scaled to have average 0 and finite-sample variance 1. The derivation below generalizes easily to any full-rank $\mathbf{X}$ without an intercept column (Reich, Hodges, and Zadnik 2006); the intercept is implicit in $\mathbf{S}$, as shown below. The error term $\boldsymbol{\epsilon}$ is n-dimensional normal with mean zero and precision matrix $\tau_e\mathbf{I}$, τ_e being the reciprocal of the error variance. Our Bayesian analysis puts a flat prior on β, but this is not necessary.

The L_2-norm improper CAR model (or prior; ICAR) on $\mathbf{S}$ can be represented as an improper n-variate normal distribution specified by its precision matrix:

$$p(\mathbf{S}|\tau_s) \propto \tau_s^{(n-G)/2} \exp(-0.5\tau_s\mathbf{S}'\mathbf{Q}\mathbf{S}), \tag{3}$$

where G is the number of islands (disconnected groups of municipalities) in the spatial map (Hodges, Carlin, and Fan 2003). The unknown τ_s controls the smoothness of $\mathbf{S}$; larger τ_s force neighboring S_i to be more similar to each other. $\mathbf{Q}$ encodes the neighbor pairs, with diagonal elements q_{ii} = number of municipality i's neighbors, and $q_{ij} = -1$ if municipalities i and j are neighbors and 0 otherwise. For the Slovenian data, $G = 1$ and we specify that municipalities i and j are neighbors if they share a boundary. Other specifications of neighbor pairs are possible but this one is common. Model (3) can be re-expressed in the pairwise-difference form

$$p(\mathbf{S}|\tau_s) \propto \tau_s^{(n-G)/2} \exp\Big(-0.5\tau_s \sum (S_i - S_j)^2\Big), \tag{4}$$

where the sum is over unique neighbor pairs (i, j).

We have written model (2) with the random effect $\mathbf{S}$ having design matrix $\mathbf{I}_n$ to emphasize that model (2) is overparameterized and is identified only because $\mathbf{S}$ is smoothed or, alternatively, constrained by the ICAR model (3). To clarify the identification issues, reparameterize (2) as follows. The neighbor-pair matrix $\mathbf{Q}$ has spectral decomposition $\mathbf{Q} = \mathbf{Z}\mathbf{D}\mathbf{Z}'$, where $\mathbf{Z}$ is $n \times n$ orthogonal and $\mathbf{D}$ is diagonal with diagonal elements $d_1 \geq \cdots \geq d_{n-G} > 0$ and $d_{n-G+1} = \cdots = d_n = 0$ (Hodges, Carlin, and Fan 2003). $\mathbf{Z}$'s columns $Z_1, \ldots, Z_n$ are $\mathbf{Q}$'s eigenvectors; $\mathbf{D}$'s diagonal elements are the corresponding eigenvalues. Reparameterize (2) as

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon}, \tag{5}$$

where $\mathbf{b} = \mathbf{Z}'\mathbf{S}$ is an n-vector with a normal distribution having mean zero and diagonal precision matrix $\tau_s\mathbf{D} = \tau_s \operatorname{diag}(d_1, \ldots, d_{n-G}, 0, \ldots, 0)$.

The spatial random effect $\mathbf{S}$ thus corresponds to a saturated collection of canonical regressors, the n columns of $\mathbf{Z}$, whose coefficients $\mathbf{b}$ are shrunk toward zero to an extent determined by $\tau_s\mathbf{D}$. The smoothing parameter τ_s controls shrinkage of all n components of $\mathbf{b}$ and the d_i control the relative degrees of shrinkage of the b_i for a given τ_s. Both the canonical regressors $\mathbf{Z}$ and the d_i are determined solely by the spatial map through $\mathbf{Q}$. The first canonical coefficient, b_1, has the largest d_i and is thus shrunk the most for any given τ_s; for the Slovenian map, $d_1 = 14.46$. The last G b_i, $b_{n-G+1}, \ldots, b_n$ are not shrunk at all because their prior precisions $\tau_s d_i$ are zero, so they are fixed effects implicit in the spatial random effect. The b_i with the smallest positive d_i, d_{n-G}, is shrunk least of all the shrunken coefficients. For the Slovenian map with $G = 1$, this is b_{193}, with $d_{193} = 0.03$, so its prior precision is smaller than b_1's by a factor of about 500 for any τ_s.

To understand the differential shrinkage of the b_i, we need to understand the columns of $\mathbf{Z}$, which the spatial map determines. $Z_{n-G+1}, \ldots, Z_n$, whose coefficients $b_{n-G+1}, \ldots, b_n$ are not shrunk at all, span the space of the means of (or intercepts for) the G islands in the spatial map. This is easily seen from (4): this distribution is flat (puts no constraint) on the G island means. Also, $\mathbf{Q}\mathbf{1}_n = 0$ for any map, so the overall intercept $\mathbf{1}_n$ always lies in the span of $Z_{n-G+1}, \ldots, Z_n$ and is thus implicit in the ICAR specification. Thus without loss of generality, we set $Z_n = \frac{1}{\sqrt{n}}\mathbf{1}_n$ so all other Z_i are contrasts, that is, $\mathbf{1}_n'Z_i = 0$.

Based on examples, Z_{n-G}, whose coefficient b_{n-G} has the smallest positive prior precision $\tau_s d_{n-G}$, can be interpreted loosely as the lowest *frequency* contrast in $\mathbf{S}$ among the shrunken contrasts. Figure 2(a) is a plot of $Z_{n-1} = Z_{193}$ for the Slovenian data, where darker and lighter colors indicate higher and lower values, respectively. Z_{193} is "low-frequency" in the sense that it is a roughly linear trend along the long axis of Slovenia's map. Again based on examples, as the value of d_i increases, the frequency of the corresponding Z_i increases as well. Figure 2(b) shows Z_1 for the Slovenian data; it is roughly the difference between the two municipalities with the most neighbors (the dark municipalities) and the average of their neighbors. For other spatial maps, the interpretations are similar. For the counties of Minnesota (Reich and Hodges 2008, figure 6), the contrast with the least-smoothed coefficient is the north-south gradient, while the contrast with the most-smoothed coefficient is roughly the difference between the county having the most neighbors and the average of those neighbors. For the spatial map representing periodontal measurement sites, the contrast with the least-smoothed coefficient is nearly linear along a dental arch, while the contrasts with the most-smoothed coefficients are the difference, on each tooth, between the average of the interproximal sites and the average of the direct sites (Reich and Hodges 2008, figure 3, which also shows other Z_i). For the Scottish lip-cancer data (Wakefield 2007 and many others), the contrast with the least-smoothed coefficient is the north-south gradient, while the contrast with the second least-smoothed coefficient is roughly quadratic along

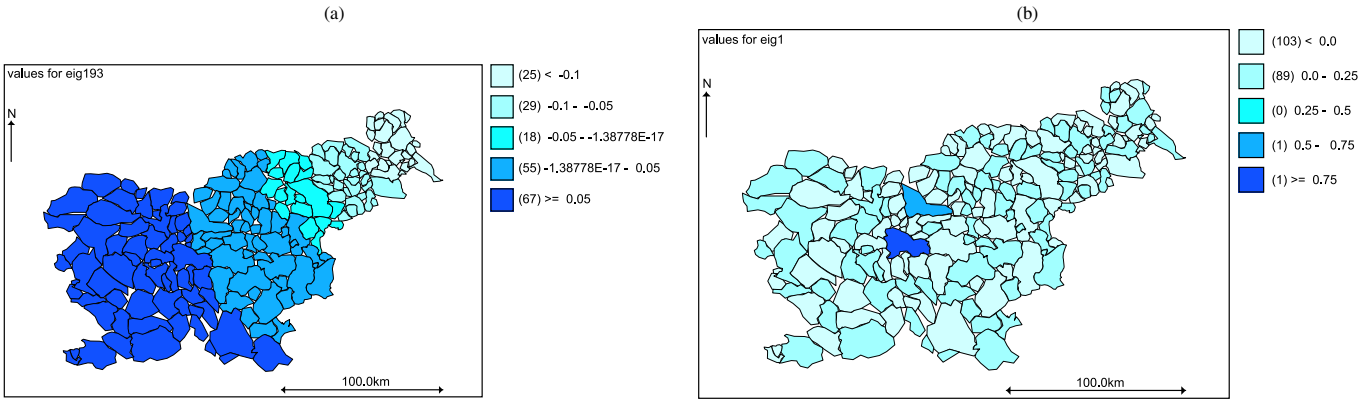

Figure 2. Two canonical regressors, columns of the matrix $\mathbf{Z}$; panel (a): Z_{193}, with the smallest positive eigenvalue d_{193} and thus the least-shrunk coefficient b_{193} among the shrunk coefficients; panel (b): Z_1, with the largest eigenvalue d_{193} and thus the most-shrunk coefficient b_1. The online version of this figure is in color.

the north-south gradient, high at the northern and southern extremes and low in the middle—just like the much-studied predictor in the Scottish lip-cancer data, AFF, which measures employment in agriculture, fisheries, and forestry.

2.2 Spatial Confounding Explained in Linear-Model Terms

For the normal-errors model (5), it is straightforward to show (Reich, Hodges, and Zadnik 2006) that the posterior mean of β conditional on the precisions τ_e and τ_s is

$$E(\beta|\tau_e,\tau_s,\mathbf{y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}E(\mathbf{b}|\tau_e,\tau_s,\mathbf{y}), \quad (6)$$

where $E(\mathbf{b}|\tau_e,\tau_s,\mathbf{y})$ is *not* conditional on β, taking the value

$$E(\mathbf{b}|\tau_e,\tau_s,\mathbf{y}) = (\mathbf{Z}'\mathbf{P}^c\mathbf{Z} + r\mathbf{D})^{-1}\mathbf{Z}'\mathbf{P}^c\mathbf{y} \quad (7)$$

for $r = \tau_s/\tau_e$ and $\mathbf{P}^c = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, the familiar residual projection matrix from linear models. These expressions are correct for full-rank $\mathbf{X}$ of any dimension.

In (6), the term $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ is the ordinary least squares estimate of β and also β's posterior mean in a fit without $\mathbf{S}$, using a flat prior on β. Thus the second term $-(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}E(\mathbf{b}|\tau_e,\tau_s,\mathbf{y})$ is the change in β's posterior mean, conditional on (τ_e,τ_s), from adding $\mathbf{S}$. Note that $\mathbf{Z}E(\mathbf{b}|\tau_e,\tau_s,\mathbf{y})$ is the fitted value of $\mathbf{S}$ given (τ_e,τ_s). Thus when $\mathbf{S}$ is added to the model, the change in β's posterior mean given (τ_e,τ_s) is -1 times the regression on $\mathbf{X}$ of the fitted values of $\mathbf{S}$.

Because $\mathbf{X}$ is centered and scaled (so $\mathbf{X}'\mathbf{X} = n-1$) and $\mathbf{Z}$ is orthogonal, the correlations ρ_i of $\mathbf{X}$ and Z_i, the ith column of $\mathbf{Z}$, can be written as $\mathbf{R} = (\rho_1,\ldots,\rho_{n-G},0,\ldots,0)' = (n-1)^{-0.5}\mathbf{Z}'\mathbf{X}$. Equation (6) can then be written as

$$E(\beta|\tau_e,\tau_s,\mathbf{y}) = \hat{\beta}_{\mathrm{OLS}} - (n-1)^{1/2}\mathbf{R}'E(\mathbf{b}|\tau_e,\tau_s,\mathbf{y}). \quad (8)$$

(The Appendix gives a more explicit expression.) From (8), if ρ_i and $E(b_i|\tau_e,\tau_s,\mathbf{y})$ are large for the same i, then adding $\mathbf{S}$ to the model can induce a large change in β's posterior mean. This happens if four conditions hold: $\mathbf{X}$ is highly correlated with Z_i; $\mathbf{y}$ has a substantial correlation with both $\mathbf{X}$ and Z_i; $r = \tau_s/\tau_e$ is small; and d_i is small. The first two conditions define confounding in ordinary linear models; the last two conditions ensure that b_i is not shrunk much toward zero. (If more than one i meets these conditions, their effects on $E(\beta|\tau_e,\tau_s,\mathbf{y})$ can cancel each other.) These necessary conditions are all present in the Slovenian data: Figure 1 shows the strong association of $\mathbf{y}$ and $\mathbf{X} = \mathrm{SE}c$, $\rho_{193} = \mathrm{correlation}(\mathrm{SE}c, Z_{193}) = 0.72$, $d_{193} = 0.03$, and $\mathbf{S}$ is not smoothed much (the effective number of parameters in the fit is $p_D = 62.3$).

This effect on β's estimate is easily understood in linear-model terms as the effect of adding a collinear regressor to a linear model. If $\mathbf{S}$ were not smoothed at all—if the coefficients $\mathbf{b}$ of the saturated design matrix $\mathbf{Z}$ were not shrunk toward zero—then β would not be identified. This corresponds to setting the smoothing precision τ_s to 0, so the smoothing ratio $r = \tau_s/\tau_e = 0$. If the smoothing ratio r is small, β is identified but the coefficients of Z_i with small d_i are shrunk very little, so if these Z_i are collinear with $\mathbf{X}$, the estimate of β is subject to the same collinearity effects as in linear models.

Consider also how β's posterior variance changes when the ICAR-distributed $\mathbf{S}$ is added to the model. Reich, Hodges, and Zadnik (2006) showed that conditional on τ_e and τ_s, adding $\mathbf{S}$ to the model multiplies the conditional posterior variance of β by

$$\left[1 - \sum \frac{\rho_i^2}{1+rd_i}\right]^{-1}, \quad (9)$$

where the sum is over i with $d_i > 0$ and ρ_i is as above. Expression (9) holds only for single-column $\mathbf{X}$; Reich, Hodges, and Zadnik (2006) gave an expression for general full-rank $\mathbf{X}$. In general, $\sum_{i=1}^{n}\rho_i^2 = 1$. Because $Z_n \propto \mathbf{1}_n$ and $\mathbf{X}$ is centered, $\rho_n = (n-1)^{-0.5}Z_n'\mathbf{X} = 0$, so if $r = 0$, that is, $\mathbf{S}$ is not smoothed, $\mathrm{var}(\beta|\tau_e,\tau_s,\mathbf{y})$ is infinite because $\sum\rho_i^2/(1+rd_i) = \sum_{i=1}^{n-1}\rho_i^2 = 1$. As r grows from zero, $\mathbf{S}$ is smoothed more and $\mathrm{var}(\beta|\tau_e,\tau_s,\mathbf{y})$ decreases. For given r, the variance inflation factor (9) is large if ρ_i is large for the smallest d_i, as in the Slovenian data. Again, this differs from the analogous result in linear-model theory only because the b_i are shrunk.

2.3 Spatial Confounding Explained in a More Spatial-Statistics Style

The linear-models explanation above seems odd to many spatial-statistics mavens, who usually think of the ICAR as a model for the error covariance. We now give a derivation more in line with this viewpoint.

Begin with the reparameterized normal-errors model (5) but rewrite it in a more spatial-statistics style as $\mathbf{y} = \mathbf{X}\beta + \psi$, where the elements of $\psi = \mathbf{Zb} + \boldsymbol{\epsilon}$ are not independent. $\mathbf{S} = \mathbf{Zb}$ does not have a proper covariance matrix under an ICAR model, so we must proceed indirectly. Partition $\mathbf{Z} = (\mathbf{Z}^{(1)}|\mathbf{Z}^{(2)})$, where $\mathbf{Z}^{(1)}$ has $n - G$ columns and $\mathbf{Z}^{(2)}$ has G columns, and partition $\mathbf{b}$ conformably as $\mathbf{b} = (\mathbf{b}^{(1)\prime}|\mathbf{b}^{(2)\prime})'$, so $\mathbf{b}^{(1)}$ is $(n - G) \times 1$ and $\mathbf{b}^{(2)}$ is $G \times 1$. Pre-multiply (5) by $\mathbf{Z}'$, so (5) becomes

$$\begin{aligned} \mathbf{Z}^{(1)\prime}\mathbf{y} &= \mathbf{Z}^{(1)\prime}\mathbf{X}\beta + \mathbf{e}_1, \\ &\quad \text{precision}(\mathbf{e}_{1i}) = \tau_e(rd_i)/(1 + rd_i) < \tau_e, \\ \mathbf{Z}^{(2)\prime}\mathbf{y} &= \mathbf{Z}^{(2)\prime}\mathbf{X}\beta + \mathbf{b}^{(2)} + \mathbf{e}_2, \\ &\quad \text{precision}(\mathbf{e}_{2i}) = \tau_e, \end{aligned} \tag{10}$$

where $\mathbf{e}_1 = \mathbf{b}^{(1)} + \mathbf{Z}^{(1)\prime}\boldsymbol{\epsilon}$ and $\mathbf{e}_2 = \mathbf{Z}^{(2)\prime}\boldsymbol{\epsilon}$, so the two rows of (10) are independent, and recall τ_e is $\boldsymbol{\epsilon}$'s error precision in (5).

Suppose $G = 1$ as in the Slovenian data; $G > 1$ is discussed below. Then $\mathbf{Z}^{(2)} \propto \mathbf{1}_n$ and $\mathbf{Z}^{(2)\prime}\mathbf{X} = 0$ because $\mathbf{X}$ is centered, so all the information about β comes from the first row of (10). Fix τ_s and τ_e. Without $\mathbf{S}$ in the model, β's posterior precision (the reciprocal of β's posterior variance) is $\mathbf{X}'\mathbf{X}\tau_e = (n - 1)\tau_e$, which is also β's information matrix (a scalar, in this case). By adding $\mathbf{S}$ to the model, the posterior precision of β decreases by $(n - 1)\tau_e \sum_{i=1}^{n-1} \rho_i^2/(1 + rd_i)$, where as before $\rho_i = \text{correlation}(\mathbf{X}, Z_i) = (n - 1)^{-1/2} Z_i'\mathbf{X}$ and $r = \tau_s/\tau_e$. The information loss is large if r is small (relatively little spatial smoothing) and ρ_i is large for i with small d_i, as in the Slovenian data. Because the information about β in the different rows of (10) may not be entirely consistent, if row i of (10) differs from the other rows and row i is effectively deleted by the combination of large ρ_i and small d_i, β's estimate can change markedly when $\mathbf{S}$ is added to the model.

If the spatial map has $G > 1$ islands, the ICAR model includes an implicit $G - 1$ degree of freedom unsmoothed fixed effect for the island means, $\mathbf{b}^{(2)}$, in addition to the overall intercept. This island-mean fixed effect may be collinear with $\mathbf{X}$ in the usual manner of linear models even if $\mathbf{S}$ is smoothed maximally within islands (i.e., τ_s is very large).

2.4 Avoiding Spatial Confounding: Restricted Spatial Regression

Spatial confounding can be interpreted in linear-model terms as a collinearity problem. This suggests adapting a simple trick sometimes used in linear models with collinearity problems.

Suppose we are fitting a linear model with $\mathbf{y}$ as the dependent variable and $\mathbf{X}_1$ and $\mathbf{X}_2$ as independent variables. If $\mathbf{X}_1$ and $\mathbf{X}_2$ are highly correlated with each other, that collinearity can produce a variety of inconvenient side effects. Many regression classes teach the following trick to get rid of this collinearity: regress $\mathbf{y}$ on $\mathbf{X}_2$ and $\mathbf{X}_1^* = (\mathbf{I} - \mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2')\mathbf{X}_1$ instead of on $\mathbf{X}_2$ and $\mathbf{X}_1$. $\mathbf{X}_1^*$ is the residuals of $\mathbf{X}_1$ regressed on $\mathbf{X}_2$, so $\mathbf{X}_1^*$ and $\mathbf{X}_2$ are orthogonal by construction and their estimated coefficients are uncorrelated. This trick does require attributing to $\mathbf{X}_2$ all variation in $\mathbf{y}$ over which $\mathbf{X}_2$ and $\mathbf{X}_1$ are competing, which can be hard to justify in particular cases.

The analog in our spatial problem is to restrict the spatial random effect $\mathbf{S}$ to the subspace of n-dimensional space orthogonal to the fixed effect $\mathbf{X}$, which we call "restricted spatial regression." We show how to do this for a one-dimensional $\mathbf{X}$, which is easily generalized to higher dimensions (Reich, Hodges, and Zadnik 2006, sec. 3; sec. 4 extends the method to nonnormal observables). This attributes to the fixed effect $\mathbf{X}$ all variation in $\mathbf{y}$ over which $\mathbf{X}$ and $\mathbf{S}$ are competing, but we argue in Section 3 that this is appropriate, indeed necessary, in a large class of situations.

The simplest way to specify a restricted spatial regression is to replace model (2) with $\mathbf{y} = \mathbf{X}\beta + \mathbf{P}^c\mathbf{S} + \boldsymbol{\epsilon}$. The design matrix in front of $\mathbf{S}$ has changed from $\mathbf{I}_n$ to $\mathbf{P}^c = \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, the residual projection matrix for a regression on $\mathbf{X}$, but otherwise the model is unchanged. Written this way, $\mathbf{S}$ has a superfluous dimension: one linear combination of $\mathbf{S}$, $(\mathbf{I}_n - \mathbf{P}^c)\mathbf{S}$, necessarily contributes nothing to the fitted values of $\mathbf{y}$ and the data provide no information about it.

For the spatial models considered here, it is easy to reformulate the restricted spatial regression so it has no superfluous dimensions. Let $\mathbf{P}^c$ have spectral decomposition $\mathbf{P}^c = (\mathbf{L}|\mathbf{K})\boldsymbol{\Phi}(\mathbf{L}|\mathbf{K})'$, where $\boldsymbol{\Phi}$ is a diagonal matrix with $n - 1$ eigenvalues of 1 and one 0 eigenvalue, $\mathbf{L}$ has n rows and $n - 1$ columns, and $\mathbf{K}$ has n rows and one column, with $\mathbf{K}$ proportional to $\mathbf{X}$ and $\mathbf{K}'\mathbf{L} = 0$. Then fit the following model:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{LS}^* + \boldsymbol{\epsilon}, \tag{11}$$

where $\mathbf{S}^*$ is $(n - 1)$-dimensional normal with mean 0 and precision matrix $\tau_s\mathbf{L}'\mathbf{QL}$, $\mathbf{Q}$ is the neighbor matrix from $\mathbf{S}$'s ICAR model, and $\boldsymbol{\epsilon}$ is iid normal with mean 0 and precision τ_e.

Using this model in either form and conditioning on (τ_s, τ_e), β has the same conditional posterior mean as in the analysis without the ICAR-distributed $\mathbf{S}$, but has larger conditional posterior variance (Reich, Hodges, and Zadnik 2006, sec. 3). Thus, restricted spatial regression discounts the sample size to account for spatial correlation without changing β's point estimate conditional on τ_s and τ_e.

2.5 Spatial Confounding Is Not an Artifact of the ICAR Model

Spatial confounding also occurs if, in model (2), $\mathbf{S}$ is given a proper multivariate normal distribution with any of several covariance matrices capturing the intuition that near regions are more similar than distant. Wakefield (2007), for example, found very similar spatial-confounding effects in the Scottish lip-cancer data using the ICAR model and a so-called geostatistical model. For the Slovenian data, we considered geostatistical models in which each municipality's X_i and y_i were treated as being measured at one point in the municipality. For the analyses described below and in Part A of the online supplement, each municipality's point had east-west coordinate the average of the farthest east and farthest west boundary points, and

analogous north-south coordinate. An example of a proper covariance matrix is $\text{cov}(\mathbf{S})$ with (i, j)th element $\sigma_s^2 \exp(-\delta_{ij}/\theta)$, with δ_{ij} being Euclidean distance between the points representing municipalities i and j and θ controlling spatial correlation. Each such covariance matrix that we considered had an unknown parameter like θ, which for now we treat as fixed and known. Applying Section 2.2's approach to such models for $\mathbf{S}$ requires only one change, arising because the precision matrix $\text{cov}(\mathbf{S})^{-1}$ now has no zero eigenvalues: we must add an explicit intercept to model (5). Therefore, holding θ fixed, spatial confounding will occur by the same mechanism as with the ICAR model for $\mathbf{S}$. (For the Slovenian data, the smallest eigenvalue of $\text{cov}(\mathbf{S})^{-1}$ has eigenvector Z_{194} which, while not constant over the map as for the ICAR model, nonetheless varies little. We have seen this in other geostatistical models; it explains why the intercept can be poorly identified in such models, but the generality of this intercept confounding is unknown.)

In the preceding paragraph, we treated as fixed the parameter θ in $\text{cov}(\mathbf{S})$. In a non-Bayesian analysis these parameters are typically estimated, while in a Bayesian analysis they are random variables. Thus in applying Section 2.2's analysis to geostatistical models for $\mathbf{S}$, the canonical regressors $\mathbf{Z}$ depend on the unknown parameter θ of $\text{cov}(\mathbf{S})$. However, for at least four common forms of $\text{cov}(\mathbf{S})$ that we explored (Part A of the online supplement), as θ varied over a wide range the eigenvector corresponding to the second-smallest eigenvalue of $\text{cov}(\mathbf{S})^{-1}$—the canonical regressor Z_i most likely to produce spatial confounding—was highly correlated with the analogous eigenvector of the ICAR model's precision matrix $\tau_s\mathbf{Q}$, the canonical regressor that *did* produce spatial confounding in this dataset. Thus, spatial confounding will occur in the Slovenian data for these four geostatistical models.

Applying the spectral approximation to geostatistical models (Part B of the online supplement) is another way to make the foregoing more concrete. For measurements taken on a regular square grid, the spectral approximation is closely analogous to Section 2.1's formulation of the ICAR model as a linear model. For this special case, the canonical regressors with least- and most-shrunk coefficients are literally trigonometric functions that are low- and high-frequency, respectively.

Penalized splines are a quite different approach to spatial smoothing; Ruppert, Wand, and Carroll (2003, chap. 13) gave a very accessible introduction. However, penalized splines produce the same confounding effect in the Slovenian data. In a class project (Salkowski 2008), a student fit a two-dimensional penalized spline to the Slovenian data, attributing each municipality's counts to a point as described above. He used the R package SemiPar (version 1.0-2; the package and documentation are at *http://www.uow.edu.au/~mwand/webspr/rsplus.html*) to fit a model in which municipality i's observed count of stomach cancers O_i was Poisson with log mean

$$\log\{E_i\} + \beta_0 + \beta_{\text{SE}c}\text{SE}c_i + \beta_A A_i + \beta_N N_i + \sum_k u_k \text{basis}_{ki}, \tag{12}$$

where A_i is municipality i's east-west coordinate, N_i is its north-south coordinate (each coordinate was centered and both were scaled by a single scaling constant to preserve the map's shape), basis_{ki} is the default basis in SemiPar (based on the Matérn covariance function), the u_k were modeled as iid normal, and the knots were SemiPar's default knots. SemiPar's default fitting method, penalized quasi-likelihood, shrank the random-effect term $\sum_k u_k \text{basis}_{ki}$ to zero but the two fixed effects implicit in the spline, A_i and N_i, remained in the model and produced a collinearity effect as in an ordinary linear model. Without the spatial spline, a simple generalized linear model fit gave an estimate for $\beta_{\text{SE}c}$ of -0.137 (standard error 0.020), essentially the same as in Zadnik's Bayesian analysis, while adding just the fixed effects A_i and N_i changed $\beta_{\text{SE}c}$'s estimate to -0.052 (SE 0.028). As the spline fit was forced to be progressively less smooth, $\beta_{\text{SE}c}$'s estimate increased monotonically and eventually became positive. (Steinberg and Bursztyn 2004, p. 415, noted in passing a similar confounding effect in a different spline.)

Thus, spatial confounding is not an artifact of the ICAR model, but arises from other, perhaps all, specifications of the intuition that measures taken at locations near to each other are more similar than measures taken at distant locations.

3. EVALUATING THE FIVE INTERPRETATIONS; IMPLICATIONS FOR PRACTICE

Before we can discuss interpretations of spatial confounding, we need to distinguish two interpretations of "random effect." One is traditional, as in the definition from Scheffé (1959) noted above: the levels of a random effect are draws from a population, and the draws are not of interest in themselves but only as samples from the larger population, which *is* of interest. In recent years, "random effect" has come to be used in a second sense, to describe effects that have the mathematical form of a Scheffé-style random effect but which are quite different. For these newer-style random effects, the levels are the entire population; or the levels are themselves of interest; or the levels are in no meaningful sense draws from a population, from which further draws could be made. The Slovenian data are an example in which the levels (municipalities) are the entire population. Hospital quality-of-care studies provide examples in which the levels (hospitals) may be considered draws from a population but are themselves of interest. The mixed-model representation of penalized splines (Ruppert, Wand, and Carroll 2003) is an example of random effects with levels that are not draws from any conceivable population. In the simplest case of a one-dimensional penalized spline with a truncated-line basis, the random effect is the changes in the fit's slope at each knot, and its distribution is simply a device for penalizing changes in the slope. The senselessness of imagining further draws from such random effects is clearest for the examples in the book by Ruppert, Wand, and Carroll (2003) in which penalized splines are used to estimate smooth functions in the physical sciences.

A full discussion of random-effect interpretations is beyond the present article's scope. We note, however, that discussions of spatial random effects are generally either unclear about their interpretation or seem to treat them as Scheffé-style random effects. Finally, it is both economical and accurate to describe all non-Scheffé-style random effects as formal devices to implement a smoother, interpreting shrinkage estimation as a kind of smoothing, so from now on we do so.

3.1 The Random Effect S Is a Formal Device to Implement a Smoother

Consider situations in which **S** is not a Scheffé-style random effect. For these situations, we have seen two interpretations of spatial confounding.

(i) Spatially-correlated errors remove bias in estimating β and are generally conservative (Clayton, Bernardinelli, and Montomoli 1993).
(ii) Spatially-correlated errors can introduce or remove bias in estimating β and are not necessarily conservative (Wakefield 2007; implicit in Reich, Hodges, and Zadnik 2006).

It is commonly argued (e.g., Clayton, Bernardinelli, and Montomoli 1993) that introducing spatially correlated errors into a model, as with $\mathbf{S}+\boldsymbol{\epsilon}$, captures the effects of spatially-structured missing covariates and thus adjusts the estimate of β for such missing covariates even if we have no idea what those covariates might be. Interpretation (i) reflects this view. We have also heard a somewhat different statement of this view, as in: "I *know* I am missing some confounders, in fact I have some specific confounders in mind that I was unable to collect, but from experience I know they have a spatial pattern. Therefore, I will add **S** to the model to try to recover them and let the data decide how much can be recovered." In some fields, it is nearly impossible to get an article published unless a random effect is included for this purpose.

We can now evaluate this view using the results of Section 2, which are a modest elaboration of linear-model theory. Indeed, the only aspect of the present problem not present in linear-model theory is that most of the canonical coefficients b_i are shrunk toward 0, although the b_i that produce spatial confounding are shrunk the least and thus deviate least from linear-model theory.

To make the discussion concrete, consider estimating β using the model $\mathbf{y}=\mathbf{1}_n\alpha+\mathbf{X}\beta+\boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ is iid normal error, then estimating β using a larger model, either $\mathbf{y}=\mathbf{1}_n\alpha+\mathbf{X}\beta+\mathbf{H}\gamma+\boldsymbol{\epsilon}$, where **H** is a supposed missing covariate, or using model (2), which adds the ICAR-distributed spatial random effect **S**. The Appendix gives explicit expressions for the adjustment in the estimate of β under either of these larger models, and for the *expected* adjustment assuming the data were generated by the model

$$\mathbf{y}=\mathbf{1}_n\alpha+\mathbf{X}\beta+\mathbf{H}\gamma+\boldsymbol{\epsilon}. \tag{13}$$

We now summarize the results in the Appendix.

There is no necessary relationship between the adjustment to β's estimate arising from adding **S** to the model and the adjustment arising from adding the supposed missing covariate **H**. This is most striking if we suppose that **H** is uncorrelated with **X**, so that adding **H** to the model would not change the estimate of β. In this case, adding the spatial random effect **S** *does* adjust the estimate of β, and in a manner that depends not on **H** but on the correlation of **X** and **y** and on the spatial map. If the data are generated by (13), the expected adjustment in β's estimate from adding **S** to the model is not zero in general and can be biased in either direction. If there *are* no missing covariates, adding **S** nonetheless adjusts β's estimate in the manner just described although the expected adjustment is zero. It is fair to describe such adjustments as haphazard.

Now suppose **H** is correlated with **X**, so that adding **H** to the model changes the estimate of β. In this case, the adjustment to β's estimate under the spatial model is again biased relative to the correct adjustment from including **H**. The bias can be large and either positive or negative, depending on the degree of smoothing (more smoothness generally implies larger bias) and depending haphazardly on **H** and on the spatial map. The bias can even be in the wrong direction, so that on average β's estimate becomes larger when it would become smaller if **H** were added to the model.

Therefore, adding spatially correlated errors is not conservative: a canonical regressor Z_i that is collinear with **X** can cause β's estimate to increase in absolute value just as in ordinary linear models. Further, in cases in which β's estimate should not be adjusted, introducing spatially-correlated errors will, nonetheless, adjust the estimate haphazardly.

From the perspective of linear-model theory, it seems perverse to use an error term to adjust for the possibility of missing confounders. The analog in ordinary linear models would be to move part of the fitted coefficients into error to allow for the possibility of as-yet-unconceived missing confounders. In using an ordinary linear model, we know that if missing confounders are correlated with included fixed effects, variation in **y** that would be attributed to the missing confounders is instead attributed to the included fixed effects. We acknowledge that possibility in the standard disclaimer that if we have omitted confounders, our coefficient estimates could be wrong. In spatial modeling, the analogy to this practice would be to use restricted spatial regression, so that all variation in **y** in the column space of included fixed effects is attributed to those included effects instead of being haphazardly reallocated to the spatial random effect.

Therefore, interpretation (i) cannot be sustained and interpretation (ii) is correct, when the random effect **S** is interpreted as a mere formal device to implement a smoother. Adding spatially-correlated errors cannot be expected to capture the effect of a spatially-structured missing covariate, but only to smooth fitted values and discount the sample size in computing standard errors or posterior standard deviations for fixed effects. Therefore, in such cases you should *always* use restricted spatial regression so the sample size can be discounted without distorting the fixed-effect estimate. If you are concerned about specific unmeasured confounders, you should add to the model a suitable explicit fixed effect, not adjust haphazardly by means of a spatially-correlated error. Finally, conclusions from such analyses should be qualified as in any other observational study; for example, we have estimated the association of our outcome with our regressors accounting for measured confounders, and if we have omitted confounders, then our estimate could be wrong.

3.2 S Is a Scheffé-Style Random Effect

For these situations, we have seen three interpretations. (We have rewritten interpretation (iii) in light of Section 2.)

(iii) The regressors $\mathbf{Z}$ implicit in the spatial effect $\mathbf{S}$ are collinear with the fixed effect $\mathbf{X}$, but neither estimate of β is biased (David B. Nelson, personal communication).
(iv) Adding the spatial effect $\mathbf{S}$ creates information loss, but neither estimate of β is biased (David B. Nelson, personal communication).
(v) Because error is correlated with the regressor $\mathbf{X}$ in the sense commonly used in econometrics, *both* estimates of β are biased (Paciorek 2011).

Interpretations (iii) and (iv) treat the fixed effect $\mathbf{X}$ as measured without error and not otherwise drawn from a probability distribution ("fixed and known"), while interpretation (v) treats $\mathbf{X}$ as drawn from a probability distribution. Interpreting spatial confounding therefore depends on whether $\mathbf{X}$ is interpreted as fixed and known or as a random variable. This is a messy business, which seems to be determined in practice less by facts than by the department in which one was trained. The present authors' training inclines us to view $\mathbf{X}$ as fixed and known as a default, while econometricians, for example, seem inclined to the opposite default.

To see the difficulty, consider an example in which the random effect is hospitals selected as a random sample from a population of hospitals, and the fixed effect is an indicator of whether a hospital is a teaching hospital. The present authors' default is to treat teaching status as fixed and known. However, if we have drawn 20 hospitals at random, then the teaching status of hospital i is a random variable determined by the sample of hospitals we happen to draw, so $\mathbf{X}$ is drawn from a probability distribution. But what if, as often happens, sampling is stratified by teaching status to ensure that (say) 10 hospitals are teaching hospitals and 10 are not? Now teaching status is fixed and known. But what if someone gives us the dataset and we do not know whether sampling was stratified by teaching status? One might argue that our ignorance disqualifies us from analyzing these data, but that argument is not compelling to, for example, people who interpret the Likelihood Principle as meaning they can ignore the sampling mechanism, or to many people who do not have a tenured, hard-money faculty position.

Again, a full discussion of this issue is beyond the present article's scope. It is also unnecessary for the present purpose, because there are unarguable instances of each kind of fixed effect. An example of a fixed and known $\mathbf{X}$ could arise in analyzing air pollution measured at many fixed monitoring stations on each day in a year. The days could be interpreted as a Scheffé-style random effect, and the elevation of each monitoring station as a fixed and known $\mathbf{X}$. For an example of $\mathbf{X}$ plainly drawn from a probability distribution, consider the hospitals example just above, where the morbidity score for each hospital's patients is a random variable for the period being studied.

So first assume $\mathbf{X}$ is fixed and known. It is then straightforward to show that both (iii) and (iv) are correct and indeed arguably identical, though we think they are worth distinguishing. Interpretation (iv) follows from Section 2.3 and the familiar fact that generalized least squares gives unbiased estimates even when the covariance matrix is specified incorrectly. For interpretation (iii), recall from (8) that the estimate of β in the spatial model (which, given τ_e and τ_s, is both the posterior mean and the usual estimate following maximization of the restricted likelihood) is

$$E(\beta|\tau_e, \tau_s, \mathbf{y}) = \hat{\beta}_{\text{OLS}} - (n-1)^{-1/2}\mathbf{R}'E(\mathbf{b}|\tau_e, \tau_s, \mathbf{y}). \quad (14)$$

By (7), $(n-1)^{-1/2}\mathbf{R}'E(\mathbf{b}|\tau_e, \tau_s, \mathbf{y})$ can be written as $\mathbf{K}\mathbf{P}^c\mathbf{y}$ where $\mathbf{K}$ is an appropriate-sized known square matrix. Recalling that $\mathbf{P}^c = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, $\mathbf{P}^c\mathbf{y} = \mathbf{P}^c(\mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon})$, which has expectation 0 with respect to $\mathbf{b}$ and $\boldsymbol{\epsilon}$. Hence the spatial and OLS estimates of β have the same expectation and are unbiased based on the aforementioned familiar fact about generalized least squares.

Now assume $\mathbf{X}$ is a random variable. Paciorek (2011) interpreted model (5) as $\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\psi}$ where the error term $\boldsymbol{\psi} = \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon}$ has a nondiagonal covariance matrix. Because $\mathbf{X}'\mathbf{Z} \neq 0$, $\mathbf{X}$ is correlated with $\boldsymbol{\psi}$, so by the standard result in econometrics, *both* the OLS estimate of β and the estimate of β using (5) are biased. Formulating the result this way is more precise than the common statement that bias arises when "the random effect is correlated with the fixed effect," because the distribution of "the random effect" depends on the parameterization: the random effect $\mathbf{b}$ in model (5) is independent of $\mathbf{X}$, but the random effect $\mathbf{S}$ in model (2) is not.

The main point of the work of Paciorek (2011), which presumes the spatial random effect $\mathbf{S}$ captures a missing covariate, is that "bias [in estimating β] is reduced only when there is variation in the covariate [$\mathbf{X}$] at a scale smaller than the scale of the unmeasured covariate [supposedly captured by $\mathbf{S}$]." We conjecture that this can be interpreted in Section 2.2's terms as meaning that bias is reduced if $\mathbf{X}$ is not too highly correlated with the low-frequency canonical regressors Z_i that have small d_i and hence little shrinkage in b_i.

Paciorek (2011) concluded that restricted spatial regression is either irrelevant to the issue of bias in estimating β or makes an overly strong assumption by attributing to $\mathbf{X}$ all of the disputed variation in $\mathbf{y}$. The latter appears to presume that the spatial random effect $\mathbf{S}$ captures an unspecified missing covariate, which, we have argued, is difficult to sustain. However, this area of research is just beginning and much remains to be developed.

4. CONCLUSION

The preceding sections laid out the mechanics by which spatial confounding occurs, expanding on the work of Reich, Hodges, and Zadnik (2006); showed briefly that this is not an artifact of the ICAR model but is far more general; gave an alternative analysis (restricted spatial regression) that removes spatial confounding; and considered proposed interpretations of spatial confounding, concluding that restricted spatial regression should be used routinely when $\mathbf{S}$ is a formal device to implement a smoother, a common situation.

The literature on spatial confounding is small but it appears many people encounter this problem in practice. Thus although the present article is not the last word on the subject, it does bring together the various approaches to this phenomenon, which should in time yield generally-accepted advice for statistical practice.

Our understanding of the mechanics of spatial confounding is underdeveloped in certain respects. In debunking the common belief that spatially-correlated errors adjust for unspecified missing covariates, our derivation (the Appendix) took the smoothing ratio $r = \tau_s/\tau_e$ as given. Although r's marginal posterior distribution is easily derived when τ_e has a gamma prior, it is harder to interpret than the posterior mean of β so its implications are as yet unclear. However, it should be possible to extract some generalizations which, with the expressions in the Appendix, will permit understanding of the situations in which spatial confounding will and will not occur. We hypothesize this work will show that whenever both $\mathbf{y}$ and $\mathbf{X}$ are highly correlated with Z_{n-G}, the canonical regressor with the least-smoothed coefficient, there will be little smoothing (r will be small) and β's estimate under the spatial model will be close to zero. In other words, we hypothesize that in any map, when both $\mathbf{y}$ and $\mathbf{X}$ show a strong trend along the long axis of the map, adding a spatially-correlated error will nullify the obvious association between $\mathbf{y}$ and $\mathbf{X}$ as it did in the Slovenian data.

The theory is particularly underdeveloped for the situation in which both $\mathbf{S}$ and $\mathbf{X}$ can be interpreted as random in Scheffé's sense. The work of Paciorek (2011) is a first step in what should be a rich area of research.

APPENDIX: HOW β'S ESTIMATE CHANGES WHEN A SUPPOSED MISSING COVARIATE OR THE RANDOM EFFECT S IS ADDED TO THE MODEL

Consider these three models:

- *Model* 0: $\mathbf{y} = \mathbf{1}_n\alpha + \mathbf{X}\beta + \boldsymbol{\epsilon}$,
- *Model* H: $\mathbf{y} = \mathbf{1}_n\alpha + \mathbf{X}\beta + \mathbf{H}\gamma + \boldsymbol{\epsilon}$,
- *Model* S: $\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon}$,

where Model S, the spatial model, is the same as (2) and (5). Assume, as before, that $\mathbf{X}$ is centered and scaled so $\mathbf{1}_n'\mathbf{X} = 0$ and $\mathbf{X}'\mathbf{X} = n - 1$; assume $\mathbf{H}$ is centered and scaled the same way; and assume $\mathbf{y}$ is centered but not scaled. Bayesian results below assume flat (improper) priors on α, β, and γ.

Under Model 0, the estimate of β—the posterior mean or least squares estimate—is

$$\hat{\beta}^{(0)} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y = \rho_{XY}(\mathbf{y}'\mathbf{y})^{0.5}, \tag{A.1}$$

where ρ_{AB} is Pearson's correlation of the vectors $\mathbf{A}$ and $\mathbf{B}$. Under Models H and S, the estimates of β given τ_s and τ_e—the conditional posterior mean for both models; for Model H this is also the least squares estimate, while for Model S it is the customary estimate after maximizing the restricted likelihood—are

$$\hat{\beta}^{(H)} = \hat{\beta}^{(0)} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{H}(\mathbf{H}'\mathbf{P}^c\mathbf{H})^{-1}\mathbf{H}'\mathbf{P}^c\mathbf{y}, \tag{A.2}$$

$$\hat{\beta}^{(S)} = \hat{\beta}^{(0)} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{P}^c\mathbf{Z} + r\mathbf{D})^{-1}\mathbf{Z}'\mathbf{P}^c\mathbf{y}, \tag{A.3}$$

where, as before, $r = \tau_s/\tau_e$, $\mathbf{P}^c = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, and $\mathbf{D} = \text{diag}(d_1, \ldots, d_{n-G}, 0, \ldots, 0)$ is the diagonal matrix containing the eigenvalues d_i of the spatial-neighbor matrix $\mathbf{Q}$. The estimate $\hat{\beta}^{(S)}$ was given in Section 2.2 and $\hat{\beta}^{(H)}$ is derived by a similar argument.

Define $\mathbf{B}_X = \mathbf{Z}'\mathbf{X}/(n-1)^{0.5}$; the entries in $\mathbf{B}_X$ are the correlations between $\mathbf{X}$ and the columns of $\mathbf{Z}$ (we called this $\mathbf{R}$ in Section 2.2). Define $\mathbf{B}_H$ analogously as $\mathbf{B}_H = \mathbf{Z}'\mathbf{H}/(n-1)^{0.5}$. Finally, define $\mathbf{B}_y = \mathbf{Z}'\mathbf{y}/[(n-1)(\mathbf{y}'\mathbf{y})]^{0.5}$; $\mathbf{B}_y$'s entries are the correlations between $\mathbf{y}$ and the columns of $\mathbf{Z}$. Then the estimates of β under Models H and S, given τ_s and τ_e, can be shown to be

$$\begin{aligned}\hat{\beta}^{(H)} &= \hat{\beta}^{(0)}\left[1 - \frac{\frac{\rho_{XH}\rho_{HY}}{\rho_{XY}} - \rho_{XH}^2}{1 - \rho_{XH}^2}\right],\\ \hat{\beta}^{(S)} &= \hat{\beta}^{(0)}\left[1 - \frac{\frac{\rho'_{XY}}{\rho_{XY}} - q}{1 - q}\right],\end{aligned} \tag{A.4}$$

where $\rho'_{XY} = \mathbf{B}'_X(\mathbf{I} + r\mathbf{D})^{-1}\mathbf{B}_y = \mathbf{X}'(I + r\mathbf{Q})^{-1}\mathbf{y}/((n-1) \times \mathbf{y}'\mathbf{y})^{0.5}$ and $q = \mathbf{B}'_X(\mathbf{I} + r\mathbf{D})^{-1}\mathbf{B}_X = \mathbf{X}'(I + r\mathbf{Q})^{-1}\mathbf{X}/(n-1)$.

In (A.4), the expressions in square brackets for $\hat{\beta}^{(H)}$ and for $\hat{\beta}^{(S)}$ have no necessary relation to each other. For example, if $\mathbf{X}$ and $\mathbf{H}$ are uncorrelated, so $\rho_{XH} = 0$, then $\hat{\beta}^{(H)} = \hat{\beta}^{(0)}$ but $\hat{\beta}^{(S)} \neq \hat{\beta}^{(0)}$. In particular, $\hat{\beta}^{(S)}$ can be larger or smaller than $\hat{\beta}^{(H)}$ in absolute value; this depends on how $(\mathbf{I} + r\mathbf{D})^{-1}$ differentially downweights specific coordinates of $\mathbf{B}_X$ and $\mathbf{B}_y$. When $\rho'_{XY} \approx \rho_{XY}$, $\hat{\beta}^{(S)} \approx 0$. This happens if the ith coordinate of $\mathbf{B}_X$, the correlation of $\mathbf{X}$ and Z_i, is large; the ith coordinate of $\mathbf{B}_y$, the correlation of $\mathbf{y}$ and Z_i, is large; and d_i and r are small, as in the Slovenian data.

If the data are generated by Model H, we can treat $\hat{\beta}^{(H)}$ and $\hat{\beta}^{(S)}$ as functions of $\mathbf{y}$, holding the precisions τ_e and τ_s fixed, and compute the expected change in β's estimate from adding either $\mathbf{H}$ or $\mathbf{S}$ to Model 0. The expected changes are

$$E\left(\hat{\beta}^{(H)} - \hat{\beta}^{(0)}|\tau_e\right) = \rho_{XH}\gamma, \tag{A.5}$$

$$\begin{aligned}&E\left(\hat{\beta}^{(S)} - \hat{\beta}^{(0)}|\tau_e, \tau_s\right)\\ &\quad= \left[\frac{\frac{\rho'_{XH}}{\rho_{XH}} - q}{1 - q}\right]\rho_{XH}\gamma, \quad \text{if } \rho_{XH} \neq 0,\end{aligned} \tag{A.6}$$

$$= \frac{\rho'_{XH}}{1 - q}\gamma, \quad \text{if } \rho_{XH} = 0, \tag{A.7}$$

where $\rho'_{XH} = \mathbf{B}'_X(\mathbf{I} + r\mathbf{D})^{-1}\mathbf{B}_H = \mathbf{X}'(I + r\mathbf{Q})^{-1}\mathbf{H}/(n-1)$. If $\gamma \neq 0$, the expected adjustment under Model S is biased, with the bias depending on how $(\mathbf{I} + r\mathbf{D})^{-1}$ differentially downweights specific coordinates of $\mathbf{B}_X$ and $\mathbf{B}_H$; the bias can be positive or negative. If $\mathbf{H} = Z_j$, then $\rho'_{XH} = \rho_{XH}/(1 + rd_j)$, so (A.6) becomes

$$E\left(\hat{\beta}^{(S)} - \hat{\beta}^{(0)}|\tau_e, \tau_s\right) = \left[\frac{(1 + rd_j)^{-1} - q}{1 - q}\right]\rho_{XH}\gamma. \tag{A.8}$$

The expression in square brackets is less than 1 when $rd_j > 0$ and becomes negative if r or d_j is large enough, that is, Model S adjusts $\hat{\beta}$ in the wrong direction from $\hat{\beta}^{(0)}$. If there is no missing covariate, $\gamma = 0$, then the expected adjustment is zero under both Model H and Model S.

SUPPLEMENTARY MATERIALS

Additional sections: *Part A*: A small exploration of spatial confounding in some common geostatistical models. *Part B*: Spatial confounding with spectral methods. (supplement.pdf)

[Received March 2010. Revised October 2010.]

REFERENCES

Besag, J., York, J. C., and Mollié, A. (1991), "Bayesian Image Restoration, With Two Applications in Spatial Statistics" (with discussion), *Annals of the Institute of Statistical Mathematics*, 43, 1–59. [325]

Clayton, D. G., Bernardinelli, L., and Montomoli, C. (1993), "Spatial Correlation in Ecological Analysis," *International Journal of Epidemiology*, 22, 1193–1201. [326,331]

Hodges, J. S., Carlin, B. P., and Fan, Q. (2003), "On the Precision of the Conditionally Autoregressive Prior in Spatial Models," *Biometrics*, 59, 317–322. [327]

Paciorek, C. J. (2011), "The Importance of Scale for Spatial-Confounding Bias and Precision of Spatial Regression Estimators," *Statistical Science*, to appear. [326,332,333]

Reich, B. J., and Hodges, J. S. (2008), "Identification of the Variance Components in the General Two-Variance Linear Model," *Journal of Statistical Planning and Inference*, 138, 1592–1604. [327]

Reich, B. J., Hodges, J. S., and Zadnik, V. (2006), "Effects of Residual Smoothing on the Posterior of the Fixed Effects in Disease-Mapping Models," *Biometrics*, 62, 1197–1206. [326-329,331,332]

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003), *Semiparametric Regression*, New York: Cambridge University Press. [330]

Salkowski, N. J. (2008), "Using the SemiPar Package," class project, available at *http://www.biostat.umn.edu/~hodges/SalkowskiRPMProject.pdf*. [330]

Scheffé, H. (1959), *The Analysis of Variance*, New York: Wiley. [326,330]

Spiegelhalter, D. M., Best, D. G., Carlin, B. P., and Van der Linde, A. (2002), "Bayesian Measures of Model Complexity and Fit" (with discussion), *Journal of the Royal Statistical Society, Ser. B*, 64, 583–639. [326]

Steinberg, D. M., and Bursztyn, D. (2004), "Data Analytic Tools for Understanding Random Field Regression Models," *Technometrics*, 46, 411–420. [330]

Wakefield, J. (2007), "Disease Mapping and Spatial Regression With Count Data," *Biostatistics*, 8, 158–183. [326,327,329,331]

Zadnik, V., and Reich, B. J. (2006), "Analysis of the Relationship Between Socioeconomic Factors and Stomach Cancer Incidence in Slovenia," *Neoplasma*, 53, 103–110. [325]