"Science Stays True Here"
Journal of Mathematics and Statistical Science (ISSN 2411-2518), Vol.5, 333-344 | Science Signpost Publishing

Correlation induced by missing spatial covariates: a connection between variance components models and kriging

Jessica Rothman[1], Monica C. Jackson[2], Kimberly F. Sellers[3], Talithia Williams[4], Subhash R. Lele[5], and Lance A. Waller[6]

[1]Department of Biostatistics, Yale University, New Haven, CT 06510

[2]Department of Mathematics and Statistics, American University, Washington, DC 20016. Corresponding author: monica@american.edu

[3]Department of Mathematics and Statistics, Georgetown University, Washington, DC 20057

[4]Department of Mathematics, Harvey Mudd College, Claremont, CA 91711

[5]Department of Mathematical and Statistical Sciences, University of Alberta, University of Alberta Edmonton, AB, Canada

[6]Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA 30322

Abstract

Residual spatial correlation in linear models of environmental data is often attributed to spatial patterns in related covariates omitted from the fitted model. We connect the nonunique decomposition of error in geostatistical models into trend and covariance components to the similarly non-unique decomposition of mixed models into fixed and random effects. We specify spatial correlation induced by missing spatial covariates as a function of the strength of association and (spatial) covariation of the missing covariates. The connection with variance components models provides insight into estimation procedures. We showed how missing covariates in spatial linear models actually induces spatial autocorrelation in the covariates. This finding was confirmed through the use of simulated data and the Binary Steve dataset.

Keywords: geostatistics, spatial regression, variable selection, semi-variogram, spatial autocorrelation

Introduction

An underpinning of much of spatial statistics is the notion of positive *spatial autocorrelation*, i.e. measurements of the same quantity in nearby locations are positively correlated. Indeed, such autocorrelation enables most geostatistical methods for prediction (e.g. kriging) since observed measurements provide information for prediction at unmeasured locations. The "First Law of Geography" (Tobler, 1970) summarizes the philosophical basis for positive spatial autocorrelation: "everything is related to everything else, but near things are more related than distant things." We consider this law in the framework of linear models with correlated error terms. In this setting, we assume spatial autocorrelation is induced by unmeasured or unmeasurable covariates, an assumption based on the idea that nearby measurements are similar due to shared local environmental factors. Such a conceptual model is appropriate in an observational setting, wherein covariate values represent observed values of random variables rather than values fixed by an experimental design. This set of assumptions has a long history in quantitative geog-

raphy, leading Ripley (2005) to remark: "Indeed, the philosophy adopted seems to have been that if 'spatial autocorrelation' is found more explanatory variables should be introduced until it disappears!" (p. 98).

Research that focuses on potential unknown spatial error in model residuals has been sparse in recent years. Previous work by Hodges and Reich (2010) proposes a restricted spatial regression model to eliminate spatial confounding in spatially-correlated error terms. Other researchers have studied the relationship between spatially misaligned data by characterizing the Berkson error induced from kriging a spatially misaligned dataset. Some authors characterize the total measurement error as part of a broader class of Berkson error models and develop an estimated generalized least squares estimator using estimated covariance parameters (Lopiano et al., 2011, 2013, 2014). Various authors have modeled known spatial error by applying a Berkson-type measurement error on the error structure. Gryparis et al. (2009) and Szpiro et al. (2011) employed a "parameter bootstrap," a computationally effective method using nonlinear optimization to solve for exposure model parameters.

An apparent argument in geostatistics is that accurate modeling of the spatial covariance structure obviates the need to include relevant covariates. Most "trends" removed in geostatistics are simple linear or quadratic functions of coordinates rather than measured covariates. Here, however, the goal is accurate prediction rather than accurate estimation of particular covariate effects. The statistical mechanism for accounting for induced spatial autocorrelation is not well understood and has received little research effort to date. We outline the connection between missing covariates and induced spatial autocorrelation in the dependent variables. Next we frame this relationship in terms of variance components and illustrate a convenient partitioning of error terms into independent error and spatially correlated components. The variance components formulation provides links to established estimation techniques, which we illustrate on the *Binary Steve* dataset.

Simple case: Autocorrelation induced by a missing spatial covariate

Consider a very simple example to fix ideas. Let Y_i denote the dependent variable measured at spatial location i, $i = 1, \ldots, n$. Let X_i denote an independent variable observed at the same location. We restrict attention to observational studies and assume that X_i is a random variable (a covariate, in the truest sense of the word), and is not a design variable set at a particular value by the experimenter. Suppose the model

$$Y_i = \mu + \beta X_i + \epsilon_i \tag{1}$$

holds, where the ϵ_i are independent, identically distributed Gaussian random variables. Note that the $Y_i \mid X_i$ are independent. Now restructure the model as follows:

$$Y_i = \mu + \phi_i, \tag{2}$$

where $\phi_i = \beta X_i + \epsilon_i$. We see that

$$\begin{aligned} Cov(Y_i, Y_j) &= Cov(\phi_i, \phi_j) \\ &= Cov(\beta X_i + \epsilon_i, \beta X_j + \epsilon_j) \\ &= \beta^2 Cov(X_i, X_j), \end{aligned} \tag{3}$$

i.e. the covariance of $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ is induced by the spatial covariance of $\boldsymbol{X} = (X_1, \ldots, X_n)^T$, modified by the strength of association between $\boldsymbol{Y}$ and $\boldsymbol{X}$ (i.e. β). Such examples evolve from correlation induced by linear filters of random sequences outlined for time series by Diggle (1990).

We can consider Equation (2) as a version of Equation (1) wherein a main (fixed) effect is recast as a random effect (with a correlation structure) in the language of mixed models. Equivalently, Equation (2) shifts the effect of covariate X from an explicit "large-scale" or trend error component to a "small-scale" or covariance error component in partitioning of errors in spatial models as described in Cressie (1993). As noted by Cressie (1993), partitioning of the variation in $\boldsymbol{Y}$ between main (trend) effects and correlation is not unique once we allow model

errors to be correlated. Here we see a conceptual connection between mixed and spatial models, a connection we formalize in the next section.

General linear case

Suppose we consider a linear model for $\boldsymbol{Y}$, the vector of dependent variables, and a $n \times p$ matrix of covariate values $\boldsymbol{X}$. Again, we limit consideration to an observational setting where the observed values of X_{ij}, $i = 1, \ldots, n$; $j = 1, \ldots, p$ are realizations of random variables observed at different spatial locations (some of which may exhibit spatial autocorrelation themselves).

Following the notation of Draper and Smith (1998), suppose we can partition the covariate matrix $\boldsymbol{X}$ and parameter vector $\boldsymbol{\beta}$ into two components (i.e. $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, and $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, respectively) such that

$$\boldsymbol{X} = (\boldsymbol{X}_1 | \boldsymbol{X}_2) \qquad \text{and} \qquad \boldsymbol{\beta} = \left(\begin{array}{c} \boldsymbol{\beta}_1 \\ \hline \boldsymbol{\beta}_2 \end{array} \right).$$

Further, suppose that one postulates the model relationship to be $\boldsymbol{Y} = \boldsymbol{X}_1\boldsymbol{\beta}_1$ instead of $\boldsymbol{Y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2$. Letting $\boldsymbol{Y}$ be normally distributed with mean $\boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2$ and variance $\sigma^2\boldsymbol{V}$, let $\boldsymbol{P}_{\boldsymbol{X}_1} = \boldsymbol{I} - \boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'$ denote the projection/hat matrix that projects onto the $n - k$ dimensional space that is orthogonal to the $k < p$ dimensional space spanned by the columns of $\boldsymbol{X}_1$. Thus, we find that the residuals stemming from misidentifying the proper model structure have the following mean and variance:

$$\begin{aligned} E(\boldsymbol{P}_{\boldsymbol{X}_1}\boldsymbol{Y}) &= \boldsymbol{P}_{\boldsymbol{X}_1}E(\boldsymbol{Y}) \\ &= \boldsymbol{P}_{\boldsymbol{X}_1}(\boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2) \\ &= \boldsymbol{P}_{\boldsymbol{X}_1}(\boldsymbol{X}_1\boldsymbol{\beta}_1) + \boldsymbol{P}\boldsymbol{X}_1(\boldsymbol{X}_2\boldsymbol{\beta}_2) \\ &= \boldsymbol{0} + \boldsymbol{P}_{\boldsymbol{X}_1}(\boldsymbol{X}_2\boldsymbol{\beta}_2) = \boldsymbol{P}_{\boldsymbol{X}_1}\boldsymbol{X}_2\boldsymbol{\beta}_2, \end{aligned}$$

and

$$Var(\boldsymbol{P}_{\boldsymbol{X}_1}\boldsymbol{Y}) = \boldsymbol{P}_{\boldsymbol{X}_1}Var(\boldsymbol{Y})(\boldsymbol{P}_{\boldsymbol{X}_1})' = \boldsymbol{P}_{\boldsymbol{X}_1}(\sigma^2\boldsymbol{V})(\boldsymbol{P}_{\boldsymbol{X}_1})' = \sigma^2\boldsymbol{P}_{\boldsymbol{X}_1}\boldsymbol{V}(\boldsymbol{P}_{\boldsymbol{X}_1})'.$$

Results

Data simulations

Simulated data were created to understand the impact of misspecifying a spatial model. To better understand such an impace, we simulated data having the form

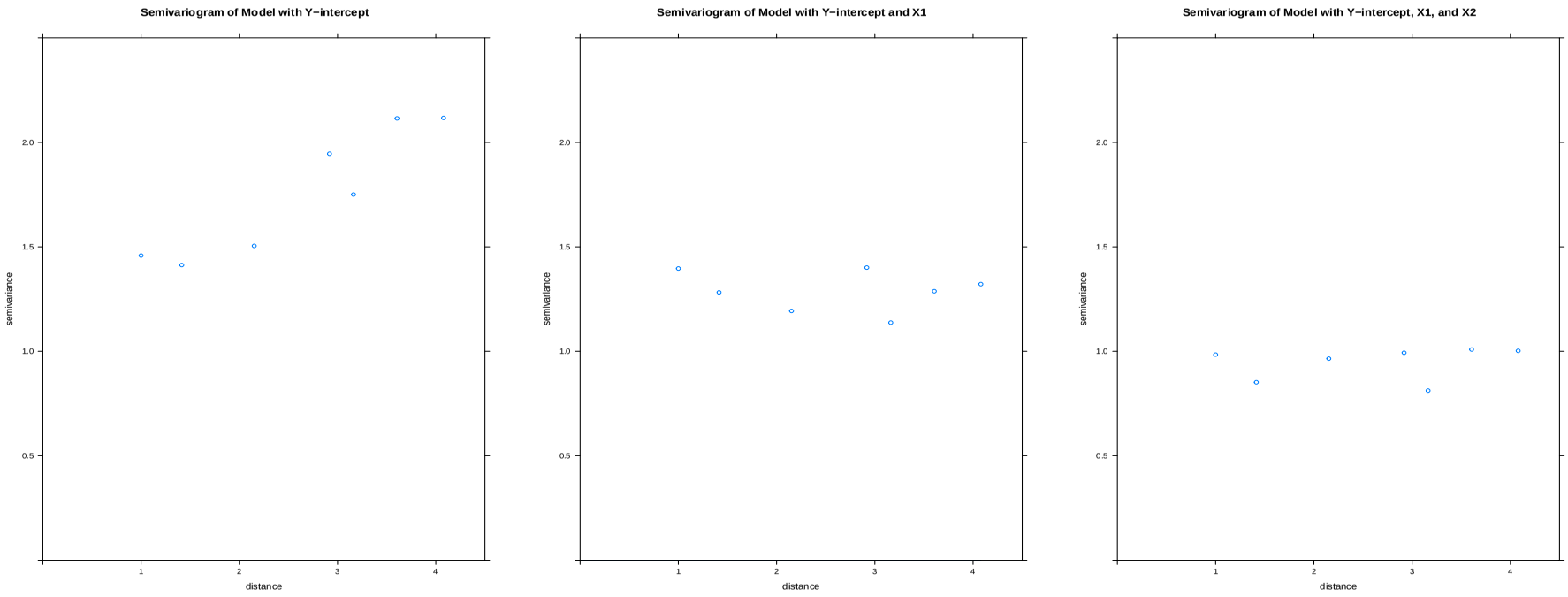
$$\boldsymbol{Y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon} \tag{4}$$

in the following manner. We created a data frame representing a 10×10 grid of 100 (i, j) point pairs where $i, j = 1, \ldots, 10$, respectively. Using this data set, we made a distance matrix of size 100×100 that represents the distance from each point $\{(i, j); i, j = 1, \ldots, 10\}$ to all point combinations $\{(k, l); k, l = 1, \ldots, 10\}$ in the 10×10 grid. Next, we simulated two covariates $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, where $\boldsymbol{X}_1$ has a forced column trend with values from 1 to 10, and $\boldsymbol{X}_2$ has multivariate normal distribution with parameters $\boldsymbol{\mu}_{100} = \boldsymbol{0}_{100}$ and $\boldsymbol{\Sigma}_{100\times100}$, which has an exponential correlation structure of the form for distance, h,

$$\gamma(h) = \begin{cases} 0, & h = 0 \\ c_0 + c_e[1 - \exp(-h/a_e)], & h > 0; \end{cases} \tag{5}$$

we simulated $\boldsymbol{\Sigma}$ where $c_0 = 0$, $c_e = 1$, $a_e = 10$. The error term, $\boldsymbol{\epsilon}$, was simulated as having a random normal $N(0, 1)$ structure. We tested different values of β_1 and β_2 to see if β_1 was biased for different values of β_2 and determined that values of 0.5 for both led to the most accurate prediction of expected values. The data were next simulated 1500 times so that the semivariograms could be fitted with the average values of the parameters from the resulting model runs. We considered three postulated linear models: one that contained only y-intercept parameter, γ_0; one that contained γ_0 and a parameter γ_1 associated with variable X_1; and one that considered the full postulated model such that the expectation of the response variable equals $\gamma_0 + \gamma_1 X_1 + \gamma_2 X_2$. After, we examined the semivariograms of the residuals from each of the models (Figure 1 (a)-(c)). We found that, as the covariates were respectively added to the model, the points on the semivariogram decreased in terms of semivariance towards values of 1.

Further, properly considering the full model resulted in a semivariogram whose points converged to a semivariance value of 1 (Figure 1(c)).



(a) (b) (c)

Figure 1: Semivariogram of residuals from various models considered on simulated data stemming from the model, $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$, where $\epsilon \sim N(0,1)$, X_1 column trend, $X_2 \sim MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}_{100} = \mathbf{0}_{100}$ is a vector of size 100, and $\boldsymbol{\Sigma}_{100\times100}$ is a 100×100 matrix that has the exponential correlation structure defined in Equation (5). Postulated models are (a) $Y = \gamma_0 + \epsilon^*$; (b) $Y = \gamma_0 + \gamma_1 X_1 + \epsilon^*$; (c) $Y = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_2 + \epsilon^*$.

Real data: Binary Steve

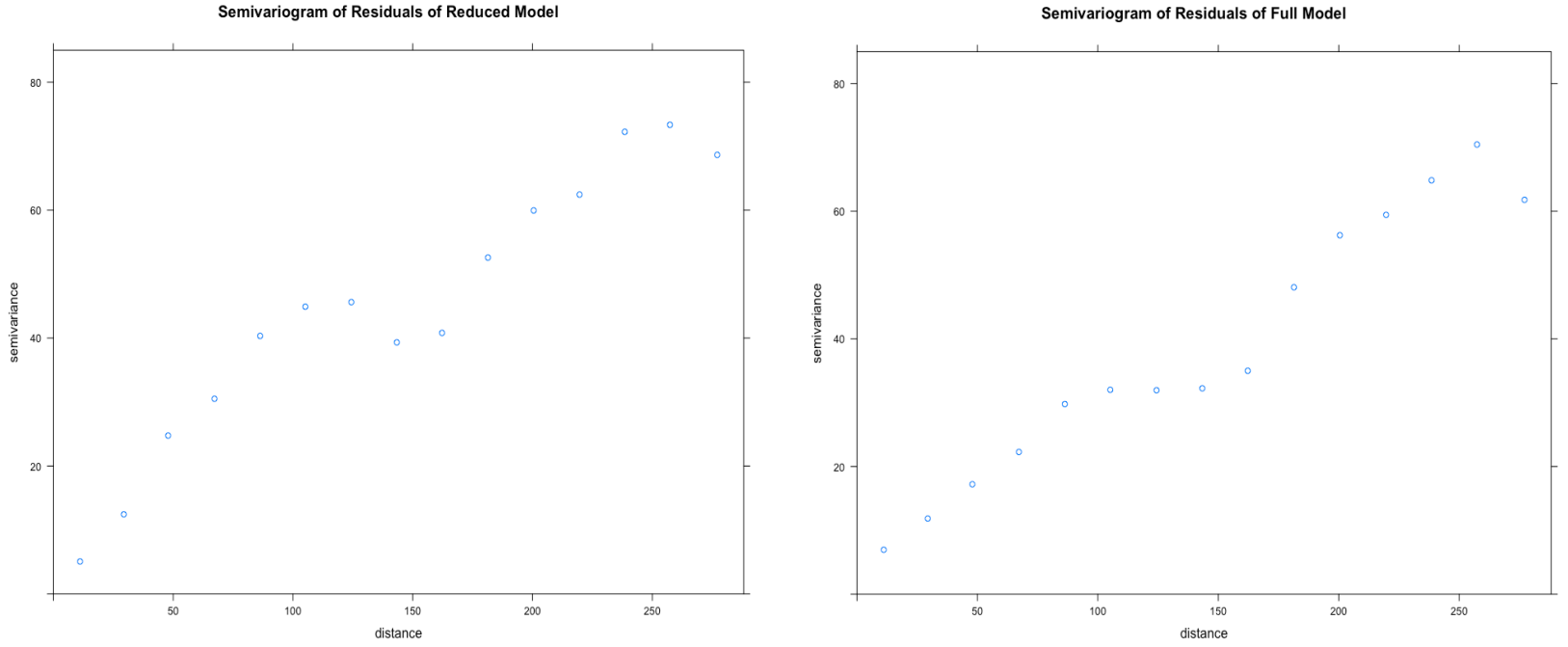
We used the Binary Steve dataset to see if the results found in the simulation were replicated in a real data example. This is an ecological dataset collected in the Negev Desert in Israel with variables related to the borrows which are the residences of the species called isopods. Some burrows can last through a generation of the isopods, while others cannot (Banerjee et al., 2003). We created a linear regression model to explain burrow survival (Y) with various variables, including slope, shrub density, rock density, and dew duration. To increase the spatial correlation in this dataset, we simulated a variable with strong spatial structure and added it to the response variable, burrow survival. We meanwhile considered the reduced model that did not include the slope nor shrub density at 15m in order to illustrate the impact of misspecifying a model that doesn't account for variables demonstrating a significant spatial component. Our full model was determined via a stepwise linear regression model, namely

$$Y = \beta_0 + \beta_1 \texttt{Dew5m} + \beta_2 \texttt{Dew15m} + \beta_3 \texttt{Slope15m} + \beta_4 \texttt{Shrub5m} + \beta_5 \texttt{Shrub15m} + \beta_6 \texttt{Rock5m} + \epsilon, \quad (6)$$

where `Dew5m` and `Dew15m`, respectively, denote the time in minutes (from 8 a.m.) to evaporation if the morning dew are five and 15 meters away from the burrows, `Shrub5m` denotes the density of shrubs five meters away from the burrows, and `Rock5m` measures the density of rocks five meters away from the burrows. Thus, removing `Slope 15m` and `Shrub 15m` from the full model, we establish the postulated model,

$$Y = \gamma_0 + \gamma_1 \texttt{Dew5m} + \gamma_2 \texttt{Dew15m} + \gamma_3 \texttt{Shrub5m} + \gamma_4 \texttt{Rock5m} + \epsilon^* \quad (7)$$

and examine the resulting semivariograms associated with the residuals from the respective models; see Figure 2 (a)-(b). Consistent with our results from the simulated data analysis, model misspecification resulted in semivariograms containing larger semivariance. Including the missing covariates (i.e. going from the reduced model to the full model) produced a semivariogram whose range of point values decreased (Figure 2).



(a) (b)

Figure 2: Semivariogram of the residuals associated with the (a) reduced model, $Y = \gamma_0 + \gamma_1 \texttt{Dew5m} + \gamma_2 \texttt{Dew15m} + \gamma_3 \texttt{Shrub5m} + \gamma_4 \texttt{Rock5m} + \epsilon^*$, and (b) full model, $Y = \beta_0 + \beta_1 \texttt{Dew5m} + \beta_2 \texttt{Dew15m} + \beta_3 \texttt{Slope15m} + \beta_4 \texttt{Shrub5m} + \beta_5 \texttt{Shrub15m} + \beta_6 \texttt{Rock5m} + \epsilon$.

Discussion

Typically, in geostatistics, the spatial covariance structure is seen to make any covariates in a linear model obsolete, thus all do not need to be accounted for. We reject this notion, and instead focus on creating spatial linear models that prioritize correct outcome prediction rather than precise covariate estimates. We showed how missing covariates in spatial linear models actually induces spatial autocorrelation in the covariates. Through the use of semivariograms of the residuals, we found that in comparing reduced models to the full postulated model, the semivariance is higher in the reduced models and decreases as the model is more accurately specified. Essentially, larger semivariance is found when misspecification of the model is present. This finding was confirmed through the use of simulated data and the Binary Steve dataset.

This work focused on linear models, but future work should include extending this to generalized linear mixed models (GLMM). Other limitations to the study include the choice of correlation, and dependence of the spatial structure and its impact on the semi-variogram. We used an exponential correlation to represent the spatial structure. As is often the case for spatial data analysis, a different correlation structure will impact the results of any such study. Further, the results in the semi-variogram may depend on the strength of the spatial structure. Future work will likewise consider variations of the above choices for correlation and spatial structure to better understand model robustness.

References

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2003). *Hierarchical Modeling and Analysis for Spatial Data.* Chapman and Hall/CRC.

Cressie, N. (1993). *Spatial Statistics, Revised Edition.* John Wiley and Sons.

Diggle, P. (1990). *Time Series: A Biostatistical Introduction.* Oxford University Press.

Draper, N. and Smith, H. (1998). *Applied Regression Analysis, Third Edition*. John Wiley and Sons.

Gryparis, A., Paciorek, C. J., Zeka, A., Schwartz, J., and Coull, B. A. (2009). Measurement error caused by spatial misalignment in environmental epidemiology. *Biostatistics*, 10(2):258–274.

Hodges, J. S. and Reich, B. J. (2010). Adding spatially-correlated errors can mess up the fixed effect you love. *The American Statistician*, 64(4):325–334.

Lopiano, K. K., Young, L. J., and Gotway, C. A. (2011). A comparison of errors in variables methods for use in regression models with spatially misaligned data. *Statistical methods in medical research*, 20:29–47.

Lopiano, K. K., Young, L. J., and Gotway, C. A. (2013). Estimated generalized least squares in spatially misaligned regression models with berkson error. *Biostatistics*, 14(4):737–751.

Lopiano, K. K., Young, L. J., and Gotway, C. A. (2014). A pseudo-penalized quasi-likelihood approach to the spatial misalignment problem with non-normal data. *Biometrics*, 70(3):648–660.

Ripley, B. (2005). *Spatial Statistics*. Wiley Series in Probability and Statistics. Wiley.

Szpiro, A. A., Sheppard, L., and Lumley, T. (2011). Efficient measurement error correction with spatially misaligned data. *Biostatistics*, 12(4):610–623.

Tobler, W. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46:234–240.