Data fusion and spatial confounding in semiparametric methods for spatial and spatio-temporal data

by

Guilherme Vieira Nunes Ludwig

A dissertation submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
(Statistics)

at the

University of Wisconsin–Madison

2016

Date of final oral examination: July 20, 2016

The dissertation is approved by the following members of the Final Oral Committee:

- Professor Jun Zhu, Department of Statistics and Department of Entomology
- Professor Murray Clayton, Department of Statistics and Department of Plant Pathology
- Professor Ronald Gangnon, Department of Population Health Sciences and Department of Biostatistics and Medical Informatics
- Professor Bret Hanlon, Department of Statistics
- Professor Brian Yandell, Department of Statistics and Department of Horticulture

Abstract

In this dissertation, we will investigate two semiparametric regression problems in spatial and spatio-temporal statistics. We first examine semiparametric spline-based regression methods for spatial data in general, and their role as an alternative to specifying a covariance structure for the errors. We discuss the method under a scenario of spatial confounding, that is, when we expect changes in the regression coefficients estimates due to multicollinearity between spatial random effects and covariates. For these situations, including a smoother, such as a spline, changes the ordinary least squares' regression coefficients estimates in ways that are difficult to anticipate. An application to a problem in precision agriculture is described, where soybean growth is studied in a regression model based on seeding rates and environmental covariates.

The second investigation is motivated by a problem in the field of occupational hygiene, where health hazards such as noise are monitored with maps for risk assessment. Recent technological advancements allow data to be collected from a combination of static and roving sensors, but the current occupational hygiene methodology is lacking in methods to produce dynamic maps that provide a fusion of the data sources. We propose a spatio-temporal model that incorporates data from the static sensors, which captures complete temporal information but is sparse in space, with data from roving sensors, which provide a rich coverage of space but can potentially be confounded with short fluctuations of the hazard in time.

Acknowledgments

First I would like to thank Professor Jun Zhu, my advisor, for proposing the problems addressed in this work, for her helpfulness, support, knowledge and patience.

I thank the members of my committee, Professor Murray Clayton, Professor Ron Gangnon, Professor Bret Hanlon and Professor Brian Yandell, for their valuable input and advice.

For the work in Chapter 2 I would like to thank the input from Professors Chun-Shu Chen and Shawn Conley. I also thank Ethan Smidt and John Gaska for the help with the soybean data example.

The work in Chapter 3 is submitted for publication in collaboration with Professors Tingjin Chu, Haonan Wang and Kirsten Koehler, whom I would like to thank as well. The data for the motivating problem in ocupational hazard mapping was provided by Professor Koehler.

I also would like to thank the financial support from the CAPES Foundation, Brazil, and my master's degree supervisors, Professor Nancy Garcia and Professor Ronaldo Dias, from the University of Campinas.

Contents

Chapter 1

Introduction

1.1 Spatial confounding in semiparametric regression

For spatial linear regression analysis, the traditional approach is to use a parametric linear mixed-effects model such that spatial dependence is captured as a spatial random effect. A common assumption is that the spatial random effect is a Gaussian process with mean zero and a parametric covariance function. Spline surfaces can be used as an alternative approach to capture spatial variability, giving rise to a semiparametric method that does not require the specification of a parametric covariance structure (see, e.g., Stroup, 2012).

Recent advances in spatial statistics have brought to light the issue of spatial confounding, where including a random spatial effect in a linear regression model changes the estimated regression coefficients (Hodges and Reich, 2010; Paciorek, 2010; Hughes and Haran, 2013). The spline term in a semiparametric method also does impact the estimation of the regression coefficients. In Chapter 2, we will investigate such an impact in spatial linear regression with spline-based spatial effects. Statistical properties of the regression coefficient estimators are established under the model assumptions of the traditional spatial linear regression. We also develop a method to choose the tuning parameter for the smoothing splines that is tailored toward drawing inference about the regression coefficients. Further, we examine the empirical properties of the regression coefficient estimators under a scenario of spatial confounding via a simulation study. A data example in precision agriculture research regarding soybean yield in relation to field conditions and seeding rates is presented for consideration.

1.2 Spatio-temporal data fusion of static and roving sensors

Rapid technological advances have drastically improved the data collection capacity in occupational exposure assessment. However, advanced statistical methods for analyzing such data and drawing proper inference remain limited. In Chapter 3 we will (1) provide new spatio-temporal methodology that combines data from both roving and static sensors for data processing and hazard mapping across space and over time in an indoor environment, and (2) compare the new methodology with the current industry practice, demonstrate the distinct advantages of the new method and the impact it may have on occupational hazard assessment and future policy making in environmental health as well as occupational health. A novel spatio-temporal model with continuous index in both space and time is proposed, and a profile likelihood based model fitting procedure is developed that allows fusion of the two types of data. To account for potential differences between the static and roving sensors, we extend the model to have non-homogenous measurement error variances. Our methodology is applied to a case study conducted in an engine test facility and dynamic hazard maps are drawn to show features in the data that would have been missed by existing approaches, but are captured by the new method.

1.3 Main contributions

The main contributions of this thesis are (1) a description in semiparametric spatial regression of the relationship between spline and the regression coefficient estimates, how spatial confounding affects the regression coefficients estimates and how the spline tuning parameter can be used to mitigate issues related to confounding; and (2) an application-driven model for the fusion of spatio-temporal data sampled with different instruments, which allows the creation of dynamic hazard maps incorporating strengths from both types of instruments.

1.4 Organization of the dissertation

In Chapter 2, we will introduce the spline-based approach to spatial linear regression (Section 2.2), including the statistical properties of the regression coefficient estimators (Section 2.3). We will also discuss a spatial confounding scenario and its impact on regression coefficients estimates via simulation studies (Section 2.6) and an application to model soybean yield in precision agriculture (Section 2.7). The Appendix 2.A extends results from Section 2.6 to some more general cases of spatial structures.

In Chapter 3, we develop a model for fusion of static and roving sensor spatio-temporal data. The context of the application is described in Section 3.1, with the model outlined in Section 3.2 and the model properties discussed in 3.3. The analysis of the data is presented in Section 3.4. The Appendix 3.A shows a simulation study to compare the performance of the proposed model with traditional approaches.

In Chapter 4 we discuss some considerations on scalability and rank of the data fusion model in Section 4.1. A discussion of spatial confounding in the context of generalized linear models and Poisson point processes is also outlined in Section 4.2.

Chapter 2

Spatial confounding in spline-based semiparametric methods

2.1 Introduction

For linear regression of spatial data, the traditional approach is to use a parametric linear mixed-effects model such that spatial dependence is captured by a spatial random effect. Spline surfaces can be used as an alternative approach to capture spatial effects, giving rise to a semiparametric method that does not require the specification of a parametric covariance structure for the spatial random effect. The use of smoothing splines in this semiparametric method, however, impacts the estimation of the regression coefficients. This is related to spatial confounding, a phenomenon seen in spatial linear regression where the inclusion of a spatial random effect can affect the estimates of the regression coefficients. Our purpose in this chapter is to describe the statistical properties of spatial linear regression with spline-based spatial effects, and spatial confounding issues that may arise from the use of splines.

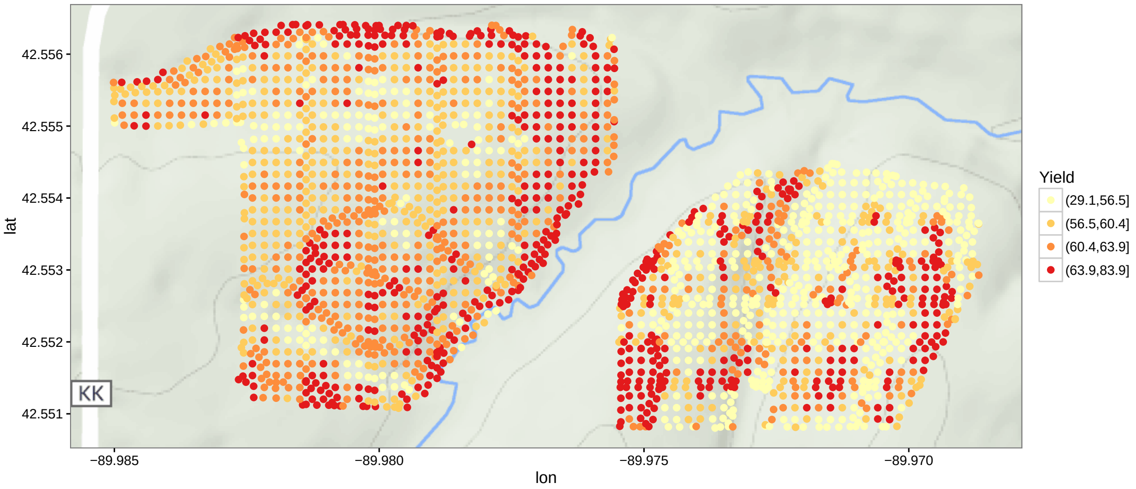

Figure 2.1: Soybean yield obtained in the experiment. The color codes are based on the quartiles from combined yield from both fields. The field to the left is called H4, and the field to the right is called Oak Creek. The region map is obtained with the `ggmap` R package (Kahle and Wickham, 2013).

A motivating example is drawn from research on precision agriculture (see, e.g., McBratney et al., 2005). Field studies were conducted in Wisconsin to evaluate the relations between soybean yield and various field conditions with the ultimate goal of improving soybean farming and management practices in the state (Smidt et al., 2016). Along with soybean yields, data were collected on soil characteristics such as soil pH, phosporus, potassium, elevation, slope and seeding rate, among others. The layout of the field, with the spatial sample locations and a color indicator of quartiles of yield, is shown in Figure 2.1. Spatial linear regression with spline-based spatial effects is an effective tool for the purpose of the study and can be readily implemented in popular statistical software such as SAS® `PROC MIXED` (SAS Institute, Inc., 2008) or R's lme4 (R Core Team, 2016; Bates et al., 2015). It is not clear, however, how the inclusion of spline-based spatial effects could affect the inference about the regression coefficients and how spatial confounding may play a role in the analysis of spatial data from such field studies.

There is well-established research on semiparametric methods with splines under specific conditions, of which the most essential are a deterministic and smooth functional component of the data model and independent errors (Ruppert et al., 2003). Rice (1986) derived the asymptotics for semiparametric unidimensional spline models, and argued that automatic smoothing (such as by cross-validation or generalized cross-validation) is invalid for the estimation of the regression coefficients, a conclusion shared by Green et al. (1985). Splines have become a popular approach to capturing spatial effects in spatial linear regression, giving rise to a semiparametric method that does not require the specification of a parametric covariance structure for the spatial random effect in a linear mixed-effects model (see, e.g., Stroup, 2012). However, the statistical properties of this semiparametric method are not well studied in theory and empirically, as well as in light of the recent findings about spatial confounding.

Different types of spatial confounding have been identified for spatial linear regression models. Hodges and Reich (2010) considered two cases: In the first case, the spatial random effects are the spatial random effects are of Scheffé style, in the sense that they are randomly drawn from a population, the draws are not of interest in themselves, and uncorrelated with the covariates. For this case the regression coefficients' estimates can be biased depending on the dependence scales of the covariates and the spatial random effect (Paciorek, 2010). A second case is when the spatial random effect is merely a formal device to implement a smoother, for example, in lieu of important but missing covariates (see, e.g., Clayton et al., 1993; Ruppert et al., 2003; Reich et al., 2006; Stroup, 2012). For this case, Hodges and Reich (2010) argued that changes in the coefficients' estimates due to the inclusion of spatial random effects can be biased, in the sense that these changes will not reflect changes in the coefficients' estimates from adding the missing covariate if it were actually available. Various insights have been provided and strategies proposed to help mitigate spatial confounding for spatial lattice data (see, e.g., Hodges and Reich, 2010; Hughes and Haran, 2013) and for geostatistical data (see, e.g., Paciorek, 2010; Hanks et al., 2015). However, the research above has focused on parametric model fitting via likelihood or Bayesian methods and, to the best of our knowledge, only Hodges (2013) has commented on the possibility of using a semiparameric approach and how spatial confounding plays out in the corresponding statistical inference about the regression coefficients.

Here we consider a common semiparametric approach to spatial linear regression and in particular the role of spline-based spatial effects. First, we establish the statistical properties of the regression coefficient estimators in terms of the impact of spline-based spatial effects on the biases and variances under the model assumptions of the traditional spatial linear regression. Next, we observe that the standard approach to tune the smoothing parameter in splines is generally tailored toward spatial interpolation (or, Kriging) (Altman, 2000; Nychka, 2000) and can be improved for the purpose of making inference about the regression coefficients. Thus, we propose a method for selecting the spline tuning parameter based on minimization of the mean squared error of the regression coefficients estimators. Further, we examine the empirical properties of the regression coefficients estimators under a scenario of spatial confounding by conducting a fairly extensive simulation study. Finally, we return to the soybean yield data example to illustrate the methods.

The remainder of this chapter is organized as follows. In Section 2.2, we present the spatial linear model for geostatistical data and the semiparametric method based on thin-plate splines for model fitting. In Section 2.3, we investigate the spline-based estimation method and the statistical properties of the resulting penalized least squares estimators in terms of biases and variances. We also develop a tuning parameter selection method that is tailored toward the inference of the regression coefficients in Section 2.4. A simulation study is conducted in Section 2.5 that empirically examines the statistical properties and the tuning parameter selection method. In Section 2.6, we study a scenario of spatial confounding in the splined-based spatial linear regression. In Section 2.7, we present the soybean data example in precision agricultural research.

2.2 Spline-based approach to spatial linear regression

2.2.1 Spatial linear regression model

Let $\mathcal{D} \subset \mathbb{R}^2$ denote a spatial domain of interest. Let n denote the sample size. The observations are denoted $Y(\boldsymbol{s}_i)$ and observed at the spatial sampling location $\boldsymbol{s}_i = (s_{i1}, s_{i2})' \in \mathcal{D}$ for $i = 1, \ldots, n$. For such a geostatistical data set, we consider a traditional spatial linear regression model:

$$Y(\boldsymbol{s}) = \beta_0 + \beta_1 x_1(\boldsymbol{s}) + \cdots + \beta_p x_p(\boldsymbol{s}) + \eta(\boldsymbol{s}) + \varepsilon(\boldsymbol{s}), \tag{2.1}$$

where $\boldsymbol{s} = (s_1, s_2)' \in \mathcal{D}$, $\beta_0, \beta_1, \ldots, \beta_p$ are the regression coefficients including the intercept and the slopes for the p covariates $x_1(\boldsymbol{s}), \ldots, x_p(\boldsymbol{s})$, $\eta(\boldsymbol{s})$ is a spatial process with mean 0 and spatial covariance function $\gamma(\boldsymbol{s}, \boldsymbol{s}^*) = \mathrm{Cov}\,(\eta(\boldsymbol{s}), \eta(\boldsymbol{s}^*))$, $\boldsymbol{s}^* \in \mathcal{D}$, $\varepsilon(\boldsymbol{s})$ is a measurement error with $\mathbb{E}(\varepsilon(\boldsymbol{s})) = 0$ and $\mathrm{Var}\,(\varepsilon(\boldsymbol{s})) = \sigma^2$.

It is common practice to assume that the measurement errors $\varepsilon(\boldsymbol{s}_i)$ are iid Gaussian random variables and are independent of the spatial random effects $\eta(\boldsymbol{s})$, while the spatial random effects $\eta(\boldsymbol{s})$ follows a Gaussian process that has mean zero and is mean squared differentiable in the sense that

$$\lim_{\|\boldsymbol{\delta}\|\to 0} \mathbb{E}\left\{\eta(\boldsymbol{s}) - \eta(\boldsymbol{s}+\boldsymbol{\delta})\right\}^2 = 0,$$

where $\|\cdot\|$ denotes the L_2 norm. Mean squared differentiability implies $\eta(\boldsymbol{s})$ has a spatial covariance function γ that is continuous. Define $\phi_k(\boldsymbol{s})$, $k = 1, 2, \ldots$ to be continuous orthonormal functions such that $\int_{\mathcal{D}} \gamma(\boldsymbol{s}, \boldsymbol{s}^*)\phi_k(\boldsymbol{s}^*)\mathrm{d}\boldsymbol{s}^* = \xi_k\phi_k(\boldsymbol{s})$, where ξ_k and $\phi_k(\boldsymbol{s})$ are referred to as the kth eigenvalue and eigenfunction, respectively. Then, the spatial covariance function γ has a spectral decomposition $\gamma(\boldsymbol{s}, \boldsymbol{s}^*) = \sum_{k=1}^{\infty} \xi_k\phi_k(\boldsymbol{s})\phi_k(\boldsymbol{s}^*)$ and $\eta(\boldsymbol{s})$ admits a Karhunen-Loève decomposition

$$\eta(\boldsymbol{s}) = \sum_{k=1}^{\infty} \xi_k^{1/2} Z_k \phi_k(\boldsymbol{s}), \tag{2.2}$$

where $\{Z_k\}_{k=1}^{\infty}$ are iid $N(0,1)$ (Loève, 1978).

The model specified in (2.1) can be viewed as a type of linear mixed-effects model, where $\eta(\boldsymbol{s})$ is a random effect to account for the spatial effects in the regression of the response on p covariates. The unknown regression coefficients are sometimes referred to as the fixed effects.

2.2.2 Thin-plate splines

Often one wishes to fit the spatial linear regression model (2.1) without prespecifying a spatial covariance function, but somehow still account for the spatial effects. In this case, smoothing splines provide a viable alternative to the traditional likelihood-based or Bayesian approach (see, e.g. Stroup, 2012). We focus on thin-plate splines (Wahba, 1990), defined to be the solution of the variational problem

$$f_\lambda = \arg\min_{g\in\mathcal{W}_2^2} \sum_{i=1}^{n} \{Y(\boldsymbol{s}_i) - g(\boldsymbol{s}_i)\}^2 + \lambda J[g], \tag{2.3}$$

where $\mathcal{W}_2^2$ is the class of functions $g : \mathbb{R}^2 \to \mathbb{R}$ that are differentiable and have bounded second derivatives, and $J[g]$ is a roughness penalty on g. Thin-plate splines are in general a combination of low-order polynomials and a linear combination of radial basis functions. For example, for a penalty based on the squared norm of the bending energy,

$$J[g] = \int_{\mathbb{R}^2} \left\{\frac{\partial^2}{\partial s_1^2} g(\boldsymbol{s})\right\}^2 + 2\left\{\frac{\partial^2}{\partial s_1 \partial s_2} g(\boldsymbol{s})\right\}^2 + \left\{\frac{\partial^2}{\partial s_2^2} g(\boldsymbol{s})\right\}^2 \mathrm{d}\boldsymbol{s}, \tag{2.4}$$

where $\boldsymbol{s} = (s_1, s_2)'$, we have a thin-plate spline that has a closed form given by $f_\lambda(\boldsymbol{s}) = \alpha_0 + \alpha_1 s_1 + \alpha_2 s_2 + \sum_{i=1}^{n} \theta_i \varphi_i(\boldsymbol{s})$, where α_0 is the intercept, α_1 and α_2 are the slopes for the two coordinates of $\boldsymbol{s}$, $\{\theta_i\}_{i=1}^{n}$ are the spline coefficients, subject to the constraints $\sum_{i=1}^{n} \theta_i = \sum_{i=1}^{n} \theta_i s_{1,i} = \sum_{i=1}^{n} \theta_i s_{2,i} = 0$, and φ_i are the collection of thin-plate spline radial basis functions given by $\varphi_i(\boldsymbol{s}) = \|\boldsymbol{s} - \boldsymbol{s}_i\|^2 \log \|\boldsymbol{s} - \boldsymbol{s}_i\|$. The thin-plate spline can also be written as $f_\lambda(\boldsymbol{s}) = \alpha_0 + \alpha_1 s_1 + \alpha_2 s_2 + \boldsymbol{\Phi}(\boldsymbol{s})\boldsymbol{\theta}$, where $\boldsymbol{\Phi}(\boldsymbol{s}) = (\varphi_1(\boldsymbol{s}), \ldots, \varphi_n(\boldsymbol{s}))$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)'$.

2.3 Model fitting and statistical properties

In this section, we will consider model fitting based on thin-plate splines and the statistical properties of the resulting regression coefficients estimates.

2.3.1 Penalized least squares estimation

To fit the spatial linear regression model (2.1), the traditional approach is to assume Gaussian distributions and base the inference on likelihood methods such as maximum likelihood or restricted maximum likelihood (see, e.g., Stroup, 2012). The estimates of the regression coefficients in (2.1) can be obtained by generalized least squares, by grouping $\eta(\boldsymbol{s})+\varepsilon(\boldsymbol{s})$ as a structured noise with a certain variance-covariance function. Alternatively, Bayesian methods can be applied for drawing inference about the model parameters, but they also require the specification of the variance-covariance function (Diggle and Ribeiro, 2007).

In an alternative semiparametric approach, the spatial linear regression model (2.1) is fitted to the spatial data with the same linear regression on the covariates but a thin-plate spline in place of the spatial random effect. Adjustments are needed, however, to ensure identifiability. Since there are two intercepts, α_0 in the thin-plate spline and β_0 in the linear regression, we set $\beta_0 \equiv 0$. Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$ denote a $n \times p$ matrix of covariates and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ a $p \times 1$ vector of regression coefficients. Further, let $\mathbf{T}$ be a $n \times 3$ matrix with each row i corresponding to $(1, s_{i,1}, s_{i,2})$, and $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2)'$. We require that $\mathbf{X}$ and $\mathbf{T}$ are not perfectly collinear with each other, and that $\mathbf{X}'\boldsymbol{\theta} = \mathbf{0}$ and $\mathbf{T}'\boldsymbol{\theta} = \mathbf{0}$. Now, let $\boldsymbol{\Phi}$ denote an $n \times n$ matrix of basis functions $(\boldsymbol{\Phi}(\boldsymbol{s}_1)', \ldots, \boldsymbol{\Phi}(\boldsymbol{s}_n)')'$, where $\boldsymbol{\Phi}(\boldsymbol{s}_i) = (\varphi_1(\boldsymbol{s}_i), \ldots, \varphi_n(\boldsymbol{s}_i))$. The regression coefficients and spline coefficients, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p), , \boldsymbol{\alpha} = (\alpha_0, \alpha_1, \alpha_2), \boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$, are estimated by minimizing

$$Q(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{T}\boldsymbol{\alpha} - \boldsymbol{\Phi}\boldsymbol{\theta}\|^2 + \lambda J[f] \tag{2.5}$$

subject to $\mathbf{X}'\boldsymbol{\theta} = \mathbf{0}$ and $\mathbf{T}'\boldsymbol{\theta} = \mathbf{0}$, where λ is a tuning parameter that controls the smoothness of the spline function and $J[f]$ is a roughness penalty. The roughness penalty for thin-plate splines can be rewritten as

$$J[f] = \boldsymbol{\theta}'\mathbf{R}\boldsymbol{\theta} \tag{2.6}$$

where $\mathbf{R}$ is an $n \times n$ matrix with entries $\mathbf{R}_{i,i'} = \varphi_i(\boldsymbol{s}_{i'})$ for $i, i' = 1, \ldots, n$ (Wahba, 1990, p. 32). For notation simplicity let $\mathbf{f} = \mathbf{T}\boldsymbol{\alpha} + \boldsymbol{\Phi}\boldsymbol{\theta}$ be a $n \times 1$ vector with the spline function evaluated at $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$. The solution of (2.5) for a fixed tuning parameter λ is

$$\begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{f}} \end{pmatrix} = \begin{pmatrix} \{\mathbf{X}'(\mathbf{I} - \mathbf{S}_\lambda)\mathbf{X}\}^{-1}\mathbf{X}'(\mathbf{I} - \mathbf{S}_\lambda)\mathbf{y} \\ \mathbf{S}_\lambda(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}), \end{pmatrix} \tag{2.7}$$

where $\mathbf{S}_\lambda$ satisfies $\mathbf{S}_\lambda(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{T}\hat{\boldsymbol{\alpha}} + \boldsymbol{\Phi}\hat{\boldsymbol{\theta}}$ and is called a smoother matrix, of dimension $n \times n$, and $\mathbf{I}$ is the $n \times n$ identity matrix. Moreover, $\hat{\boldsymbol{\alpha}} = \left\{\mathbf{T}'(\mathbf{I} - \mathbf{H}_{\boldsymbol{\theta}})\mathbf{T}\right\}^{-1}\mathbf{T}'(\mathbf{I} - \mathbf{H}_{\boldsymbol{\theta}})(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$, where $\mathbf{H}_{\boldsymbol{\theta}} = \boldsymbol{\Phi}\left\{\boldsymbol{\Phi}'\boldsymbol{\Phi} + \lambda\mathbf{R}\right\}^{-1}\boldsymbol{\Phi}'$, and $\hat{\boldsymbol{\theta}} = \left\{\boldsymbol{\Phi}'\boldsymbol{\Phi} + \lambda\mathbf{R}\right\}^{-1}\boldsymbol{\Phi}'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{T}\hat{\boldsymbol{\alpha}})$. We will refer to $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{f}}$ as the penalized least squares (PLS) estimates. From (2.7), the regression coefficients estimate $\hat{\boldsymbol{\beta}}$ can be viewed as weighted least squares, with weights based on the complement of the smoother matrix. The estimator $\hat{\mathbf{f}}$ is obtained by applying the smoother matrix to the detrended data $\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$.

2.3.2 Statistical properties of regression coefficients estimators

To study the role of spatial effects play in the spline-based estimation of the regression coefficients, first we consider the interpretation of the spatial effects $\eta(\boldsymbol{s})$ in the spatial linear regression model (2.1). One interpretation is that $\eta(\boldsymbol{s})$ represents spatial correlation, or a structured noise, among the observations. A more difficult problem is when important covariates are unavailable (or available but unaccounted for). In the first interpretation, the spatial effects $\eta(\boldsymbol{s})$ are understood to be random effects and thus, if replicates of the data were available, a different realization of $\eta(\boldsymbol{s})$ could be obtained. In the second interpretation, the spatial effects $\eta(\boldsymbol{s})$ are used to account for missing covariate information and thus, it is more suitable to treat $\eta(\boldsymbol{s})$ as a fixed realization of some spatial process. In practical applications, the role for the spatial effects in spatial linear regression may be ambiguous and often is overlooked, given that spatial data are rarely replicated. However, the two possible interpretations of $\eta(\boldsymbol{s})$ impact the notion of bias and variance of the PLS estimates $\hat{\boldsymbol{\beta}}$ for the regression coefficients. In the following, we will examine bias of the PLS estimators $\hat{\boldsymbol{\beta}}$ in terms of the unconditional and conditional expectations of $\hat{\boldsymbol{\beta}}$ for the interpretation of spatial random effects.

Let $\mathbf{y} = (y(\boldsymbol{s}_1), \ldots, y(\boldsymbol{s}_n))'$ and $\boldsymbol{\eta} = (\eta(\boldsymbol{s}_1), \ldots, \eta(\boldsymbol{s}_n))'$. If we understand $\boldsymbol{\eta}$ as a random spatial effect, $\mathbb{E}(\boldsymbol{\eta}) = \mathbf{0}$ and $\mathbb{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$. It is straightforward to see that

$$\mathbb{E}(\hat{\boldsymbol{\beta}}) = \left\{\mathbf{X}'(\mathbf{I} - \mathbf{S}_\lambda)\mathbf{X}\right\}^{-1} \mathbf{X}'(\mathbf{I} - \mathbf{S}_\lambda)\mathbb{E}(\mathbf{y}) = \boldsymbol{\beta}.$$

Further, we have $\mathbb{E}(\hat{\mathbf{f}}) = \mathbf{S}_\lambda\{\mathbb{E}(\mathbf{y}) - \mathbf{X}\mathbb{E}(\hat{\boldsymbol{\beta}})\} = \mathbf{0}$. Thus, under the spatial linear regression model (2.1), for any fixed choice of the tuning parameter λ, the PLS estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{f}}$ in (2.7) are unbiased.

If, on the other hand, we condition the PLS estimator $\hat{\boldsymbol{\beta}}$ on the spatial random effects $\boldsymbol{\eta}$, we obtain

$$\mathbb{E}(\hat{\boldsymbol{\beta}}|\boldsymbol{\eta}) = \left\{\mathbf{X}'(\mathbf{I} - \mathbf{S}_\lambda)\mathbf{X}\right\}^{-1} \mathbf{X}'(\mathbf{I} - \mathbf{S}_\lambda)\mathbb{E}(\mathbf{y}|\boldsymbol{\eta}),$$

which is $\boldsymbol{\beta} + \left\{\mathbf{X}'(\mathbf{I} - \mathbf{S}_\lambda)\mathbf{X}\right\}^{-1} \mathbf{X}'(\mathbf{I} - \mathbf{S}_\lambda)\boldsymbol{\eta}$. Due to the Karhunen-Loève expansion of $\eta(\boldsymbol{s})$ in (2.2), the conditional expectation can be written as

$$\mathbb{E}(\hat{\boldsymbol{\beta}}|\boldsymbol{\eta}) = \boldsymbol{\beta} + \sum_{k=1}^{\infty} \xi_k^{1/2} Z_k \boldsymbol{\psi}_{k,\lambda}, \tag{2.8}$$

where $\boldsymbol{\psi}_{k,\lambda} = \left\{\mathbf{X}'(\mathbf{I} - \mathbf{S}_\lambda)\mathbf{X}\right\}^{-1} \mathbf{X}'(\mathbf{I} - \mathbf{S}_\lambda)\boldsymbol{\phi}_k$ is a $p \times 1$ vector of the kth eigenvector of $\mathrm{Var}\,(\boldsymbol{\eta})$ regressed on $\mathbf{X}$, and $\boldsymbol{\phi}_k = (\phi_k(\boldsymbol{s}_1), \ldots, \phi_k(\boldsymbol{s}_n))$ is the $n \times 1$ kth eigenvector of $\mathrm{Var}\,(\boldsymbol{\eta})$. Further, we have

$$\mathbb{E}(\hat{\mathbf{f}}|\boldsymbol{\eta}) = \mathbf{S}_\lambda\{\mathbb{E}(\mathbf{y}|\boldsymbol{\eta}) - \mathbf{X}\mathbb{E}(\hat{\boldsymbol{\beta}}|\boldsymbol{\eta})\} = \mathbf{S}_\lambda[\mathbf{I} - \mathbf{X}\left\{\mathbf{X}'(\mathbf{I} - \mathbf{S}_\lambda)\mathbf{X}\right\}^{-1} \mathbf{X}'(\mathbf{I} - \mathbf{S}_\lambda)]\boldsymbol{\eta}. \tag{2.9}$$

Thus, under the spatial linear regression model (2.1), for any fixed choice of the tuning parameter λ, and conditionally on the spatial random effects $\boldsymbol{\eta}$, the PLS estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{f}}$ in (2.7) are biased. The conditional bias in $\hat{\boldsymbol{\beta}}$ can be seen as a deviation from $\boldsymbol{\beta}$ caused by the correlation of the eigenvectors $\boldsymbol{\phi}_k$ of $\text{Var}\,(\boldsymbol{\eta})$ with the covariates. Thus, the bias is present as long as $\boldsymbol{\phi}_k$ and the columns of $\mathbf{X}$ display any collinearity, and independence of $\mathbf{X}$ and $\boldsymbol{\eta}$ is not sufficient for $\hat{\boldsymbol{\beta}}$ to be conditionally unbiased.

Next, we will study the variance of the PLS estimators $\hat{\boldsymbol{\beta}}$, as the inference about the regression coefficients $\boldsymbol{\beta}$ is of primary interest. Let $\mathcal{P}_{\mathbf{X}} = \{\mathbf{X}'(\mathbf{I}-\mathbf{S}_\lambda)\mathbf{X}\}^{-1}\mathbf{X}'(\mathbf{I}-\mathbf{S}_\lambda)$. Under the spatial linear regression model (2.1) and due to (2.2), we have $\text{Var}\,(\mathbf{y}) = \sigma^2\mathbf{I} + \sum_{k=1}^{\infty}\xi_k\boldsymbol{\phi}_k\boldsymbol{\phi}_k'$. Thus,

$$\begin{aligned}\text{Var}\,(\hat{\boldsymbol{\beta}}) &= \mathcal{P}_{\mathbf{X}}\text{Var}\,(\mathbf{y})\mathcal{P}_{\mathbf{X}}' = \sigma^2\mathcal{P}_{\mathbf{X}}\mathcal{P}_{\mathbf{X}}' + \mathcal{P}_{\mathbf{X}}\left(\sum_{k=1}^{\infty}\xi_k\boldsymbol{\phi}_k\boldsymbol{\phi}_k'\right)\mathcal{P}_{\mathbf{X}}' \\ &= \sigma^2\{\mathbf{X}'(\mathbf{I}-\mathbf{S}_\lambda)\mathbf{X}\}^{-1}\mathbf{X}'(\mathbf{I}-\mathbf{S}_\lambda)^2\mathbf{X}\{\mathbf{X}'(\mathbf{I}-\mathbf{S}_\lambda)\mathbf{X}\}^{-1} + \sum_{k=1}^{\infty}\xi_k\boldsymbol{\psi}_{k,\lambda}\boldsymbol{\psi}_{k,\lambda}'.\end{aligned} \tag{2.10}$$

Consider the decomposition of $\text{Var}\,(\hat{\boldsymbol{\beta}})$ into two additive components:

$$\text{Var}\,(\hat{\boldsymbol{\beta}}) = \text{Var}_{\mathbf{X}}(\hat{\boldsymbol{\beta}}) + \text{Var}_{\boldsymbol{\eta}}(\hat{\boldsymbol{\beta}}),$$

where the first component is the variability of $\hat{\boldsymbol{\beta}}$ attributed to the columns of $\mathbf{X}$,

$$\text{Var}_{\mathbf{X}}(\hat{\boldsymbol{\beta}}) = \mathbb{E}\left\{\text{Var}\,(\hat{\boldsymbol{\beta}}|\boldsymbol{\eta})\right\} = \sigma^2\{\mathbf{X}'(\mathbf{I}-\mathbf{S}_\lambda)\mathbf{X}\}^{-1}\mathbf{X}'(\mathbf{I}-\mathbf{S}_\lambda)^2\mathbf{X}\{\mathbf{X}'(\mathbf{I}-\mathbf{S}_\lambda)\mathbf{X}\}^{-1}, \tag{2.11}$$

and the second component is the variability of $\hat{\boldsymbol{\beta}}$ attributed to the spatial random effects $\boldsymbol{\eta}$,

$$\text{Var}_{\boldsymbol{\eta}}(\hat{\boldsymbol{\beta}}) = \text{Var}\left\{\mathbb{E}(\hat{\boldsymbol{\beta}}|\boldsymbol{\eta})\right\} = \text{Var}\left(\sum_{k=1}^{\infty}\xi_k^{1/2}Z_k\boldsymbol{\psi}_{k,\lambda}\right) = \sum_{k=1}^{\infty}\xi_k\mathbb{E}(Z_k^2)\boldsymbol{\psi}_{k,\lambda}\boldsymbol{\psi}_{k,\lambda}'. \tag{2.12}$$

We note the connection of (2.8) and (2.12), which shows that, conditional on the realization of $\boldsymbol{\eta}$, the bias of $\hat{\boldsymbol{\beta}}$ is a function of $\boldsymbol{\psi}_{k,\lambda}$, and unconditionally the effect of $\boldsymbol{\psi}_{k,\lambda}$ is present as an inflating factor in $\mathrm{Var}\,(\hat{\boldsymbol{\beta}})$. As such, we will now investigate the impact of selection of λ on the bias and variance of $\hat{\boldsymbol{\beta}}$.

2.4 Selection of penalty tuning parameter

In this section we will discuss the behavior of the matrix $\mathbf{S}_\lambda$ as a function of λ, the so called smoothness tuning parameter in spline literature, and in particular how it affects $\boldsymbol{\psi}_{k,\lambda}$, a key component in describing the relation between $\boldsymbol{\eta}$ and the regression coefficients $\hat{\boldsymbol{\beta}}$. Then we will propose an approach to selection of the tuning parameter λ when the researcher interest lies primarily in obtaining the estimates of $\boldsymbol{\beta}$.

2.4.1 The matrix $\mathbf{S}_\lambda$ and the smoothness property

Since $\mathbf{S}_\lambda$ is symmetric, it has a spectral decomposition $\mathbf{S}_\lambda = \mathbf{QLQ}' = \sum_{i=1}^n \ell_i \mathbf{q}_i \mathbf{q}_i'$, where $\mathbf{L}$ is a diagonal matrix of eigenvalues $\ell_1(\lambda) \geq \ldots \geq \ell_n(\lambda)$ and $\mathbf{Q}$ is an orthogonal matrix of eigenvectors $\mathbf{q}_1, \ldots, \mathbf{q}_n$. Further, the eigenvalues of $\mathbf{S}_\lambda$ are in $[0, 1]$ (Hastie and Tibshirani, 1990; Wahba, 1990), with the first 3 eigenvalues equal to 1 for the eigenvectors corresponding to $\mathbf{T}$ and eigenvalues shrinking toward 0 as a function of λ for the eigenvectors corresponding to $\boldsymbol{\Phi}$. The behavior of the matrix $\mathbf{S}_\lambda$ for $\mathbf{S}_\lambda(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$ is therefore, according to Hastie and Tibshirani (1990), of shrinking by ℓ_i the projection of $\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ along $\mathbf{q}_i$. In the limiting case of $\lambda = 0$, the smoother matrix interpolates the data, i.e., $\mathbf{S}_0(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$. On the other hand, for $\lambda \to \infty$, the limiting estimate is equivalent to the regression of the data on the columns of $\mathbf{T}$, or in notation $\mathbf{S}_\lambda \xrightarrow{\lambda\to\infty} \mathbf{T}(\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}' = \mathbf{H}_\mathbf{T}$.

The behaviour of $\mathcal{P}_\mathbf{X}$ can be described based on $\mathbf{S}_\lambda$. The case $\lambda \to \infty$ is simple, since $\text{rank}(\mathbf{H}_\mathbf{T}) = \text{rank}(\mathbf{T}) = 3$ and $\mathbf{H}_\mathbf{T}$ is idempotent. This implies that $\boldsymbol{\psi}_{k,\lambda} \xrightarrow{\lambda\to\infty}$ $\left\{\mathbf{X}'(\mathbf{I}-\mathbf{H}_\mathbf{T})\mathbf{X}\right\}^{-1}\mathbf{X}'(\mathbf{I}-\mathbf{H}_\mathbf{T})\boldsymbol{\phi}_k$ and $\text{Var}_\mathbf{X}(\hat{\boldsymbol{\beta}}) \xrightarrow{\lambda\to\infty}= \sigma^2\left\{\mathbf{X}'(\mathbf{I}-\mathbf{H}_\mathbf{T})\mathbf{X}\right\}^{-1}$. The results shown are, in a certain way, similar to introducing extra explanatory variables (in this case, $\mathbf{T}$) to the regression matrix $\mathbf{X}$ (see, e.g., Seber and Lee, 2003, p. 54). The case when $\lambda = 0$ is difficult, however, since for a generic $n \times 1$ vector $\mathbf{z}$, $\mathbf{S}_0\mathbf{z} = \mathbf{I}\mathbf{z} = \mathbf{z}$ and therefore $(\mathbf{I}-\mathbf{S}_0)\mathbf{z} = \mathbf{0}$. Moreover, $\lim_{\lambda\to 0}\left\{\mathbf{X}'(\mathbf{I}-\mathbf{S}_\lambda)\mathbf{X}\right\}^{-1}$ is undefined. In practice, however, software implementations overcome these difficulties by considering a lower rank implementation of $\mathbf{S}_\lambda$ based on its spectral decomposition, i.e., using only the first M largest eigenvalues (Wood, 2003). For example, as explained in `R`'s `mgcv` package documentation, the default action is to use the first $M = 30$ eigenvalues for two dimensional data. As a consequence, computationally the limit $\lim_{\lambda\to 0}\left\{\mathbf{X}'(\mathbf{I}-\mathbf{S}_\lambda)\mathbf{X}\right\}^{-1}$ does exist, and is equal to $\left\{\mathbf{X}'\mathbf{Q}'\mathbf{L}^*\mathbf{Q}\mathbf{X}\right\}^{-1}$, where $\mathbf{L}^*$ is a diagonal matrix with entries equal zero corresponding to the first M largest eigenvalues in $\mathbf{L}$, and $\ell_i^* = 1$ otherwise, for $i = M+1,\ldots,n$. The associated eigenvectors of $\mathbf{S}_\lambda$ are those that display the most "wiggliness" (see, e.g., Ruppert et al., 2003, p. 79, for the smoothing spline case, which is analogous). Borrowing from signal processing literature, we will hereafter refer to the eigenvectors of $\mathbf{S}_\lambda$ associated with large eigenvalues as "low frequency" components, and the eigenvectors associated with lower eigenvalues as "high frequency" (this connection is pointed out in, e.g., Wahba, 1990, p.145).

To illustrate the eigenvectors and eigenvalues of $\mathbf{S}_\lambda$, we have a diagram similar to the diagram available in page 79 of Ruppert et al. (2003), except that instead of

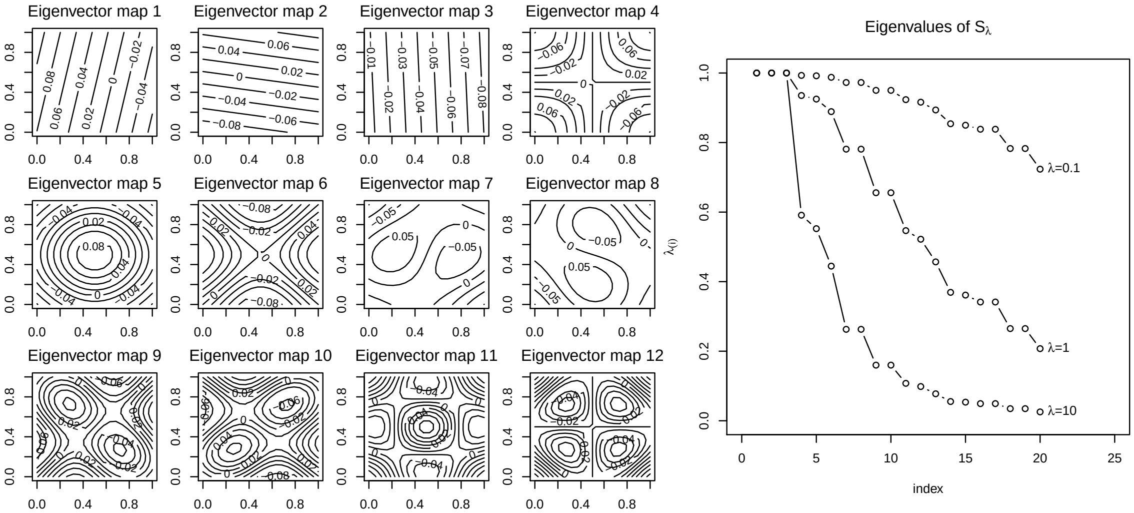

Figure 2.2: Eigenvectors and eigenvalues of the matrix $\mathbf{S}_\lambda$. The first 12 eigenvectors (left) are shown as contour plots for the corresponding spatial coordinates. The first 20 eigenvalues (right) are shown as connected points for different values of λ.

the unidimensional spline smoother, we consider the thin-plate spline instead. To draw this diagram, we picked a regular spatial grid of 400 equallly spaced points in $[0, 1] \times [0, 1]$, and evaluated the thin-plate spline smoother matrix $\mathbf{S}_\lambda$ for values of λ equal to 0.1, 1 and 10. The first 12 eigenvectors are displayed in Figure 2.2, to the left, as well as connected scatterplots showing the first 20 eigenvalues. Notice the first three eigenvectors span $\mathbf{T}$, and the corresponding first three eigenvalues are equal to one, regardless of λ. We may also observe that the eigenvalues ℓ_i decrease as a function of λ, shrinking the spatial data along the corresponding eigenvector $\mathbf{q}_i$.

A practical implication of selection of λ is in, for example, the components $\text{Var}_{\mathbf{X}}(\hat{\boldsymbol{\beta}})$ and $\text{Var}_{\boldsymbol{\eta}}(\hat{\boldsymbol{\beta}})$. When $\lambda \to \infty$, $\text{Var}_{\mathbf{X}}(\hat{\boldsymbol{\beta}})$ becomes equivalent to the variance of ordinary least squares $\hat{\boldsymbol{\beta}}$ if $\mathbf{T}$ were included as additional covariates. On the other hand, when $\lambda \to 0$, then $\text{Var}_{\mathbf{X}}(\hat{\boldsymbol{\beta}})$ behaves as if the columns of $\mathbf{X}$ were projected

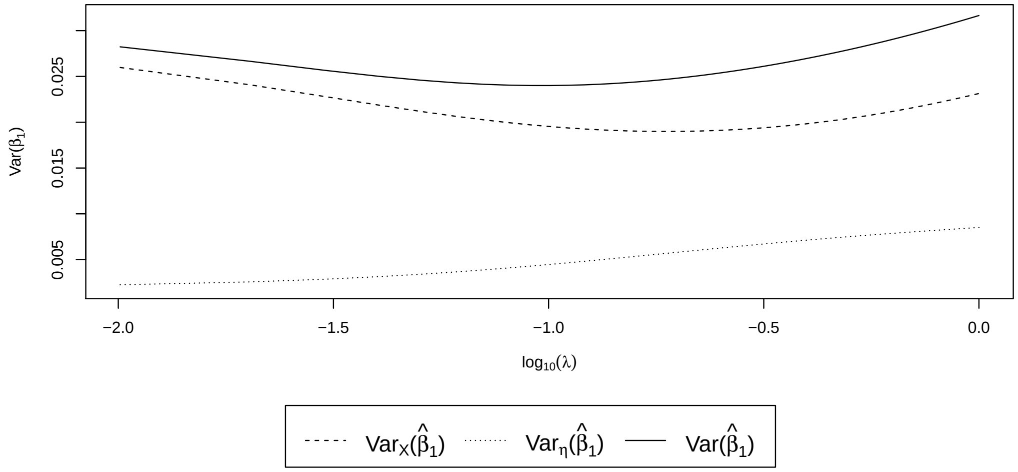

Figure 2.3: An illustration of the two components of $\mathrm{Var}\,(\hat{\boldsymbol{\beta}})$, given here by $\mathrm{Var}_{\mathbf{X}}(\hat{\beta}_1)$ (variance attributed to the design matrix $\mathbf{X}$) and $\mathrm{Var}_{\boldsymbol{\eta}}(\hat{\beta}_1)$ (variance attributed to the spatial process $\eta(\boldsymbol{s})$). $\boldsymbol{\eta}$ is independent of $\mathbf{X}$.

along the eigenvectors of $\mathbf{S}_\lambda$ corresponding to the smallest eigenvalues. The interpretation of $\mathrm{Var}_{\boldsymbol{\eta}}(\hat{\boldsymbol{\beta}})$ is similar, except that we consider the regression of $\boldsymbol{\phi}_k$ onto $\mathbf{X}$. An illustration of the balance between $\mathrm{Var}_{\mathbf{X}}(\hat{\boldsymbol{\beta}})$ and $\mathrm{Var}_{\boldsymbol{\eta}}(\hat{\boldsymbol{\beta}})$ is shown in Figure 2.3 for a single covariate $x_1(\boldsymbol{s})$ that is uncorrelated with the spatial random effect $\boldsymbol{\eta}$. There is a local minimum for $\mathrm{Var}_{\mathbf{X}}(\hat{\beta}_1)$ as a function of λ, indicating that in this case a moderate amount of smoothing on $\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ is benefitial to reduce $\mathrm{Var}_{\mathbf{X}}(\boldsymbol{\beta})$, but to minimize $\mathrm{Var}_{\boldsymbol{\eta}}(\boldsymbol{\beta})$ a very small tuning parameter λ would be preferrable, i.e. close to interpolation of the structured noise. Conditionally on $\boldsymbol{\eta}$, we understand that small λ would minimize the conditional bias.

2.4.2 An algorithm for selection of λ

A tuning parameter that minimizes some function of the variance $\mathrm{Var}\,(\hat{\boldsymbol{\beta}})$ would constitute a reasonable choice, as the primary interest lies in the estimation of the regression coefficients. In addition, the degrees of freedom of a spline, given by $\mathrm{tr}(\mathbf{S}_\lambda)$, are decreasing as a function of λ (Ruppert et al., 2003) and therefore such minimization can help to balance model complexity. We now, therefore, consider the selection of the tuning parameter λ. Selection of λ is usually based on cross validation, which approximates the prediction squared error (PSE) (Hastie and Tibshirani, 1990). In other words, let

$$\mathrm{PSE}(\lambda) = n^{-1} \sum_{i=1}^{n} \mathbb{E}\{y^*(\boldsymbol{s}_i) - \hat{y}_\lambda(\boldsymbol{s}_i)\}^2,$$

where $y^*(\boldsymbol{s}_i)$ is a new observation, $\hat{y}_\lambda(\boldsymbol{s}_i) = \mathbf{x}(\boldsymbol{s}_i)'\hat{\boldsymbol{\beta}} - \hat{f}_\lambda(\boldsymbol{s}_i)$ is the predicted value of $y^*(\boldsymbol{s}_i)$ and $\hat{f}_\lambda(\boldsymbol{s}_i)$ is the thin-plate spline evaluated at spatial sampling location $\boldsymbol{s}_,i$ for $i = 1, \ldots, n,$. Also, define the cross validation (CV) as

$$\mathrm{CV}(\lambda) = n^{-1} \sum_{i=1}^{n} \mathbb{E}\{y(\boldsymbol{s}_i) - \hat{y}_\lambda^{-i}(\boldsymbol{s}_i)\}^2,$$

where $\hat{y}_\lambda^{-i}(\boldsymbol{s}_i)$ is the predicted value from a fitted model that excludes the ith observation. We have $\mathbb{E}\{\mathrm{CV}(\lambda)\} \approx \mathrm{PSE}(\lambda)$. A related well-known criterion, with lower computational cost, is the generalized cross validation (GCV), given by

$$\mathrm{GCV}(\lambda) = n^{-1} \sum_{i=1}^{n} \left\{ \frac{y(\boldsymbol{s}_i) - \hat{y}_\lambda(\boldsymbol{s}_i)}{1 - (p + \mathrm{tr}(\mathbf{S}_\lambda))/n} \right\}^2,$$

where $\mathrm{tr}(\mathbf{S}_\lambda)$ is also commonly called the degrees of freedom of the smoother (Ruppert et al., 2003). The λ^* that minimizes $\mathrm{CV}(\lambda)$, or $\mathrm{GCV}(\lambda)$, is not necessarily the best when the primary interest is in the regression coefficients that relate the response variable to covariates.

Here we develop a different approach to select the tuning parameter λ. It follows from (2.10) that the mean squared error (MSE) of $\hat{\boldsymbol{\beta}}$ is given by

$$\begin{aligned}\mathrm{MSE}(\hat{\boldsymbol{\beta}}) &= \mathrm{tr}(\mathrm{Var}\,(\hat{\boldsymbol{\beta}})) \\ &= \sigma^2 \mathrm{tr}\left(\mathbf{X}'(\mathbf{I}-\mathbf{S}_\lambda)^2\mathbf{X}\{\mathbf{X}'(\mathbf{I}-\mathbf{S}_\lambda)\mathbf{X}\}^{-2}\right) + \sum_{k=1}^{\infty} \xi_k \|\boldsymbol{\psi}_{k,\lambda}\|^2,\end{aligned} \tag{2.13}$$

which is a function of the tuning parameter λ. When the focus is on the regression coefficients and not on prediction, a natural way to choose the tuning parameter is to minimize the MSE of $\hat{\boldsymbol{\beta}}$. That is, let

$$\lambda_{\mathrm{opt}} = \arg\min_{\lambda \geq 0} \mathrm{tr}(\mathrm{Var}\,(\hat{\boldsymbol{\beta}})).$$

While the first term on the right-hand-side of (2.13) is known except for σ^2, the second term can be a challenge to approximate. We propose a plug-in estimate based on the following argument: By the Karhunen-Loève decompostion (2.2), we have

$$\begin{aligned}\|\eta(\boldsymbol{s})\|^2 &= \int_D \left\{\sum_{k=1}^{\infty} \xi_k^{1/2} Z_k \phi_k(\boldsymbol{s})\right\}^2 \mathrm{d}\boldsymbol{s} \\ &= \int_D \left\{\sum_{k=1}^{\infty} \xi_k Z_k^2 \phi_k^2(\boldsymbol{s}) + 2\sum_{k=1}^{\infty}\sum_{j>k} \xi_k^{1/2}\xi_j^{1/2} Z_k Z_k \phi_k(\boldsymbol{s})\phi_j(\boldsymbol{s})\right\} \mathrm{d}\boldsymbol{s} \\ &= \sum_{k=1}^{\infty} \xi_k Z_k^2 \int_D \phi_k^2(\boldsymbol{s})\mathrm{d}\boldsymbol{s} + 2\sum_{k=1}^{\infty}\sum_{j>k} \xi_k^{1/2}\xi_j^{1/2} Z_k Z_k \int_D \phi_k(\boldsymbol{s})\phi_j(\boldsymbol{s})\mathrm{d}\boldsymbol{s}.\end{aligned}$$

Since $\{\phi_k(\boldsymbol{s})\}_{k=1}^{\infty}$ are orthogonal functions, we have

$$\|\eta(\boldsymbol{s})\|^2 = \sum_{k=1}^{\infty} \xi_k Z_k^2 \|\phi_k(\boldsymbol{s})\|^2 \quad \text{and} \quad \mathbb{E}\left(\|\eta(\boldsymbol{s})\|^2\right) = \sum_{k=1}^{\infty} \xi_k \|\phi_k(\boldsymbol{s})\|^2. \tag{2.14}$$

Consequently, we have

$$\sum_{k=1}^{\infty} \xi_k \|\boldsymbol{\psi}_{k,\lambda}\|^2 = \mathbb{E}\left(\|\mathcal{P}_{\mathbf{X}}\boldsymbol{\eta}\|^2\right) \tag{2.15}$$

The right-hand-side of (2.15) and thus the second term on the right-hand-side of (2.13) can be estimated by $\|\mathcal{P}_{\mathbf{X}}\hat{\mathbf{f}}\|^2$. The criterion we propose, which aims to minimize

MSE($\hat{\boldsymbol{\beta}}$), is given by an estimated MSE,

$$\text{eMSE}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2 \text{tr}\left(\mathbf{X}'(\mathbf{I}-\mathbf{S}_\lambda)^2\mathbf{X}\{\mathbf{X}'(\mathbf{I}-\mathbf{S}_\lambda)\mathbf{X}\}^{-2}\right) + \|\mathcal{P}_{\mathbf{X}}\hat{\mathbf{f}}\|^2 \tag{2.16}$$

where $\hat{\sigma}^2 = (n - p - 2\text{tr}(\mathbf{S}_\lambda) + \text{tr}(\mathbf{S}_\lambda^2))^{-1}\sum_{i=1}^n (y_i - \hat{y}_i)^2$ is an estimate of the measurement error variance (Ruppert et al., 2003, p. 83). We will refer to PLS estimates of $\boldsymbol{\beta}$ obtained with the tuning parameter selected by minimizing the estimated MSE (PLS-MSE). Since there are no replicates of $\hat{\mathbf{f}}$, $\mathbb{E}\left(\|\mathcal{P}_{\mathbf{X}}\boldsymbol{\eta}\|^2\right)$ is estimated by one realization of $\|\mathcal{P}_{\mathbf{X}}\hat{\mathbf{f}}\|^2$.

2.5 A simulation study of the statistical properties of PLS

2.5.1 Simulation setup

We first conducted a simulation study to evaluate the statistical properties of the PLS estimates of the regression coefficient developed in Section 2.3, as well as to evaluate the performance of the selection procedure for the tuning parameter developed in Section 2.4.

The spatial domain is the unit square $[0,1]^2$. A total of $n = 50$ locations were selected randomly over the unit square and were used as the sampling sites $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$. The responses $Y(\boldsymbol{s}_i)$ for $i = 1, \ldots, n$ were simulated from the spatial linear regression model (2.1) with $p = 2$:

$$Y(\boldsymbol{s}) = \beta_0 + \beta_1 x_1(\boldsymbol{s}) + \beta_2 x_2(\boldsymbol{s}) + \eta(\boldsymbol{s}) + \varepsilon(\boldsymbol{s}),$$

where the spatial random effects $\eta(\boldsymbol{s})$ follow a Gaussian process with mean zero and a Matérn covariance function

$$\gamma(d;\rho,\kappa,\sigma_\eta^2) = \frac{\sigma_\eta^2}{2^{\kappa-1}\Gamma(\kappa)}\left(\sqrt{2\kappa}\frac{d}{\rho}\right)^\kappa K_\kappa\left(\sqrt{2\kappa}\frac{d}{\rho}\right), \tag{2.17}$$

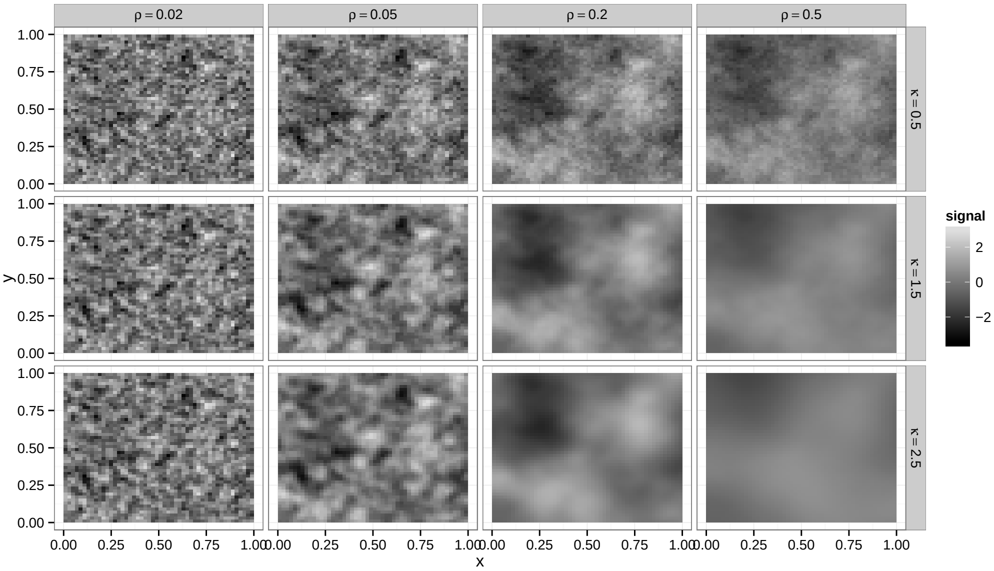

Figure 2.4: Examples of Matérn covariance processes for different parameters: each panel shows the realization of a single Gaussian process with choices of parameters $\kappa = 0.5, 1.5, 2.5,\ \rho = 0.02, 0.05, 0.2, 0.5$.

where d is a spatial lag distance and K_κ is the modified Bessel function of the second kind. The parameter κ controls the smoothness of the process, as the process with Matérn covariance is m-times differentiable for $m < \kappa$ (Stein, 1999). The parameter σ_η^2 is the variance of $\eta(\boldsymbol{s})$ and the parameter ρ controls the range of spatial dependence. Figure 2.4 shows a realization of a Gaussian process, simulated with the same random seed, but with different choices of ρ and κ for the Matérn covariance.

The two covariates, $x_1(\boldsymbol{s}_i)$ and $x_2(\boldsymbol{s}_i)$ for $i = 1, \ldots, n$, were generated as iid $N(0, 1)$ and independent of each other, before the simulations; that is, they were treated as fixed terms when $Y(\boldsymbol{s}_i)$ were generated repeatedly. The regression coefficients were set to $\beta_1 = \beta_2 = 1$. Within each simulation, the measurement errors $\varepsilon(\boldsymbol{s}_i)$ were generated as iid $N(0, 1)$ and the spatial random effects $\eta(\boldsymbol{s}_i)$ were generated from a Gaussian process with mean zero and the Matérn covariance function, for

$i = 1, \ldots, n$. The smoothness and range parameters of the Matérn covariance function (2.17) were set to $\kappa = 0.5, 1.5$ or 2.5 and $\rho = 0.02, 0.05, 0.2$ or 0.5, while the variance parameter was fixed at $\sigma_\eta^2 = 1$. The error variance is $\sigma^2 = 1$. The total number of simulations was $S = 500$ for each of the 12 combinations of κ and ρ values.

The semiparametric thin-plate method by the PLS estimation in Section 2.3 was applied with the tuning parameter selected by either GCV or eMSE (PLS-GCV or PLS-MSE, respectively). For comparison, we considered two alternative approaches. One approach ignored the spatial random effects and fitted the standard linear regression model by ordinary least squares (OLS). The other approach fitted the true spatial linear regression model (2.1), which is also a linear mixed-effects model with a spatial random effect, by restricted maximum likelihood (REML) in the `geoR` package (Ribeiro Jr. and Diggle, 2015).

2.5.2 Simulation results

The box plots of the $S = 500$ regression coefficient estimates are shown in Figure 2.5 for different κ and ρ values and fitted by four different methods (OLS, REML, PLS-GCV, and PLS-MSE). The smoothness parameter κ does not affect much the regression coefficient estimates of $\boldsymbol{\beta}$, while the range parameter ρ does. Consider a measure of relative efficiency, given by $S^{-1}\sum_{\ell=1}^{S}(\hat{\beta}_j^{(s)} - \beta_j)^2$, where $\hat{\beta}_j^{(s)}$ is the estimate of β_j in the sth simulation. The relative efficiency suggests that the OLS is the most efficient for the smallest range parameter value $\rho = 0.02$. This is not surprising since the spatial effects are weak and the spatial process $\eta(\boldsymbol{s})$ is nearly independent for this value of range parameter. When the range parameter increases to $\rho = 0.05$, the REML and spline-based methods give comparable results and both are about 5%

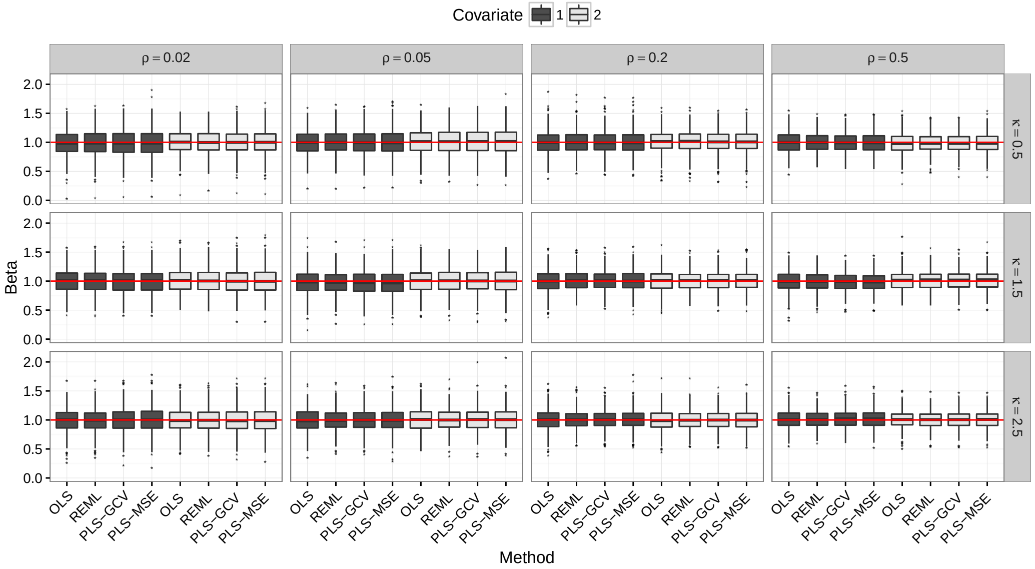

Figure 2.5: Boxplots of regression coefficients estimates ($\hat{\beta}_1$ and $\hat{\beta}_2$) under the standard linear model fitted by ordinary least squares (OLS), the linear mixed-effects model fitted by restricted maximum likelihood (REML), the spline-based approach fitted by PLS with the tuning parameter selected by generalized cross validation (PLS-GCV) or mean squared errors (PLS-MSE). The dark and light boxes are for $\hat{\beta}_1$ and $\hat{\beta}_2$, respectively. The spatial random effects in the true model has a Matérn covariance function with various smoothness parameters $\kappa = 0.5, 1.5, 2.5$ (in rows) and range parameters $\rho = 0.02, 0.05, 0.2, 0.5$ (in columns). The horizontal solid line in each subfigure indicates the true value of $\beta_1 = \beta_2 = 1$.

to 10% more efficient than OLS. For the larger range parameters ($\rho = 0.2, 0.5$), the REML and spline-based methods are about 10% to 15% more efficient than OLS. Further, the results are comparable for the two approaches, PLS-GCV and PLS-MSE, for tuning parameter selection.

We further consider the methods effectiveness to predict the response y. In Figure 2.6, we have the mean squared prediction error (MSPE), that is, the mean squared differences between the estimated $\hat{\mathbf{y}}$ and the ground truth $\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\eta}$, for each simulation step, broken down by the spatial process parameters. The results suggest that including a thin-plate spline component improves the MSPE in general, and more so, when the scale parameter ρ is larger. Overall, the semiparametric methods perform better as the scale parameter increases, yielding MSPEs that are smaller and varying less across simulated replications as ρ increases. The OLS also yields smaller MSPE as ρ increases, but the MSPE is always larger than the semiparametric methods. Furthermore, the variability of the observed MSPE for OLS increases with ρ. The performance of the REML approach is similar to OLS, in terms of MSPE. Since the data are observed at sampling locations only, the smoothness parameter κ does not seem to affect the spline model substantially. The predictions based on PLS-MSE are in general worse than the ones obtained by PLS-GCV, with a slight increase in the MSPE.

2.6 Spatial confounding simulation study

The simulation study conducted in Section 2.5 was conducted under the assumption that the covariates, x_1 and x_2, are each realizations of a process that has no spatial dependence and are independent of the spatial random effects $\eta(\boldsymbol{s})$. In this section, we explore a case of spatial confounding, and the impact on the regression coefficient estimates, via further simulations. We continue the simulation setup in Section 2.5, except that we alter the nature of the second covariate $x_2(\boldsymbol{s})$. In addition, we examine spatial confounding under different signal-to-noise ratios, as well as for nonstationary $\eta(\boldsymbol{s})$, in the Appendix 2.A.

For $i = 1, \ldots, n$, $x_2(\boldsymbol{s}_i)$ were generated from a spatial process with mean zero and a Matérn covariance function. The variance parameter is $\sigma^2_{x_2} = 1$, the smoothness parameter is $\kappa = 2.5$, and the range parameter is $\rho_{x_2} = 0.2$ or 0.5. This scenario, in

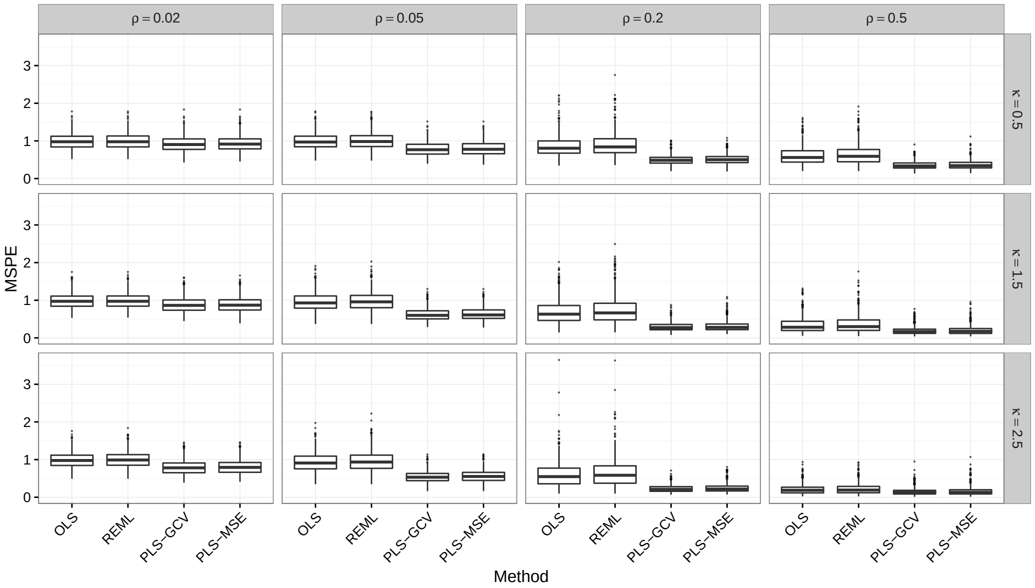

Figure 2.6: Mean squared prediction error (MSPE) for 400 simulations. OLS refers to the standard linear model, REML refers to restricted maximum likelihood estimation, PSL-GCV is the semiparametric thin-plate spline model with tuning parameter obtained by minimizing GCV and PLS-MSE is the PLS model with λ obtained by the algorithm from Section 2.4. The spatial process has a Matérn covariance function with parameters $\kappa = 0.5, 1.5, 2.5$ and $\rho = 0.02, 0.05, 0.2, 0.5$.

which covariates are independent of the spatial process but have strong dependence scales themselves, resembles the spatial confounding problem seen in Paciorek (2010) and, more closely, in Hanks et al. (2015). We interpret this scenario as close to the "Scheffé-style random effects" scenario of Hodges and Reich (2010).

We illustrate the behavior of estimates (2.7) in a similar way to Figure 2.3. Figure 2.7 has two sets of curves, one corresponding to $\hat{\beta}_1$ for reference, and the other corresponding to $\hat{\beta}_2$, the coefficient for $x_2(\boldsymbol{s})$. The variance of $\hat{\beta}_2$ is greater than that of $\hat{\beta}_1$, but as the tuning parameter λ increases, both the variability attributed to $\mathbf{X}$ and the variability attributed to $\boldsymbol{\eta}$ decrease. This can be understood since, because $x_2(\boldsymbol{s})$ is generated as a Matérn process, the resulting eigenvectors of $\mathbf{x}_2$ are collinear with

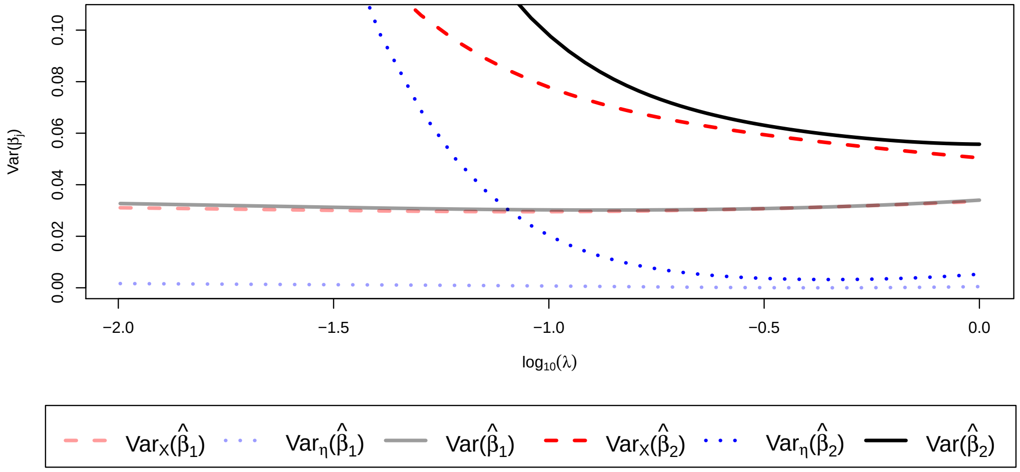

Figure 2.7: Illustration of the two components of Var $(\hat{\boldsymbol{\beta}}_j)$, $j = 1, 2$, given by Var$_{\mathbf{X}}(\hat{\boldsymbol{\beta}}_j)$ (variance attributed to the design matrix $\mathbf{X}$) and Var$_{\boldsymbol{\eta}}(\hat{\boldsymbol{\beta}}_j)$ (variance attributed to the spatial process $\eta(\boldsymbol{s})$). In this case, $\eta(\boldsymbol{s})$ is independent of $\mathbf{X}$ but $x_2(\boldsymbol{s})$ is a Gaussian random field with Matèrn covariance and parameters $\kappa = 2.5$, $\rho = 0.2$.

the eigenvectors of $\boldsymbol{\eta}$, and both have strong spatial dependence, therefore with large associated eigenvalues for the low frequency components of variation. As such, for smaller values of λ, the matrix $\mathbf{I} - \mathbf{S}_\lambda$ captures only the high frequency components of variation of $\mathbf{X}$, attributing the low frequency components of $\mathbf{x}_2$ to $\hat{\mathbf{f}}$ and modifying the $\hat{\beta}_2$ estimate in ways that are difficult to anticipate. On the other hand, for larger λ, the low frequency components of $\mathbf{X}$ are preserved and the estimate variance decreases. Note that Var$_{\boldsymbol{\eta}}(\hat{\beta}_1)$ decreases only up to a point. We understand that increasing λ too much would attribute the variation of $\boldsymbol{\eta}$ to $\mathbf{x}_2$ instead.

The box plots of the $S = 500$ regression coefficient estimates of β_1, β_2 are shown in Figure 2.8 in a similar arrangement to Figure 2.5, but with two different range parameters for $x_2(\boldsymbol{s})$, either $\rho_{x_2} = 0.2$ or 0.5. Unlike the scenario in Section 2.5 with no spatial confounding, the variability of $\hat{\beta}_2$ is higher when the smoothness parameter

κ is smaller. The estimated $\hat{\beta}_2$ with the spline-based approach are comparable to OLS and REML when the spatial random effects $\eta(\boldsymbol{s})$ have a stronger spatial dependence than $x_2(\boldsymbol{s})$ and the smoothness parameter κ is larger. Otherwise, the variability is larger than OLS and REML especially for the weaker spatial dependence of the spatial random effects ($\rho = 0.02$ and $\rho = 0.05$). The relative efficiency of the PLS-MSE method for selecting the tuning parameter is about 10% better than the PLS-GCV, providing moderate evidence that the PLS-MSE method helps to mitigate the spatial confounding in this case.

2.7 Case study: soybean yield

Field trials were conducted at multiple locations in 2013–2014 across Wisconsin to study the effectiveness of precision agriculture. Here we consider the two fields, known as H4 and Oak Creek, and shown in Figure 2.1. Fields were divided into strips of similar size and aligned so they were not parallel to either the orientation of the dominant soil types or direction of travel of the planter. The fields were planted during the month of May when conditions allowed. Three seeding rates, high, medium, and low, were randomly assigned to these strips. Variable rate technology (VRT) prescriptions were uploaded into the planter monitors and the fields were mechanically seeded by the growers in a certain row width.

Yield data were recorded by GPS-equipped yield monitoring systems on their commercial harvesters (Figure 2.1). The data were gridded to a 18 m by 18 m resolution. For this example, in addition to seeding rate, we consider a spatial regression model which also includes elevation (Figure 2.9).

First, a standard linear regression model assuming independent errors was fitted

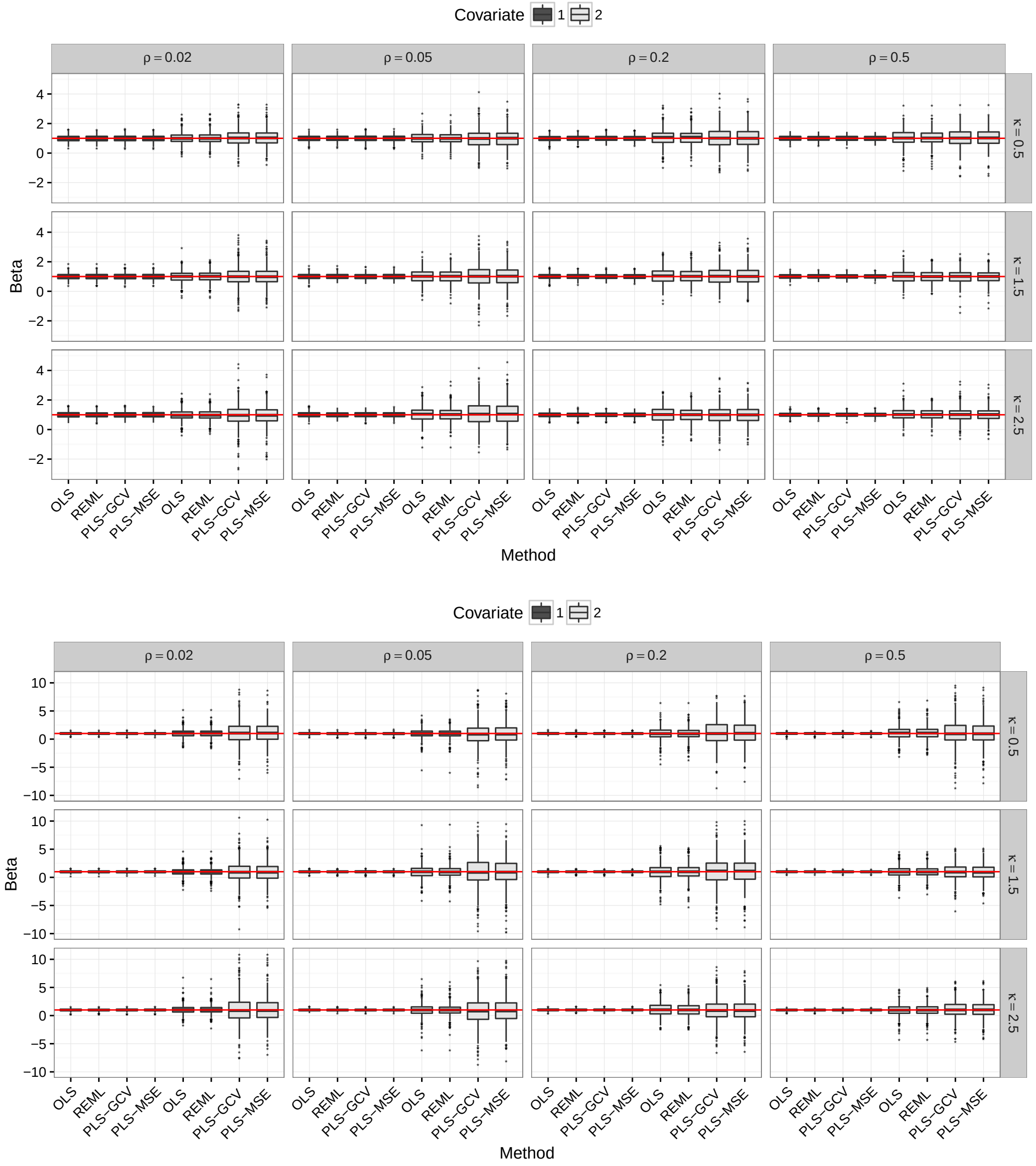

Figure 2.8: Spatial confounding scenario with the range parameters for x_2 $\rho_{x_2} = 0.2$ (top half) and $\rho_{x_2} = 0.5$ (bottom half): Boxplots of $\hat{\beta}_1, \hat{\beta}_2$ under the standard linear model fitted by ordinary least squares (OLS), the linear mixed-effects model fitted by restricted maximum likelihood (REML), the spline-based approach fitted by PLS with the tuning parameter selected by generalized cross validation (PLS-GCV) or mean squared errors (PLS-MSE). The dark and light boxes are for $\hat{\beta}_1$ and $\hat{\beta}_2$, respectively. The spatial random effects in the true model has a Matérn covariance function with various smoothness parameters $\kappa = 0.5, 1.5, 2.5$ (in rows) and range parameters $\rho = 0.02, 0.05, 0.2, 0.5$ (in columns). The horizontal solid line in each subfigure indicates the true value of $\beta_1 = \beta_2 = 1$.

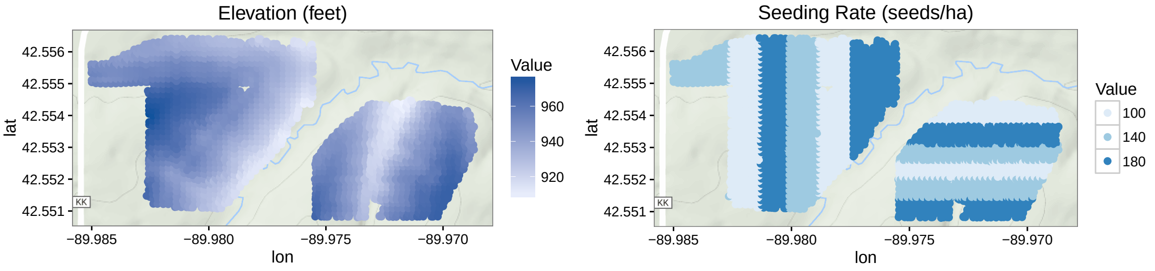

Figure 2.9: Two covariates used in the soybean yield study, elevation and seeding rate.

by the ordinary least squares to regress soybean yield as the response on all the covariates. The OLS estimates of the regression coefficients are given in Table 2.1. We remark that a visual inspection of Figure 2.1 shows that in the H4 field, the higher values of yield seem to be associated with the lower values of elevation in Figure 2.9; on the other hand, for the field Oak Creek higher values of yield are associated with higher elevations. We therefore expect a negative regression coefficient for elevation in H4, and a positive regression coefficient for elevation in Oak Creek, and those are the values shown in Table 2.1 for the OLS method. The residuals are mapped in Figure 2.10 after a thin-plate spline smoothing and show strong spatial dependence that is not fully captured by the covariates.

We believe the elevation covariate has characteristics that might cause spatial confounding. Compare the empirical semivariograms of standardized elevation against the semivariograms of the standardized residuals, obtained with the `geoR` package (Ribeiro Jr. and Diggle, 2015), shown in Figure 2.11. We notice that elevation has a stronger spatial dependence range, and is therefore akin to the spatial confounding scenario where the covariates have stronger spatial dependence than the spatial effect

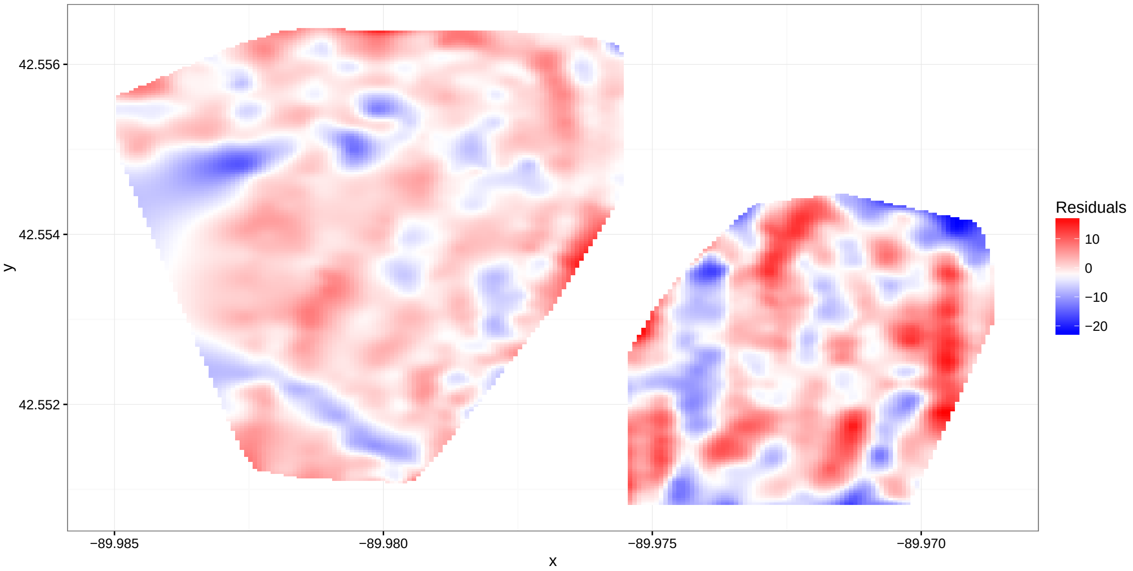

Figure 2.10: The residuals after the ordinary least squares fitting of a standard linear regression model. The residuals are smoothed by thin-plate splines, using the `Tps` function from R `fields` package (Nychka et al., 2014).

itself.

We applied the spline-based methods developed in Sections 2.3 and 2.4. The PLS estimates of the regression coeffients are shown in Table 2.1 with the tuning parameter selected by either GCV or MSE. For comparison, we also fitted a spatial linear regression model (2.1) by assuming that the spatial random effects follow a Matérn covariance function with smoothness parameter $\kappa = 2.5$, and by REML, using the `geoR` package (Ribeiro Jr. and Diggle, 2015). We observe the issue of spatial confounding by noticing how large is the change in coefficient estimates, displayed in Table 2.1, for the GCV-based PLS and REML. Our PLS-MSE approach however does produce estimates that are close to the OLS.

The results from OLS and PLS-MSE are similar to each other, and different than

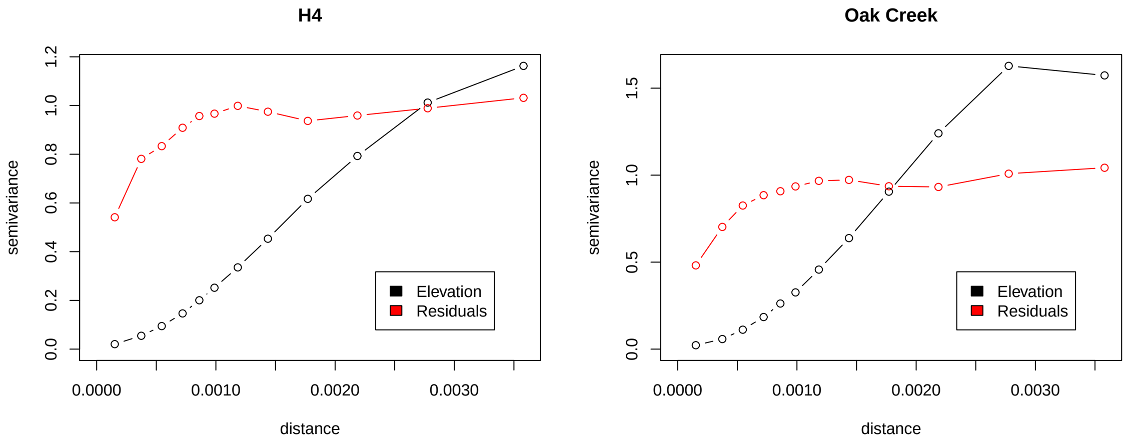

Figure 2.11: Empirical semivariograms of the standardized elevation covariate and the standardized residuals of the OLS method, for H4 and Oak Creek, respectively.

Table 2.1: Regression coefficient estimates by the ordinary least squares fit (OLS) of a standard linear regression, the spline-based approach by penalized least squares where the tuning parameters are selected by either generalized cross validation (PLS-GCV) or mean squared errors (PLS-MSE), and restricted maximum likelihood (REML) for spatial linear regression using the Matérn covariance function. H4 is the field to the left of the map, Oak Creek is the field to the right. Asterisks mark the coefficients with p-values ≤ 0.05 (using the asymptotic test for REML).

	H4				Oak Creek			
	Elevation		Seeding Rate		Elevation		Seeding Rate	
OLS	-0.109	*	0.028	*	0.125	*	0.022	*
PLS-GCV	0.288	*	0.009		0.178		0.003	
PLS-MSE	-0.109	*	0.028	*	0.125	*	0.022	*
REML	0.097	*	0.013	*	0.162	*	0.033	*

PLS-GCV and REML, whereas the latter two seem to agree with each other, to a lesser extent. Seeding Rate shares the same signs of regression coefficient estimates across all methods. In particular, higher seeding rates are associated with greater soybean yield. However, in both fields the seeding rate changes from a non-statistically significant

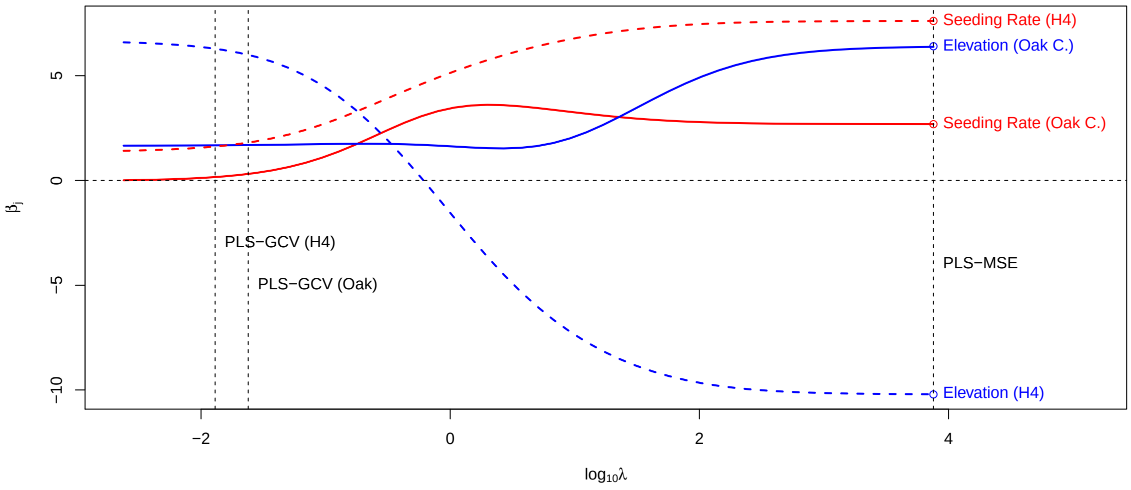

Figure 2.12: The regression coefficient estimates $\hat{\beta}_j$, divided by the corresponding $s.e.(\hat{\beta}_j)$, as a function of the tuning parameter $\log_{10}(\lambda)$. The points on the right end correspond to the ordinary least squares fit, which is approximately equivalent to the case of $\lambda \to \infty$. The vertical dashed lines indicate the optimal λ selected by either generalized cross validation (GCV) or mean squared error (MSE).

effect (using PLS-GCV) to a statistically significant positive effect (using PLS-MSE). Figure 2.12 further delineates the regression coefficient estimates from the spline-based approach as a function of the tuning parameter λ, displaying the coefficients divided by the corresponding standard error.

2.8 Conclusion and discussion

In this chapter, we have considered a semiparametric method based on smoothing splines for spatial linear regression. We have derived the statistical properties of PLS estimates of the regression coefficients under the traditional spatial linear regression model. We discussed how the spline and the spatial random effect are connected in terms of their spectral decomposition. Moreover, an alternative method to choose the tuning parameter for the thin-plate splines, tailored toward drawing inference about the regression coefficients, is proposed. Simulation studies have been conducted to evaluate the statistical properties of regression coefficient estimates.

Further, we have examined the empirical properties of the regression coefficient estimators under a scenario of spatial confounding via a simulation study. A data example in precision agricultural research regarding soybean yield in relation to field conditions is presented for illustration. We have found that the use of thin-plate splines for spatial linear regression with spatially confounded data can change coefficient estimates substantially and needs careful consideration. When covariates are uncorrelated with the spatial random effects but have smaller dependence range than the spatial effect, using splines to capture the spatial effect works well. But when the covariates have similar (or stronger) ranges of spatial dependence, the spline might compete with the covariate to capture the strongest effect.

Appendix

2.A Extensions to the spatial confounding simulation study

2.A.1 Signal-to-noise ratios

We investigated changes in σ_η^2, which lead to different signal-to-noise ratios (SNR) for the spatial process $\eta(\boldsymbol{s})$. Hereafter we will show only the estimates of β_2. Figure 2.A.1 shows the case in which there is moderate spatial confounding, and σ_η^2 is set to $0.1\sigma^2$, $0.5\sigma^2$, $2\sigma^2$, or $10\sigma^2$. We can observe that for different signal-to-noise ratios, the PLS method has a worse performance than the REML using the Matérn covariance, when the SNR is larger. Aside from the increased variability of the $\hat{\beta}_2$ estimates, the results are consistent with what was previously observed.

2.A.2 Nonstationarity: anisotropy and variable dependence range

In all the simulation studies above, the spatial random effects $\eta(\boldsymbol{s})$ are assumed to be stationary and isotropic. We now consider generating data from processes that are either anisotropic or have location variable dependence ranges based on transformations of the Matérn covariance function.

To simulate anisotropic spatial data, we let $\eta(\boldsymbol{s}^*)$ denote a spatial process with

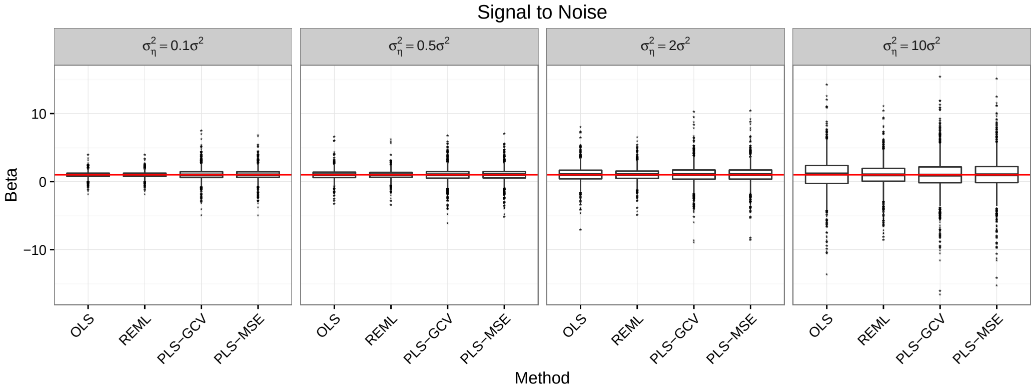

Figure 2.A.1: Comparison of $\hat{\beta}_2$ when σ^2_η is set to $0.1\sigma^2$, $0.5\sigma^2$, $2\sigma^2$, or $10\sigma^2$. In this case, $\eta(\boldsymbol{s})$ is a Matérn covariance Gaussian process with parameters $\nu = 2.5$, $\rho = 0.2$.

a Matérn covariance function (2.17) and generated the random effects at locations $\boldsymbol{s}^*_1, \ldots, \boldsymbol{s}^*_n$. Then we obtain the spatial random effects $\eta(\boldsymbol{s}_i)$ by rotating and rescaling the coordinates of the locations $\boldsymbol{s}^*_i$ with a deformation matrix such that $\boldsymbol{s}^*_i = \mathbf{P}\boldsymbol{s}_i$ for $i = 1, \ldots, n$ (see, e.g., Wackernagel, 2003). In particular, we used $\theta = \pi/4, a = 1, b = 8$ in the transformation matrix $\mathbf{P}$:

$$\mathbf{P} = \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & cos(\theta) \end{pmatrix} \begin{pmatrix} \sqrt{a} & 0 \\ 0 & \sqrt{b} \end{pmatrix}.$$

To simulate the nonstationary with variable dependence range spatial data, we again used the Matérn covariance model (2.17), but let the range parameter ρ vary in the spatial domain by employing a weighting function $w(\boldsymbol{s})$ on the coordinates. Thus, $\operatorname{Cov}(\eta(\boldsymbol{s}), \eta(\boldsymbol{s}')) = \gamma(w(\boldsymbol{s})\boldsymbol{s}, w(\boldsymbol{s}^*)\boldsymbol{s}^*)$. We chose $w(\boldsymbol{s}) = 0.3 + (1 - 0.3)/[1 + \exp\{(s_1 - 0.4)/0.05\}]$ so that the resulting spatial dependence of $\eta(\boldsymbol{s})$ transitions along the first coordinate of $\boldsymbol{s}$. Figure 2.A.2 shows the realizations of the said anisotropic and nonstationary spatial random effects for different values of the range parameter ρ.

Figure 2.A.3 provides the box plots of $S = 500$ regression coefficient estimates

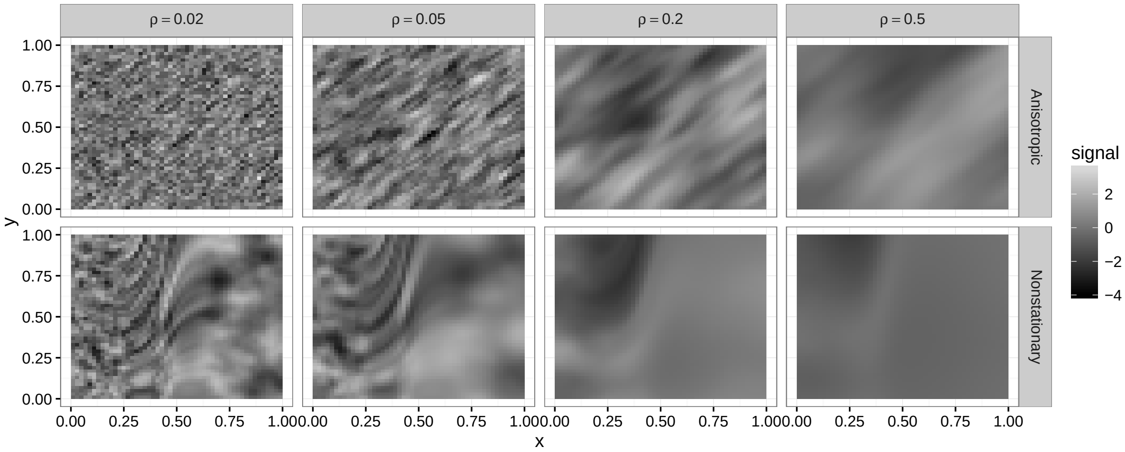

Figure 2.A.2: Examples of nonstationary covariance processes for different covariance functions: each facet shows the realization of a single Gaussian process. The anisotropic case is a deformation of a Matérn process with choices of parameters $\kappa = 2.5$ and $\rho = 0.02, 0.05, 0.2, 0.5$. The nonstationary case has a weighting function applied to the ρ coefficient, so that the dependence range between $Y(\boldsymbol{s})$ and $Y(\boldsymbol{s}^*)$ changes depending on whether $s_1, s_1^* > 0.4$.

$\hat{\beta}_2$ for the anisotropic or nonstationary spatial random effects $\boldsymbol{\eta}$, based on the transformations of a Matérn covariance function with $\kappa = 2.5$, $\rho = 0.02, 0.05, 0.2$, and 0.5, and $\sigma^2 = 1$. In this case of REML, the model is misspecified to have stationary spatial random effects following a Matérn covariance function.

In this case, the results for the anisotropic spatial random effects are similar to those of the stationary case, except that outlying regression coefficient estimates seem to occasionaly appear. For the nonstationary varying dependence range case, however, a comparison with Figure 2.8 suggests that the behavior of the regression coefficient estimates is dictated by the strongest dependence range in the region, since the simulated distribution of the PLS-GCV and PLS-MSE becomes comparable to REML at lower values of ρ.

The PLS-MSE does not improve upon PLS-GCV regression coefficient estimates in the nonstationary cases, in general.

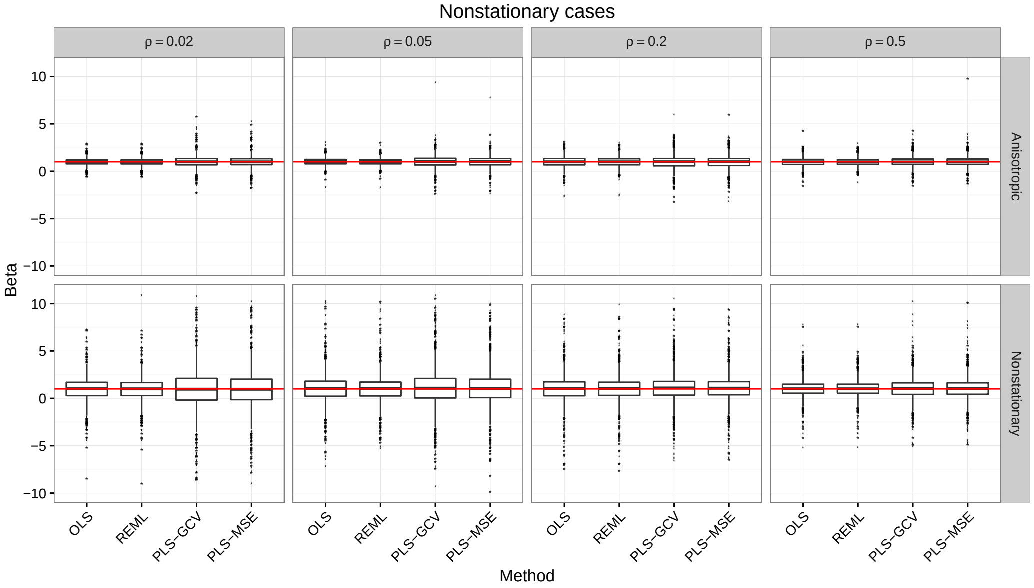

Figure 2.A.3: Comparison of $\hat{\beta}_2$ for the Anisotropic and Nonstationary cases. This figure is analogous to Figure 2.8 top half, in which $x_2(\boldsymbol{s})$ has a spatial dependence of 0.2. The horizontal solid line indicates the true value of $\beta_2 x$.

Chapter 3

A semiparametric model for fusion of static and roving sensor spatio-temporal data

3.1 Introduction

Occupational exposure assessment refers to assessment of the level of contaminants an employee is exposed to during their work-shift. The traditional method for occupational exposure assessment is personal monitoring using lightweight devices that can be worn by the workers. Personal exposure estimates are typically sought because they can be compared against regulatory standards to ensure compliance with existing laws. However, personal monitoring is generally expensive and requires workers to carry equipment with them during their work. As such, it is common for a small number of measurements, on a small number of employees, to be collected (Tornero-Velez et al., 1997; Cherrie, 2003), resulting in small sample sizes that cannot accurately capture true levels of contamination. Additionally, without the ability to track worker location, there is little ability to apportion exposures to different areas or tasks.

Occupational hazard maps, contour plots of contaminant concentration over the two-dimensional floor plan of the workplace, have gained popularity as a method to overcome some of the limitations of the traditional personal sampling that is generally expensive with small sample sizes (Koehler and Peters, 2013; Peters et al., 2012; Evans et al., 2008; Peters et al., 2006; Ologe et al., 2006). Hazard maps are commonly produced by industrial hygienists or researchers using direct-reading instruments (DRIs) to capture contaminant concentrations at high spatial resolution following a pre-determined grid throughout the facility of interest (hereafter, roving sensors). Such maps are powerful tools to communicate risk in an easily understood format and to guide decisions on control strategies aimed at reducing worker exposures (O'Brien, 2003).

Hazard maps that rely on roving monitor data alone, while cost-effective to produce and conceptually simple, likely fail to represent the temporal variability in concentrations present in many occupational settings (Koehler and Volckens, 2011; Lake et al., 2015). Augmenting the data with static sensors that collect time series data but at a few locations, can allow practitioners to expand the temporal and spatial coverage of data collection (Lake et al., 2015). As DRIs become more affordable and accessible, these types of exposure data (from static and roving sensors with known spatial information) are expected to become more abundant but rigorous statistical methods for analyzing data and drawing proper inference remain limited. The current hazard mapping approach to occupational exposure assessment, although novel, represents several challenges. Maps that are created from roving sensor data alone are often either collected over a short temporal interval or aggregated over time and neglect the temporal variability in the dataset. As such, temporal variability can be mistakenly displayed as spatial variability. In our previous work, we compared maps created using the roving sensor data and static sensor data separately (Lake et al., 2015). The method employed was somewhat ad hoc because a comprehensive statistical methodology was lacking to combine the datasets (static and roving) to provide a representation of exposures across space and time. The maps should give not only the most representative indication of the mean value, accounting for both data types, but also an indication of the variability in concentrations, as a function of both time and space.

Statistical methods for integrating different sources of data in space and/or time have been researched in the past. For example, Isaacson and Zimmerman (2000) developed methodology for combining environmental data that are temporally correlated and from two measurement systems. Their autoregressive moving-average models allowed a common time trend, system-specific measurement errors, and missing data, for which the inference was conducted by both frequentist and Bayesian approaches. Cowles et al. (2002) extended Isaacson and Zimmerman (2000) to temporally correlated data from multiple measurement systems that are measured at distinct sites in space. A Bayesian approach was taken to estimate the long-term trend and evaluate differences among the measurement systems. Further, Smith and Cowles (2007) considered an integrated model for correlating point-referenced radon and areal uranium data for quantifying a common spatial process using also a Bayesian approach, whereas Sahu et al. (2010) fused point-referenced and areal wet deposition data in space and time. Such prior research is illuminating, but none considered the possibility of roving sensors and thus is not directly applicable for the hazard mapping under consideration. The objectives of this chapter are (1) to develop new spatio-temporal methodology that combines data from both roving and static sensors for data processing and hazard mapping across space and over time in an indoor environment and (2) to compare the new method with the current industry practice, demonstrating the distinct advantages of the new method and the impact on occupational hazard assessment and future policy making in environmental health as well as occupational health.

Combining data from static and roving sensors in a statistically sound way is challenging. First of all, while the roving sensors expand the spatial coverage of data, the observations are sparse in time at any given location. This is in contrast to the static sensors that are at a smaller number of sampling locations but observations are denser in time at each sampling location. An ad hoc approach would be to analyze the two types of data separately but there is potential benefit to be gained by developing statistical methodology that pools the two data sources and takes full advantage of their respective strengths. In addition, inaccurate and missing data can be a thorny issue in such data analysis due to different measurement systems, instrumentation failures, and uneven or asynchronous monitoring times, etc. To address these challenges, we propose a novel spatio-temporal process that has continuous index in both space and time. That is, in the spatial domain of interest, the sampling locations can occur anywhere in which sense the modeling is geostatistical (see, e.g., Cressie, 1993), whereas within the temporal window of interest, the sampling can occur at any time and thus the modeling may be viewed as functional (see, e.g., Ramsay and Silverman, 2005). We then develop a model fitting procedure that allows the fusion of the two types of data based on profile likelihood accompanied by a fast computational algorithm. Further, to account for potential differences between the static and roving sensors, we extend the spatio-temporal model to allow for inhomogenous measurement error variances. Finally, we compare our new methodology with the current industrial standard/practice which does not model temporal variability and generates hazard maps from data averaged in time.

As we will demonstrate in a case study conducted in an engine test facility, the dynamic hazard maps that interpolate across space and over time are far more informative and representative of the evolution of hazard levels in space and time. This finding can impact the way occupational hazards are to be mapped in the future and move the industry and regulation forward to more accurate assessment of environmental hazard. Moreover, while the methodology developed here is geared toward spatio-temporal hazard mapping, we believe that other scientific disciplines might benefit from our approach for fusing data with very different spatial and/or temporal scales, such as data collected by individuals with personal devices versus data collected at stationary monitoring stations used to study exposure and adverse health effects in environmental epidemiology (see, e.g., Hall and McMullen, 2004).

3.2 Data and model specification

3.2.1 Data specification

We now specify the notation for the spatio-temporal process of a generic hazard. Let $D \subset \mathbb{R}^2$ denote a two-dimensional spatial domain of interest and let $[0, T]$ denote a temporal window of interest where $T > 0$. Let $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_{n_\mathrm{S}}$ denote the locations of the static sensors, where $\boldsymbol{s}_i = (s_{i,x}, s_{i,y})'$ is the location of the ith static sensor and n_{S} is the number of static sensors. For the ith static sensor, let there be p_i sampling time points, denoted as $t_{k,i}$, for $k = 1, \ldots, p_i$, $i = 1, \ldots, n_{\mathrm{S}}$. In contrast, let n_{R} denote the number of roving sensors. For the jth roving sensor, let there be q_j sampling time points. A roving sensor generally has different sampling locations at different sampling time points and therefore, the sampling locations are denoted by $\boldsymbol{r}_{1,j}, \ldots, \boldsymbol{r}_{q_j,j}$, where $\boldsymbol{r}_{l,j} = (r_{l,j,x}, r_{l,j,y})'$ is associated with sampling time $t_{l,j}$, for $l = 1, \ldots, q_j$, $j = 1, \ldots, n_{\mathrm{R}}$. Let $y_{\boldsymbol{s}}(t)$ denote the intensity of a hazard at a given point in time t in $[0, T]$ and a given spatial location $\boldsymbol{s} = (s_x, s_y)' \in D$. We will denote the samples collected by the static sensors as $y_{\boldsymbol{s}_i}(t_{k,i})$, for $k = 1, \ldots, p_i$, $i = 1, \ldots, n_{\mathrm{S}}$ and the samples collected by the roving sensors as $y_{\boldsymbol{r}_{l,j}}(t_{l,j})$, for $l = 1, \ldots, q_j$, $j = 1, \ldots, n_{\mathrm{R}}$.

3.2.2 Model specification

To model the static sensor data $\{y_{\boldsymbol{s}_i}(t_{k,i})\}$ and roving sensor data $\{y_{\boldsymbol{r}_{l,j}}(t_{l,j})\}$, we consider a spatio-temporal model

$$y_{\boldsymbol{s}}(t) = \mu_{\boldsymbol{s}}(t) + \eta_{\boldsymbol{s}}(t) + \varepsilon_{\boldsymbol{s}}(t), \tag{3.1}$$

where $\mu_{\boldsymbol{s}}(t)$ is a deterministic mean function, $\eta_{\boldsymbol{s}}(t)$ is a random spatio-temporal process with $\mathbb{E}(\eta_{\boldsymbol{s}}(t)) = 0$ and covariance function $\gamma(t, \boldsymbol{s}, t', \boldsymbol{s}') = \mathrm{Cov}\left(\eta_{\boldsymbol{s}}(t), \eta_{\boldsymbol{s}'}(t')\right)$, and $\varepsilon_{\boldsymbol{s}}(t)$ is a measurement error process with $\mathbb{E}(\varepsilon_{\boldsymbol{s}}(t)) = 0$, constant variance $\sigma^2 = \mathrm{Var}\left(\varepsilon_{\boldsymbol{s}}(t)\right)$, and zero correlation (Cressie and Wikle, 2011). Further, we assume that the spatio-temporal process $\eta_{\boldsymbol{s}}(t)$ is square integrable and the spatio-temporal covariance function of $\eta_{\boldsymbol{s}}(t)$ at location $\boldsymbol{s} \in D$ satisfies

$$\gamma(t, \boldsymbol{s}, t', \boldsymbol{s}) = \gamma_0(t, t'), \tag{3.2}$$

where $\gamma_0(t,t') = \gamma(t,\boldsymbol{s}_0,t',\boldsymbol{s}_0)$ is a temporal covariance function at any spatial location $\boldsymbol{s}_0 \in D$. That is, the temporal correlation function $\eta_{\boldsymbol{s}}(t)$ and $\eta_{\boldsymbol{s}}(t')$ is invariant in space.

The spatio-temporal process $\eta_{\boldsymbol{s}}(t)$ has a type of Karhunen-Loève decomposition (see, e.g., Gromenko and Kokoszka, 2013):

$$\eta_{\boldsymbol{s}}(t) = \sum_{\ell=1}^{\infty} \xi_\ell(\boldsymbol{s})\varphi_\ell(t), \tag{3.3}$$

where $\{\varphi_\ell(t)\}_{\ell=1}^{\infty}$ is a sequence of deterministic orthogonal temporal functions and $\{\xi_\ell(\boldsymbol{s})\}_{\ell=1}^{\infty}$ is a sequence of zero-mean random spatial processes that are uncorrelated with each other. The decomposition (3.3) represents the spatio-temporal process as a linear combination of the temporal basis functions $\varphi_\ell(t)$ (based on the temporal covariance function) with the random spatial processes $\xi_\ell(\boldsymbol{s})$ as coefficients.

We assume that the spatial covariance function of $\xi_\ell(\boldsymbol{s})$ takes on the form

$$\mathrm{Cov}\left(\xi_\ell(\boldsymbol{s}), \xi_\ell(\boldsymbol{s}')\right) = \lambda_\ell \rho_\ell(\|\boldsymbol{s}-\boldsymbol{s}'\|; \boldsymbol{\theta}_\ell),$$

where $\lambda_\ell = \mathrm{Var}\left(\xi_\ell(\boldsymbol{s})\right)$ is the variance of $\xi_\ell(\boldsymbol{s})$, $\rho_\ell(\cdot;\boldsymbol{\theta}_\ell)$ is a correlation function parameterized by $\boldsymbol{\theta}_\ell$, and $\|\cdot\|$ denotes the Euclidean distance. From (3.3), we can write the spatio-temporal covariance function $\gamma(t,\boldsymbol{s},t',\boldsymbol{s}')$ of $\eta_{\boldsymbol{s}}(t)$ as

$$\gamma(t,\boldsymbol{s},t',\boldsymbol{s}') = \sum_{\ell=1}^{\infty} \mathrm{Cov}\left(\xi_\ell(\boldsymbol{s}), \xi_\ell(\boldsymbol{s}')\right)\varphi_\ell(t)\varphi_\ell(t'). \tag{3.4}$$

In the case $\boldsymbol{s} = \boldsymbol{s}'$, (3.4) is reduced to the temporal covariance function

$$\gamma_0(t,t') = \sum_{\ell=1}^{\infty} \lambda_\ell \varphi_\ell(t)\varphi_\ell(t').$$

This makes clear that $\{\varphi_\ell(t)\}_{\ell=1}^{\infty}$ are analogous to the eigenfunctions of $\gamma_0(t,t')$ with the corresponding eigenvalues $\{\lambda_\ell\}_{\ell=1}^{\infty}$. It is based on (3.4) that we will devise a semiparametric likelihood approach to fitting the spatio-temporal model (3.1) to the static and roving sensor data, as well as mapping the true spatio-temporal process of the hazard $\mu_{\boldsymbol{s}}(t)+\eta_{\boldsymbol{s}}(t)$, while taking into account the spatial and temporal variability.

Our modeling approach is tailored toward the distinct features of static and roving sensor data. The spatial index is continuous in the spatial domain and the temporal index is continuous within the time window. Thus, the sensors can be placed anywhere in the study area and do not need to be on a regular grid. Further, sampling can occur at any point in time and no regular time intervals are required. In addition, our modeling framework is semiparametric and flexible. The specification of the deterministic mean function $\mu_{\boldsymbol{s}}(t)$ is nonparametric, while the specification of the spatio-temporal process is semiparametric in the sense that the spatial covariance function $\operatorname{Cov}(\xi_\ell(\boldsymbol{s}), \xi_\ell(\boldsymbol{s}'))$ in (3.4) is parametric but the temporal covariance function $\gamma_0(t, t')$ in (3.2) is nonparametric. The nonparametric specification allows for capturing different sources for a hazard, some of which are unexpected to be present, or are present at unexpected time intervals, such as the outside noise near sensor #18 in the case study.

The class of spatio-temporal covariance functions (3.4) is broad, encompassing processes that are nonstationary and nonseparable in space and time with the separable case corresponding to $\lambda_\ell = 0$ for $\ell \geq 2$. These properties can be contrasted to spatio-temporal kriging (Cressie and Wikle, 2011, p. 321), which in many practical scenarios require the specification of the spatio-temporal covariance function, and the time series of spatial process approach (Cressie and Wikle, 2011, p. 336), which requires the temporal coordinates to be sampled at regular intervals.

3.3 Statistical inference

3.3.1 Profile likelihood estimation

Parameter estimation by maximum likelihood can be challenging due the large number of parameters in the model (3.1). Thus, we use the idea of functional principal components and develop a profile likelihood approach to estimating the model parameters (Ramsay and Silverman, 2005). Further, only finitely many eigenvalues are estimable from a sample covariance matrix in the nonparametric specification of the covariance function and therefore, it is necessary that for some $L \leq \min_i\{p_i\}$ (see Wahba, 1990, p.5),

$$\eta_{\boldsymbol{s}}(t) \approx \sum_{\ell=1}^{L} \xi_\ell(\boldsymbol{s})\varphi_\ell(t). \tag{3.5}$$

Let $\boldsymbol{y}_{\boldsymbol{s}_i} = (y_{\boldsymbol{s}_i}(t_1), \ldots, y_{\boldsymbol{s}_i}(t_{p_i}))'$ and $\boldsymbol{\mu}_{\boldsymbol{s}_i} = (\mu_{\boldsymbol{s}_i}(t_1), \ldots, \mu_{\boldsymbol{s}_i}(t_{p_i}))'$ denote the vector of the data from the ith static sensor at time points t_k for $k = 1, \ldots, p_i$ and the corresponding mean vector, for $i = 1, \ldots, n_{\mathrm{S}}$. Let $\boldsymbol{y}_{\boldsymbol{r}_j} = (y_{\boldsymbol{r}_{1,j}}(t_{1,j}), \ldots, y_{\boldsymbol{r}_{q_j,j}}(t_{q_j,j}))'$ and $\boldsymbol{\mu}_{\boldsymbol{r}_j} = (\mu_{\boldsymbol{r}_{1,j}}(t_{1,j}), \ldots, \mu_{\boldsymbol{r}_{q_j,j}}(t_{q_j,j}))'$ denote the vector of the data from the jth roving sensor at spatial locations $\boldsymbol{r}_{l,j}$ and time points $t_{l,j}$ for $l = 1, \ldots, q_j$ and the corresponding mean vector, for $j = 1, \ldots, n_{\mathrm{R}}$. Also, let $\boldsymbol{y} = (\boldsymbol{y}'_{\boldsymbol{s}_1}, \ldots, \boldsymbol{y}'_{\boldsymbol{s}_{n_{\mathrm{S}}}}, \boldsymbol{y}'_{\boldsymbol{r}_1}, \ldots, \boldsymbol{y}'_{\boldsymbol{r}_{n_{\mathrm{R}}}})'$ and $\boldsymbol{\mu} = (\boldsymbol{\mu}'_{\boldsymbol{s}_1}, \ldots, \boldsymbol{\mu}'_{\boldsymbol{s}_{n_{\mathrm{S}}}}, \boldsymbol{\mu}'_{\boldsymbol{r}_1}, \ldots, \boldsymbol{\mu}'_{\boldsymbol{r}_{n_{\mathrm{R}}}})'$ denote the vector of all the observations and the corresponding mean vector.

Let

$$\boldsymbol{\Phi} = \begin{pmatrix} \mathrm{diag}\{\boldsymbol{A}_i\}_{i=1}^{n_{\mathrm{S}}} & & & \\ & \mathrm{diag}\{\boldsymbol{b}_{l,1}\}_{l=1}^{q_1} & & \\ & & \ddots & \\ & & & \mathrm{diag}\{\boldsymbol{b}_{l,n_{\mathrm{R}}}\}_{l=1}^{q_{n_{\mathrm{R}}}} \end{pmatrix}$$

denote a block diagonal matrix, where

$$\boldsymbol{A}_i = \begin{pmatrix} \varphi_1(t_{1,i}) & \cdots & \varphi_1(t_{p_i,i}) \\ \vdots & \ddots & \vdots \\ \varphi_L(t_{1,i}) & \cdots & \varphi_L(t_{p_i,i}) \end{pmatrix} \text{ and } \boldsymbol{b}_{l,j} = \begin{pmatrix} \varphi_1(t_{l,j}) \\ \vdots \\ \varphi_L(t_{l,j}) \end{pmatrix}.$$

Let

$$\boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_{\text{S,S}} & \boldsymbol{\Lambda}_{\text{S,R}} \\ \boldsymbol{\Lambda}_{\text{R,S}} & \boldsymbol{\Lambda}_{\text{R,R}} \end{pmatrix}$$

with

$$\boldsymbol{\Lambda}_{\text{S,S}} = \begin{pmatrix} \text{diag}\{\lambda_\ell \rho_\ell(\|\boldsymbol{s}_1 - \boldsymbol{s}_1\|; \boldsymbol{\theta}_\ell)\}_{\ell=1}^L & \cdots & \text{diag}\{\lambda_\ell \rho_\ell(\|\boldsymbol{s}_1 - \boldsymbol{s}_n\|; \boldsymbol{\theta}_\ell)\}_{\ell=1}^L \\ \vdots & \ddots & \vdots \\ \text{diag}\{\lambda_\ell \rho_\ell(\|\boldsymbol{s}_n - \boldsymbol{s}_1\|; \boldsymbol{\theta}_\ell)\}_{\ell=1}^L & \cdots & \text{diag}\{\lambda_\ell \rho_\ell(\|\boldsymbol{s}_n - \boldsymbol{s}_n\|; \boldsymbol{\theta}_\ell)\}_{\ell=1}^L \end{pmatrix}$$

where $\lambda_\ell \rho_\ell(\|\boldsymbol{s} - \boldsymbol{s}^*\|) = \text{Cov}\,(\xi_\ell(\boldsymbol{s}), \xi_\ell(\boldsymbol{s}^*))$, $\lambda_\ell = \text{Var}\,(\xi_\ell(\boldsymbol{s}))$ and $\rho_\ell(\cdot)$ is a spatial correlation function that may be modeled by the Matérn class (Stein, 1999). Note $\boldsymbol{\Lambda}_{\text{S,S}}$ is a block matrix with blocks corresponding to distinct spatial locations. The submatrices $\boldsymbol{\Lambda}_{\text{S,R}}$ and $\boldsymbol{\Lambda}_{\text{R,R}}$ are defined analogously; however, for a given roving sensor, each distinct spatial location corresponds to its own block. This illustrates that the covariance structure is more complex than a sampling scheme that involves only static sensors, showing that roving sensors play a role in both spatial and temporal dependence. The rank of $\boldsymbol{\Lambda}$ can be as large as $L(n_S + \sum_{j=1}^{n_R} q_j)$.

Suppose $\eta_{\boldsymbol{s}}(t)$ and $\varepsilon_{\boldsymbol{s}}(t)$ are Gaussian processes. Then $\xi_\ell(\boldsymbol{s})$ are Gaussian processes and $\boldsymbol{y}$ follows a multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance

$$\boldsymbol{\Sigma} = \boldsymbol{\Phi}' \boldsymbol{\Lambda} \boldsymbol{\Phi} + \sigma^2 \boldsymbol{I}_N, \tag{3.6}$$

where $N = \sum_{i=1}^{n_\text{S}} p_i + \sum_{j=1}^{n_\text{R}} q_j$ is the total sample size combining static and roving sensor data and $\boldsymbol{I}_N$ is the N-dimensional identity matrix.

Estimation of all the components is not always possible depending on the choices made for the parameters $\boldsymbol{\theta}_\ell$ in the spatial correlation function $\rho_\ell(\cdot)$ and those made for the shape of the temporal process $\varphi_\ell(\cdot)$. We now develop a profile likelihood approach to parameter estimation. At initialization, we estimate the mean function $\mu_{\boldsymbol{s}}(t)$ by ordinary least squares (OLS) and denote the estimated mean function as $\hat{\mu}_{\boldsymbol{s}}(t)$. For example, a fitted mean function could comprise both a linear model spatially and a nonparametric model temporally,

$$\hat{\mu}_{\boldsymbol{s}}(t) = \hat{\beta}_0 + \hat{\beta}_x s_x + \hat{\beta}_y s_y + \sum_{k=1}^{K} \hat{\beta}_k B_k(t) \tag{3.7}$$

where $\boldsymbol{s} = (s_x, s_y)'$, $B_k(\cdot)$ are cubic spline basis functions, K is the number of basis functions that controls the smoothness of $\hat{\mu}$, and $\hat{\beta}_{\cdot}$ are the OLS estimates of the coefficients. Let $\hat{\boldsymbol{\mu}}_{\mathrm{OLS}}$ denote the vector of $\hat{\mu}_{\boldsymbol{s}}(t)$ at all sampling locations and time points. Let $\hat{\boldsymbol{y}} = \boldsymbol{y} - \hat{\boldsymbol{\mu}}_{\mathrm{OLS}}$ denote the detrended data comprising $\hat{\boldsymbol{y}}_{\boldsymbol{s}_i} = (\hat{\boldsymbol{y}}_{\boldsymbol{s}_i}(t_{i,k}) : k = 1, \ldots, p_i)'$ for $i = 1, \ldots, n_{\mathrm{S}}$ and $\hat{\boldsymbol{y}}_{\boldsymbol{r}_j} = (\hat{\boldsymbol{y}}_{\boldsymbol{r}_{j,l}}(t_{j,l}) : l = 1, \ldots, q_j)'$ for $j = 1, \ldots, n_{\mathrm{R}}$.

Next, we estimate λ_ℓ and $\varphi_\ell(t)$ by applying a functional principal component analysis to the data from the static sensors (Ramsay and Silverman, 2005, pp. 178–182). In this step, we estimate the temporal covariance function $\hat{\gamma}_0(t, t')$ from vectors $\hat{\boldsymbol{y}}_{\boldsymbol{s}_i}$ expanded in B-spline basis functions, obtaining a functional estimate of $y_{\boldsymbol{s}_i}(t) - \mu_{\boldsymbol{s}_i}(t)$ denoted by $\hat{y}_{\boldsymbol{s}_i}(t)$. Thus

$$\hat{\gamma}_0(t, t') = n_{\mathrm{S}}^{-1} \sum_{i=1}^{n_{\mathrm{S}}} \hat{y}_{\boldsymbol{s}_i}(t) \hat{y}_{\boldsymbol{s}_i}(t'). \tag{3.8}$$

The estimate of the first temporal function $\hat{\varphi}_1(\cdot)$ is the maximizer of

$$\max_{\|f(t)\|_\zeta = 1} \int_0^T \int_0^T f(t) \hat{\gamma}_0(t, t') f(t') \mathrm{d}t \mathrm{d}t', \tag{3.9}$$

where $\|f(t)\|_\zeta = 1$ in the L^ζ norm, $\langle f, g \rangle_\zeta = \int_0^T f(t) g(t) \mathrm{d}t + \zeta \int_0^T f''(t) g''(t) \mathrm{d}t$ is an inner product, and ζ is a tuning parameter that controls the smoothness of the estimates. The estimates of the subsequent temporal functions $\hat{\varphi}_\ell(\cdot)$ for $\ell \geq 2$ are the maximizer of (3.9) under the constraint of orthogonality of $\hat{\varphi}_\ell(\cdot)$ and $\hat{\varphi}_{\ell'}(\cdot)$ (in the $\langle \cdot, \cdot \rangle_\zeta$ sense) for all $\ell' < \ell$. Finally, λ_ℓ is estimated by

$$\hat{\lambda}_\ell = \int_0^T \int_0^T \hat{\varphi}_\ell(t)\hat{\gamma}_0(t,t')\hat{\varphi}_\ell(t')\mathrm{d}t\mathrm{d}t'$$

for $\ell = 1, \ldots, L$.

Unlike static sensors, it is not possible to obtain a full time series of data at a fixed location $\boldsymbol{r}_{j,l}$ for the roving sensor. Therefore, estimates of $\varphi_\ell(t)$ are based on static sensors only and the values of $\hat{\varphi}_\ell(t)$ for the roving sensors are interpolated by evaluating the estimates $\hat{\varphi}_\ell(t)$ at time points $t_{l,j}$, for $l = 1, \ldots, q_j$, $j = 1, \ldots, n_{\mathrm{R}}$. The smoothness of $\varphi_\ell(t)$ estimates is influenced by the number of basis functions K, the tuning parameter ζ, and the number of functional principal components L.

Given $\hat{\boldsymbol{\mu}}$, $\hat{\lambda}_\ell$, and $\hat{\varphi}_\ell(t)$, we minimize the negative profile log-likelihood of $\boldsymbol{\theta}$ and σ^2 defined as

$$\begin{aligned} f(\boldsymbol{\theta}, \sigma^2) &= -2\mathcal{L}(\boldsymbol{\theta}, \sigma^2) \\ &\propto \hat{\boldsymbol{y}}'\boldsymbol{\Sigma}_1(\boldsymbol{\theta}, \sigma^2)^{-1}\hat{\boldsymbol{y}} + \log[\det\{\boldsymbol{\Sigma}_1(\boldsymbol{\theta}, \sigma^2)^{-1}\}], \end{aligned} \tag{3.10}$$

where $\hat{\boldsymbol{y}} = \boldsymbol{y} - \hat{\boldsymbol{\mu}}$ is the detrended data vector and $\boldsymbol{\Sigma}_1(\boldsymbol{\theta}, \sigma^2) = \boldsymbol{\Sigma}(\boldsymbol{\theta}, \sigma^2; \hat{\lambda}_\ell, \hat{\varphi}_\ell(t))$ is the covariance (3.6) parameterized by $\boldsymbol{\theta}$ and σ^2 and evaluated at $\hat{\lambda}_\ell$ and $\hat{\varphi}_\ell(t)$. The solutions to (3.10) can be obtained by using box-constrained optimization to ensure that all estimated parameters are positive. The optimization in (3.10) is obtained numerically using the gradients

$$\begin{aligned} \frac{\partial}{\partial \theta_\ell} f(\boldsymbol{\theta}, \sigma^2) &= -\hat{\boldsymbol{y}}' \boldsymbol{\Sigma}_1(\boldsymbol{\theta}, \sigma^2)^{-1} \boldsymbol{\Phi}' \left\{ \frac{\partial}{\partial \theta_\ell} \boldsymbol{\Lambda}(\boldsymbol{\theta}) \right\} \boldsymbol{\Phi} \boldsymbol{\Sigma}_1(\boldsymbol{\theta}, \sigma^2)^{-1} \hat{\boldsymbol{y}} \\ &\quad + \operatorname{tr} \left\{ \boldsymbol{\Sigma}_1(\boldsymbol{\theta}, \sigma^2)^{-1} \boldsymbol{\Phi}' \frac{\partial}{\partial \theta_\ell} \boldsymbol{\Lambda}(\boldsymbol{\theta}) \boldsymbol{\Phi} \right\} \qquad (3.11) \\ \frac{\partial}{\partial \sigma^2} f(\boldsymbol{\theta}, \sigma^2) &= -\hat{\boldsymbol{y}}' \boldsymbol{\Sigma}_1(\boldsymbol{\theta}, \sigma^2)^{-2} \hat{\boldsymbol{y}} + \operatorname{tr} \left\{ \boldsymbol{\Sigma}_1(\boldsymbol{\theta}, \sigma^2)^{-1} \right\}. \end{aligned}$$

The solutions to (3.10) can be obtained by using box-constrained optimization to ensure that all estimated parameters are positive. Further, the coefficients in the mean function (3.7) are updated by

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}' \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{X} \right)^{-1} \mathbf{X} \hat{\boldsymbol{\Sigma}}^{-1} \mathbf{y},$$

where $\hat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}, \hat{\sigma}^2; \hat{\lambda}_\ell, \hat{\varphi}_\ell(t))$ is the covariance (3.6) evaluated at the parameter estimates and $\mathbf{X}$ is the design matrix for the covariates in (3.7).

3.3.2 Uncertainty quantification

For static data only, asymptotics were derived in Chu et al. (2016). But to quantify the uncertainty of estimated parameters and evaluate the overall prediction accuracy of the models, we employ a leave-one-sensor-out cross-validation procedure. For each static sensor $i = 1, \ldots, n_{\mathrm{S}}$, we remove the corresponding data $\boldsymbol{y}_{\boldsymbol{s}_i}$, and fit the model with the remaining data $\boldsymbol{y}^{-i}$.

To obtain confidence intervals and standard errors for $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\theta}}$, and $\hat{\sigma}^2$, a cross-validated empirical distribution of the respective parameters can be used. Let $\hat{\alpha}$ denote a generic parameter estimate. By removing the data for one static sensor at a time, we obtain n_{S} different estimates $\hat{\alpha}^{(-i)}$ for $i = 1, \ldots, n_{\mathrm{S}}$. The standard error of $\hat{\alpha}$ is obtained by

$$\text{s.e.}(\hat{\alpha}) = \left\{ n_{\mathrm{S}}^{-1} \sum_{i=1}^{n_{\mathrm{S}}} (\hat{\alpha}^{(-i)} - \hat{\alpha})^2 \right\}^{1/2}.$$

In addition, the $(1-\alpha)100\%$ confidence interval can be constructed by the $(\alpha/2)$th and $(1-\alpha/2)$th empirical quantiles from the empirical cumulative distribution function

$$\widehat{F}(\eta) = n_{\mathrm{S}}^{-1} \sum_{i=1}^{n_{\mathrm{S}}} \mathcal{I}\{\hat{\alpha}^{(-i)} \leq \eta\},$$

where $\mathcal{I}(\cdot)$ is the indicator function.

To obtain the mean squared prediction error (MSPE), the model fitted without $\boldsymbol{y}_{\boldsymbol{s}_i}$ is used to produce a predicted value $\hat{\boldsymbol{y}}_{\boldsymbol{s}_i}^{(-i)}$ and thus,

$$\mathrm{MSPE} = n_{\mathrm{S}}^{-1} \sum_{i=1}^{n_{\mathrm{S}}} \sum_{k=1}^{p_i} \{y_{\boldsymbol{s}_i}(t_k) - \hat{y}_{\boldsymbol{s}_i}^{(-i)}(t_k)\}^2.$$

A similar approach is taken to tune the smoothness of $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\varphi}}$ estimators by searching over a grid of K and ζ values for K_0 and ζ_0 that minimize the MSPE. It is of interest to find optimal tuning parameters without resorting to cross validation, although tuning principal functional components automatically is an open problem by itself (Ramsay and Silverman, 2005, p.179).

3.3.3 Spatio-temporal Kriging and prediction of spatial loadings

To predict $y_{\boldsymbol{s}_0}(t)$ at an unsampled location $\boldsymbol{s}_0$ and time t_0, we use

$$\hat{y}_{\boldsymbol{s}_0}(t_0) = \hat{\mu}_{\boldsymbol{s}_0}(t_0) + \hat{\eta}_{\boldsymbol{s}_0}(t_0) = \hat{\mu}_{\boldsymbol{s}_0}(t_0) + \hat{\boldsymbol{\Sigma}}_{\boldsymbol{s}_0,t_0} \hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{y} - \hat{\boldsymbol{\mu}}) \tag{3.12}$$

where $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{s}_0,t_0} = \hat{\boldsymbol{\Phi}}(t_0)'\hat{\boldsymbol{\Lambda}}(\boldsymbol{s}_0)\hat{\boldsymbol{\Phi}}(t_0)$, $\hat{\boldsymbol{\Phi}}(t_0)$ is $\boldsymbol{\Phi}$ evaluated at t_0, and $\hat{\boldsymbol{\Lambda}}(\boldsymbol{s}_0) = \begin{pmatrix} \hat{\boldsymbol{\Lambda}}_{\boldsymbol{s}_0,\mathrm{S}} & \hat{\boldsymbol{\Lambda}}_{\boldsymbol{s}_0,\mathrm{R}} \end{pmatrix}$, with $\boldsymbol{\Lambda}_{\boldsymbol{s}_0,\mathrm{S}} = \{\mathrm{diag}\{\lambda_\ell \rho_\ell(\|\boldsymbol{s}_0 - \boldsymbol{s}_i\|; \boldsymbol{\theta}_\ell\}_{\ell=1}^{L}\}_{i=1}^{n_{\mathrm{S}}}$, and $\boldsymbol{\Lambda}_{\boldsymbol{s}_0,\mathrm{R}} = \{\mathrm{diag}\{\lambda_\ell \rho_\ell(\|\boldsymbol{s}_0 - \boldsymbol{r}_{l,j}\|; \boldsymbol{\theta}_\ell\}_{\ell=1}^{L}\}_{l=1}^{q_j}{}_{,j=1}^{n_{\mathrm{R}}}$. Equation (3.12) is used over a fine spatial grid to generate the hazard maps over time. The prediction standard error is given by $\hat{\sigma}_{\boldsymbol{s}_0}(t_0)$, where

$$\begin{aligned} \hat{\sigma}_{\boldsymbol{s}_0}^2(t_0) = \hat{\sigma}^2 &+ \sum_{\ell=1}^{L} \hat{\lambda}_\ell \hat{\varphi}_\ell^2(t_0) - \hat{\boldsymbol{\Sigma}}'_{\boldsymbol{s}_0,t_0} \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{s}_0,t_0} \\ &+ (\mathbf{x} - \mathbf{X}'\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{s}_0,t_0})'(\mathbf{X}'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X})^{-1}(\mathbf{x} - \mathbf{X}'\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{s}_0,t_0}), \end{aligned}$$

and $\mathbf{x}$ are the covariates of in (3.7) evaluated at $\boldsymbol{s}_0$ and t_0.

We can also predict the spatial loadings $\xi_\ell(\boldsymbol{s})$ using conditional expectations, in an approach similar to Yao et al. (2005) for non-spatial data. Let $\boldsymbol{s}_1^*, \ldots,$ $\boldsymbol{s}_m^* \in D$ be the unsampled locations of interest. The ℓth loading evaluated at locations $\boldsymbol{s}_1^*, \ldots,$ $\boldsymbol{s}_m^*$ is a linear predictor of $(\xi_\ell(\boldsymbol{s}_1^*), \ldots, \xi_\ell(\boldsymbol{s}_m^*))$. Assuming $\{\xi_\ell(\boldsymbol{s}_1^*), \ldots,$ $\xi_\ell(\boldsymbol{s}_m^*)\}_{\ell=1}^\infty$ and $\varepsilon_{\boldsymbol{s}}(t)$ are jointly Gaussian, we have

$$\begin{aligned}(\hat{\xi}_\ell(\boldsymbol{s}_1^*), \ldots, \hat{\xi}_\ell(\boldsymbol{s}_m^*))' &= \hat{\mathbb{E}}\left((\xi_\ell(\boldsymbol{s}_1^*), \ldots, \xi_\ell(\boldsymbol{s}_m^*))' \mid \mathbf{y}\right) \\ &= \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\xi}_\ell, \mathbf{y}} \hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{y} - \hat{\boldsymbol{\mu}}),\end{aligned}$$

where $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\xi}_\ell, \mathbf{y}}$ is the plug-in estimate of $\boldsymbol{\Sigma}_{\boldsymbol{\xi}_\ell, \mathbf{y}}$, and $\boldsymbol{\Sigma}_{\boldsymbol{\xi}_\ell, \mathbf{y}}$ is the sample covariance of $(\boldsymbol{\xi}_\ell(\boldsymbol{s}_1^*),$ $\ldots,$ $\boldsymbol{\xi}_\ell(\boldsymbol{s}_m^*))'$ and $\mathbf{y}$, given by

$$\boldsymbol{\Sigma}_{\boldsymbol{\xi}_\ell, \mathbf{y}} = \lambda_\ell \begin{pmatrix} \varphi_\ell(t_{1,1})\rho_\ell(\|\boldsymbol{s}_1^* - \boldsymbol{s}_1\|; \boldsymbol{\theta}_\ell) & \cdots & \varphi_\ell(t_{q_{n_\mathrm{R}}, n_\mathrm{R}})\rho_\ell(\|\boldsymbol{s}_1^* - \boldsymbol{r}_{q_{n_\mathrm{R}}, n_\mathrm{R}}\|; \boldsymbol{\theta}_\ell) \\ \vdots & \ddots & \vdots \\ \varphi_\ell(t_{1,1})\rho_\ell(\|\boldsymbol{s}_m^* - \boldsymbol{s}_1\|; \boldsymbol{\theta}_\ell) & \cdots & \varphi_\ell(t_{q_{n_\mathrm{R}}, n_\mathrm{R}})\rho_\ell(\|\boldsymbol{s}_m^* - \boldsymbol{r}_{q_{n_\mathrm{R}}, n_\mathrm{R}}\|; \boldsymbol{\theta}_\ell) \end{pmatrix}$$

and $\hat{\boldsymbol{\Sigma}}$, $\mathbf{y}$ and $\hat{\boldsymbol{\mu}}$ are as defined in Section 3.3.1.

3.3.4 Inhomogeneous variances

The model given in (3.1) assumes that the measurement error variance is the same for the static and roving sensors. In practice, however, this assumption does not always hold. In the following we extend the data model to accommodate the situation that the measurement error variance for the static sensors is different from the roving sensors.

Consider model (3.1) again

$$y_{\boldsymbol{s}}(t) = \mu_{\boldsymbol{s}}(t) + \eta_{\boldsymbol{s}}(t) + \varepsilon_{\boldsymbol{s}}(t), \tag{3.13}$$

but with measurement error variances $\mathrm{Var}\,(\varepsilon_{\boldsymbol{s}}(t)) = \sigma^2_{\mathrm{S}}$ for static sensors and $\mathrm{Var}\,(\varepsilon_{\boldsymbol{s}}(t)) = \sigma^2_{\mathrm{R}}$ for roving sensors. The model (3.13) will be referred to as the *inhomogeneous variance* case. The model (3.1) is a special case where $\sigma^2_{\mathrm{S}} = \sigma^2_{\mathrm{R}}$ and will be referred to as the *homogeneous variance* case. In the inhomogeneous variance case, the covariance is a general case of (3.6) and the estimation algorithm is modified by using

$$\boldsymbol{\Sigma} = \boldsymbol{\Phi}'\boldsymbol{\Lambda}\boldsymbol{\Phi} + \sigma^2_{\mathrm{S}}\boldsymbol{D}_{\mathrm{S}} + \sigma^2_{\mathrm{R}}\boldsymbol{D}_{\mathrm{R}}, \tag{3.14}$$

where $\boldsymbol{D}_{\mathrm{S}}, \boldsymbol{D}_{\mathrm{R}}$ are diagonal matrices with diagonal entries equal to 1 for static and roving sensors, respectively, and 0 otherwise. The model (3.13) allows the fusion of the two sources of data with varying degrees of spatio-temporal resolutions and drawing inference about the true underlying process $\mu_{\boldsymbol{s}}(t) + \eta_{\boldsymbol{s}}(t)$. The profile-likelihood approach to parameter estimation is modified as follows. The data are detrended with initial estimates of $\boldsymbol{\mu}$ using (3.7) and the λ_ℓ and $\varphi_\ell(t)$ terms are estimated using functional principal component analysis over the static sensors as before. Given $\hat{\boldsymbol{\mu}}$, $\hat{\lambda}_\ell$, and $\hat{\varphi}(t)$, we minimize the negative profile log-likelihood of $\boldsymbol{\theta}$, σ^2_{S}, and σ^2_{R} defined as

$$\begin{aligned} f(\boldsymbol{\theta}, \sigma^2_{\mathrm{S}}, \sigma^2_{\mathrm{R}}) = -2\mathcal{L}(\boldsymbol{\theta}, \sigma^2_{\mathrm{S}}, \sigma^2_{\mathrm{R}}) &= \hat{\boldsymbol{y}}'\boldsymbol{\Sigma}_2(\boldsymbol{\theta}, \sigma^2_{\mathrm{S}}, \sigma^2_{\mathrm{R}})^{-1}\hat{\boldsymbol{y}} \\ &+ \log[\det\{\boldsymbol{\Sigma}_2(\boldsymbol{\theta}, \sigma^2_{\mathrm{S}}, \sigma^2_{\mathrm{R}})\}] + N\log(2\pi), \end{aligned} \tag{3.15}$$

where $\hat{\boldsymbol{y}} = \boldsymbol{y} - \hat{\boldsymbol{\mu}}$ and $\boldsymbol{\Sigma}_2(\boldsymbol{\theta}, \sigma^2_{\mathrm{S}}, \sigma^2_{\mathrm{R}}) = \boldsymbol{\Sigma}(\boldsymbol{\theta}, \sigma^2_{\mathrm{S}}, \sigma^2_{\mathrm{R}}; \hat{\lambda}_\ell, \hat{\varphi}_\ell(t))$ given in (3.14) is parameterized by $\boldsymbol{\theta}$, σ^2_{S}, and σ^2_{R} evaluated at $\hat{\lambda}_\ell$ and $\hat{\varphi}_\ell(t)$. The coefficients in the mean function are updated by

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{y},$$

where $\hat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}, \hat{\sigma}^2_{\mathrm{S}}, \hat{\sigma}^2_{\mathrm{R}}; \hat{\lambda}_\ell, \hat{\varphi}(t))$ and $\mathbf{X}$ is the design matrix for the covariates in (3.7).

The optimal values of $\boldsymbol{\theta}, \sigma^2_{\mathrm{S}}$, and σ^2_{R}, given $\hat{\boldsymbol{\mu}}$, $\hat{\lambda}_\ell$, and $\hat{\varphi}(t)$, are obtained analogously with the gradient given by

$$\begin{aligned}
\frac{\partial}{\partial \theta_\ell} f(\boldsymbol{\theta}, \sigma_{\mathrm{S}}^2, \sigma_{\mathrm{R}}^2) &= -\hat{\boldsymbol{y}}' \boldsymbol{\Sigma}_2(\boldsymbol{\theta}, \sigma_{\mathrm{S}}^2, \sigma_{\mathrm{R}}^2)^{-1} \boldsymbol{\Phi}' \left\{ \frac{\partial}{\partial \theta_\ell} \boldsymbol{\Lambda}(\boldsymbol{\theta}) \right\} \boldsymbol{\Phi} \boldsymbol{\Sigma}_2(\boldsymbol{\theta}, \sigma_{\mathrm{S}}^2, \sigma_{\mathrm{R}}^2)^{-1} \hat{\boldsymbol{y}} \\
&\quad + \operatorname{tr} \left\{ \boldsymbol{\Sigma}_2(\boldsymbol{\theta}, \sigma_{\mathrm{S}}^2, \sigma_{\mathrm{R}}^2)^{-1} \boldsymbol{\Phi}' \frac{\partial}{\partial \theta_\ell} \boldsymbol{\Lambda}(\boldsymbol{\theta}) \boldsymbol{\Phi} \right\}, \\
\frac{\partial}{\partial \sigma_{\mathrm{S}}^2} f(\boldsymbol{\theta}, \sigma_{\mathrm{S}}^2, \sigma_{\mathrm{R}}^2) &= -\hat{\boldsymbol{y}}' \boldsymbol{\Sigma}_2(\boldsymbol{\theta}, \sigma_{\mathrm{S}}^2, \sigma_{\mathrm{R}}^2)^{-1} \boldsymbol{D}_{\mathrm{S}} \boldsymbol{\Sigma}_2(\boldsymbol{\theta}, \sigma_{\mathrm{S}}^2, \sigma_{\mathrm{R}}^2)^{-1} \hat{\boldsymbol{y}} \\
&\quad + \operatorname{tr} \left\{ \boldsymbol{\Sigma}_2(\boldsymbol{\theta}, \sigma_{\mathrm{S}}^2, \sigma_{\mathrm{R}}^2)^{-1} \boldsymbol{D}_{\mathrm{S}} \right\}, \\
\frac{\partial}{\partial \sigma_{\mathrm{R}}^2} f(\boldsymbol{\theta}, \sigma_{\mathrm{S}}^2, \sigma_{\mathrm{R}}^2) &= -\hat{\boldsymbol{y}}' \boldsymbol{\Sigma}_2(\boldsymbol{\theta}, \sigma_{\mathrm{S}}^2, \sigma_{\mathrm{R}}^2)^{-1} \boldsymbol{D}_{\mathrm{R}} \boldsymbol{\Sigma}_2(\boldsymbol{\theta}, \sigma_{\mathrm{S}}^2, \sigma_{\mathrm{R}}^2)^{-1} \hat{\boldsymbol{y}} \\
&\quad + \operatorname{tr} \left\{ \boldsymbol{\Sigma}_2(\boldsymbol{\theta}, \sigma_{\mathrm{S}}^2, \sigma_{\mathrm{R}}^2)^{-1} \boldsymbol{D}_{\mathrm{R}} \right\}.
\end{aligned} \tag{3.16}$$

This computation can be potentially more expensive than evaluating (3.11) because it requires explicit evaluation of $\boldsymbol{\Sigma}_2(\boldsymbol{\theta}, \sigma_{\mathrm{S}}^2, \sigma_{\mathrm{R}}^2)^{-1}$ in order to calculate the partial traces $\operatorname{tr}\{\boldsymbol{\Sigma}_2(\boldsymbol{\theta}, \sigma_{\mathrm{S}}^2, \sigma_{\mathrm{R}}^2)^{-1} \boldsymbol{D}_{\mathrm{S}}\}$ and $\operatorname{tr}\{\boldsymbol{\Sigma}_2(\boldsymbol{\theta}, \sigma_{\mathrm{S}}^2, \sigma_{\mathrm{R}}^2)^{-1} \boldsymbol{D}_{\mathrm{R}}\}$. In the homogeneous variance case, this trace can be calculated faster using only the eigenvalues of $\boldsymbol{\Sigma}_1(\boldsymbol{\theta}, \sigma^2)$.

The algorithms in Sections 3.3.1 and 3.3.4 are implemented in C++ and incorporated into `R` using the `Rcpp` (Eddelbuettel and François, 2011) and `RcppEigen` packages (Bates and Eddelbuettel, 2013). A generic implementation of the algorithm is available in `www.github.com/guiludwig/STDF`, with additional examples in the documentation.

3.4 Case study: spatio-temporal occupational hazard mapping

3.4.1 Application and data

A study was conducted in the spring of 2013 in an engine test facility located in Colorado to evaluate occupational exposure (Lake et al., 2015). The facility has two rooms, both rectangular in shape (14.8 m by 6.5 m and 14.8 m by 33.7 m, respectively), separated by a sliding door and encompassing a combined area of about 595 m^2. A floor plan is shown in Figure 3.1(a). In one experiment, for example, an active engine, located in the upper left corner of the facility (black square in Figure 3.1(a)), was operating between 10:00:00 am and 11:10:00 am, while the sliding door was open. Measurements of noise intensity were collected by 18 static sensors and 2 roving sensors. The locations of the static sensors are given in Figure 3.1(a), whereas the pathways of the roving sensors are shown in Figure 3.1(b). The static sensors started collecting data at 9:45:00 am and ended at 11:23:20 am when they were turned off. The operation of the first roving sensor started at 10:28:45 am and that of the second roving sensor started at 10:52:45 am. Both roving sensors were in operation until the end of the experiment, but not continuously. Static sensor measurements were collected at every minute, while roving sensors measured hazard levels with a resolution of 20 seconds. There are thus 100 sampled points for each of the 18 static sensors, 105 sampled points for the first roving sensor and 72 sampled points for the second roving sensor, for a total of 1977 observations.

The measurements sampled over time are plotted in Figure 3.2. For each static sensor, a time series of noise intensity is plotted. The static sensors near the active engine in the upper left corner of the facility had higher intensity (gray solid lines) than those further away (dashed lines). One static sensor (#18) was far away from the active engine but had high noise intensity, due to an unexpected noise source outside the facility (black solid line). The roving sensors are also displayed as filled circles (sensor 1) or open circles (sensor 2).

The static sensors all have dense sampling points in time and thus relatively

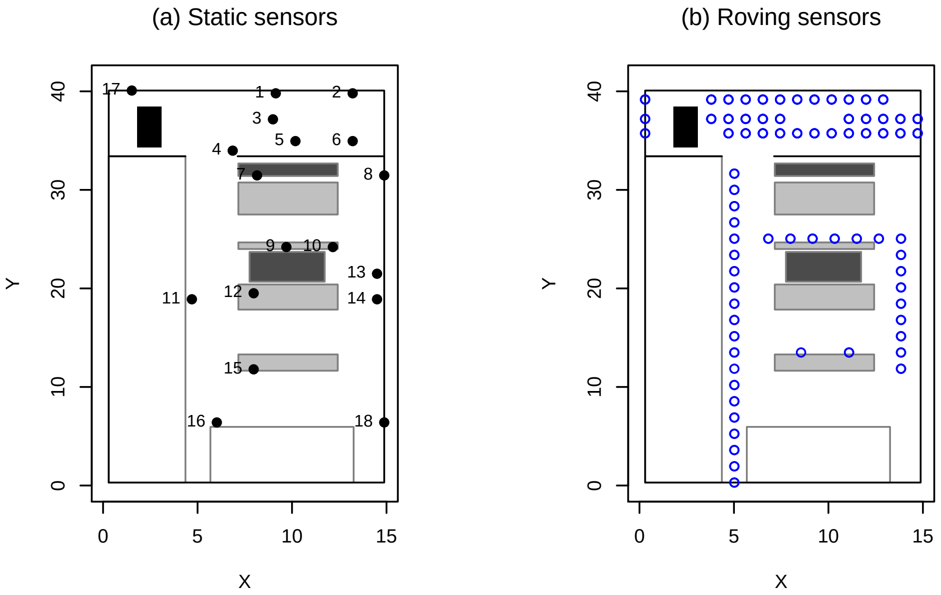

Figure 3.1: (a) Floor plan of the engine test facility. The black rectangle in the upper left corner is the source of noise, white rectangles are offices, gray rectangles are inactive engines, and dark gray rectangles are floor openings. The locations of static sensors are numbered from 1 to 18. (b) Pathway of the first roving sensor is drawn in open circles. The pathway of the second roving sensor is similar thus omitted.

complete profiles of the temporal processes at the sampling locations (Figure 3.2). However, the spatial coverage by the static sensors is limited to the 18 sampling locations where they were installed (Figure 3.1(a)). In contrast, each roving sensor has a wider spatial coverage (Figure 3.1(b)), but the information at any given location is sparse in time (Figure 3.2). The sampling time points for the roving sensors are also irregularly spaced, with occasional breaks that vary from about 20 seconds to about 5 minutes.

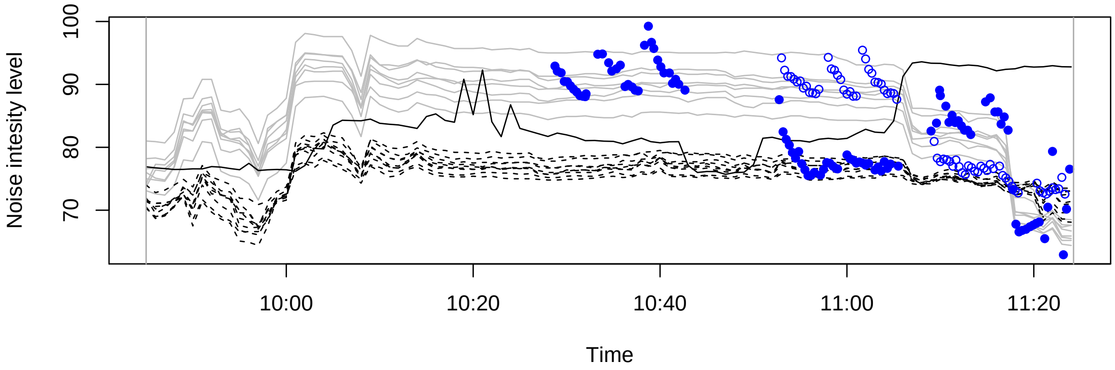

Figure 3.2: Observed noise intensity over time from 9:45:00 am to 11:23:20 am. Gray solid lines are time series for static sensors near one noise source (#1 through 7, and 17). Black solid line is near the secondary noise source (#18). Dashed lines are for the remainder static sensors. Filled and open circles are samples taken by the first roving sensor that started at 10:28:45 am and the second roving sensor that started at 10:52:20 am, respectively.

3.4.2 Current practice of hazard mapping

In a recent review of hazard mapping approaches, Koehler and Peters (2013) noted that relatively simple methods have been used to construct maps, often averaging the sensors in time before employing a spatial interpolation technique, or interpolating across space for a fixed time and averaging the maps. We show static maps commonly obtained by industrial hygienists for the case study in Figure 3.3, after averaging over all data in time at each unique sampling location. The map labeled "Roving" is based on kriging estimates and uses the roving sensors exclusively. It corresponds to early practice with DRIs and still is the most common approach. The map labeled "Roving and Static" incorporates both sources of data. The aggregated

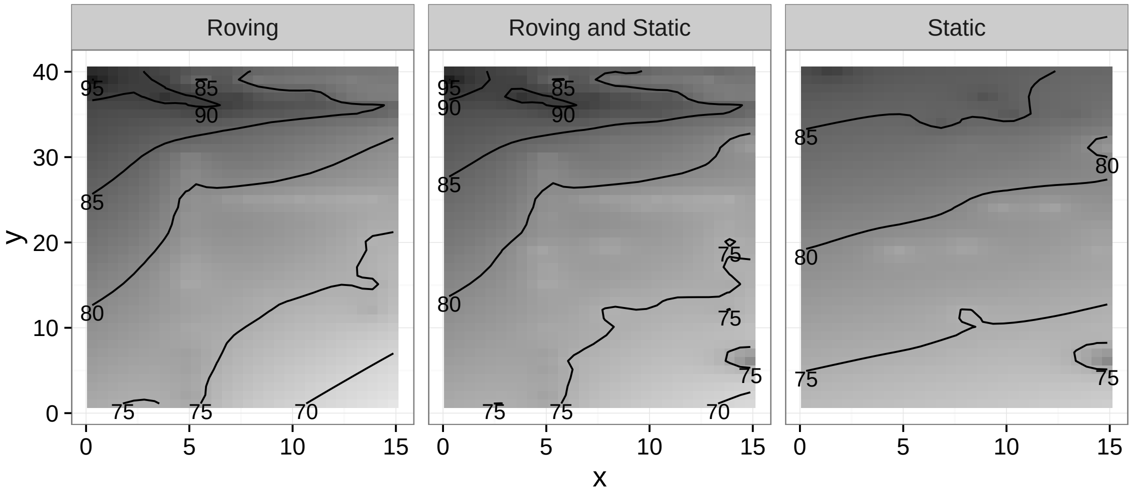

Figure 3.3: Static maps of the noise intensity obtained by kriging using the roving sensor data only (left), the roving and static sensor data (center) and the static sensor data only (right), averaging data at the same location in time.

map gives equal weights to observations available in the roving path (only one or two data points) and the ones from the static sensors (a hundred observations over the data collection period). Both maps tend to show local features that are not necessarily informative. In addition, the map labeled “Static” uses only static sensor data. It loses the local features provided by roving sensors, due to a fairly limited spatial coverage. All these static maps naturally fail to capture the evolution of the hazard level in time and, as we will demonstrate, misrepresent the intensity of a secondary noise source in the southeastern part of the facility. Health effects of short duration but high-level exposures are unclear, and the static maps in current practice have limited capacity for studying these events.

3.4.3 Model fitting

Before fitting an inhomogeneous variance model for the noise data, we selected the tuning parameters by a leave-one-sensor-out cross validation approach. More specifically, we considered a grid of values for the number of deterministic spline basis functions K, the number of temporal principal components L, and the principal components smoothing parameters ζ, and searched for a minimizer of the estimated mean squared prediction error (MSPE).

A preliminary step in the model fitting is to determine the number of basis functions K, the tuning parameter ζ, and the number of functional principal components L. We employed the leave-one-static-sensor-out algorithm to estimate the MSPE. More specifically, we considered a grid of values for the number of deterministic spline basis functions ($K = 4, 8, 12, 16$ and 32), the number of temporal principal components $L = 2, 3, 4$, and the principal components smoothing parameters $\zeta = 0, 1, 10, 100, 10^3$, and 10^4. The results for the inhomogeneous variance case are given in Table 3.1. It also displays estimates $\hat{\sigma}^2_{\mathrm{S}}$ and $\hat{\sigma}^2_{\mathrm{R}}$. There is a modest improvement from incorporating a larger L number of components, particularly at lower smoothness values for both the deterministic mean function and the random spatio-temporal effects, with diminishing returns. The deterministic spline smoothness is optimal at mid-range values for K (either 8 or 12) and the random effects do not need to be smoothed (a value of $\zeta = 0$ seems best). The minimum MSPE corresponds to $K = 12$, $\zeta = 0$, and $L = 4$. However, we decided to go with $L = 3$ functional components, as the corresponding MSPE (11.8) is close to the minimum (11.2).

There are two remarks to be made. First, the choice of smoothing reflects on the estimates of σ^2_{S} and σ^2_{R}. The estimate of σ^2_{S} changes the most as a function of the tuning parameter ζ when $\zeta > 10^2$. Second, the spline temporal trend and principal functional components temporal effects compete with each other. For example, consider the MSPE when $L = 3$ for Table 3.1. A small K minimizes MSPE when ζ is small, but conversely, allowing 32 basis functions for the trend spline requires more smoothing of the functional component ($\zeta \approx 10^4$) to minimize the MSPE.

We tuned the parameters for smoothness in the homogeneous variance case similarly. In this case, the estimated MSPE values were close, so we decided to keep the same choices for K, ζ and L for comparison. We also fitted the model without using roving sensors at all and kept the choices of tuning parameters and number of components. The selected parameters are $L = 3$, $K = 12$ and $\zeta = 0$. The estimated deterministic component is

$$\hat{\mu}_{\boldsymbol{s}}(t) = 83.64 - 0.40 s_x + 0.31 s_y + \hat{S}(t),$$

where the spline term $\hat{S}(t)$ is shown in Figure 3.7(a). The estimates of the spatial process variances are $\hat{\lambda}_1 = 13.20$, $\hat{\lambda}_2 = 8.07$, and $\hat{\lambda}_3 = 0.13$. The estimates for spatial range parameters, assuming an exponential spatial covariance model $\rho_\ell(\|\boldsymbol{s}_i - \boldsymbol{s}_{i'}\|; \boldsymbol{\theta}_\ell) = \exp\{-\|\boldsymbol{s}_i - \boldsymbol{s}_{i'}\|/\theta_\ell\}$, for $\ell = 1, 2, 3$, are $\hat{\theta}_1 = 22.34$, $\hat{\theta}_2 = 10.83$, and $\hat{\theta}_3 = 40.34$. The nonparametric estimates of the temporal functions $\varphi_\ell(\cdot)$ are shown in Figure 3.7(b). The estimated measurement error variances are $\hat{\sigma}^2_{\mathrm{S}} = 1.49$ and $\hat{\sigma}^2_{\mathrm{R}} = 1.05$. The parameter estimates for the homogeneous variance case and for the case in which only the static sensor data are used are given in Table 3.2. The inhomogeneous variance case is denoted by STDF (which stands for spatio-temporal data fusion), the homogeneous variance case by STDFh, and the scenario with static sensors only by STDF*.

Table 3.1: Choosing the best tuning parameters for the STDF algorithm: estimates of MSPE, σ^2_{S} and σ^2_{R} based on choice of smoothness for deterministic trend (df), functional principal component smoothness (ζ) and number of principal components (L).

	K	MSPE						$\hat{\sigma}^2_{\mathrm{S}}$						$\hat{\sigma}^2_{\mathrm{R}}$					
ζ		0	1	10	10^2	10^3	10^4	0	1	10	10^2	10^3	10^4	0	1	10	10^2	10^3	10^4
$L=2$	4	14.2	14.6	16.4	18.6	18.0	18.3	3.5	3.6	3.9	4.4	4.5	4.9	1.3	1.4	1.3	1.0	1.1	1.5
	8	12.0	12.0	12.8	21.4	17.4	18.1	2.1	2.1	2.2	2.7	3.9	4.3	1.9	2.1	2.4	2.8	1.1	1.5
	12	13.1	12.6	12.8	19.3	16.5	16.9	2.0	2.0	2.2	2.7	3.2	3.7	1.9	2.1	2.2	1.8	1.0	1.6
	16	13.8	14.2	16.6	34.6	16.5	17.2	1.5	1.5	1.6	2.1	3.2	3.7	1.1	1.2	1.5	1.9	0.9	1.5
	32	37.5	37.1	35.8	33.7	32.8	29.7	1.5	1.5	1.6	1.9	2.6	3.7	0.4	0.4	0.4	0.4	0.7	5.0
$L=3$	4	12.8	13.0	13.0	13.6	17.0	18.1	2.2	2.3	2.4	2.8	4.1	4.8	1.4	1.5	1.6	1.6	1.1	1.5
	8	12.1	12.0	12.6	19.6	17.2	17.7	1.6	1.6	1.8	2.4	3.8	4.2	1.4	1.5	1.7	2.6	1.0	1.5
	12	11.8	11.9	12.4	18.6	16.1	16.3	1.6	1.6	1.7	2.5	3.1	3.6	1.2	1.4	1.6	1.7	0.9	1.6
	16	12.9	13.0	15.2	34.5	16.1	16.7	1.3	1.3	1.5	2.0	3.1	3.6	1.0	1.1	1.3	1.8	0.8	1.5
	32	37.6	37.1	35.5	33.0	32.3	29.5	1.4	1.4	1.5	1.8	2.4	3.6	0.3	0.3	0.3	0.4	0.6	5.1
$L=4$	4	11.5	11.4	11.8	14.0	17.0	18.0	1.7	1.8	1.9	2.5	4.1	4.8	1.5	1.7	1.9	2.1	1.0	1.5
	8	11.4	11.5	12.4	19.3	17.3	17.6	1.5	1.5	1.7	2.3	3.7	4.2	1.2	1.4	1.5	2.5	0.9	1.5
	12	11.2	11.4	12.2	18.0	16.3	16.3	1.4	1.4	1.6	2.4	3.1	3.6	1.1	1.3	1.4	1.5	0.8	1.5
	16	12.7	13.0	15.1	34.4	16.2	16.7	1.3	1.3	1.4	1.9	3.1	3.6	0.9	1.0	1.2	1.8	0.7	1.4
	32	38.0	37.7	36.3	33.3	32.9	29.4	1.3	1.3	1.4	1.8	2.4	3.6	0.3	0.3	0.3	0.3	0.3	5.1

A series of dynamic hazard maps for the predicted noise intensity using our STDF model in Figure 3.4 show overall low intensity levels at the beginning and near the end of the study (panels 09:50 am and 11:20 am). They identify peaks in most time transects around the spatial coordinates $s_x = 2.5$ and $s_y = 35$, where the noise source is located. The noise intensity decreases as the readings are made further away from the noise source. In addition, a secondary noise source located near sensor #18 is captured (see also Figure 3.1(a)). The standard errors of the predicted noise intensity across time and space are plotted in Figure 3.5. As expected, lower standard errors are found near the static sensors and along the trajectories of the roving sensors, shown in Figure 3.1. The prediction standard errors are larger in areas without sensors at the end of the experiment (at around 11:20 am). This corresponds to the time period in which the main source of noise was turned off, and the secondary noise source increased in intensity, shown in Figure 3.2.

Figure 3.6 maps the estimated $(\hat{\xi}_\ell(\boldsymbol{s}_1^*), \ldots, \hat{\xi}_\ell(\boldsymbol{s}_m^*))', \ell = 1, 2, 3$ over a fine grid of spatial locations $\boldsymbol{s}_1^*, \ldots, \boldsymbol{s}_m^*$, with each ℓ loading standardized to have zero mean and standard deviation one. The temporal components $\hat{\varphi}_\ell(t), \ell = 1, 2, 3$ are shown in Figure 3.7(b). The interpretation of the the temporal functions $\varphi_\ell(\cdot)$ can be made in light of Figure 3.2, as well as the Karhunen-Loève expansion based on (3.5). The first temporal component $\hat{\varphi}_1$ is nearly constant around 1 (Figure 3.7). The corresponding $\hat{\xi}_1(\boldsymbol{s})$ shown in Figure 3.6 shows higher noise intensity values in the northern room, and near the secondary source around sensor #18. For the second temporal component, since $\hat{\varphi}_2$ is negative until approximately 11:10 am, it subtracts the $\hat{\xi}_2(\boldsymbol{s})$ effect from $\hat{\xi}_1(\boldsymbol{s})$, but after 11:10 am it adds the effect of the secondary noise source. Thus, in the dynamic hazard maps (Fig. 3.4), the noise intensity is higher in the northern room

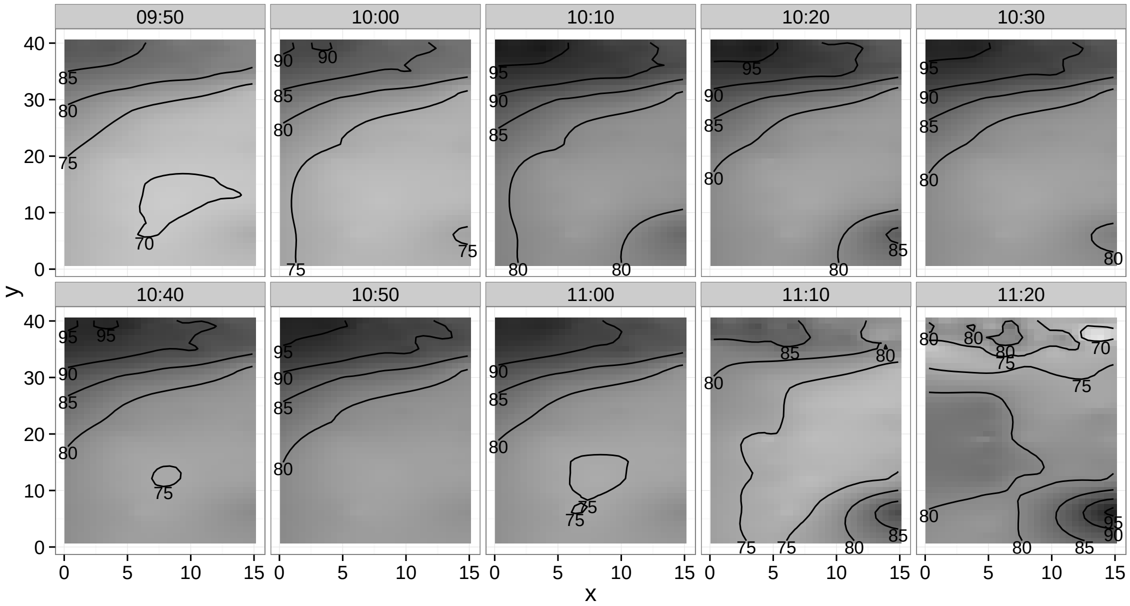

Figure 3.4: Dynamic hazard maps with contour lines obtained from the spatio-temporal data fusion (STDF) method; each panel corresponds to a point in time from 9:50 am to 11:20 am at 10-minute intervals.

before 11:10 am, and afterwards the noise intensity is higher near sensor #18. The third temporal component $\hat{\varphi}_3$ is associated with a much smaller variance ($\hat{\lambda}_3 = 0.13$), and does not impact the hazard map as much as the first two components. The roles of $\hat{\varphi}_3$ and $\hat{\xi}_3(\boldsymbol{s})$ seem to be compensating $\hat{\xi}_1(\boldsymbol{s})$ and $\hat{\xi}_2(\boldsymbol{s})$ at the very beginning and the very end of the experiment, in order for the hazard map to be closer to background noise.

The standard errors for the parameter estimates are obtained via cross validation by leaving one sensor out at each time, as detailed in Appendix 3.A, and are shown in Table 3.2. While the main objective of Table 3.2 is to quantify the uncertainty regarding the parameter estimates, we note that the large range parameter estimate

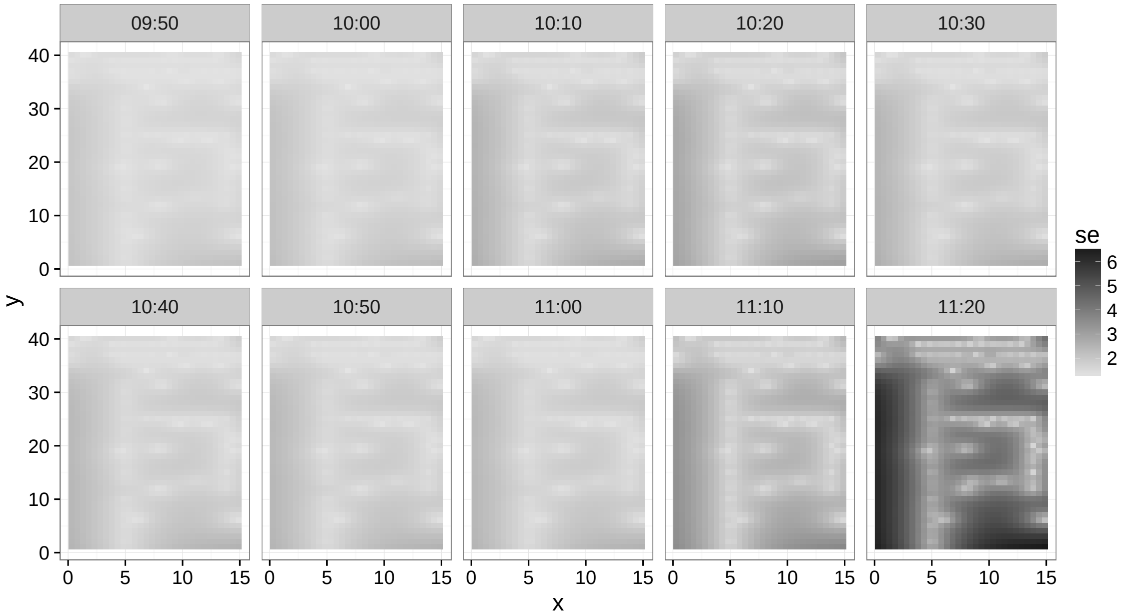

Figure 3.5: Prediction standard error maps for the dynamic hazard maps given in Figure 3.4; each panel corresponds to a point in time from 9:50 am to 11:20 am at 10-minute intervals.

$\hat{\theta}_3$ reflects a relatively weak spatial dependence. Consequently the standard error for $\hat{\theta}_3$ also might be inflated by static sensors within a certain radius of each other being removed during the cross-validation step. Table 3.2 also displays how the roving sensors affect the estimates of the spatial dependence coefficients. When only the static sensors are used, the range parameter estimates $\hat{\theta}_2$ and $\hat{\theta}_3$ are similar. However, when the roving sensors are included, the estimate $\hat{\theta}_2$ becomes quite a bit smaller than $\hat{\theta}_3$ and is more informative, in the sense that the spatial effect of the secondary noise source becomes more prominent owing to the roving sensors near this secondary source toward the end of the experiment. This illustrates that the inclusion of roving sensors adds information about the spatial dependence at finer scales.

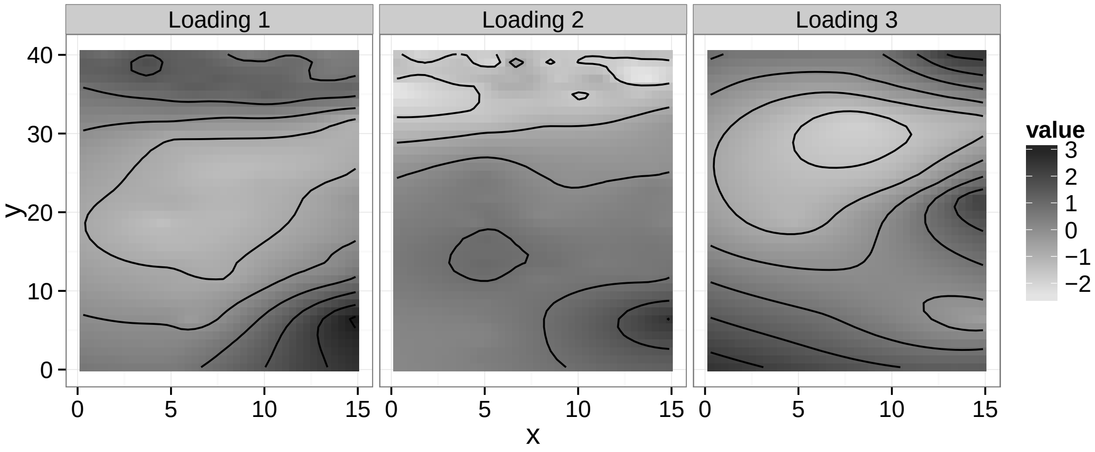

Figure 3.6: Estimated $(\hat{\xi}_\ell(\boldsymbol{s}_1^*), \ldots, \hat{\xi}_\ell(\boldsymbol{s}_m^*))', \ell = 1, 2, 3$. Each ℓ loading is standardized to have zero mean and standard deviation one. The associated variability coefficients are $\hat{\lambda}_1 = 13.20$, $\hat{\lambda}_2 = 8.07$ and $\hat{\lambda}_3 = 0.13$.

In addition to maps displaying the intensity of a hazard, an informative representation of the hazard can be made by plotting the probability of a hazard exposure exceeding a threshold. The marginal probabilities can be obtained by computing the z-scores using the hazard map and standard error maps, and evaluating the probability of exceedance. To illustrate, Figure 3.8 shows the probability that the hazard levels exceed 85 dB, with contour lines at 0.95.

3.4.4 Scientific implications

Both static (Figure 3.3) and dynamic (Figure 3.4) hazard maps capture the high noise intensity in the northern room near the working engine. However, the dynamic hazard maps also show that the average intensity of hazard exposure in the southern room increases during the period between 10:00 am and 11:00 am, when the engine in the northern room is turned on. Further, the outside noise source is detected in the southeastern corner, after 11:10 am. In particular, the intensity levels exceed 85 dB,

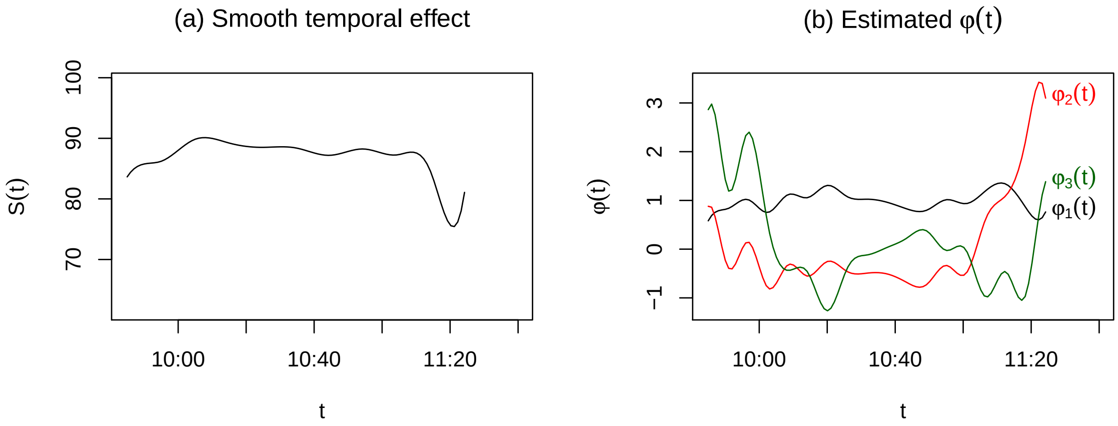

Figure 3.7: (a) Deterministic effect for the temporal component estimated with a spline function (b) $\hat{\varphi}_\ell(t)$, $\ell = 1, 2, 3$ components, corresponding to a random effect in the temporal component.

which is generally viewed as a harmful level for exposures longer than 8 hours. While the static maps do detect a small noise peak near the same location, the predicted intensity levels are understated.

The problem of interpolating hazard maps in time and space from discrete sampled observations was discussed in Koehler and Peters (2013). While kriging is well accepted for spatial interpolation in the occupational hygiene literature, often the observations or the hazard maps are averaged in time to produce estimates at unsampled time points and thus the temporal aspect of the data is ignored. To compare our method with the existing approaches, we consider kriging using the roving data only and kriging while incorporating roving and static sensor data, both averaging the observations in time. We also consider fixed-time universal kriging (UK), fixed-time thin-plate spline (TPS), and fixed-time simple linear regression on the spatial coordinates (LM). By fixed-time we mean that a time point is fixed and the spatial map is

Table 3.2: Parameter estimates, and cross-validated standard errors (in parenthesis). STDF denotes spatio-temporal data fusion for the inhomogeneous variance case, STDFh for the homogeneous variance case, and STDF* for only the static sensors.

Coefficient	STDF	STDFh	STDF*
β_0	83.64 (4.52)	83.53 (4.49)	82.23 (3.59)
β_x	-0.40 (0.08)	-0.40 (0.08)	-0.38 (0.11)
β_y	0.31 (0.04)	0.31 (0.04)	0.31 (0.04)
θ_1	22.34 (2.93)	22.20 (2.95)	13.17 (1.83)
θ_2	10.93 (2.13)	12.08 (1.88)	30.99 (2.57)
θ_3	40.34 (11.15)	40.34 (11.10)	32.56 (8.42)
σ^2_{S}	1.49 (0.16)	1.48 (0.15)	1.46 (0.12)
σ^2_{R}	1.05 (0.29)	–	–

constructed for the data corresponding to the time transect selected, thus preserving some of the temporal structure from the data.

To compare the methods globally, the MSPE values for our spatio-temporal data fusion method are obtained using the leave-one-sensor-out cross-validation described Appendix 3.A. We generate prediction maps at every time point in which the static sensors were sampled, and averaged the values across space and over time. When including the roving sensors, the inhomogeneous variance case (STDF) has an MSPE of 11.82, and the homogeneous version (STDFh) has an MSPE of 11.66. When using only the static sensors (STDF*), the MSPE is 14.56. The MSPE for the static map using roving sensors only is 37.42, and the MSPE for the static map incorporating roving and static sensor data is 38.33. This shows a clear advantage of our method over the current practice of using static maps. For the other alternative approaches, the MSPE for fixed-time universal kriging (UK) is 24.4, thin-plate spline regression (TPS) is 14.99, and simple linear regression model (LM) is 25.14, all of which are outperformed by our method. We observe that the homogeneous variance case gives

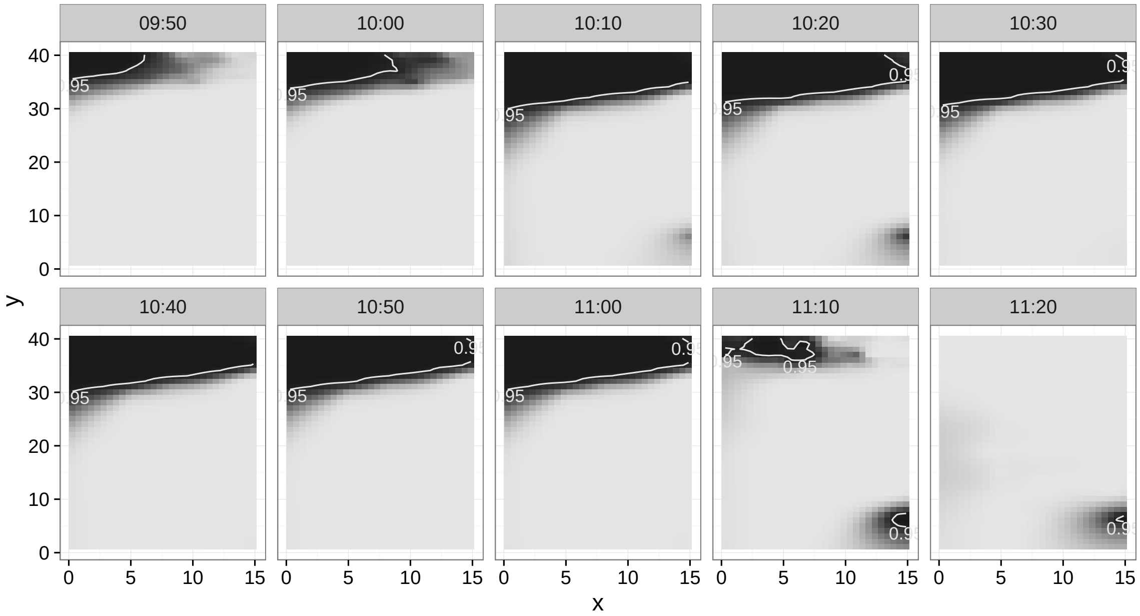

Figure 3.8: Probability of noise exceeding 85 dB map; darker areas indicate higher probability, and the contour lines at 0.95 probability are included. Each panel corresponds to a point in time from 9:50 am to 11:20 am at 10-minute intervals.

similar predictions to the inhomogeneous variance case.

In Figure 3.9 we focus on a small-scale example from the dataset, to illustrate the possibility of detecting short duration hazard intensity peaks. The static sensors are sampled every minute (Section 3.4.1) and thus the measurements are available between 11:04 am and 11:05 am. For the UK and TPS generated maps, we use data from 11:04 am and 11:05 am, and interpolate linearly the values from the maps at times 11:04:20 and 11:04:40. We can see in Figure 3.9 that UK maps do not capture the secondary noise source in the southeastern part of the facility. This is because the method underestimates the range of spatial dependence, and produces prediction that resemble a plane except where the static sensors are located. On the other hand, the TPS method over-smooths the data, losing local features such as the sharper distinction between the northern and southern rooms in the facility. We omit the linear model estimates, which are planes only. In addition, Figure 3.10 shows the corresponding standard error maps for each method in Figure 3.9. We observe that the STDF maps generally have lower standard errors than those of UK. The TPS only has comparable standard errors when very near the static sensors, but otherwise much higher standard errors than either STDF or UK.

A final remark is that we can consider cases in which the spatial covariance function is not exponential. For example, we repeated the analysis using the Matérn class of covariance functions, with known smoothness parameter $\nu = 2.5$. The resulting hazard maps (not shown) are similar except for slightly smoother contour lines. We anticipate that our method is robust to the choice of a covariance function in the Matérn class.

3.5 Discussion

In this chapter, we have developed a spatio-temporal static and roving data fusion model, with each data sensor having potentially different instrument variances. The approach to model fitting and statistical inference has been applied to produce hazard maps that capture dependence across space and over time in indoor environments. Modeling the spatio-temporal dependence structure allows the hazard maps to capture features that are missed by the current practice in occupational hazard assessment. Furthermore, our approach enables continuous-time prediction of hazard, which the existing approaches are unable to produce.

With the semiparametric model specification, our method is able to detect unexpected hazard sources that occur sporadically during a study. A sudden fluctuation of intensity, such as the secondary noise source in the southeastern corner of the facility, are undetected or underestimated when using current practices but can be detected by our method. Moreover, health effects of short duration but high-level exposures are unclear, and our method provides a way to better capture such transient exposures.

Cross validation shows that our methodology outperforms the traditional methods in the scientific application, a conclusion that is corroborated by the simulation study given in Appendix 3.A. The simulations have also demonstrated that our method is robust to different instrument variances, while the traditional approaches tend to provide less accurate prediction.

While the height of the sensors is not accounted for directly, the model with heterogeneous measurement error variances may accommodate possibly different heights for different sensors. It would be interesting, however, to examine this third dimension more closely, as well as to consider three-dimensional hazard maps when data are collected at different heights (see, e.g., Tracey et al., 2014).

Other covariance modeling allows for nonseparability, although stationarity in time is generally assumed (Gneiting, 2002; Ma, 2003; Quick et al., 2015). Stein (2005) proposed models that are asymmetric in time, allowing for different smoothness degrees in space over time. It may be of interest to extend such models and develop estimation methods for fusion of static and roving sensor data. For datasets of much larger sample size with more sampling locations and/or time points, the proposed profile likelihood approach would need to be improved for it to be more computationally feasible. It may be helpful to utilize some form of approximation, such as blocking or tapering, in the covariance matrix inversion. We leave this for future research as well.

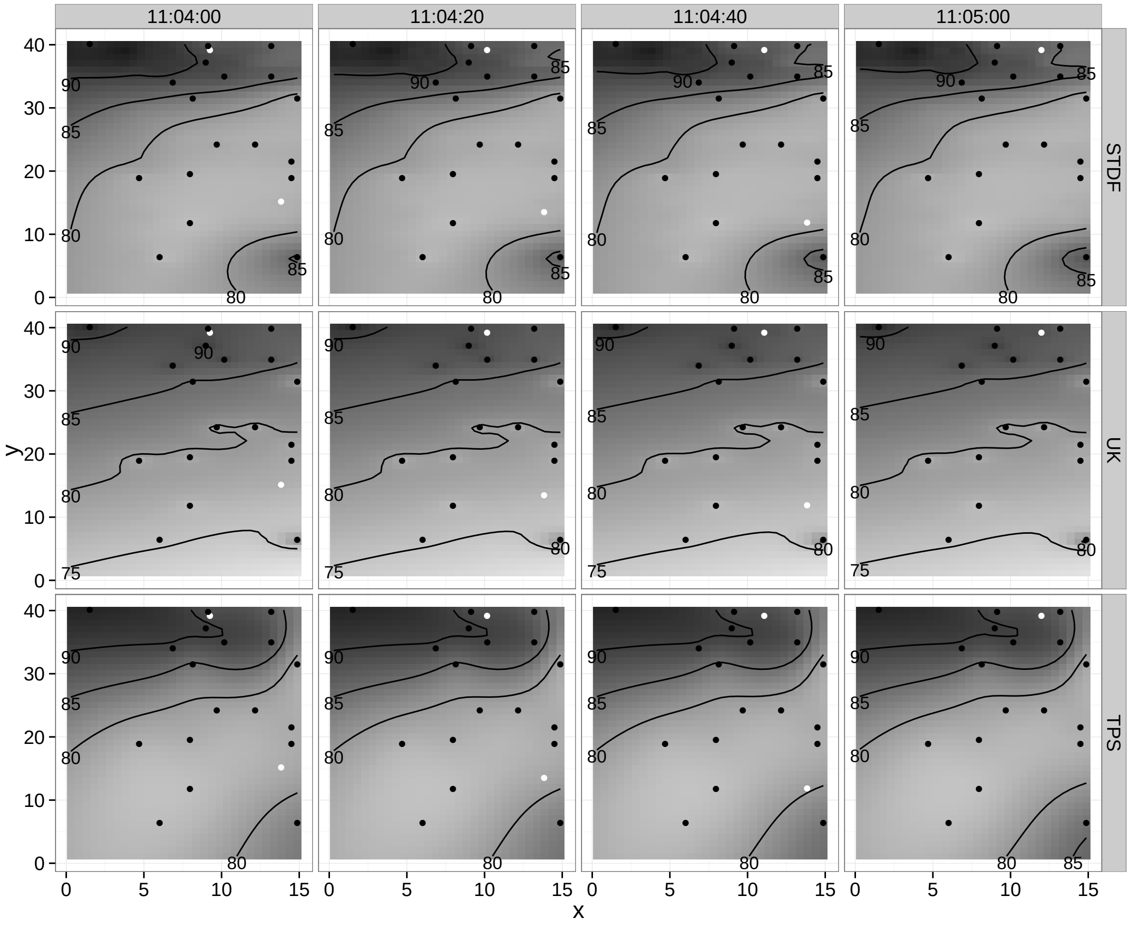

Figure 3.9: Comparison of STDF hazard map with maps created by universal kriging (UK) and thin-plate splines (TPS) methods. For the UK and TPS, data are from 11:04:00 and 11:05:00, and interpolated linearly between 11:04:20 and 11:04:40. Black points mark the static sensor locations, while white points mark the roving sensor locations at the corresponding time.

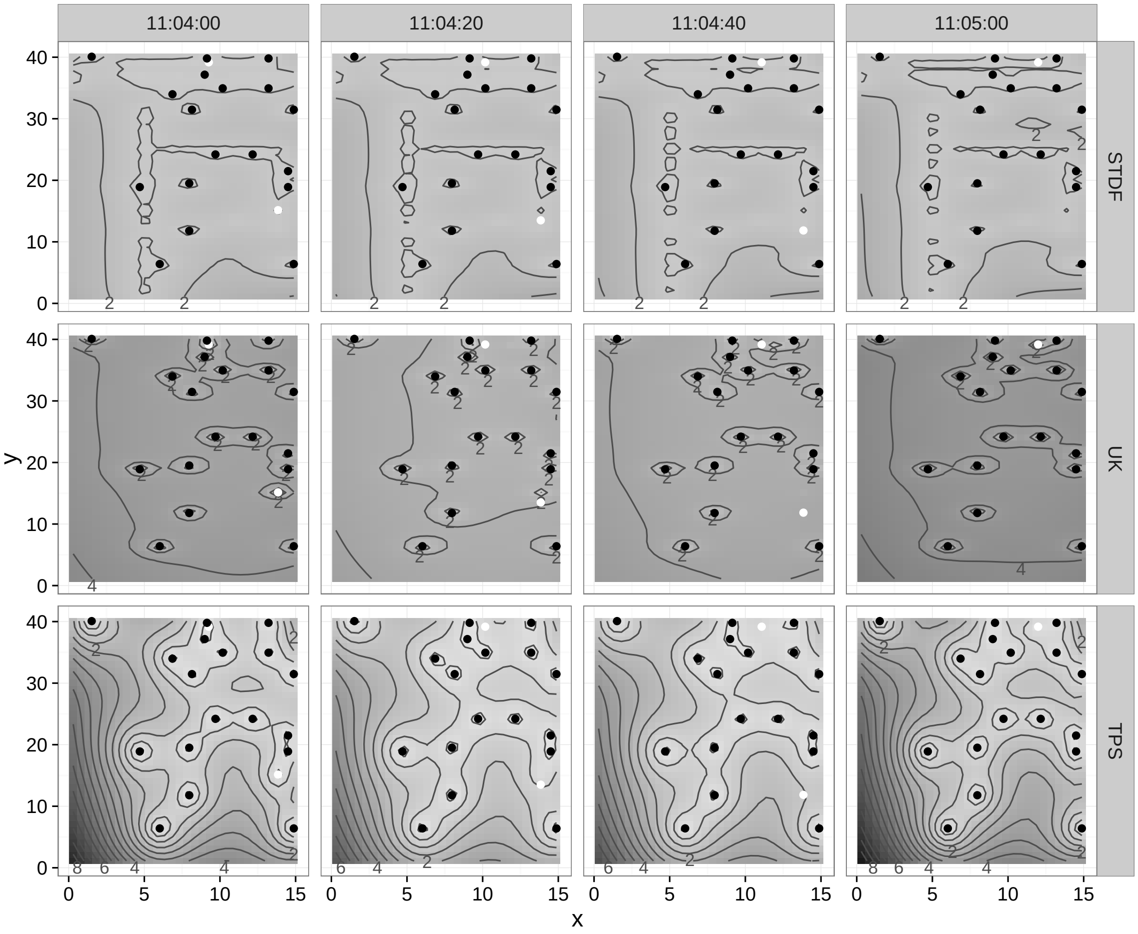

Figure 3.10: Comparison of standard errors of STDF hazard maps with standard error of universal kriging (UK) and thin-plate splines (TPS) methods. For UK and TPS, data are from 11:04:00 and 11:05:00, and interpolated linearly between 11:04:20 and 11:04:40. Black points mark the static sensor locations, while white points mark the roving sensor locations at the corresponding time.

Appendix

3.A Simulation study

Our purpose in this simulation is to study the quality of mapping, the effect of fusing roving and static sensor data in terms of prediction, and to compare our methodology with some of the existing approaches (Koehler and Peters, 2013). There are two models for the deterministic component $\mu_{\boldsymbol{s}}(t)$, three models for the covariance functions of $\eta_{\boldsymbol{s}}(t)$ and six configurations of sensors: a combination of either 6 or 18 static sensors, and either 0, 1 or 2 roving sensors. The location of the sensors (roving, static and those reserved for MSPE calculation) are displayed in Figure 3.A.1; the static sensor and prediction sensor locations were once generated randomly, but fixed across independent experiments. We included three scenarios for the measurement error variances $(\sigma^2_{\mathrm{S}}, \sigma^2_{\mathrm{R}})$ of the static and roving sensors: the first in which they are equal, a second in which $\sigma^2_{\mathrm{R}} = \sigma^2_{\mathrm{S}}/4$ and a third in which $\sigma^2_{\mathrm{R}} = 4\sigma^2_{\mathrm{S}}$. Each sensor was observed for 60 units of time with complete observations for all sensors, and each experiment was replicated, independently, 200 times.

In all scenarios our STDF method was fitted with the same tuning parameters. The number of basis functions used for the deterministic mean function is 3, and the number of basis functions for the random spatio-temporal effects part is 10, with

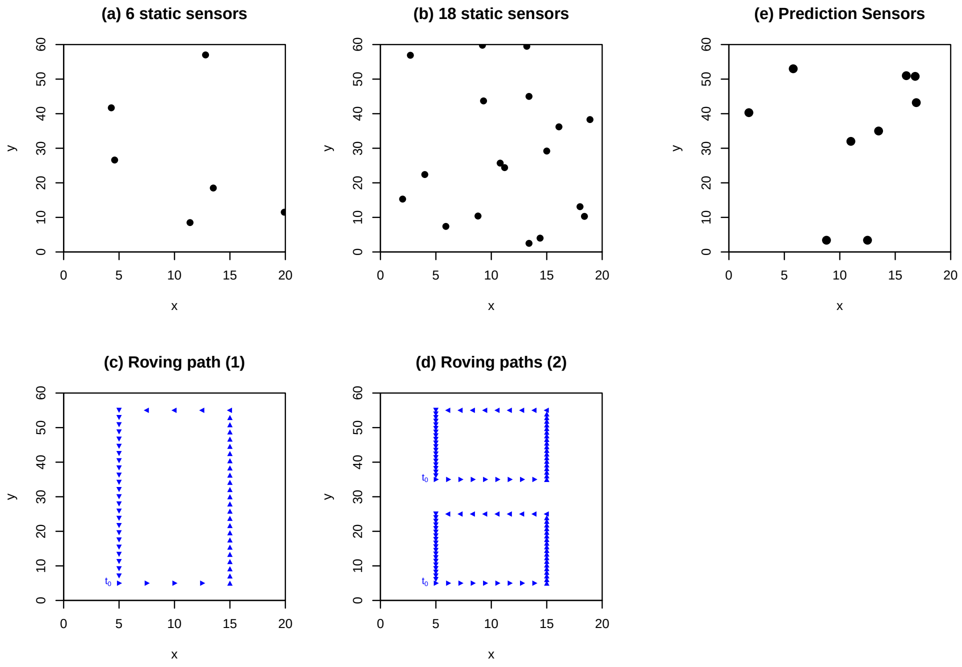

Figure 3.A.1: Sensor configurations used in the simulation. (a) and (b): Static sensors, either 6 or 18. (c) and (d): Roving sensors, either 0 (no picture), 1 (top) or 2 (bottom). (e): Static points for which the mean squared prediction error (MSPE) was calculated.

smoothness parameter $\zeta = 0$. We also evaluated the effect of fitting models with homogeneous and inhomogeneous variances, the former of which is denoted by STDFh.

The mean functions used were either linear in time and space, with

$$\mu_A(\boldsymbol{s}, t) = 40 - s_x + s_y/2 - t/5,$$

or linear in space, nonlinear in time, with

$$\mu_B(\boldsymbol{s}, t) = 40 - s_x + s_y/2 - t/5 + 8\sin(2\pi t/60).$$

We allowed the deterministic time to be (potentially) nonlinear by including a spline term in the least squares detrending step.

For the spatio-temporal components, we used $\varepsilon_{\boldsymbol{s}}(t) \sim N(0, 1^2)$, $\varepsilon_{\boldsymbol{r}}(t) \sim N(0, \gamma_k 1^2)$,

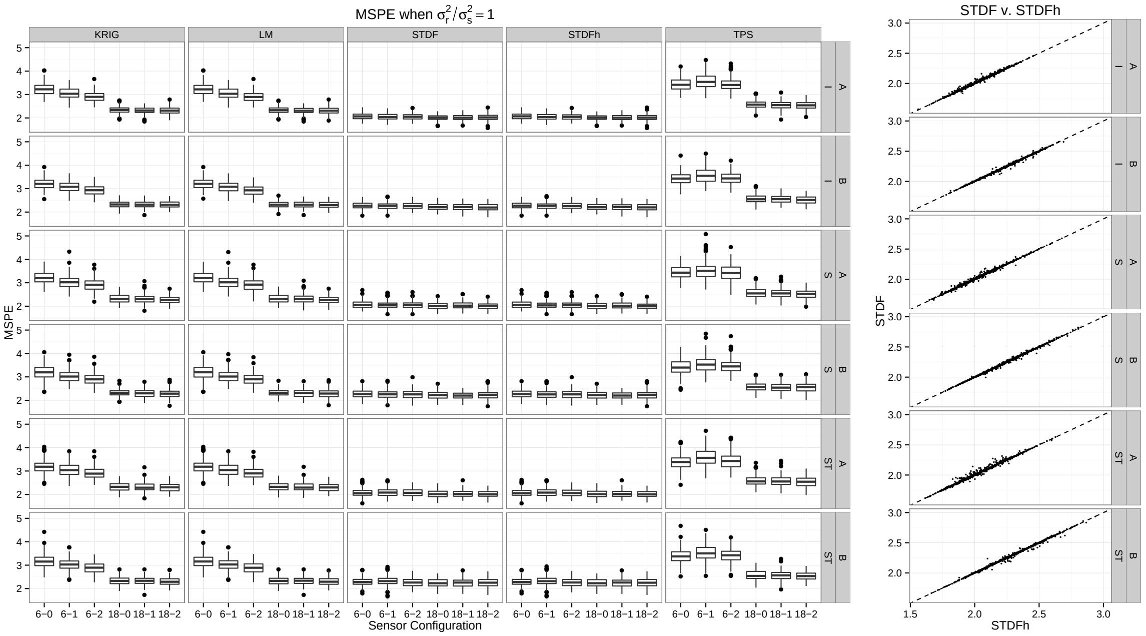

Figure 3.A.2: Left: Comparison of MSPE averaged in space and time in simulation study; x-axis is labeled according to static-roving sensor combination; strips in the top part of the panel indicate the model fitted, and strips in the right part of the panel indicate which deterministic mean and covariance model was used. In all cases, $\sigma^2_{\mathrm{R}} = \sigma^2_{\mathrm{S}}$. Right: Scatterplot of the MSPE results for the STDF model and the STDFh (STDF homogeneous) model. The observed ratio of MSPEs is 0.9950 on average.

$\gamma_k = 1/4, 1, 4$, and $\eta_{\boldsymbol{s}}(t)$ is a Gaussian process with mean 0 and three choices of covariance function $\sigma(\boldsymbol{s}, t, \boldsymbol{s}', t')$:

(I) Independent case, in which $\eta(\boldsymbol{s}, t) = 0$.

(S) Spatially correlated case, in which $\sigma_\eta(\boldsymbol{s}, t, \boldsymbol{s}', t') = \sigma^2_\eta e^{-\|\boldsymbol{s}-\boldsymbol{s}'\|/\theta_{\boldsymbol{s}}}\delta(t = t')$, where δ is the indicator function.

(ST) Spatio-temporal case, with $\sigma_\eta(\boldsymbol{s}, t, \boldsymbol{s}', t') = \sigma^2_\eta e^{-\|\boldsymbol{s}-\boldsymbol{s}'\|/\theta_{\boldsymbol{s}}} e^{-|t-t'|/\theta_t}$.

The traditional methods considered here were universal kriging (UK), thin-plate spline (TPS), and linear regression (LM) at transects in time. More specifically, we

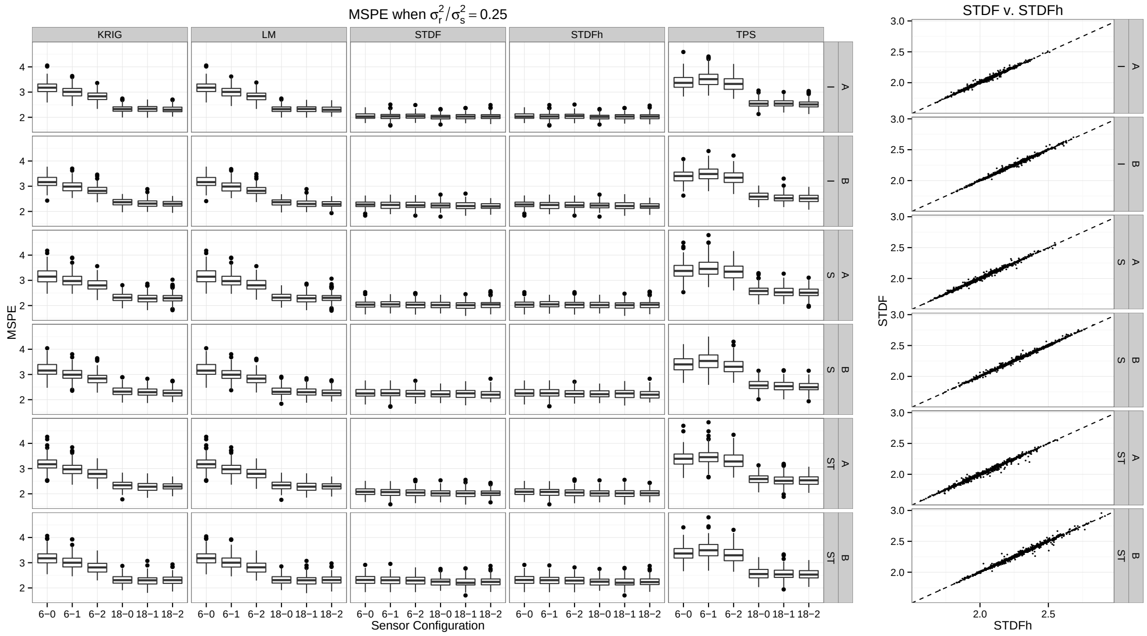

Figure 3.A.3: Left: Comparison of MSPE averaged in space and time in simulation study; x-axis is labeled according to static-roving sensor combination; strips in the top part of the panel indicate the model fitted, and strips in the right part of the panel indicate which deterministic mean and covariance model was used. In all cases, $\sigma^2_{\mathrm{R}} = 0.25\sigma^2_{\mathrm{S}}$. Right: Scatterplot of the MSPE results for the STDF model and the STDFh (STDF homogeneous) model. The observed ratio of MSPEs is 0.9994 on average.

fix points in time and fit the traditional methods to the corresponding observations. Predictions are made based only on the observations at the same time point. The first two were fitted using the default specification in the `fields` R package (Nychka et al., 2014). That is, for each fixed time, each of the methods was applied. The reported mean squared prediction errors (MSPE) are averages in time.

The MSPE results are compared across models in Figures 3.A.2, 3.A.3 and 3.A.4, with the former corresponding to the case when $\sigma^2_{\mathrm{S}} = \sigma^2_{\mathrm{R}}$, the second with the case when $0.25\sigma^2_{\mathrm{S}} = \sigma^2_{\mathrm{R}}$ and the latter when $4\sigma^2_{\mathrm{S}} = \sigma^2_{\mathrm{R}}$. The results show that the STDF

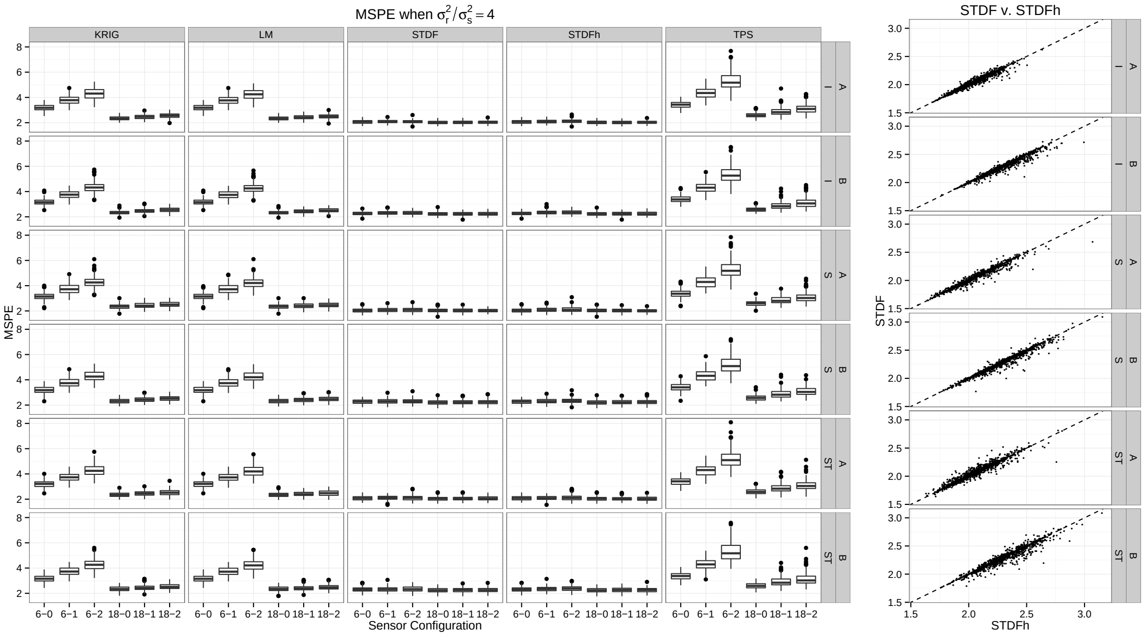

Figure 3.A.4: Left: Comparison of MSPE averaged in space and time in simulation study; x-axis is labeled according to static-roving sensor combination; strips in the top part of the panel indicate the model fitted, and strips in the right part of the panel indicate which deterministic mean and covariance model was used. In all cases, $\sigma_{\mathrm{R}}^2 = 4\sigma_{\mathrm{S}}^2$. Right: Scatterplot of the MSPE results for the STDF model and the STDFh (STDF homogeneous) model. The observed ratio of MSPEs is 0.9471 on average.

method outperforms the traditional methods in basically all scenarios. The results also reveal that our method is robust when roving sensors have variance greater than the static sensors (Figure 3.A.4). The STDF method has similar performance for the homogeneous and inhomogeneous specifications when the measurement error variances are the same for static and roving sensors. The inhomogeneous case performs better when the roving sensor variances are larger, as shown in the right scatterplot panel for Figure 3.A.4.

Chapter 4

Discussion and future work

4.1 Scalability and computational issues in data fusion problems

An important question about the spatio-temporal data fusion problem is of scalability. It is possible to envision applications in which the hazard maps must be obtained very fast, perhaps in real time, and using massive datasets. While the approach described in Chapter 3 does have a fast implementation in the optimization step, it still is a bottleneck for the procedure. Research for scalable spatio-temporal models often involves approaches such as low rank approximations and tapering (see, e.g., Sang and Huang, 2012). On the other hand, Stein (2014) criticizes low rank approximations of spatial covariance functions for their tendency to smooth out important spatial features of the data. We believe our spatial-temporal data fusion approach can be improved in terms of theory-driven computational efficiency, while being insulated from Stein (2014)'s criticism due to the spatially rich coverage of the roving sensors.

4.1.1 Advantages and issues with static data only maps

Consider the case where only static data is collected. The model for Σ is

$$\boldsymbol{\Sigma} = \boldsymbol{\Phi}'\boldsymbol{\Lambda}\boldsymbol{\Phi} + \sigma^2\boldsymbol{I}_N, \tag{4.1}$$

where $N = \sum_{i=1}^{n_\mathrm{S}} p_i$ is the total sample size for static sensor data, n_S is the number of static sensors, and $\boldsymbol{I}_N$ is the N-dimensional identity matrix. Here $\boldsymbol{\Phi}$ is a $(L \cdot n_S) \times N$ matrix, and $\boldsymbol{\Lambda}$ is $(L \cdot n_S) \times (L \cdot n_S)$. We refer to the definitions in Chapter 3. We observe that $\boldsymbol{\Lambda}$ is of rank no greater than $L \cdot n_S$ (exactly $L \cdot n_S$ as long as all static sensors have distinct locations and no $\theta_\ell = 0, \ell = 1, \dots, L$) and since the number of static sensors is small, then $(L \cdot n_S) \ll N$. To show the $\boldsymbol{\Lambda}$ structure explicitly, if we let $L = 2$ and $n_S = 2$, then

$$\boldsymbol{\Lambda} = \begin{pmatrix} \lambda_1 & 0 & \lambda_1\rho_1(\|\boldsymbol{s}_1 - \boldsymbol{s}_2\|; \theta_1) & 0 \\ & \lambda_2 & 0 & \lambda_2\rho_2(\|\boldsymbol{s}_1 - \boldsymbol{s}_2\|; \theta_2) \\ & & \lambda_1 & 0 \\ & & & \lambda_2 \end{pmatrix}. \tag{4.2}$$

As such, it is clear that if we consider only static sensors, $\boldsymbol{\Lambda}$ will be necessarily a low rank matrix. While this is desirable, for instance, as it allows the fast evaluation of $\boldsymbol{\Sigma}^{-1}$, on the other hand the spatial smoothing from low rank models discussed in Stein (2014) becomes present.

Stein (2014) considered matrices of structure

$$\boldsymbol{\Sigma} = \mathbf{A}'\mathbf{B}\mathbf{A} + \sigma^2\mathbf{D}, \tag{4.3}$$

where $\mathbf{B}$ is low rank (or rather, low dimensional with complete rank) and $\mathbf{D}$ is sparse (of complete rank). It is possible to obtain $\boldsymbol{\Sigma}^{-1}$ explicitly by using Sherman-Morrison-Woodbury formula.

$$\boldsymbol{\Sigma}^{-1} = \sigma^{-2}\mathbf{D}^{-1} - \sigma^{-4}\mathbf{D}^{-1}\mathbf{A}'\left(\mathbf{B}^{-1} + \sigma^{-2}\mathbf{A}\mathbf{D}^{-1}\mathbf{A}'\right)^{-1}\mathbf{A}\mathbf{D}^{-1}$$

which is (computationally) easy to calculate since $\mathbf{D}$ is sparse and $\mathbf{B}$ is low dimensional. The cost of evaluating $\mathbf{D}^{-1}$ is negligible so the cost of evaluating $\boldsymbol{\Sigma}^{-1}$ is roughly the same of evaluating $\mathbf{B}^{-1}$. Stein (2014) argued that for matrices of form (4.3), if we consider $r_n = \text{rank}(\mathbf{B})$, three cases of interest arise:

1. $r_n = Dn + O(1)$ as $n \to \infty$ for $0 < D < 1$.
2. $r_n = Dn^{\delta} + O(1)$ as $n \to \infty$ for $D > 0$ and $0 < \delta < 1$.
3. $r_n = r_0$ as $n \to \infty$ (i.e. constant).

Case 1 is not of a low rank smoother. Cases 2 and 3, however, do show worse performance (in terms of Kullback-Leibler divergence) in comparison to using an independent blocks approximation, for massive datasets.

For the model (4.1), with static sensors only, first assume we have fixed L. If we increase sampling in time but keep the number of static sensors constant, there is a trade-off: our model is scalable as long as the number of static sensors is small, but we fall into Stein's worse case scenario. However, if the number of observations in time is fixed and we increase the number of sensors, then the model belongs to Stein's case 1. However this makes the model harder to scale (since the dimension of $\boldsymbol{\Lambda}$ is $L \cdot n_S$). It is also unrealistic in the context of the application.

A further complication is that we are tuning L via cross-validation. It is an open problem to determine the large sample behavior of the cross-validation choice of L.

4.1.2 Roving and static data

Having roving sensors in the model changes matters substantially. Now, we have that

$$\boldsymbol{\Sigma} = \boldsymbol{\Phi}'\boldsymbol{\Lambda}\boldsymbol{\Phi} + \sigma_{\mathrm{S}}^2\boldsymbol{D}_{\mathrm{S}} + \sigma_{\mathrm{R}}^2\boldsymbol{D}_{\mathrm{R}}, \tag{4.4}$$

where $N = \sum_{i=1}^{n_{\mathrm{S}}} p_i + \sum_{j=1}^{n_{\mathrm{R}}} q_j$ is the total sample size combining static and roving sensor data and $\boldsymbol{I}_N$ is the N-dimensional identity matrix. While $\sigma_{\mathrm{S}}^2\boldsymbol{D}_{\mathrm{S}} + \sigma_{\mathrm{R}}^2\boldsymbol{D}_{\mathrm{R}}$ is still a sparse matrix (it is diagonal), $\boldsymbol{\Lambda}$ is now fairly complicated, with rank no greater than $L \cdot n_S + L\sum_{j=1}^{n_{\mathrm{R}}} q_j$. This happens since a static sensor has a different location for each point in time. Considering the example in (4.2), let now also $n_R = 1$ and $q_1 = 2$; observe that

$$\boldsymbol{\Lambda} = \left(\begin{array}{ccccc}
\lambda_1 & 0 & \lambda_1\rho_1(\|\boldsymbol{s}_1 - \boldsymbol{s}_2\|;\theta_1) & 0 & \lambda_1\rho_1(\|\boldsymbol{s}_1 - \boldsymbol{r}_{1,1}\|;\theta_1) \\
 & \lambda_2 & 0 & \lambda_2\rho_2(\|\boldsymbol{s}_1 - \boldsymbol{s}_2\|;\theta_2) & 0 \\
 & & \lambda_1 & 0 & \lambda_1\rho_1(\|\boldsymbol{s}_2 - \boldsymbol{r}_{1,1}\|;\theta_1) \\
 & & & \lambda_2 & 0 \\
 & & & & \lambda_1 \\
 & & & & \\
 & & & & \\
 & & & &
\end{array}\right. \cdots$$

$$\cdots \left.\begin{array}{ccc}
0 & \lambda_1\rho_1(\|\boldsymbol{s}_1 - \boldsymbol{r}_{2,1}\|;\theta_1) & 0 \\
\lambda_2\rho_2(\|\boldsymbol{s}_1 - \boldsymbol{r}_{1,1}\|;\theta_2) & 0 & \lambda_2\rho_2(\|\boldsymbol{s}_1 - \boldsymbol{r}_{2,1}\|;\theta_2) \\
0 & \lambda_1\rho_1(\|\boldsymbol{s}_2 - \boldsymbol{r}_{2,1}\|;\theta_1) & 0 \\
\lambda_2\rho_2(\|\boldsymbol{s}_2 - \boldsymbol{r}_{1,1}\|;\theta_2) & 0 & \lambda_2\rho_2(\|\boldsymbol{s}_2 - \boldsymbol{r}_{2,1}\|;\theta_2) \\
0 & \lambda_1\rho_1(\|\boldsymbol{r}_{1,1} - \boldsymbol{r}_{2,1}\|;\theta_1) & 0 \\
\lambda_2 & 0 & \lambda_2\rho_2(\|\boldsymbol{r}_{1,1} - \boldsymbol{r}_{2,1}\|;\theta_2) \\
 & \lambda_1 & 0 \\
 & & \lambda_2
\end{array}\right)$$

We can argue that $\boldsymbol{\Lambda}$ rank is based on the number of unique locations visited by the roving sensors. It is clear that rank($\boldsymbol{\Lambda}$) increases also as a function of time. This can be developed to formalize the statement that "including more roving sensors increase the spatial coverage of data". Under some mild assumptions – we conjecture that having space-filling movement patterns for the roving sensors within the region of interest could be sufficient – it might be possible to show that the method belongs to case 1 of Stein's taxonomy. We also conjecture that the constant "D" in Stein's asymptotics is, in this case, given by the proportion of roving sensor data in the dataset. However, this comes at the cost of (easy) scalability, since Sherman-Morrison-Woodbury formula does not help in evaluating $\hat{\mathbf{\Sigma}}^{-1}$ if the rank of $\mathbf{\Lambda}$ is large. To determine an appropriate method for scalability of the data constitutes a question for future research.

4.2 Spatial confounding in non-Gaussian regression models

4.2.1 Spatial confounding in generalized linear models

Spline-based semiparametric generalized linear models (GLM) were discussed in Green and Yandell (1985), O'Sullivan et al. (1986) and Green (1987). In generalized linear models, we have observations $y(\boldsymbol{s}) = (y(\boldsymbol{s}_1), \ldots, y(\boldsymbol{s}_n))'$ of some random variable with an exponential family distribution at locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$, such that

$$g(\mathbb{E}(y(\boldsymbol{s}))) = \mathbf{x}(\boldsymbol{s})'\boldsymbol{\beta} + \eta(\boldsymbol{s}) \tag{4.5}$$

where $g(\cdot)$ is a link function, $\mathbf{x}(\boldsymbol{s}) = (x_1(\boldsymbol{s}), \ldots, x_p(\boldsymbol{s}))'$ is a vector of covariates with coefficients $\boldsymbol{\beta}$, and $\eta(\boldsymbol{s})$ is a spatial random process that is not observed directly. Semiparametric generalized linear models can be fitted by optimizing the penalized log-likelihood function

$$\ell(\boldsymbol{\beta}, \mathbf{f}; \mathbf{y}, \mathbf{X}) + \lambda J[\mathbf{f}]$$

which is analogous to (2.5), in the regression context. We believe that in the context of semiparametric generalized linear models, spatial confounding can be challenging due to the non-linearity nature of the data spatial dependence.

4.2.2 Spatial confounding in spatial point processes regression

Another extension of Chapter 2 to be pursued is to spatial point processes. From Gaetan and Guyon (2010), let D be a closed subset of $\mathbb{R}^2$. Let

$$x = \{x_1, x_2, \ldots\}, \quad x_i \in S;$$

x is a *locally finite* subset of points if $x \cap B$ is finite for any Borel set B.

Definition 4.1 (Point process). *A point process on D is a mapping X from a probability space $\{\Omega, \mathcal{B}, \mathbb{P}\}$ to the set $\mathcal{N}_D$ of locally finite configurations such that for every bounded Borel set B, the number of points $N(B)$ of X falling in B is a random variable.*

Denote the number of points within $B \subset D$ by $N(B)$. The first order moment measure $\mu(B)$ is

$$\mu(B) = \mathbb{E}(N(B)) = \mathbb{E}\left(\sum_{\mathbf{y} \in Y} 1\{y \in B\}\right)$$

We assume the existence of an intensity function λ such that

$$\mu(B) = \int_B \lambda(\boldsymbol{s}) d\boldsymbol{s}.$$

Definition 4.2 (Poisson process). *Assume that*

- *For any bounded Borel set $B \subset D$, $\mu(B) \in (0, \infty)$ and $N(B) \sim$ Poisson$(\mu(B))$.*

- *The joint density f of event locations $\mathbf{y}_1, \dots, \mathbf{y}_n$, conditioned on $N(B) = n$, satisfies*

$$f(\mathbf{y}_1, \dots, \mathbf{y}_n | N(B) = n) \propto \prod_{i=1}^{n} \lambda(\mathbf{y}_i)$$

then Y is a Poisson process on D with intensity function λ.

Regression models for inhomogeneous Poisson point process are based on the log-intensity function

$$\log\{\lambda(\boldsymbol{s})\} = \beta_0 + \sum_{j=1}^{p} \beta_j x_j(\boldsymbol{s})$$

where $x_j(\boldsymbol{s})$, $j = 1, \dots, p$ are covariates, and $\beta_0, \dots, \beta_p$ are parameters. Let $y = \{y_1, \dots, y_n\}$ be a Poisson process realization on a window W. The log-likelihood for the Poisson process where λ is a function of parameters $\boldsymbol{\beta}$ is

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log \lambda(y_i; \boldsymbol{\beta}) - \int_W \lambda(\boldsymbol{s}; \boldsymbol{\beta}) \mathrm{d}\boldsymbol{s}$$

An algorithm by Baddeley and Turner (2000) can be used to obtain $\hat{\boldsymbol{\beta}}$. The algorithm is based on the Poisson GLM model. However, homogeneous Poisson point processes correspond to the hypothesis of complete spatial randomness. That is, the locations are distributed independently. While inhomogeneous Poisson point processes offer some modeling flexibility when covariates are available, often spatial processes show clustering behavior that cannot be entirely explained by the covariates alone (see, e.g., Møller and Waagepetersen, 2003). Here, we will consider a class of clustering processes called the Neyman-Scott process.

Definition 4.3 (Neyman-Scott process)**.** *We will define the Neyman-Scott process by the algorithm that generates a point pattern of such process.*

- *Generate parent process $C(\boldsymbol{s})$ as an homogeneous Poisson process with constant intensity function κ.*
- *For each point of $\mathbf{c}$, generate a child process $Y_{\mathbf{c}}(\boldsymbol{s})$ as an inhomogeneous Poisson process with intensity function*

  $$\lambda_{\mathbf{c}}(\mathbf{s}; \boldsymbol{\beta}, \sigma) = h(\mathbf{s} - \mathbf{c}, \sigma) \exp\{\mathbf{x}(\boldsymbol{s})'\boldsymbol{\beta}\}$$

  $h(\mathbf{s} - \mathbf{c}, \sigma)$ *is a density function with parameter* σ*.*

Usually, full maximum-likelihood estimation in spatial processes with clustering is computationally expensive. Waagepetersen (2007) showed that

$$\frac{\partial}{\partial \boldsymbol{\beta}} \ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{x}(y_i) - \int_W \mathbf{x}(\boldsymbol{s}) \lambda(\boldsymbol{s}; \boldsymbol{\beta}) \mathrm{d}\boldsymbol{s} = 0$$

is an unbiased estimating equation of $\boldsymbol{\beta}$ for Neyman-Scott processes. In other words, using the Poisson likelihood yield good estimates for $\boldsymbol{\beta}$, although Guan (2008) showed that it is possible to improve on these estimates by weighting the estimating equations. Our conjecture is: Instead of fitting a model

$$\log \lambda(\boldsymbol{s}; \boldsymbol{\beta}) = \mathbf{x}(\boldsymbol{s})'\boldsymbol{\beta}$$

if we choose to include a spline term

$$\log \lambda(\boldsymbol{s}; \boldsymbol{\beta}) = \mathbf{x}(\boldsymbol{s})'\boldsymbol{\beta} + S(\boldsymbol{s})$$

where $S(\boldsymbol{s})$ is a nonparametric component (such as a thin-plate spline), there would be an improvement over the convergence rates, properties of estimators, etc. However, we also expect to see changes in the regression coefficient estimates $\hat{\boldsymbol{\beta}}$ akin to spatial confounding.

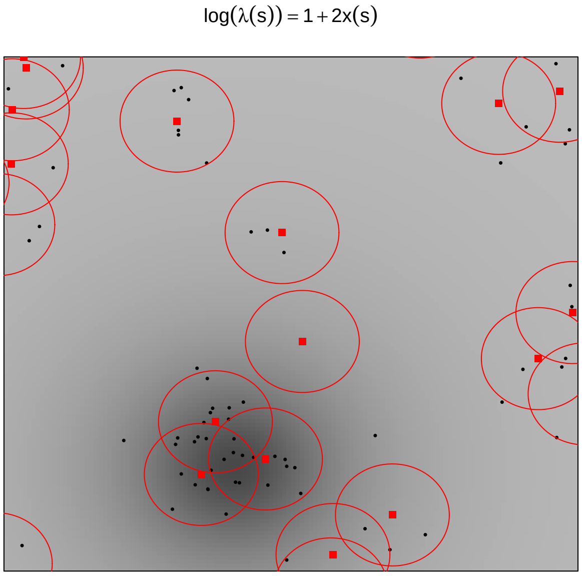

Figure 4.1: A realization of a Thomas' point process (TPP). The squared points are cluster centers, generated as a Poisson point process with intensity $\kappa = 20$. The circles around the parents are 2σ, a suggestion of the expected cluster radius. The background shade is the log-intensity $\log(\lambda(\boldsymbol{s}))$ for the child processes. The small points are the observed realization of a TPP.

We will present a small simulation to illustrate the problem. Consider a Neyman-Scott processes where $h(\boldsymbol{s} - \mathbf{c}, \sigma)$ is the Gaussian kernel with standard deviation σ. This process is called a Thomas' point process. We have generated point patterns with parent intensity function $\kappa = 20$ on a $[0, 1]^2$ window, and log-intensity function given by $\log(\lambda(\boldsymbol{s})) = 1 + 2x_1(\boldsymbol{s})$, where $x_1(\boldsymbol{s})$ is a Cauchy kernel centered on coordinates $(4/10, 2/10)$. The log-intensity function, along with a realization of the Thomas' process is shown in Figure 4.1.

Figure 4.2 show the resulting estimates of β_1 for $S = 100$ replicates of the point patterns, with different values of σ, the cluster radius. We have used the Baddeley-

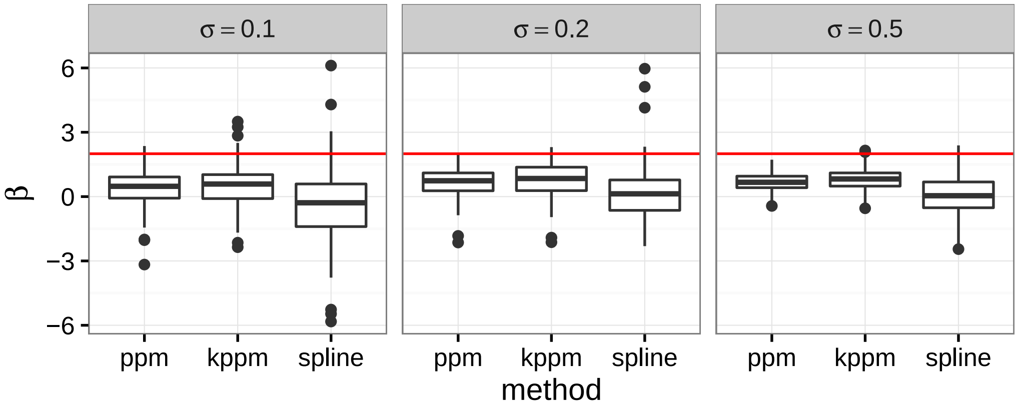

Figure 4.2: A comparison of regression coefficient estimates $\hat{\beta}_1$ for clustered point processes when using the Poisson likelihood as an estimating equation (`ppm`), weighted estimating equations (`kppm`) and the estimating equations with a thin-plate spline (`spline`)

Turner algorithm for the Poisson likelihood (denoted by `ppm`), Guan (2008) weighted estimating equations (denoted by `kppm`) and Baddeley-Turner with a thin-plate spline component (denoted by `spline`). The `ppm` and `kppm` functions are available in `R`'s `spatstat` package Baddeley et al. (2005). We see that the regression coefficient is under-estimated across all methods and clustering levels, and the spline-based estimate have a much higher variability across methods. On the other hand, the `ppm` method estimates are systematically under the true value of β_1 when $\sigma = 0.5$. We plan to keep investigating this problem, to determine when do splines work in point processes regression, and what sort of modifications can be done to improve on it.

Conclusion

This thesis addressed two questions in spatial statistics. The first question is qualitative, describing the relationship of splines and spatial random effects, in light of recent findings on spatial confounding, and whether/when the inclusion of a spline can be helpful in estimating the regression coefficients $\boldsymbol{\beta}$. We described a case of spatial confounding, and how the spectral behavior of the smoothing matrix $\mathbf{S}_\lambda$ as well as control of the tuning parameter λ can help in understanding the changes in the regression coefficients' estimates.

The second question is a data-driven problem, in which static and roving sensor spatio-temporal data can be combined to produce dynamic hazard maps. The model captures the strengths of both types of sampling devices, namely, the temporal profile from static sensors and the spatial coverage of roving sensors. The proposed model is flexible and can be applicable in other data fusion problems.

Bibliography

Altman, N. (2000), "Theory & methods: krige, smooth, both or neither?" *Australian & New Zealand Journal of Statistics*, 42, 441–461.

Baddeley, A. and Turner, R. (2000), "Practical maximum pseudolikelihood for spatial point patterns," *Australian & New Zealand Journal of Statistics*, 42, 283–322.

Baddeley, A., Turner, R., et al. (2005), "Spatstat: an R package for analyzing spatial point patterns," *Journal of statistical software*, 12, 1–42.

Bates, D. and Eddelbuettel, D. (2013), "Fast and elegant numerical linear algebra using the RcppEigen package," *Journal of Statistical Software*, 52, 1–24.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015), "Fitting Linear Mixed-Effects Models Using lme4," *Journal of Statistical Software*, 67, 1–48.

Cherrie, J. (2003), "Commentary: the beginning of the science underpinning occupational hygiene," *Annals of Occupational Hygiene*, 47, 179–185.

Chu, T., Wang, H., and Zhu, J. (2016), "On semiparametric inference of geostatistical models via local Karhunen-Loève expansion," *to appear.*

Clayton, D. G., Bernardinelli, L., and Montomoli, C. (1993), "Spatial correlation in ecological analysis," *International Journal of Epidemiology*, 22, 1193–1202.

Cowles, M. K., Zimmerman, D. L., Christ, A., and McGinnis, D. L. (2002), "Combining snow water equivalent data from multiple sources to estimate spatio-temporal trends and compare measurement systems," *Journal of Agricultural, Biological, and Environmental Statistics*, 7, 536–557.

Cressie, N. (1993), *Statistics for Spatial Data*, New York: Wiley, 2nd ed.

Cressie, N. and Wikle, C. K. (2011), *Statistics for Spatio-Temporal Data*, New York: Wiley.

Diggle, P. J. and Ribeiro, P. J. (2007), *Model-based Geostatistics*, New York: Springer.

Eddelbuettel, D. and François, R. (2011), "Rcpp: seamless R and C++ integration," *Journal of Statistical Software*, 40, 1–18.

Evans, D. E., Heitbrink, W. A., Slavin, T. J., and Peters, T. M. (2008), "Ultrafine and respirable particles in an automotive grey iron foundry," *Annals of Occupational Hygiene*, 52, 9–21.

Gaetan, C. and Guyon, X. (2010), *Spatial Statistics and Modeling*, New York: Springer.

Gneiting, T. (2002), "Nonseparable, stationary covariance functions for space–time data," *Journal of the American Statistical Association*, 97, 590–600.

Green, P. J. (1987), “Penalized likelihood for general semi-parametric regression models,” *International Statistical Review/Revue Internationale de Statistique*, 55, 245–259.

Green, P. J., Jennison, C., and Seheult, A. (1985), “Analysis of field experiments by least squares smoothing,” *Journal of the Royal Statistical Society. Series B*, 47, 299–315.

Green, P. J. and Yandell, B. S. (1985), “Semi-parametric generalized linear models,” in *Generalized Linear Models: Proceedings of the GLIM 85 Conference*, eds. Gilchrist, R., Francis, B., and Whittaker, J., New York: Springer, pp. 44–55.

Gromenko, O. and Kokoszka, P. (2013), “Nonparametric inference in small data sets of spatially indexed curves with application to ionospheric trend determination,” *Computational Statistics & Data Analysis*, 59, 82–94.

Guan, Y. (2008), “On consistent nonparametric intensity estimation for inhomogeneous spatial point processes,” *Journal of the American Statistical Association*, 103, 1238–1247.

Hall, D. L. and McMullen, S. A. (2004), *Mathematical techniques in multisensor data fusion*, Boston: Artech House.

Hanks, E. M., Schliep, E. M., Hooten, M. B., and Hoeting, J. A. (2015), “Restricted spatial regression in practice: geostatistical models, confounding, and robustness under model misspecification,” *Environmetrics*, 26, 243–254.

Hastie, T. J. and Tibshirani, R. J. (1990), *Generalized Additive Models*, vol. 43, Baton Rouge: Chapman & Hall.

Hodges, J. S. (2013), *Richly Parameterized Linear Models: Additive, Time Series, and Spatial Models using Random Effects*, Baton Rouge: Chapman & Hall.

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

Hughes, J. and Haran, M. (2013), "Dimension reduction and alleviation of confounding for spatial generalized linear mixed models," *Journal of the Royal Statistical Society: Series B*, 75, 139–159.

Isaacson, J. D. and Zimmerman, D. L. (2000), "Combining temporally correlated environmental data from two measurement systems," *Journal of agricultural, biological, and environmental statistics*, 398–416.

Kahle, D. and Wickham, H. (2013), "ggmap: Spatial Visualization with ggplot2," *The R Journal*, 5, 144–161.

Koehler, K. A. and Peters, T. M. (2013), "Influence of analysis methods on interpretation of hazard maps," *Annals of Occupational Hygiene*, 57, 558–570.

Koehler, K. A. and Volckens, J. (2011), "Prospects and pitfalls of occupational hazard mapping: between these lines there be dragons," *Annals of Occupational Hygiene*, 55, 829–840.

Lake, K., Zhu, J., Wang, H., Volckens, J., and Koehler, K. A. (2015), "Effects of data sparsity and spatiotemporal variability on hazard maps of workplace noise," *Journal of Occupational and Environmental Hygiene*, 12, 256–265.

Loève, M. (1978), *Probability Theory*, vol. II, New York: Springer, 4th ed.

Ma, C. (2003), "Families of spatio-temporal stationary covariance models," *Journal of Statistical Planning and Inference*, 116, 489–501.

McBratney, A., Whelan, B., Ancev, T., and Bouma, J. (2005), "Future directions of precision agriculture," *Precision Agriculture*, 6, 7–23.

Møller, J. and Waagepetersen, R. P. (2003), *Statistical Inference and Simulation for Spatial Point Processes*, Baton Rouge: Chapman & Hall.

O'Brien, D. M. (2003), "Aerosol mapping of a facility with multiple cases of hypersensitivity pneumonitis: demonstration of mist reduction and a possible dose/response relationship," *Applied Occupational and Environmental Hygiene*, 18, 947–952.

Nychka, D., Furrer, R., and Sain, S. (2014), *fields: Tools for Spatial Data*, r package version 7.1.

Nychka, D. W. (2000), "Spatial-process estimates as smoothers," in *Smoothing and regression: approaches, computation, and application*, ed. Schimek, M. G., New York: Wiley, pp. 393–424.

Ologe, F. E., Akande, T. M., and Olajide, T. G. (2006), "Occupational noise exposure and sensorineural hearing loss among workers of a steel rolling mill," *European Archives of Oto-Rhino-Laryngology*, 263, 618–621.

O'Sullivan, F., Yandell, B. S., and Raynor Jr, W. J. (1986), "Automatic smoothing of regression functions in generalized linear models," *Journal of the American Statistical Association*, 81, 96–103.

Paciorek, C. J. (2010), "The importance of scale for spatial-confounding bias and precision of spatial regression estimators," *Statistical Science*, 25, 107–125.

Peters, T. M., Anthony, T. R., Taylor, C., Altmaier, R., Anderson, K., and T OShaughnessy, P. (2012), "Distribution of particle and gas concentrations in swine gestation confined animal feeding operations," *Annals of Occupational Hygiene*, 56, 1080–1090.

Peters, T. M., Heitbrink, W. A., Evans, D. E., Slavin, T. J., and Maynard, A. D. (2006), "The mapping of fine and ultrafine particle concentrations in an engine machining and assembly facility," *Annals of Occupational Hygiene*, 50, 249–257.

Quick, H., Banerjee, S., and Carlin, B. P. (2015), "Bayesian modeling and analysis for gradients in spatiotemporal processes," *Biometrics*, 71, 575–584.

R Core Team (2016), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Ramsay, J. O. and Silverman, B. W. (2005), *Functional data analysis*, New York: Springer.

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.

Ribeiro Jr., P. J. and Diggle, P. J. (2015), *geoR: Analysis of Geostatistical Data*, r package version 1.7-5.1.

Rice, J. (1986), “Convergence rates for partially splined models,” *Statistics & Probability Letters*, 4, 203–208.

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

Sahu, S. K., Gelfand, A. E., and Holland, D. M. (2010), “Fusing point and areal level space-time data with application to wet deposition,” *Journal of the Royal Statistical Society: Series C*, 59, 77–103.

Sang, H. and Huang, J. Z. (2012), “A full scale approximation of covariance functions for large spatial data sets,” *Journal of the Royal Statistical Society: Series B*, 74, 111–132.

SAS Institute, Inc. (2008), *SAS/STAT® 9.2 User’s Guide*, Cary, NC: SAS Institute, Inc.

Seber, G. A. and Lee, A. J. (2003), *Linear Regression Analysis, Second Edition*, New York: Wiley.

Smidt, E. R., Conley, S. P., Zhu, J., and Arriaga, F. J. (2016), “Identifying Field Attributes that Predict Soybean Yield Using Random Forest Analysis,” *Agronomy Journal*, 108, 637–646.

Smith, B. J. and Cowles, M. K. (2007), “Correlating point-referenced radon and areal uranium data arising from a common spatial process," *Journal of the Royal Statistical Society: Series C*, 56, 313–326.

Stein, M. L. (1999), *Interpolation of Spatial Data: Some Theory for Kriging*, New York: Springer.

— (2005), "Space–time covariance functions," *Journal of the American Statistical Association*, 100, 310–321.

— (2014), "Limitations on low rank approximations for covariance matrices of spatial data," *Spatial Statistics*, 8, 1–19.

Stroup, W. W. (2012), *Generalized Linear Mixed Models: Modern Concepts, Methods and Applications*, Baton Rouge: Chapman & Hall.

Tornero-Velez, R., Symanski, E., Kromhout, H., Yu, R. C., and Rappaport, S. M. (1997), "Compliance versus risk in assessing occupational exposures," *Risk Analysis*, 17, 279–292.

Tracey, J. A., Sheppard, J., Zhu, J., Wei, F., Swaisgood, R. R., and Fisher, R. N. (2014), "Movement-based estimation and visualization of space use in 3D for wildlife ecology and conservation," *PLoS ONE*, 9, e101205.

Waagepetersen, R. P. (2007), "An estimating function approach to inference for inhomogeneous Neyman–Scott processes," *Biometrics*, 63, 252–258.

Wackernagel, H. (2003), *Multivariate Geostatistics*, Springer Science & Business Media.

Wahba, G. (1990), *Spline Models for Observational Data*, vol. 59, Philadelphia: SIAM.

Wood, S. N. (2003), “Thin plate regression splines,” *Journal of the Royal Statistical Society: Series B*, 65, 95–114.

Yao, F., Müller, H.-G., Wang, J.-L., et al. (2005), “Functional linear regression analysis for longitudinal data,” *The Annals of Statistics*, 33, 2873–2903.