BIOMETRICS 72, 678–686
September 2016

DOI: 10.1111/biom.12474

Spatial Regression with Covariate Measurement Error: A Semiparametric Approach

Md Hamidul Huque,[1,*] Howard D. Bondell,[2] Raymond J. Carroll,[3] and Louise M. Ryan[1]

[1]School of Mathematical and Physical Sciences, University of Technology Sydney, Australia, 15 Broadway, Ultimo, NSW, 2007, Australia.
[2]Department of Statistics, North Carolina State University, 2311 Stinson Drive, Campus Box 8203, Raleigh, NC 27695-8203 USA
[3]Department of Statistics, 447 Blocker Building, Texas A&M University College Station, TX 77843-3143 USA
[*]*email:* MdHamidul.Huque@student.uts.edu.au

SUMMARY. Spatial data have become increasingly common in epidemiology and public health research thanks to advances in GIS (Geographic Information Systems) technology. In health research, for example, it is common for epidemiologists to incorporate geographically indexed data into their studies. In practice, however, the spatially defined covariates are often measured with error. Naive estimators of regression coefficients are attenuated if measurement error is ignored. Moreover, the classical measurement error theory is inapplicable in the context of spatial modeling because of the presence of spatial correlation among the observations. We propose a semiparametric regression approach to obtain bias-corrected estimates of regression parameters and derive their large sample properties. We evaluate the performance of the proposed method through simulation studies and illustrate using data on Ischemic Heart Disease (IHD). Both simulation and practical application demonstrate that the proposed method can be effective in practice.

KEY WORDS: Bivariate smoothing; Geoadditive models; Penalized least squares; Regression calibration; Socio-economic indexes for areas; Spatial linear model.

1. Introduction

With the rapid growth of Geographic Information Systems (GIS), it is now common for epidemiologists to incorporate spatially indexed data into their studies (Elliott and Wartenberg, 2004). Analysis of such data, however, is complicated by correlations among neighboring observations. Although there are well-known statistical methods to adjust for spatial correlation, relatively little has been done in the context of spatial modeling when the covariate of interest is measured with error. In the case study that motivates this study, Australian researchers explored the relationship between the SEIFA index (an area-based measure of socio-economic status produced by the Australian Bureau of Statistics) and acute hospitalization for Ischemic Heart Disease (IHD) in New South Wales, Australia (Burden et al., 2005). Multivariate regression models suggest a significantly negative association between SEIFA and IHD, implying that heart disease rates increase with social disadvantages. However, the strength of association might be attenuated due to the fact that the SEIFA index is constructed using principal component analysis, therefore, is highly likely to be measured with error (Huque et al., 2014).

Many articles have appeared in the literature over the years on covariate measurement error in the context of independent data (Fuller, 1987; Carroll et al., 2006). However, relatively few have addressed the specific context of spatial modeling. Bernadinelli et al. (1997) and Xia and Carlin (1998) presented a spatio-temporal analysis of spatially correlated data with errors in covariates, in the context of disease mapping. They empirically studied several alternative measurement error models using a Gibbs algorithm. Li et al. (2009) derived asymptotic bias expressions for estimated regression coefficients in the context of a spatial linear mixed model. They showed that the regression estimates obtained from naive use of an error-prone covariate are attenuated, while variance component estimates are inflated.

Recently, Huque et al. (2014) confirmed the findings of Li et al. (2009) showing that the amount of attenuation depends on the degree of spatial correlation in both the covariate of interest and the assumed random error from the regression model and derived expressions for the bias when measurement error is ignored. They proposed two different strategies for obtaining consistent estimates: (i) correcting the estimate using an estimated attenuation factor; and (ii) using an appropriate transformation of the error-prone covariate. They showed that both bias correction methods work reasonably well, however, the standard error is underestimated in the case when measurement error variances are estimated from the data. Moreover, their approach is fully parametric. Indeed, Ruppert et al. (2009) argued that penalized splines are the most effective method for correcting the covariate measurement error in case of independent data. So it is of natural interest to extent the spatial regression model with measurement error to a semiparametric framework.

In this article, we propose a joint modeling approach to assess the relationship between a covariate with measurement error and a spatially correlated outcome in a semiparametric regression context. Our approach contrasts with what is commonly assumed in the measurement error context, namely that some form of validation data are available. Underlying our approach is the critical assumption that the true, but unobserved covariate is smooth and that any random fluctuations from this smooth surface represent measurement error. This assumption makes our model identifiable by representing the unknown true covariate with a linear combination of spline basis functions (Yu and Ruppert, 2002; Xun et al., 2013). We use penalized least squares which makes the estimation of parameters and inference straightforward. We develop asymptotic theory for the estimated parameters and provide both model-based and simulation-based standard error estimates. Our simulation results reveal that the proposed method works well in obtaining consistent estimates of the true regression coefficient in the presence of measurement error. Our approach is computationally efficient and stable and can be implemented using standard nonlinear least squares software.

The structure of the article is as follows: Section 2 describes our model formulation, estimation, and inference procedures. Section 3 presents the data-generation process and results from the simulation study. In Section 4, we present an application of the proposed method to data on Ischemic Heart Disease (IHD). We conclude with general discussion in Section 5. The Web Appendix (`http:www.tibs.org/biometrics`) gives detailed proofs, as needed.

2. Model

Suppose that X_i represents the true covariate of interest measured at geographical location, $S_i \in \mathbb{R}^2$, $i = 1, ..., n$ and suppose that X_i is related to an outcome Y_i, according to a spatial linear model:

$$Y_i = \beta_0 + \beta_1 X_i + G_1(S_i) + \epsilon_i, \tag{1}$$

where $\boldsymbol{\epsilon} = (\epsilon_1, ...\epsilon_n)^T \sim N(0, \sigma_\epsilon^2)$ and $\{G_1(S_i) : S_i \in \mathbb{R}^2\}$ is an unknown function that captures the spatial correlation, for now kept arbitrary. Further assume that ϵ_i and $G_1(S_i)$ are independent of each other and of the true covariate X_i (Cressie, 1993). In practice, the outcome might also be related to other covariates and it is straight forward to extent model (1) to include these. However, for simplicity, we only consider a single covariate in model (1).

In the presence of measurement error, measurements on the true covariate X are not observed directly, instead an error-contaminated version is available. Let W_i be the observed covariate for location $S_i \in \mathbb{R}^2$, $i = 1, ..., n$, related to the true covariate X_i according to a classical measurement error model:

$$W_i = X_i + U_i, \tag{2}$$

where $U_i \sim N(0, \sigma_u^2)$. Note that a consistent estimate of the true regression coefficient β_1 can be obtained if either the measurement error variance is known or can be estimated using a validation data set on the true covariate (X) without measurement error (Carroll et al., 2006). However, in the spatial epidemiology setting such validation data are relatively rare. We develop an alternative approach assuming that the true covariate X is smooth and can be modeled by a second smooth function, $G_2(S_i)$.

Many different choices of smoothers have been discussed in the literature, including locally weighted running line smoothers (loess), Kernel smoothers or splines (Hastie and Tibshirani, 1990). In general, techniques based on regression splines are robust in approximating the true underling smooth functions and are relatively straight forward from a computational perspective, but have rigorous mathematical properties (Ruppert et al., 2003; Wood, 2006). In this article, we also adopt such a technique, specifically, cubic thin plate splines (Wood, 2006).

Within this framework, the unknown smooth functions, $G_j(S_i)$, for $j = 1, 2$ are represented by linear combination of thin plate spline basis functions, i.e., $G_j(S_i) = B_j^{\mathrm{T}}(S_i)\theta_j$. Here, $B_1(S_i)$ and $B_2(S_i)$ are two sets of thin plate splines basis functions with dimensions $(q_1 + 3) \times 1$ and $(q_2 + 3) \times 1$, respectively, where q_1 and q_2 are the corresponding number of knots and θ_1 and θ_2 are vectors of corresponding basis coefficients.

Under the above specifications model (1) and (2) can be rewritten as

$$Y_i = B_2^{\mathrm{T}}(S_i)\theta_2\beta_1 + B_1^{\mathrm{T}}(S_i)\theta_1 + \epsilon_i; \tag{3}$$

$$W_i = B_2^{\mathrm{T}}(S_i)\theta_2 + U_i. \tag{4}$$

Since these equations are linear with respect to a set of unknown parameters, we use penalized least squares techniques for estimation (Yu and Ruppert, 2002; Xun et al., 2013). In this method, the data, (Y, W), are fitted to two different sets of spline basis functions $B_1(S_i)$ and $B_2(S_i)$ by least squares where parameters are estimated by minimizing the usual sum of squares plus roughness penalties. That is, we minimize

$$\boldsymbol{J}(\beta, \theta_1) = n^{-1}\textstyle\sum_{i=1}^n \{Y_i - B_2^{\mathrm{T}}(S_i)\theta_2\beta_1 - B_1^{\mathrm{T}}(S_i)\theta_1\}^2 + \delta_1\theta_1^{\mathrm{T}} D_1\theta_1; \tag{5}$$

$$\boldsymbol{J}(\theta_2) = n^{-1}\textstyle\sum_{i=1}^n \{W_i - B_2^{\mathrm{T}}(S_i)\theta_2\}^2 + \delta_2\theta_2^{\mathrm{T}} D_2\theta_2, \tag{6}$$

where the terms $\delta_1\theta_1^{\mathrm{T}} D_1\theta_1$ and $\delta_2\theta_2^{\mathrm{T}} D_2\theta_2$ are roughness penalties associated with models (3) and (4). These involve unknown regression coefficients θ_j, j = 1,2, penalty parameters δ_j and penalty matrices D_j of dimension $(q_j + 3) \times (q_j + 3)$. The penalty matrices map the spline basis functions to the data, whereas the penalty parameters control the amount of smoothing (Ruppert et al., 2003; Wood, 2006). Given knot locations $\{x_{j(i)}^* : 1, 2, ..., q_j\}$, penalty matrices have zeroes everywhere except in its lower right $q_j \times q_j$ block with $D_{j(ik)} = \|x_{j(i)}^* - x_{j(k)}^*\|^2 \log \|x_{j(i)}^* - x_{j(k)}^*\|$, for $i, k \leq q_j$.

Note that the intercept term β_0 in the model (1) is set to 0 in (3), because it is not identifiable in the presence of a nonparametric function $G_1(\cdot)$. Even so, the parameters of these models are not completely identifiable without some additional assumptions outlined in the next section.

2.1. *Identifiability*

From the above models (3) and (4), it is evident that if $B_1(\cdot) \equiv B_2(\cdot)$, then these models are not identifiable because in this case (3) becomes

$$Y_i = B_2^{\mathrm{T}}(S_i)(\theta_2\beta_1 + \theta_1) + \epsilon_i.$$

Thus, we can identify only θ_2 and $\theta_2\beta_1 + \theta_1$, and cannot separate out β_1 and θ_1. To make these models identifiable, we assume that the asymptotic variability, Λ_1 and Λ_2 of two sets of basis functions $B_1(.)$ and $B_2(.)$, respectively, is different. The asymptotic variability Λ_j for j = 1, 2, are the limiting values of Λ_{nj}, where

$$\Lambda_{nj} = \{n^{-1}\textstyle\sum_{i=1}^{n} B_j(S_i)B_j^{\mathrm{T}}(S_i) - \delta_j D_j\}^{-1}. \quad (7)$$

In practice, this requirement can be easily achieved by ensuring that the numbers of knots q_1 and q_2 are unequal.

2.2. *Parameter Estimation*

In addition to the assumption that $\Lambda_1 \neq \Lambda_2$, we also assume that the penalty parameters are small relative to the sample size, i.e., $n^{1/2}\delta_j \to 0$ for $j = 1, 2$. This means that with large sample sizes, the estimated regression coefficients obtained using penalized least squares will be close to the OLS estimates. Thus, minimizing the penalized sum of squares (6) and solving for θ_2, we have

$$\widehat{\theta}_2 = \Lambda_{n2} n^{-1}\textstyle\sum_{i=1}^{n} B_2(S_i)W_i, \quad (8)$$

where Λ_{n2} is defined in equation (7). A detailed derivation of $\widehat{\theta}_2$ along with it's asymptotic distribution is given in Web Appendix A.1. Similarly, we can estimate θ_1 and β_1 by minimizing the corresponding penalized sum of squares (5). This yields (see the Web Appendix A.2 and A.3)

$$\widehat{\theta}_1 = V_n - R_n\widehat{\theta}_2\widehat{\beta}_1 \quad (9)$$

$$\widehat{\beta}_1 = \frac{n^{-1}\sum_{i=1}^{n} Y_i\{B_2^{\mathrm{T}}(S_i) - B_1^{\mathrm{T}}(S_i)R_n\}\widehat{\theta}_2}{\widehat{\theta}_2^{\mathrm{T}}(\mathcal{T}_n - R_n^{\mathrm{T}}\Lambda_{n1}^{-1}R_n)\widehat{\theta}_2}, \quad (10)$$

where

$$\begin{aligned} V_n &= \Lambda_{n1} n^{-1}\textstyle\sum_{i=1}^{n} B_1(S_i)Y_i; \\ R_n &= \Lambda_{n1} n^{-1}\textstyle\sum_{i=1}^{n} B_1(S_i)B_2^{\mathrm{T}}(S_i); \\ \mathcal{T}_n &= n^{-1}\textstyle\sum_{i=1}^{n} B_2(S_i)B_2^{\mathrm{T}}(S_i). \end{aligned}$$

Although the above estimator of β_1 was estimated using pseudolikelihood, it is consistent for β_1. In the next section, we will establish the asymptotic properties of the estimator.

2.3. *Asymptotic Theory*

Asymptotic theory for the estimators $\widehat{\beta}_1$ is based on treating the spatial locations $S_i \in \mathbb{R}^2$ as fixed constants. Following Yu and Ruppert (2002), if $\delta_j \to 0$ as $n \to \infty$, then the bias also tends to 0 and consistency can be established. Asymptotic normality is established by the following theorem, whose proof appears in Web Appendix A.4.

THEOREM 1. *Assume that the smoothing parameters are small relative to the sample size, i.e.,* $n^{1/2}\delta_j \to 0$, *and the spatial correlation* $G_1(.)$ *and unknown covariate* X *are correctly represented by a finite number of splines basis functions. Then, the estimate of* β_1 *is consistent and asymptotically normally distributed with*

$$n^{1/2}\left(\widehat{\beta}_1 - \beta_1\right) \xrightarrow{d} N\left(0, \sigma^2\right), \quad (11)$$

where

$$\begin{aligned} \sigma^2 &= \lim_{n\to\infty} n^{-1}\textstyle\sum_{i=1}^{n}(\sigma_\epsilon^2\mathcal{G}_{ni}^2 + \sigma_u^2\mathcal{H}_{ni}^2); \\ \mathcal{G}_{ni} &= \mathcal{D}_{ni}(\theta_2^{\mathrm{T}}\mathcal{C}_n\theta_2)^{-1}; \\ \mathcal{H}_{ni} &= \mathcal{A}_n\Lambda_{n2}B_2(S_i)(\theta_2^{\mathrm{T}}\mathcal{C}_n\theta_2)^{-1} - \mathcal{A}_n\theta_2\mathcal{F}_{ni}(\theta_2^{\mathrm{T}}\mathcal{C}_n\theta_2)^{-2}; \\ \mathcal{A}_n &= n^{-1}\textstyle\sum_{i=1}^{n}\{G_2(S_i)\beta_1 + G_1(S_i)\}\{B_2(S_i) - R_n^{\mathrm{T}}B_1(S_i)\}^{\mathrm{T}}; \\ \mathcal{C}_n &= \mathcal{T}_n - R_n^{\mathrm{T}}\Lambda_{n1}^{-1}R_n; \\ \mathcal{D}_{ni} &= \{B_2(S_i) - R_n^{\mathrm{T}}B_1(S_i)\}^{\mathrm{T}}\theta_2; \\ \mathcal{F}_{ni} &= \theta_2^{\mathrm{T}}\mathcal{C}_2\Lambda_{n2}B_2(S_i) + B_2^{\mathrm{T}}(S_i)\Lambda_{n2}\mathcal{C}_n\theta_2. \\ R_n &= \Lambda_{n1}n^{-1}\textstyle\sum_{i=1}^{n} B_1(S_i)B_2^{\mathrm{T}}(S_i); \\ \mathcal{T}_n &= n^{-1}\textstyle\sum_{i=1}^{n} B_2(S_i)B_2^{\mathrm{T}}(S_i). \end{aligned} \quad (12)$$

Using this asymptotic expression, we can also estimate the standard error of the estimated regression coefficient $\widehat{\beta}_1$. The next section will discuss two such options.

2.4. *Estimating the Standard Error of* $\widehat{\boldsymbol{\beta}}_1$

We first consider a model-based estimate of the standard error of $\widehat{\beta}_1$ using the asymptotic theorem discussed in the previous section and then suggest a more robust estimate of standard error using simulation.

2.4.1. Model-based standard error. The model-based standard errors of the estimated $\widehat{\beta}_1$ can be estimated by substituting corresponding consistent estimates of σ_ϵ^2 and σ_u^2 (defined below) into expression (12). Specifically,

$$\widehat{\sigma}_\epsilon^2 = \frac{\sum_{i=1}^{n}\{Y_i - B_2(S_i)\widehat{\theta}_2\widehat{\beta}_1 - B_1(S_i)\widehat{\theta}_1\}^2}{n - 2\mathrm{trace}\{L_1(\delta_1,\delta_2)\} + \mathrm{trace}\{L_1(\delta_1,\delta_2)L_1^{\mathrm{T}}(\delta_1,\delta_2)\}}$$

$$\widehat{\sigma}_u^2 = \frac{\sum_{i=1}^{n}\{W_i - B_2(S_i)\widehat{\theta}_2\widehat{\beta}_1\}^2}{n - 2\mathrm{trace}\{L_2(\delta_2)\} + \mathrm{trace}\{L_2(\delta_2)L_1^{\mathrm{T}}(\delta_2)\}},$$

where the denominators are the residual degrees of freedom associated with model (3) and model (4) with smoother matrices $L_1(\delta_1, \delta_2)$ and $L_2(\delta_2)$, respectively (Ruppert et al., 2003). Define $\boldsymbol{B}_j = \{B_j(S_1), ..., B_j(S_n)\}^{\mathrm{T}}$ for j = 1,2 and $\boldsymbol{D}_n = \{D_{n1}, ..., D_{nn}\}^{\mathrm{T}}$. Then, the smoother matrices have the following expressions (see Web Appendix A.5)

$$L_1(\delta_1,\delta_2) = n^{-1}\left\{\boldsymbol{D}_n\boldsymbol{D}_n^{\mathrm{T}}(\widehat{\theta}_2^{\mathrm{T}}\mathcal{C}_n\widehat{\theta}_2)^{-1} + \boldsymbol{B}_1\Lambda_{n1}\boldsymbol{B}_1^{\mathrm{T}}\right\} \quad (13)$$

$$L_2(\delta_2) = n^{-1}\boldsymbol{B}_2\Lambda_{n2}\boldsymbol{B}_2^{\mathrm{T}}. \quad (14)$$

2.4.2. Simulated standard error. From (10), the expression for $\widehat{\beta}_1$ can be written as (see the Web Appendix A.4)

$$\widehat{\beta}_1 = \frac{\mathcal{A}_n\theta_2 + n^{-1}\sum_{i=1}^{n}\{\mathcal{A}_n\Lambda_{n2}B_2(S_i)U_i + \mathcal{D}_{ni}\epsilon_i\}}{\theta_2^{\mathrm{T}}\mathcal{C}_n\theta_2 + n^{-1}\sum_{i=1}^{n}\mathcal{F}_{ni}U_i} + o_p(n^{-1/2}),$$

where ϵ_i and U_i are the random errors defined in models (1) and (2). Since these quantities are not directly observed, we can estimate the variance of $\widehat{\beta}_1$ by a residual bootstrap (Carroll et al., 2006).

Let M be a fairly large number, say 100, and for $b = 1, \ldots, M$, generate independent random samples $\epsilon_{bi} \sim \text{Normal}(0, \widehat{\sigma}_\epsilon^2)$ and $U_{bi} \sim \text{Normal}(0, \widehat{\sigma}_u^2)$ for i = 1, 2,...n. Define the b'th bootstrap estimates of β_1 as

$$\widehat{\beta}_1^b = \frac{\widehat{\mathcal{A}}_n \widehat{\theta}_2 + n^{-1}\sum_{i=1}^n \{\widehat{\mathcal{A}}_n \Lambda_{n2} B_2(S_i) U_{bi} + \widehat{\mathcal{D}}_{ni}\epsilon_{bi}\}}{\widehat{\theta}_2^{\mathrm{T}} \widehat{\mathcal{C}}_n \widehat{\theta}_2 + n^{-1}\sum_{i=1}^n \widehat{\mathcal{F}}_{ni} U_{bi}},$$

where $\widehat{\mathcal{A}}_n, \widehat{\mathcal{D}}_n, \widehat{\mathcal{C}}_n$, and $\widehat{\mathcal{F}}_{ni}$ can be estimated by substituting the appropriate quantities into expression (12). These estimated quantities preserve the underlying spatial structure. Therefore, the sample variance of $\widehat{\beta}_1^1, ..., \widehat{\beta}_1^M$ is a consistent estimate of the variance of $\widehat{\beta}_1$ (Efron and Tibshirani, 1993).

2.5. *Smoothing Parameter Selection*

Our main objective is to obtain a consistent estimate of the regression parameter β_1 such that it accounts for the measurement error in the covariate. However, selecting a suitable combination of the smoothing parameters (δ_1, δ_2) is a prerequisite to a good model fit. All discussion so far has assumed that these parameters are fixed and known.

To choose smoothing parameters that attempt to minimize the mean square error (prediction error), three common approaches have been discussed in the literature (Ruppert et al., 2003) (i) Generalized Cross Validation (GCV); (ii) Mallow's C_p; and (iii) Akaike Information Criterion (AIC). Among these methods, minimization of GCV scores is more attractive because of being invariant and computationally efficient (Wood, 2006). We use the GCV criterion to estimate the smoothing parameters (δ_1, δ_2) in a two-step procedure (Wood, 2006). We first obtain an optimum value of δ_2 by minimizing the GCV score based on model (2) and then substitute this estimated value of δ_2 into (8) to obtain an estimate of θ_2. We then use these estimates of $\widehat{\delta}_2$ and $\widehat{\theta}_2$ in (13) to obtain an expression for the smoothing matrix, $L_1(\delta_1, \widehat{\delta}_2)$. Finally, we minimize the following GCV score associated with the outcome model to get an optimum value of δ_1:

$$\text{GCV}(\delta_1) = \frac{n^{-1}\sum_{i=1}^n \{Y_i - \widehat{Y}_i\}^2}{\{1 - n^{-1}\text{trace}\{L_1(\delta_1, \widehat{\delta}_2)\}\}^2},$$

where L_1 is defined in Section 2.4.

3. Simulation Study

In this section, we discuss a simulation study designed to evaluate the finite sample properties of our proposed method in the presence of covariate measurement error in spatial linear regression.

3.1. *Data Generation*

We simulate n sample locations randomly within a square, where n is the sample size. Specifically, the ith random sample location S_i is generated by simulating two coordinates (e.g., latitude and longitude) from a Uniform[0,1] distribution. Given a set of simulated S_i's, the unobserved true covariate X is generated using a bivariate bump function. Specifically, the bivariate bump function is generated using the product of two univariate bump functions generated separately for each coordinate. That is, for each coordinate, k, we generate $X_{ik} = \frac{1}{1+a_{ik}} + 3e^{-50(a_{ik}-0.3)^2} + 2e^{-25(a_{ik}-0.7)^2}$, $k = 1, 2$, where a_{i1} and a_{i2} are the first and second coordinates of simulated ith sample location, respectively. The observed error-contaminated versions, W, of the true covariate are generated by adding independent Gaussian noise with varying the measurement error variance σ_U^2 as 0, 0.25, and 0.50 to X. The contour plot associated with the true and error-prone covariate is given in Figure 1.

As shown in the Figure 1, presence of measurement error adds noises to the true distribution of the smooth covariate. As a result, the underlying true covariate distribution becomes obscured for higher degrees of measurement error.

The smooth spatial surface, $G_1(S_i)$, is generated to have a normal distribution with mean 0 and variance–covariance matrix $\sigma_{G_1}^2 \boldsymbol{R}$, where $\sigma_{G_1}^2 = 0.2$ and $\boldsymbol{R}$ has an exponential correlation structure with range parameter τ_{G_1} (Pinheiro and Bates, 2000). This implies that the correlation between two observations with distance h units apart is $\exp(-h/\tau_{G_1})$. We considered three different range parameters (τ_{G_1}= 0.1, 0.3, and 0.5) resulting in minimal, moderate, and high correlation among the values of G_1's.

Outcome data, Y, were then generated according to equation (1), with intercept and slope parameters are $(\beta_0, \beta_1)^T = (1, 2)^T$ and the variance parameter for the independent residual error assumed to be 0.5. We used the *nlme* package (Pinheiro et al., 2013) in ***R*** to generate exponential spatial correlation for our simulated data and in model fitting. The ***R*** code for the simulation and implementation of the proposed method is available with this article at the *Biometrics* website on Wiley Online Library.

3.2. *Generating Bi-Variate Splines Basis Functions*

We now describe the steps used to fit our proposed semiparametric model. We generated two sets of basis functions $B_1(\cdot)$ and $B_2(\cdot)$ using bivariate thin plate spline regression basis with 125 and 150 knots for response and covariate models, respectively. We choose thin plate splines because they are not sensitive to knot locations, perform reasonably well for a basis of any given lower rank, are computationally efficient and more importantly rotationally invariant (Ruppert et al., 2003; Wood, 2006). Unequal number of knots were chosen for $B_1(.)$ and $B_2(.)$ to make the model identifiable, (see Section 2.1). The number of knots for the response model (1) was analogous to the default number of knots [max{20,min(n/4,150)}] suggested by Ruppert et al. (2003). For the covariate model (2), we increased the default number of knots by 20%. Knot positions were automatically selected using the cluster separation method *"clara"* (Kaufman and Rousseeuw, 2005) in *R* (R Core Team, 2013).

Of course, one could select the number of knots by another algorithm such as space filling algorithm (Nychka and Saltzman, 1998). However, implementation of this algorithm is computationally intensive. Nychka and Saltzman (1998,page-169) argued that the number of knots is flexible in the context of geo-spatial model and one needs to select large enough

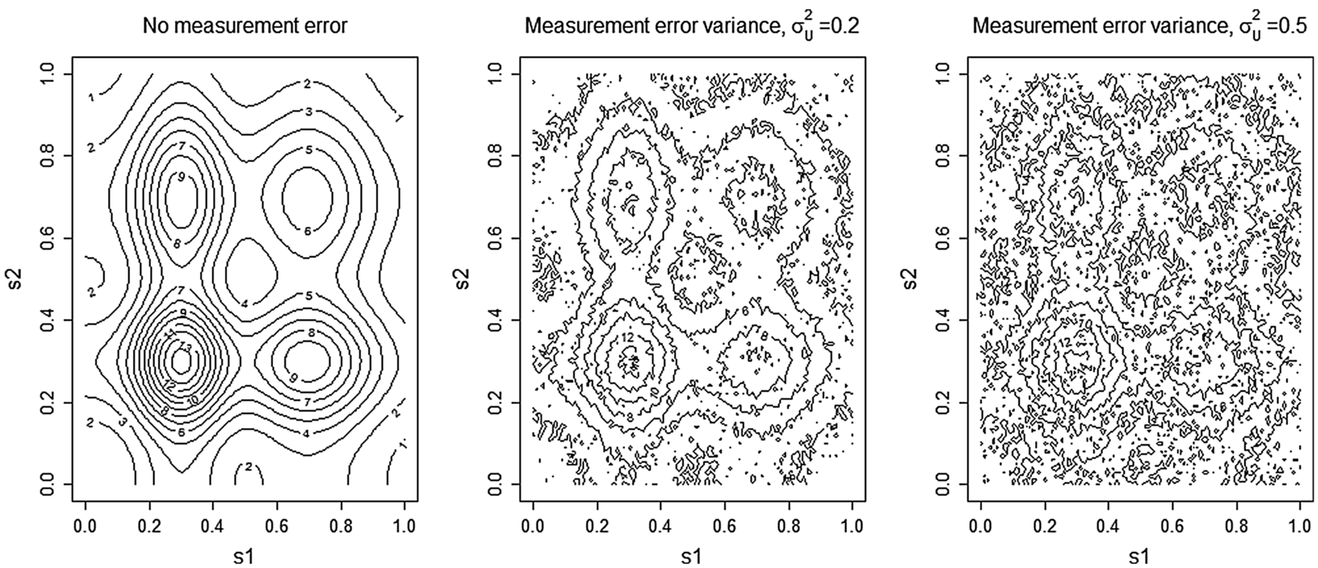

Figure 1. Contour plots of covariates (***X*** and ***W***) with different specification of measurement error variance.

knots to accurately represent the underlying function while keeping the computational burden as low as possible. Furthermore, Ruppert (2002) suggest that given the GCV criteria, the number of knots is not crucial for penalized regression splines once it reaches a certain minimum value.

3.3. *Simulation Results*

The average of estimated regression coefficients along with their estimated standard errors based on 1000 simulation runs is presented in Table 1, assuming a sample size of 500 and varying the measurement error variance σ_U^2 between 0, 0.25, and 0.50. We estimated three different standard errors of the estimated regression coefficients, including, (i) empirical standard errors obtained by taking the standard deviation of the 1000 simulated regression coefficient estimates, (ii) the average of model-based standard errors, and (iii) the average of simulated standard errors defined in Section (2.4). We considered three different range parameters ($\tau_{G_1} = 0.1$, 0.3, and 0.5) to represent minimal, moderate, and high level of spatial correlation in $G_1(S_i)$. The first column of Table 1 specifies the range parameter used in that particular simulation. The next four columns list the estimated regression coefficient using ordinary least squares (OLS), linear mixed models with spatial correlation structure (LME), generalized additive models (GAM), and our proposed method when the true covariate is measured without error. The second and thirds sets of four columns also list estimates obtained using the above four methods (OLS, LME, GAM, and proposed method) with measurement error variances 0.25 and 0.50, respectively. Except for our proposed method, all of these methods produce naive estimates of regression coefficient.

In the absence of measurement error, OLS, LME, GAM, and our method all give similar answers. As the degree of measurement error increases, OLS, LME, and GAM all exhibit bias, though the degree of bias varies. All naive standard error estimates ignoring covariate measurement error severely underestimate the empirical standard errors. In contrast, our proposed bias correction method performs well even if the degree of bias for generalized additive model with error-prone covariate varies (range: 0.99–1.32) with the strength of the spatial correlation structure. Both model-based and simulation-based estimates of the standard error appear to be working well. In all cases, the average of the estimated measurement error variances is very similar to the true values (not shown in the table).

To evaluate the performance of the proposed method under small sample settings, we also conducted simulations with sample sizes of 250 and 100 assuming a measurement error variance σ_U^2 of 0.5. The results are given in Table 2.

With the size of 250 samples, our proposed method still provides reliable estimates of the true regression coefficient. However, with small sample sizes (say, $n = 100$) the variance of estimated regression coefficients tends to be slightly inflated. To explore the impact of number of knots on our proposed method, we conducted additional simulation study by varying the number of knots for covariate model as 130, 140, and 170 with measurement error 0.025, sample size of 500, and varying range parameters, where the number of knots for the residual error model was fixed as 125. The results are presented in the Web Table 1 in the Supplementary Materials available at the *Biometrics* website on Wiley Online Library. These results indicate that the proposed methods are robust for the selection of number of knots for covariates models.

4. Application

4.1. *Analysis of Ischemic Heart Disease Data*

We applied our proposed methodology to reanalyze data on Ischemic Heart Disease (IHD). One of the key objectives of the analysis is to assess the relationship between IHD rates and area level measures of socio-economic status. These data were collected from all hospitals in New South Wales, Australia between July 1, 1994 to June 30, 2002. A detailed description of the data has been given elsewhere (Burden et al., 2005). Briefly, patients who were admitted to the hospitals

Table 1

Simulation results using different combinations of range parameters and measurement error variance. Reported numbers are averaged over 1000 simulations with 500 observations per simulation.

Range*	No measurement error				Measurement error variance, $\sigma_u^2 = 0.25$				Measurement error variance, $\sigma_u^2 = 0.5$			
(τ_{G_1})	OLS	LME	GAM	Proposed	OLS	LME	GAM	Proposed	OLS	LME	GAM	Proposed
					Estimated coefficient							
0.1	2.001	2.001	2.002	1.991	1.928	1.439	1.332	2.066	1.858	1.274	0.986	2.034
0.3	1.999	1.999	1.999	1.988	1.927	1.574	1.327	2.096	1.858	1.343	0.987	2.036
0.5	2.001	2.001	2.001	1.991	1.926	1.599	1.312	2.064	1.857	1.389	0.999	2.035
					Empirical standard error							
0.1	0.029	0.028	0.040	0.029	0.032	0.609	0.216	0.056	0.035	0.751	0.216	0.069
0.3	0.035	0.030	0.030	0.032	0.036	0.554	0.211	0.045	0.038	0.730	0.219	0.058
0.5	0.031	0.027	0.026	0.029	0.035	0.543	0.223	0.051	0.039	0.712	0.210	0.052
					Average of estimated standard errors							
0.1	0.014	0.021	0.030	0.015	0.022	0.040	0.058	0.051	0.027	0.037	0.057	0.053
0.3	0.014	0.020	0.026	0.014	0.022	0.035	0.057	0.041	0.027	0.035	0.056	0.052
0.5	0.014	0.018	0.023	0.014	0.022	0.034	0.057	0.049	0.026	0.034	0.056	0.051
					Average of simulated standard errors							
0.1	—	—	—	0.015	—	—	—	0.050	—	—	—	0.068
0.3	—	—	—	0.014	—	—	—	0.041	—	—	—	0.052
0.5	—	—	—	0.014	—	—	—	0.049	—	—	—	0.051

τ_{G_1}: values of the range parameter following exponential correlation in $G_1(s_i)$.

Table 2

Simulation results using different combinations of range parameters and sample sizes. Reported numbers are averaged over 1000 simulations with measurement error variance 0.5.

Range*	Sample size 250				Sample Size 100			
(τ_{G_1})	OLS	LME	GAM	Proposed	OLS	LME	GAM	Proposed
				Estimated coefficient				
0.1	1.860	1.511	0.976	1.952	1.859	1.831	1.037	1.947
0.3	1.861	1.495	0.975	1.951	1.859	1.824	1.045	1.948
0.5	1.860	1.522	0.980	1.950	1.860	1.831	1.036	1.949
				Empirical standard error				
0.1	0.045	0.536	0.217	0.046	0.066	0.088	0.344	0.069
0.3	0.047	0.541	0.207	0.048	0.067	0.099	0.349	0.072
0.5	0.046	0.530	0.209	0.046	0.066	0.095	0.342	0.068
				Average of estimated standard errors				
0.1	0.038	0.051	0.083	0.046	0.061	0.064	0.132	0.099
0.3	0.038	0.051	0.081	0.045	0.060	0.064	0.130	0.099
0.5	0.037	0.050	0.081	0.045	0.060	0.063	0.130	0.098
				Average of simulated standard errors				
0.1	—	—	—	0.046	—	—	—	0.101
0.3	—	—	—	0.046	—	—	—	0.101
0.5	—	—	—	0.045	—	—	—	0.099

τ_{G_1}: values of the range parameter following exponential correlation in $G_1(s_i)$.

Table 3
Analysis of Ischemic Heart Disease Data under different specification of measurement error

	Estimates for SEIFA Model based		Simulated
Methods	$\hat{\boldsymbol{\beta}}$	se($\hat{\boldsymbol{\beta}}$)	se($\hat{\boldsymbol{\beta}}$)
Ordinary Least Squares (OLS)	−0.062	0.014	—
Generalized Additive Model (GAM)	−0.145	0.014	—
Proposed semiparametric approach	−0.273	0.045	0.045
Huque et al. (2014) approach			
Method I: method of moments	−0.377	0.041	—
Method II: transformation of covariate	−0.278	0.015	—

via the emergency room and discharged with IHD were defined as acute IHD cases. Data also includes patient age, gender, and geographic location reported via postcode of residence. Data from 579 postcodes were included in the analysis. IHD event data were linked with the Census data which contains age and gender-specific population counts. SEIFA (Socio-Economic Indexes For Areas) scores and centroid coordinates (latitude and longitude) for each postcode were obtained from Australian Bureau of Statistics. We calculated age-sex adjusted standardized incidence ratios (SIR) by dividing the observed number of IHD cases by the age-sex adjusted expected number of IHD cases (Breslow and Day, 1987).

The results of our analysis are given in Table 3.

The naive analysis ignoring spatial correlation and measurement error suggests a significant protective effect associated with higher SEIFA values ($\hat{\boldsymbol{\beta}}_{\mathrm{OLS}} = -0.062$, $\mathrm{SE} = 0.014$). Similarly, analysis via a Generalized Additive model ignoring measurement error but accounting for spatial correlation also suggests that the effect is very strong ($\widehat{\beta}_{\mathrm{GAM}} = -0.145, \mathrm{SE} = 0.014$). Our proposed semiparametric approach that accounts for measurement error in the covariates results in an estimated slope parameter β_1 of -0.273 with measurement error variance estimated as 0.52. We choose 145 knots to represent the spatial correlation in the outcome model and 180 knots to represent the covariate model. The model- and simulation-based standard errors were estimated as 0.045 and 0.045, respectively. Thus, accounting for the measurement error in the covariate reflects a high magnitude of protective effect of higher SEIFA scores on IHD rates, compared with naive analysis.

5. Discussion

In this article, we develop a semiparametric framework to obtain a consistent estimate of the true regression coefficients when covariates are measured with error in spatial regression modeling settings. Asymptotic theory establishes that our approach yields consistent, asymptotically normal estimates for the regression coefficient. The theory provides both model-based and simulation-based standard error estimates. Our empirical simulation results confirm that ignoring measurement error and conducting naive analysis using both generalized additive model and linear mixed model attenuates the estimated regression coefficient toward the null hypothesis of no effect. Our results also confirm the results of Huque et al. (2014) that the degree of measurement error bias depends on the assumed correlation structure. It is interesting that the bias appears to be least with OLS. This is likely because the covariate spatial structure and residual spatial structure compete to explain the variability in the response (Waller and Gotway, 2004). Our proposed semiparametric bias correction method performs very well and provides comparable estimates of the regression parameters to the parametric methods described by Huque et al. (2014) when applied to Ischemic Heart Disease (IHD) data. Our approach is computationally efficient and stable because it involves direct estimation using least squares and can be implemented using standard nonlinear least squares software.

Although Huque et al. (2014) and Li et al. (2009) reported similar results for the bias associated with covariate measurement error in spatial regression settings, their approaches require correct specification of the true covariate measurement error variance. In addition, Huque et al. (2014) reported under estimation of standard error when measurement error variances are estimated from the data. In contrast, our approach is robust because it neither assumes that the covariate measurement error is known nor depends on any particular kind of spatial correlation structure. Our method is analogous to the popular regression calibration method where we estimate the true underlying covariate following smoothing assumption and replace the error-prone covariate with this estimate in the outcome model.

Measurement error theory makes it very clear that without some kind of information regarding the magnitude of measurement error, models will not be identifiable. Broadly speaking there are two possibilities: (i) measurement error variance is known or can be estimated using some form of validation data; (ii) assumptions are made regarding the nature of the measurement error process. By assuming that the true unobserved covariate is smooth, our article is using the second approach. Because our approach is assumption based and not an empirical measurement error adjustment, our solution will not be robust to this particular assumption. Nevertheless, because we use a semiparametric approach to quantifying the spatial correlation in our regression model, our approach should be more robust than parametric alternatives, such as those proposed by Huque et al. (2014). In practice, there will often be situations where it makes sense that spatially defined covariates are smooth. Air pollution epidemiology might be a good example. In general, however, we recommend that our proposed method be used in the spirit of sensitivity analysis to assess the impact of measurement error.

One of the additional assumptions required by our approach is that the basis functions for the covariate and the spatial residual term are unequal. In practice, this can be achieved through ensuring more knots for the basis function representing covariate than the spatial residuals. This ensures estimation of variability in covariate in a smaller scale than the residual error. In many spatial epidemiology contexts, measurement error becomes an increasing concern at small scales because of limitations in measurement resources. As a result, the covariate measurement bias reduction relies in estimating variability in covariate at scale smaller than the residual error (Paciorek, 2010).

In our simulation, we have considered only a single covariate measured with error in a spatial linear mixed model with Gaussian error. It would be of interest to explore the effect of covariate measurement error in the presence of multiple covariates and also omitted covariates. Future work should also consider extensions of our formulation to the setting of spatial-generalized linear mixed model with non-Gaussian outcomes. However, such explorations are beyond the scope of this present article.

Our heart disease example demonstrated a substantial increase in the rates of IHD as the level of SEIFA measured at the postcode level decreased, with the magnitude of the effect increasing after adjustment for measurement error. Our results are consistent with broader literature suggesting a relationship between low socio-economic status and adverse health outcomes (see systematic review by Pickett and Pearl, 2001).

Because the SEIFA Index is measured at a group level, it is tempting to think that Berkson measurement error theory should be in operation. However, this argument does not apply since we are considering measurement error in a group-level covariate applied at a group-level analysis. It is also important to note that our results can only be interpreted at a group level. Interpretation at the individual level may result in ecological bias (Sheppard, 2003). While it might be ideal to use individual-level data, in many research areas, group-level data are the only available source for analysis. Air pollution epidemiology provides a classic example, because individual measurements of air pollution studies are rarely collected, instead, they are estimated based on neighborhood monitoring and other sources (Sheppard et al., 2012). Consequently, air pollution exposures are typically measured with error.

In spatial data settings, for example, in environmental epidemiology, with the increasing popularity of the semiparametric/multilevel models to account for the observed data correlations, it is important that practitioners be aware of the consequences of measurement error. Furthermore, it is useful to quantify its potential effect on the estimating exposure–outcome relationship. The approach presented in this article provides one way of achieving this.

6. Supplementary Materials

Web Appendix A, referenced in Section 2, Web Table 1, referenced in Section 3.3, and a version of ***R*** codes for implementing the proposed method are available with this article at the *Biometrics* website on Wiley Online Library.

Acknowledgements

The authors thank the coeditor and the associated editor for constructive comments which led to considerable improvement of the manuscript. HDB was partially supported as a visitor at the School of Mathematical and Physical Sciences, University of Technology Sydney, and by grants NSF DMS-1308400 and NIH P01-CA142538. RC was partially supported by the National Cancer Institute grant U01-CA057030. LR and HH were supported by the University of Technology Sydney and by the ARC Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS). The authors thank the NSW Ministry of Health for making the data available.

References

Bernadinelli, L., Pascutto, C., Best, N., and Gilks, W. (1997). Disease mapping with errors in covariates. *Statistics in Medicine* **16**, 741–752.

Breslow, N. and Day, N. (1987). *Statistical Methods in Cancer Research. Volume II–The Design and Analysis of Cohort Studies*. International Agency for Research on Cancer. New York: Oxford University Press.

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

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*. Florida: Chapman and Hall/CRC.

Cressie, N. A. C. (1993). *Statistics for Spatial Data*. Wiley series in probability and mathematical statistics: Applied probability and statistics. New York: John Wiley & Sons.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Florida: Chapman & Hall.

Elliott, P. and Wartenberg, D. (2004). Spatial epidemiology: Current approaches and future challenges. *Environmental Health Perspectives* **112**, 998–1006.

Fuller, W. (1987). *Measurement Error Models*. Wiley Series in Probability and Statistics. New York: John Wiley & Sons.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Florida: Chapman and Hall/CRC.

Huque, M. H., Bondell, H. D., and Ryan, L. M. (2014). On the impact of covariate measurement error on spatial regression modelling. *Environmetrics* **25**, 560–570.

Kaufman, L. and Rousseeuw, P. (2005). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley Series in Probability and Statistics. New Jersey: John Wiley & Sons.

Li, Y., Tang, H., and Lin, X. (2009). Spatial linear mixed models with covariate measurement errors. *Statistica Sinica* **19**, 1077–1093.

Nychka, D. and Saltzman, N. (1998). Design of air-quality monitoring networks. In *Case Studies in Environmental Statistics*, D. Nychka, W. Piegorsch, and L. Cox (eds), volume 132 of *Lecture Notes in Statistics*, 51–76. Springer, U.S.A.

Paciorek, C. J. (2010). The importance of scale for spatial-confounding bias and precision of spatial regression estimators. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics* **25**, 107–125.

Pickett, K. E. and Pearl, M. (2001). Multilevel analyses of neighbourhood socioeconomic context and health outcomes: A critical review. *Journal of Epidemiology & Community Health* **55**, 111–122.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., and R Core Team (2013). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-109.

Pinheiro, J. C. and Bates, D. M. (2000). *Mixed-Effects Models in S and S-PLUS*. New York: Springer Science & Business Media.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics* **11**, 735–757.

Ruppert, D., Wand, M., and Carroll, R. J. (2009). Semiparametric regression during 2003–2007. *Electronic Journal of Statistics* **3**, 1193–1256.

Ruppert, D., Wand, P., and Carroll, R. (2003). *Semiparametric Regression*. Cambridge Series in Statistical and Probabilistic Mathematics. New York: Cambridge University Press.

Sheppard, L. (2003). Insights on bias and information in group-level studies. *Biostatistics* **4**, 265–278.

Sheppard, L., Burnett, R. T., Szpiro, A. A., Kim, S.-Y., Jerrett, M., Pope III, C. A., et al. (2012). Confounding and exposure measurement error in air pollution epidemiology. *Air Quality, Atmosphere & Health* **5**, 203–216.

Waller, L. A. and Gotway, C. A. (2004). *Applied Spatial Statistics for Public Health Data*, volume 368. New Jersey: John Wiley & Sons.

Wood, S. (2006). *Generalized Additive Models: An Introduction with R*. Florida: Chapman and Hall/CRC.

Xia, H. and Carlin, B. P. (1998). Spatio-temporal models with errors in covariates: Mapping Ohio lung cancer mortality. *Statistics in Medicine* **17**, 2025–2043.

Xun, X., Cao, J., Mallick, B. K., Maity, A., and Carroll, R. J. (2013). Parameter estimation of partial differential equation models. *Journal of the American Statistical Association* **108**, 1009–1020.

Yu, Y. and Ruppert, D. (2002). Penalized spline estimation for partially linear single-index models. *Journal of the American Statistical Association* **97**, 1042–1054.

Received May 2015. *Revised November* 2015.
Accepted November 2015.