DISSERTATION

ACCOUNTING FOR SPATIAL CONFOUNDING IN LARGE SCALE EPIDEMIOLOGICAL STUDIES

Submitted by

Maddie J. Rainey

Department of Statistics

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Spring 2025

Doctoral Committee:

Advisor: Kayleigh Keller

Ander Wilson
Yawen Guan
Brooke Anderson

ABSTRACT

ACCOUNTING FOR SPATIAL CONFOUNDING IN LARGE SCALE EPIDEMIOLOGICAL STUDIES

Epidemiological analyses of environmental risk factors often include spatially-varying exposures and outcomes. Unmeasured, spatially-varying factors can lead to confounding bias in estimates of associations. In this dissertation, I present a comparison of existing and new methods that use thin plate regression splines to mitigate spatial confounding bias for both cross-sectional and longitudinal analyses. I also introduce a metric to quantify the spatial smoothing induced by thin plate regression splines in varying geographic domains. I first investigate cross-sectional data, directly comparing existing approaches based on information criteria and cross-validation metrics and additionally introduce a hybrid method to selection that combines features from multiple existing approaches. Based on a simulation study, I make a recommendation for the best approach for different settings and demonstrate their use in a study of environmental exposures on birth weight in a Colorado cohort.

Next, I develop an effective bandwidth metric that quantifies the relationship between spatial splines and the range of implied spatial smoothing. I present an R Shiny application, spconfShiny, that provides a user-friendly platform to compute the metric. spconfShiny can be accessed at https://g2aging.shinyapps.io/spconfShiny/. We illustrate the procedure to compute the effective bandwidth and demonstrate its use for different numbers of spatial splines across England, India, Ireland, Northern Ireland, and the United States.

Finally, I extend two cross-sectional methods for spatial confounding adjustment to model longitudinal and time-to-event data. The additional temporal component existing in the data requires an additional selection of which coordinates to use to create thin-plate regression splines basis: the spatial coordinates, temporal coordinates, or both the spatial and temporal coordinates. I demon-

strate these methods for mixed models, generalized estimating equation models, and a proportional hazard regression framework. I demonstrate the application of these methods in a study of tropical cyclone wind exposures on preterm birth in a North Carolina cohort.

ACKNOWLEDGEMENTS

Over the past six years, I am so thankful to have had the guidance and support of many friends, colleagues, and of course, my family. My first, and greatest, thanks go to my advisor, Kayleigh Keller. Thank you for your mentorship, encouragement, and support throughout my journey as a Ph.D. student. Thank you for all the questions and then patience as I built the foundation of becoming a statistician. Thank you to Ander Wilson, Yawen Guan, and Brooke Anderson for serving on my Ph.D. committee and the feedback to enhance my dissertation. I would be remiss if I did not thank my parents, Anne Rainey and John Meyer, for their unwavering support and all the 'whales' along the way. A big thank you goes to my partner, Zach Neumann, for your support, love, and encouragement throughout the last couple of years. And finally, thank you to all of my friends and colleagues that I have met along the way, from my support system built during my undergrad, to my cohort at Colorado State, and all of the other people who have encouraged me along the way.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

The connections between environmental factors that individuals are exposed to throughout their lives and various health outcomes have been studied extensively (Tétreault et al., 2016; McCormick, 2017; McDougall et al., 2020; Svechkina et al., 2020; Jia et al., 2021; Keller et al., 2022; Zhang et al., 2023). Environmental epidemiology studies have found associations between various environmental exposures such as ambient air pollution or extreme weather events and health outcomes ranging from birth weight at the beginning of one's life to the cognitive health at the end of one's life (Peng et al., 2008; Havard et al., 2009; Di et al., 2017; Vaneckova et al., 2010; Stieb et al., 2012; Kaufman et al., 2016; Zhang et al., 2023). Frequently, both environmental exposures and health outcomes have spatial components to their variation due to societal and built environment factors, and in many epidemiology studies, temporal variation in either (or both) the exposures and health outcomes may also be present. Parameter estimates for the association between the health outcome and environmental exposure may be biased if not all spatially-varying factors are accounted for, i.e. there exists an unmeasured spatially-varying factor (Paciorek, 2010). This phenomenon, referred to as *spatial confounding*, and the mitigation of its effects will be the focus of this dissertation.

Over the past decade, many methods have been introduced to aid in the reduction of spatial confounding bias using various statistical techniques. These methods can be categorized into two broad categories: (i) approaches that remove the confounded spatial variation from the exposure, or (ii) methods that directly model the confounded spatial structure in the outcome. Techniques such as incorporating spatial splines (Dupont et al., 2022; Thaden and Kneib, 2018), spectral filtering (Guan et al., 2023), or kriging (Wiecha et al., 2024) have been investigated to remove the confounded spatial variation in the exposure. Methods to directly model the confounded structure in the outcome include spatial spline regression (Bobb et al., 2022; Keller and Szpiro, 2020), and Gaussian Processes (Marques et al., 2022; Schnell and Papadogeorgou, 2020; Hodges and Reich,

2010). In this dissertation, I will focus on comparing and developing semiparametric methodologies that leverage spatial splines to either remove the confounded spatial variation from the exposure or directly model the confounded structure, with a focus on methods that are widely applicable and easy to implement for researchers. Spatial splines are able to model multi-dimensional space while fitting within standard regression modeling framework. This greatly lowers the complexity and computational cost compared to other methods.

Over the course of the rest of this chapter, I will give a brief introduction to various topics to further motivate the methods presented in Chapters 2, 3, and 4. These topics include a general introduction to environmental epidemiological studies, confounding and spatial confounding, thin-plate regression splines (TPRS), Gaussain Processes, modeling longitudinal data, and restricted spatial regression, one of the first widely used methods to mitigate the bias due to spatial confounding. I conclude this chapter by providing a brief overview of the remaining chapters.

1.1 Environmental Epidemiology

Environmental epidemiology is the study of the effects the environment (e.g. air pollution, light at night, access to blue and green spaces, temperature, extreme weather events) has on human health (NRC, 1991). Distinguishing itself from other epidemiology subfields, environmental epidemiology focuses on the impact of the environment in which an individual lives on their health, rather than the individual's lifestyle or personal characteristics (Pekkanen and Pearce, 2001). Due to the ethical concerns of subjecting individuals to harmful environments, most environmental epidemiological studies are observational and quantify associations, rather than causal effects, between the environmental exposure of interest and the health outcome.

A common modeling strategy used in environmental epidemiology analyses, when studying location-specific exposures and/or outcomes, is two-stage modeling (Szpiro et al., 2011; Kaufman et al., 2016; Di et al., 2017; Adar et al., 2018). A two-stage approach separates the analysis into two modeling steps. In the first stage, exposures are predicted or assigned at the residential locations using statistical methods, machine learning, or remote sensing methods (Sampson et al., 2013;

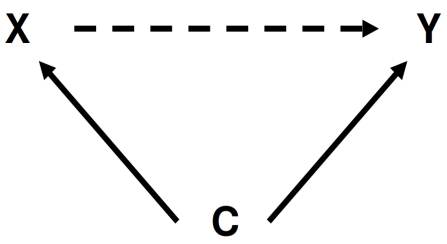


Figure 1.1: Diagram depicting the causal relationship between an exposure, X, and outcome, Y, while a confounder, C is present.

Keller et al., 2015; Di et al., 2020; McDuffie et al., 2021; Van Donkelaar et al., 2021). Then, in the second stage, an outcome model is fit using the predicted exposures along with additional measured covariates. Many times the same exposure predictions are used for multiple cohorts or outcomes. This framework means that jointly modeling the exposure and the health outcome is infeasible.

1.2 Confounding and Spatial Confounding

One of the primary concerns with observational studies is the potential presence of confounding. Confounding occurs when a factor influences both the exposure and response (outcome) variables, but is not accounted for in the the study. As shown in Figure 1.1, if the confounding variable, C, affects both the exposure, X, and outcome, Y, and we ignore C in the analysis, we might observe an association between X and Y that is biased. While confounding is a conceptual relationship between variables in an analysis, we can see its impact through a mathematical analysis of model misspecification. Let's say we fit the simple linear regression model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$; but, the true model is $y_i = \beta_0 + \beta_1 x_i + \delta c_i + \varepsilon_i$. If the values of x_i and c_i are centered and $\mathrm{E}[\varepsilon_i] = 0$, then

$$\text{Bias}(\hat{\beta}_1) = \frac{\sum_{i=1}^{n} x_i c_i}{\sum_{i=1}^{n} x_i^2} \delta. \tag{1.1}$$

Thus, c_i not being included in the model leads to bias in $\hat{\beta}_1$. These computations can be extended to the multivariate models, described in Appendix A.

Spatial confounding is defined as when a spatially-varying variable that influences a spatially-varying exposure and spatially-varying outcome is not accounted for in the study (Paciorek, 2010). In the scenario of spatial confounding, X and C are correlated due to the underlying spatial variability present. Thus, the term $\sum_{i=1}^{n} x_i c_i$ in (1.1) is non-zero, biasing $\hat{\beta}_1$. For environmental epidemiology studies, exposures vary over space and due to the structure of how we live as a society, health concerns also contain spatial variation. Thus, spatial confounding is a concern for environmental epidemiological studies.

For example, say a researcher was interested in the association between ambient air pollution levels and memory scores of an aging population, but does not account for socioeconomic status in their analysis. Air pollution levels contain spatial variation due to the distribution of sources and meteorology, among other factors. One potential source of spatial variation in memory scores is the distance individuals live from health care facilities. The closer one lives to health care facilities, the easier one can access treatments to both ascertain health status and help mitigate the negative health effects. Finally, socioeconomic status is connected to where one lives. Areas where lower income individuals live tend to have higher pollution levels and be further from health care facilities. So, when socioeconomic status is unmeasured, and ambient air pollution, memory score, and socioeconomic status all vary spatially, spatial confounding should be a concern for the researcher in this example.

1.3 Thin-Plate Regression Splines

Recently, a common approach to account for spatial confounding in modeling health effects in environmental epidemiology studies is the inclusion of thin-plate regression splines (TPRS). TPRS, introduced by Wood (2003), provide a computationally efficient basis for modeling multi-dimensional data, via a low rank approximation of thin plate smoothing splines (TPS). Introduced by Duchon (1977), TPS find the optimal function $\boldsymbol{g} = (g(\boldsymbol{x}_1), g(\boldsymbol{x}_2), \ldots, g(\boldsymbol{x}_n))$ to estimate a

smooth function f in the equation $y_i = f(\boldsymbol{x}_i) + \varepsilon_i$, by minimizing

$$||\boldsymbol{y} - \boldsymbol{g}||^2 + \lambda J(g).$$

The fit of the data is balanced with a penalty function, $J(\cdot)$, controlling the smoothness of $\boldsymbol{g}$, through a smoothing parameter, λ. The optimal solution is a radial basis representation. Given a set of center points $\boldsymbol{c}_j, j = 1, \ldots, K$ for K radial basis functions, we can define

$$g(\boldsymbol{x}) = \sum_{j=1}^{K} \omega_j \phi(||\boldsymbol{x} - \boldsymbol{c}_j||) \tag{1.2}$$

where ω_j are mapping coefficients and $\phi(r) = r^2 \ln(r)$ (Duchon, 1977; Donato and Belongie, 2002).

One of the primary obstacles of implementing TPS is the computation cost. Except in the case in 1-dimension, which will not happen when using TPS to model 2-dimensional space, modeling TPS requires $O(n^3)$ operations, making fitting TPS infeasible with large datasets. The large computation cost arises from estimating $n + 1$ parameters, one for each observation and the additional smoothing parameter (Wood, 2003).

One way to decrease the computational complexity of implementing TPS is to use regression splines (Wood, 2003). Regression splines implement knots spaced throughout the modeling space and a polynomial function is fit between each knot, without the wiggliness penalty (Friedman, 1991). However, the placement of the knots raises its own concerns, especially across multiple dimensions (Hastie et al., 2001).

Thin-plate regression splines decrease the computational complexity of TPS and remove the knot placement issues of regression splines (Wood, 2003). TPRS use a truncated eigenbasis to approximate the radial basis of the TPS. By using the truncation, TPRS now require $O(kn^2)$ operations, where k is the rank of the truncated eigenbasis, greatly improving on the implementation for larger data sets (Wood, 2003). TPRS are easily calculated through the `mgcv` package in `R` (Wood, 2011). A plot of the first 10 TPRS across a $(0, 10) \times (0, 10)$ grid is shown in Figure 1.2.

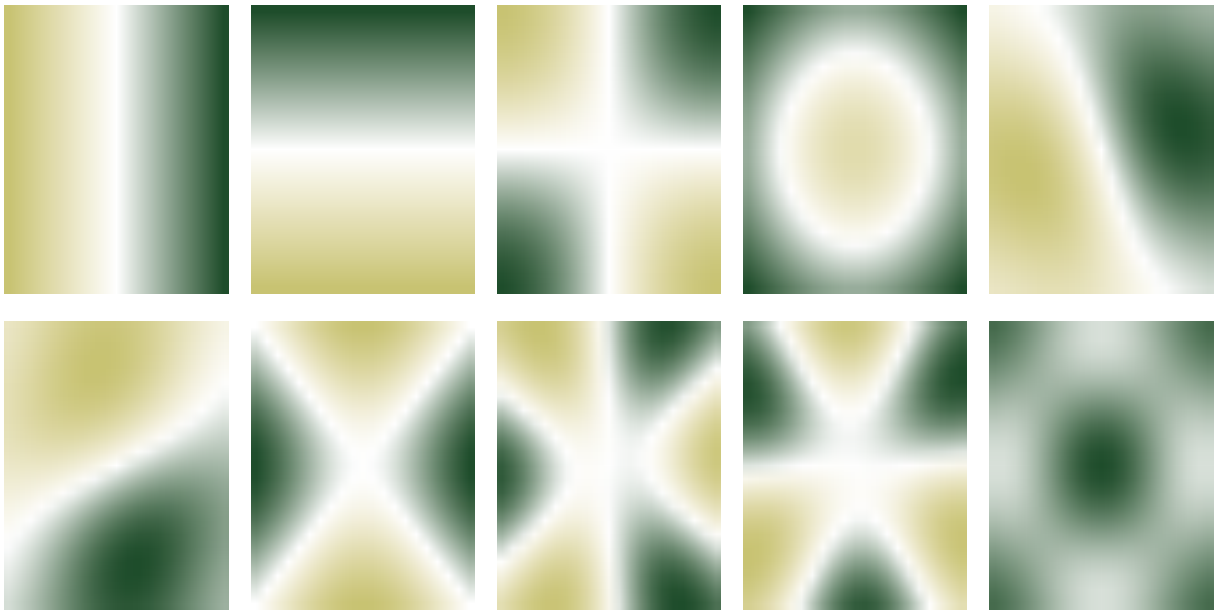


Figure 1.2: The first 10 splines in a TPRS basis computed across a $(0, 10) \times (0, 10)$ grid. The green shading in each grid depict more positive values and the yellow shading depicts more negative values.

As seen in Figure 1.2, an important characteristic of TPRS is the ordering of the splines by variation complexity. The least complex splines are the first splines in the basis and as more splines are added to the basis, the more variable the splines become. Thus, models with fewer splines will model smoother variation than models with more splines and we can take advantage of this property when implementing model selection procedures.

1.4 Gaussian Processes

Spatial data is commonly modeled using a Gaussian Process (GP) (Banerjee et al., 2015). A GP is a stochastic process for which any realization is a collection of random variables with a multivariate normal distribution. This means a GP can be completely defined by its mean and covariance functions. While a variety of covariance structures are possible, the Matèrn family of covariance functions is one of the most widely used for GP's applied to spatial data. The Matèrn covariance function is defined as

$$C_\nu(d) = \sigma^2 \frac{2^{\nu-1}}{\Gamma(\nu)} \left(\sqrt{2\nu} \frac{d}{\rho} \right)^\nu K_\nu \left(\sqrt{2\nu} \frac{d}{\rho} \right) \tag{1.3}$$

where d is a measure of distance between two points, ν is a smoothing parameter, ρ is a range parameter, Γ is the gamma function, and K_ν is the modified Bessel function of the second kind (Rasmussen and Williams, 2008). Setting $\nu = \frac{1}{2}$ simplifies (1.3) to the exponential covariance function:

$$C_{\frac{1}{2}}(d) = \sigma^2 \text{exp}\left(-\frac{d}{\rho}\right). \tag{1.4}$$

Visualizations of the Matèrn covariance function are shown in Figure 1.3. Figure 1.3a and Figure 1.3b plot the covariance against the Euclidean distance between points comparing a Matèrn covariance with low spatial variability to a covariance with high spatial variability. Figure 1.3c and Figure 1.3d simulate two spatial fields over a (0,10)$\times$(0,10) grid using a zero-mean GP with a Matèrn covariance structure, one simulated from the covariance function with low spatial variability and one simulated from the covariance function with high spatial variability. When simulating a GP with low spatial variability, the resulting spatial field will be able to model broad spatial patterns but will not have finer scale spatial details compared to a simulated GP with high spatial variability, which is able to model finer spatial details.

1.5 Modeling Longitudinal Data

When collecting data over a period of time, individuals may have measurements taken at different time points across the study. Thus, each individual can have several observations over a given period, creating longitudinal data. A mixed model or a generalized estimating equation (GEE) model are commonly used when analyzing longitudinal data. Mixed models add a random effect, commonly a random intercept or both random intercept and random slope for each individual, to account for correlation in the multiple observations per individual (Diggle et al., 2002). For a linear mixed model, the equation is

$$y_{ij} = \beta_0 + \beta_x x_{ij} + a_i + b_i x_{ij} + \varepsilon_{ij}$$

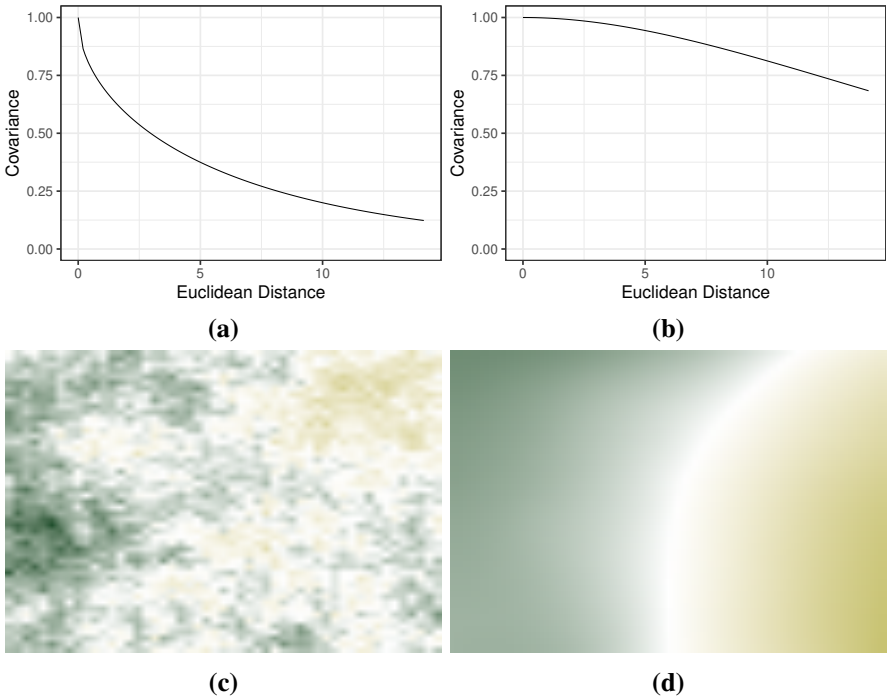


Figure 1.3: Visualization of the Matérn covariance function with $\sigma^2 = 1$ and $\rho = 10$ comparing high spatial variability, $\nu = 0.25$ [(a) and (c)] to low spatial variability, $\nu = 2$ [(b) and (d)]. Figures in (a) and (b) plot the Matérn covariance against the Euclidean distance between two points. Figures (c) and (d) plot a simulated mean-zero GP with the corresponding covariance function over a (0,10)×(0,10) grid.

where $a_i \sim N(0, \sigma_a^2)$ and $b_i \sim N(0, \sigma_b^2)$ are the random intercept and slope, respectively, and $\varepsilon_{ij} \sim N(0, \sigma_y^2)$ is additional random variation. By adding the random effect, the coefficient β_x in a mixed model has a conditional interpretation, representing the average difference in outcome for a unit difference in exposure within a single individual. GEE models assume various correlation structures among the observations (Zeger et al., 1988). Thus, we modeling an outcome as

$$y_{ij} = \beta_0 + \beta_x x_{ij} + \varepsilon_{ij} \tag{1.5}$$

where $\varepsilon_{ij} \sim N(0, \sigma_y^2 \boldsymbol{V})$ and $\boldsymbol{V}$ is the correlation structure. In Chapter 4, we assume $\boldsymbol{V}$ in (1.5) is exchangable for each individual, meaning that all observations for an individual are equally correlated. Because there are no individual random effects in (1.5), parameters in a GEE model have a marginal interpretation. This means the estimates of the exposure-outcome association can be interpreted as population-averaged values.

1.6 Restricted Spatial Regression

Restricted spatial regression (RSR) was one of the first widely-used methods to deal with spatial confounding (Hodges and Reich, 2010). Under the RSR framework, spatial confounding is defined as random effect collinearity between the spatial covariates of interest, $\boldsymbol{X}$, and the spatial random effect, $\boldsymbol{\eta}$, in a model $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\eta} + \boldsymbol{\varepsilon}$ and is mitigated by restricting $\boldsymbol{\eta}$ to be orthogonal to $\boldsymbol{X}$. This definition, however, varies from the definition presented in Section 1.2 because it prioritizes estimation of the unconditional effect of the exposure on the outcome. That is, it estimates the association between the exposure and outcome that attributes all possible unmeasured spatial variation to the exposure. In environmental epidemiology studies, the conditional estimate, which captures the association between the exposure and outcome conditional upon the unmeasured factor, is what is of primary concern to researchers. The difference between the estimates that are computed also motivates the rest of this work away from comparisons to RSR. Additional concerns of the coverage properties of RSR estimators investigated by both Zimmerman and Ver Hoef (2022) and Khan and Calder (2022) also motivate this work away from comparisons to RSR. Such comparisons are briefly considered in Chapter 2 but will not be compared in Chapter 4. Additional explanation of the RSR methodology is given in Appendix A.

1.7 Overview

1.7.1 Mitigating Spatial Confounding in Large Cohorts

In Chapter 2, I compare four methods that implement TPRS to mitigate spatial confounding bias in a cross-sectional analysis (i.e. analyses using data from a single point in time) and develop a hybrid methodology, combining two existing methodologies. I compare methods introduced by Thaden and Kneib (2018); Keller et al. (2022); Dupont et al. (2022) and Bobb et al. (2022) In my proposed methodology, I use the underlying structure of Dupont et al. (2022)'s method; however, I use a model selection approach inspired by Keller and Szpiro (2020)'s in the first step of the model.

In Chapter 4, I propose methods which extend the cross-sectional methodology introduced by Dupont et al. (2022) and Keller and Szpiro (2020) to include temporal, along with the existing

spatial, variation. Most environmental epidemiology studies are conducted over the span of many years with multiple follow up appointments. The studies are more commonly modeled in a longitudinal framework where an individual in a study will have multiple observations recorded in the data or a survival framework where the outcome of interest is now a time-to-event, such as the onset of dementia in an individual or, more commonly for survival studies, death. For longitudinal data, I extend the cross-sectional spatial confounding methods to be applicable to mixed models and generalized estimating equation (GEE) models; and for time-to-event data, I extend the methods to a proportional hazards regression model.

1.7.2 Tools for Implementation of Thin-Plate Regression Splines

In Chapter 3, I introduce a variant of the effective bandwidth proposed by Keller and Szpiro (2020) to give researchers a metric that connects the number of spatial splines to a physical distance. The metric is implemented in an R package, `spconf`, available on CRAN (Keller and Rainey, 2024). I developed an R Shiny application, `spconfShiny` deployed at https://g2aging.shinyapps.io/spconfShiny/, to provide a user interface to the `spconf` package. The development of the effective bandwidth metric gives a practical procedure for connecting the number of splines included in a model to the spatial distances across a region and therefore fills a gap in understanding what exactly adding a given number of splines is doing in a model. The contents of Chapter 3 are published in PLOS ONE with the title *spconfShiny: An R Shiny application for calculating the spatial scale of smoothing splines for point data* (Rainey and Keller, 2024).

Chapter 2

Semiparametric approaches for mitigating spatial confounding in large environmental epidemiology cohort studies

2.1 Introduction

Large-scale epidemiology studies of the health effects of environmental exposures frequently analyze variables that vary over space (e.g. Di et al., 2017; Vaneckova et al., 2010). Environmental factors such as air pollution concentrations and temperature along with different health outcomes such as the onset of dementia or birth weights have a spatial component to their variation (Zhang et al., 2023; Havard et al., 2009; Di et al., 2020; Vaneckova et al., 2010; Stieb et al., 2012). Thus, estimated associations between spatially-varying health outcomes and environmental exposures may be biased if there exist unmeasured spatially-varying factors that are not accounted for (Paciorek, 2010). We refer to this phenomenon as *spatial confounding*.

Different approaches have been developed to estimate the conditional or unconditional effect of an exposure on a health outcome in the presence of spatial confounding (Khan and Berrett, 2023). Many of the approaches to estimate the conditional effect are semiparametric and use thin plate regression splines (TPRS) to model the spatial variation (Wood, 2003). When fitting a model with TPRS, three important aspects to consider are: (i) the number of splines needed to accurately model the spatial variation; (ii) whether to include penalization on the basis when fitting the model; and (iii) whether TPRS should first be used to model the exposure separately or only included in the outcome model (Dupont et al., 2023). Differences in these choices lead to a variety of different modeling strategies for spatial confounding adjustment. The geoadditive structural equation model, gSEM, selects a large, fixed number of TPRS to include in separate models for the exposure and the outcome, and then obtains residuals that should no longer contain spatial

variation (Thaden and Kneib, 2018). The final model regresses the residuals of the outcome on the residuals of the exposure. The Spatial+ model introduced by Dupont et al. (2022) uses similar ideas of gSEM except only obtains the residuals from the exposure, then regressing the outcome of interest on the residuals of the exposure and the TPRS basis. Both gSEM and Spatial+ require a choice of using either penalization by generalized cross validation (GCV) or unpenalized TPRS. Keller and Szpiro (2020) proposed adding unpenalized TPRS to the outcome regression model, with the number of basis functions selected through using a selection criterion such as the Akaike information criterion (AIC) or Bayesian information criterion (BIC) applied to an outcome model without the exposure. In contrast, Bobb et al. (2022) proposed a model for the exposure using penalized splines to obtain a smoothing parameter that is used to model the outcome. Guan et al. (2023) introduced similar methodology but based in the spectral domain rather than using a TPRS adjustment. The selection of amount of spatial smoothing to include can impact the magnitude of bias reduction (Dupont et al., 2023; Reich et al., 2022; Papadogeorgou, 2022; Keller and Szpiro, 2020). Additionally, the assumed form of the data generating mechanism can further affect the bias mitigation (Marques and Kneib, 2022; Keller and Szpiro, 2020).

Alternative approaches for addressing spatial confounding exist that do not use TPRS. Wiecha et al. (2024) introduced a variant of the gSEM approach that uses Kriging to model latent functions of space in the exposure and outcome and obtains residuals to be used in the second stage of the methodology. Schnell and Papadogeorgou (2020) introduced an approach, both in frequentist and Bayesian frameworks, that jointly modeled the exposure and unobserved confounder using a conditional autoregressive structure. Marques et al. (2022) introduced a Bayesian approach that uses a multivariate Gaussian random field prior to directly model the spatial variation. Gilbert et al. (2024) introduces a double machine learning approach to mitigate the bias due to spatial confounding in a causal inference framework. While highly flexible, all of these methods have a higher complexity to implement than the semiparametric models discussed previously, which can make them challenging for large studies and applied researchers.

Different approaches can also target the unconditional effect of the predictor, motivated by random effect collinearity (Hodges and Reich, 2010; Hughes and Haran, 2013; Hanks et al., 2015). The recommended solution to spatial confounding in this framework is typically "restricted spatial regression" (RSR), which targets the marginal association between the exposure and the outcome. RSR limits the spatial random effect to be orthogonal to the exposure and any other fixed effects (Khan and Calder, 2022). However, Khan and Calder (2022) demonstrated that when considering a discrete spatial structure, a non-spatial model was preferred to using a RSR model in terms of coverage rates. Hanks et al. (2015) also noted that for when continuous spatial structure is assumed, RSR also had poorer coverage rates than a spatial model not accounting for spatial confounding. Furthermore, for most epidemiological applications, the unconditional effect is not of interest.

All of these methods can reduce the bias induced by spatial confounding; however, applied researchers still must choose which approach is best to mitigate the bias in different contexts. In this work, we focus on the approaches that use TPRS to model the latent spatial behavior and provide a recommendation for epidemiologists for modeling choice, based on simulations, to best reduce this bias from spatial confounding in settings that are typical for large-scale environmental epidemiology cohort studies.

The investigation of various methods to mitigate spatial confounding using TPRS that follows is motivated by analyses of the impacts of fine particulate matter ($PM_{2.5}$) at different stages of an individual's life. Particularly, we are motivated by the potential associations with the incidence of dementia in the aging population and with low birth weights at the beginning of one's life (Zhang et al., 2023; Rainey and Keller, 2024; Dematteis et al., 2024; Mork et al., 2024). Since both $PM_{2.5}$ and these health outcomes have underlying spatial variation, there is potential for spatial confounding. In this work, we obtained a large administrative data set from the Colorado Department of Public Heath and Environment (CDPHE) containing all live births between 2007 and 2018. We investigated the association between third trimester $PM_{2.5}$ and maximum daily temperatures and birth weight. Mork et al. (2024) and Dematteis et al. (2024) also used a similar Colorado birth cohort to analyze the relationship between environmental exposures and low birth weights; however,

both studies subsetted the cohort to only contain births in lower elevations in attempts to reduce the bias due to spatial confounding, greatly reducing the number of births in their analyses. Here we apply TPRS-based methods to directly account for this confounding.

The rest of the manuscript is structured as follows. In Section 2.2, we introduce the data generating model, outline existing approaches designed to reduce the bias from spatial confounding and introduce a new approach. Section 2.3 summarizes a simulation study comparing the methods. In Section 2.4, we apply the methods discussed in Section 2.2 to the CDPHE cohort. Finally, Section 2.5 completes the manuscript with a discussion.

2.2 Methods

2.2.1 Notation and Data Generating Model

We define $y(\boldsymbol{s}_i)$ as the measured outcome, $x(\boldsymbol{s}_i)$ the measured covariate of interest which we will refer to as the exposure, $\boldsymbol{s}_i \in \mathbb{R}^2$ the spatial locations, $\boldsymbol{w}_i \in \mathbb{R}^p$ the vector of measured covariates, and $f(\boldsymbol{s}_i)$ the unmeasured spatial confounder for each observation i. We note that while we will refer to $f(\boldsymbol{s}_i)$ as a single unmeasured confounder, it may be a composition of multiple unmeasured factors. We assume the exposure and unmeasured spatial confounder are derived from a combination of three independent spatial fields, $z_1(\boldsymbol{s})$, $z_2(\boldsymbol{s})$, and $z_3(\boldsymbol{s})$, and non-spatial error, ϵ_x such that

$$x(\boldsymbol{s}_i) = z_1(\boldsymbol{s}_i) + z_2(\boldsymbol{s}_i) + \epsilon_x \tag{2.1}$$

$$f(\boldsymbol{s}_i) = z_1(\boldsymbol{s}_i) + z_3(\boldsymbol{s}_i). \tag{2.2}$$

The spatial correlation between $x(\boldsymbol{s}_i)$ and $f(\boldsymbol{s}_i)$ comes from the shared spatial field, $z_1(\boldsymbol{s}_i)$. The mean of the health outcome is given by

$$\mu(\boldsymbol{s}_i) = \beta_0 + \beta_x x(\boldsymbol{s}_i) + \boldsymbol{w}_i^{\mathsf{T}}\boldsymbol{\gamma} + \beta_f f(\boldsymbol{s}_i), \tag{2.3}$$

where $\beta_0, \beta_x, \beta_f \in \mathbb{R}$ and $\boldsymbol{\gamma} \in \mathbb{R}^p$ are the regression coefficients. For a continuous outcome, we assume $y(\boldsymbol{s}_i) \sim N(\mu(\boldsymbol{s}_i), \sigma_y^2)$; and for a binary outcome, we assume $y(\boldsymbol{s}_i) \sim$ Bernoulli(p_i) where $p_i = 1/(1 + e^{-\mu(\boldsymbol{s}_i)})$. Our inferential goal is to estimate β_x while minimizing the bias created from the unmeasured spatial confounding. We will continue by suppressing both the index for each observation (i) and the location dependence $(\boldsymbol{s}_i)$ for notational convenience.

2.2.2 Existing Semiparametric Approaches

We present the mathematical detail for several semiparametric methods that have been introduced to mitigate bias induced by spatial confounding. The following exposition will be written in terms of a continuous outcome, but the methods apply in an analogous way to discrete outcomes in a generalized linear model framework.

Thin Plate Regression Splines

The approaches we compare represent spatial variation using thin plate regression splines (TPRS). TPRS are low-rank approximations of thin plate splines that provide a computationally efficient spatial basis (Wood, 2003). We will denote individual splines in the following models as $h_j(\boldsymbol{s})$. Each methodology we consider selects the number of TPRS basis functions to be included to model the spatial variability differently. Models can fit finer-scale spatial details with a larger number of basis functions, however including too many splines can reduce power and potentially amplify bias (Keller and Szpiro, 2020). For models without penalization, the number of basis functions is equal to the degrees of freedom (df); for models with penalization, the effective degrees of freedom is less than the number of basis functions. All methods that we describe here make an assumption that the included spatial splines can fully model the spatial structure of the unmeasured confounder. However, since spatial splines can capture flexible spatial structures, this assumption is mild (Wood, 2003; Dupont et al., 2022) But, if the assumption is violated and the spatial splines cannot fully model the unmeasured confounder's spatial structure, there is no guarantee that the point estimate for β_x will be unbiased.

Spatially-Unadjusted Model

As a base for comparison, we first consider an unadjusted model. That is, we fit the model $y^{\text{l}} = \beta_0^{\text{l}} + \beta^{\text{l}} x + \boldsymbol{w}^{\mathsf{T}} \boldsymbol{\gamma}^{\text{l}} + \varepsilon_y^{\text{l}}$ and assume $\varepsilon_y^{\text{l}} \sim N(0, (\sigma_y^{\text{l}})^2)$. Because this model does not account for f, the ordinary least squares estimator $\hat{\beta}^{\text{l}}$ will be biased for β_x.

Spatial+

The Spatial+ approach (Dupont et al., 2022) follows a two-step procedure. The first step fits a linear model with k_1 TPRS basis functions to the exposure using the known spatial locations of each observation, with k_1 chosen to be large. Thus, it fits the exposure model, $x = \delta_0^{\text{s.pl}} + \sum_{j=1}^{k_1} \delta_j^{\text{s.pl}} h_j + \varepsilon_x^{\text{s.pl}}$, where $\varepsilon_x^{\text{s.pl}} \sim N(0, (\sigma_x^{\text{s.pl}})^2)$ represents the non-spatial variation in the exposure. The fitted values, $\hat{x}^{\text{s.pl}} = \hat{\delta}_0^{\text{s.pl}} + \sum_{j=1}^{k_1} \hat{\delta}_j^{\text{s.pl}} h_j$, are obtained to calculate the residuals, $\hat{r}_x^{\text{s.pl}} = x - \hat{x}^{\text{s.pl}}$. The second step fits the outcome to the residuals and the same number of TPRS used in the first step, $y = \beta_0^{\text{s.pl}} + \beta^{\text{s.pl}} \hat{r}_x^{\text{s.pl}} + \boldsymbol{w}^{\mathsf{T}} \boldsymbol{\gamma}^{\text{s.pl}} + \sum_{j=1}^{k_1} \alpha_j^{\text{s.pl}} h_j + \varepsilon_y^{\text{s.pl}}$, where $\varepsilon_y^{\text{s.pl}} \sim N(0, (\sigma_y^{\text{s.pl}})^2)$ and from which the goal is to estimate $\beta^{\text{s.pl}}$. When fitting the exposure and the outcome using a TPRS, Dupont et al. (2022) proposed two different methods for selecting the degrees of freedom for the Spatial+ approach: one that implements a smoothing penalty on the coefficients using generalized-cross validation (GCV) and one that does not, which we denote s.pl and s.pl-fx respectively.

Geoadditive Structural Equation Model

The geoadditive structural equation model (gSEM) proposed by Thaden and Kneib (2018) follows a two-step methodology similar to the Spatial+ approach. The first step of the gSEM method fits a linear model with k_2 TPRS basis functions to both the exposure and the outcome, separately. Thus, the exposure and outcome models are, $x = \delta_0^{\text{xg}} + \sum_{j=1}^{k_2} \delta_j^{\text{xg}} h_j + \varepsilon_x^{\text{xg}}$ and $y = \delta_0^{\text{yg}} + \sum_{j=1}^{k_2} \delta_j^{\text{yg}} h_j + \varepsilon_y^{\text{yg}}$, respectively, where $\varepsilon_x^{\text{xg}} \sim N(0, (\sigma_x^{\text{xg}})^2)$ and $\varepsilon_y^{\text{yg}} \sim N(0, (\sigma_x^{\text{yg}})^2)$ represent the non-spatial variation in the exposure and outcome, repsectively. Similar to the Spatial+ methodology, the residuals, $\hat{r}_x^{\text{g}}$ and $\hat{r}_y^{\text{g}}$, are then obtained. The second step of the gSEM method fits a linear model on the outcome residuals using the exposure residual and other measured covariates with no intercept, $\hat{r}_{\text{y}}^{\text{g}} = \beta^{g} \hat{r}_{\text{x}}^{\text{g}} + \boldsymbol{w}^{\mathsf{T}} \boldsymbol{\gamma}^{\text{g}} + \varepsilon_{\text{r}}^{\text{g}}$ where $\varepsilon_{\text{r}}^{\text{g}} \sim N(0, (\sigma_{\text{r}}^{\text{g}})^2)$.

Approach of Keller and Szpiro

The method proposed by Keller and Szpiro (2020), which we will denote KS, starts by selecting the degrees of freedom from a model for the outcome variable with unpenalized TPRS. That is, the outcome model $y = \delta_0^{\text{KS}} + \sum_{j=1}^{k_3} \delta_j^{\text{KS}} h_j + \boldsymbol{w}^\mathsf{T}\boldsymbol{\gamma}^{\text{KS}} + \varepsilon_y^{\text{KS}}$ is fit, where $\varepsilon_y^{\text{KS}} \sim N(0, (\sigma_y^{\text{KS}})^2)$. Then, either AIC or BIC is used to select the number of basis functions, $\hat{k}_3$, included in the final model which we will denote as either KS-AIC and KS-BIC, respectively. The final model regresses the outcome on the exposure, TPRS, and measured adjustment variables with the selected $\hat{k}_3$ basis functions:

$$y = \beta_0^{\text{KS}} + \beta^{\text{KS}} x + \boldsymbol{w}^\mathsf{T}\tilde{\boldsymbol{\gamma}}^{\text{KS}} + \sum_{j=1}^{\hat{k}_3} \alpha_j^{\text{KS}} h_j + \varepsilon_y^{\text{KS}}, \tag{2.4}$$

again, where $\varepsilon_y^{\text{KS}} \sim N(0, (\sigma_y^{\text{KS}})^2)$ and with the goal of estimating β^{KS}.

Exposure-Penalized Splines

The first step in the exposure-penalized splines (E-PS) model (Bobb et al., 2022) is similar to the Spatial+ approach using GCV penalization, except instead of computing the residuals, the estimated smoothing parameter, $\hat{\lambda}$, of the penalization is obtained. E-PS assumes $x = \delta_0^{\text{E-PS}} + \sum_{j=1}^{k_4} \delta_j^{\text{E-PS}} h_j + \varepsilon_x^{\text{E-PS}}$, where $\varepsilon_x^{\text{E-PS}} \sim N(0, (\sigma_x^{\text{E-PS}})^2)$. After $\hat{\lambda}$ is computed from the first model, the outcome is fit to the exposure, TPRS basis with $\hat{\lambda}$ penalization, and measured covariates: $y = \beta_0^{\text{E-PS}} + \beta^{\text{E-PS}} x + \boldsymbol{w}^\mathsf{T}\boldsymbol{\gamma}^{\text{E-PS}} + \sum_{j=1}^{k_4} \alpha_j^{\text{E-PS}} h_j + \varepsilon_y^{\text{E-PS}}$, again, where $\varepsilon_y^{\text{E-PS}} \sim N(0, (\sigma_y^{\text{E-PS}})^2)$ and with the goal of estimating $\beta^{\text{E-PS}}$. When fitting the E-PS model in this work, we follow the code of Bobb et al. (2022) which uses penalization parameter that differs by a small scale factor in the first and second models.

2.2.3 Degrees of Freedom Selected Spatial+ (df-Spatial+)

We also consider a novel approach, which we refer to as df-Spatial+, that combines ideas from the Spatial+ and KS methods similar to E-PS method. df-Spatial+ uses the two-step methodology and structure from the Spatial+ model without the smoothing penalty applied; but df-Spatial+ also uses the degrees of freedom selection methodology from the KS method. The assumed model

for the exposure is $x = \delta_0^{\text{df.pl}} + \sum_{j=1}^{k_5} \delta_j^{\text{df.pl}} h_j + \varepsilon_x^{\text{df.pl}}$, where $\varepsilon_x^{\text{df.pl}} \sim N(0, (\sigma_x^{\text{df.pl}})^2)$. Multiple models are fit using df, then we select the df, $\hat{k}_5$, by minimizing AIC or BIC. This approach is denoted df.pl-AIC and df.pl-BIC respectively. The exposure is fit using a linear model with $\hat{k}_5$ basis surfaces, $\hat{x}^{\text{df.pl}} = \hat{\delta}_0^{\text{df.pl}} + \sum_{j=1}^{\hat{k}_5} \hat{\delta}_j^{\text{df.pl}} h_j$, to obtain the residuals: $\hat{r}_x^{\text{df.pl}} = x - \hat{x}^{\text{df.pl}}$. The final model fits the outcome using the residuals from the first stage with $\hat{k}_5$ basis functions: $y = \beta_0 + \beta^{\text{df.pl}} \hat{r}^{\text{df.pl}} + \boldsymbol{w}^{\mathsf{T}} \boldsymbol{\gamma}^{\text{df.pl}} + \sum_{j=1}^{\hat{k}_5} \alpha_j^{\text{df.pl}} h_j + \varepsilon_y^{\text{df.pl}}$, where $\varepsilon_y^{\text{df.pl}} \sim N(0, (\sigma_y^{\text{df.pl}})^2)$ and a goal to estimate $\beta^{\text{df.pl}}$. This approach is similar to the "pre-adjustment" approach described in Keller and Szpiro (2020), but allows for the TPRS coefficients in the outcome model ($\alpha_j^{\text{df.pl}}$) to be re-estimated rather than fixed. The approach differs from the approach described in Bobb et al. (2022) due to the use of the residuals in the second model and the use of AIC or BIC criteria to determine the df of the TPRS basis.

2.3 Simulations

We compared the unadjusted, KS, gSEM, E-PS, Spatial+, and df-Spatial+ methods in a simulation to evaluate their performance under different scenarios. Following the literature, we consider a continuous outcome such as birth weights or blood pressure (Bosetti et al., 2010; Stieb et al., 2012; Chan et al., 2015; Keller and Szpiro, 2020). We also consider a binary outcome, allowing for analyses of binary health outcomes, such as dementia or prevalent disease (Zhang et al., 2023).

2.3.1 Setup

Data Creation

To simulate the data, we first created a spatial grid on the domain $[0, 10] \times [0, 10]$ with an incremental step size of 0.2 and randomly sampled $n = 1000$ locations for the continuous outcome setting. For a binary outcome setting we used an incremental step size of 0.1 and randomly sampled $n = 2000$ locations. We sampled double the locations for the binary outcome due to the reduced information available in a dichotomous outcome variable. From the selected locations, we created realizations of three spatial fields, $\boldsymbol{z}_1 = (z_{11}, \cdots, z_{1n})^{\mathsf{T}}$, $\boldsymbol{z}_2 = (z_{21}, \cdots, z_{2n})^{\mathsf{T}}$,

and $\boldsymbol{z}_3 = (z_{31}, \cdots, z_{3n})^{\mathsf{T}}$. Each field was sampled from a multivariate normal distribution, $\boldsymbol{z}_l \sim N(\mathbf{0}, \boldsymbol{\Sigma}_l)$, where $\boldsymbol{\Sigma}_l$ is exponential covariance structure with a range parameter ϕ_l so that $(\boldsymbol{\Sigma}_l)_{ii'} = \exp(-||\boldsymbol{s}_i - \boldsymbol{s}_{i'}||/\phi_l)$.

Using the three spatial fields; $\boldsymbol{z}_1$, $\boldsymbol{z}_2$, and $\boldsymbol{z}_3$; we computed the exposure and unmeasured confounder using Equations (2.1) and (2.2), respectively. The non-spatial error component for the exposure and the continuous outcome are $\epsilon_{x_i} \sim N(0, \sigma_x^2)$ and $\epsilon_{y_i} \sim N(0, \sigma_y^2)$. The continuous outcome is generated as $y_i = \mu_i + \epsilon_{y_i}$, from (2.3) with $\beta_0 = 0$ and $\beta_x = \beta_f = 1$. For a discrete outcome, we sampled from a Binomial distribution with probability $p_i = 1/(1 + e^{-\mu_i})$ from (2.3) and set $\beta_0 = \text{logit}(0.15)$ and $\beta_x = \beta_f = 0.5$.

Settings Compared

We compared results across different levels of non-spatial error and differing range parameters for the spatial fields. We compared when there was no non-spatial variability in the exposure ($\sigma_x = 0$), a small amount of non-spatial variability in the exposure ($\sigma_x = 0.1$), and, when the outcome was continuous, a large amount of non-spatial variability in the exposure ($\sigma_x = 2$). We compared a moderate or large amount of non-spatial variability in the continuous outcome, corresponding to $\sigma_y = 1$ or 10, respectively.

We evaluated combinations of $\phi_1 \in \{5, 50, 150\}$ when $\phi_2 = 10$, and $\phi_3 = 100$. For the binary outcome with a small amount of non-spatial variability in the exposure ($\sigma_x = 0.1$), we also evaluated combinations of $\phi_1 \in \{5, 50, 150\}$ when $\phi_2 = 1$ and $\phi_3 = 100$. Scenarios where $\phi_1 < \phi_2$ were designed to be settings where the long-range spatial variation that is unique to the exposure, $\boldsymbol{z}_2$, is overwhelmed by the shared finer-scale spatial variation in $\boldsymbol{z}_1$ and the exposure cannot be reliably distinguished from the confounder (Gilbert et al., 2024). In the cases where $\phi_1 > \phi_2$, there exist finer-scale spatial details unique to the exposure that can be distinguished from the confounder. In these cases, we expected our estimate of β_x to have an overall lower bias.

For models with AIC or BIC selection, we considered a grid between 3 and 300 df for the continuous outcome and between 3 and 200 df for the binary outcome. For the Spatial+ and gSEM methods without penalization, we set the df at 300 for a continuous outcome and 200 for a binary

outcome. For the KS models with fixed df, we set it at 10 df. See Table B.1 in the supplement for additional details of the candidate df. For each combination set of σ_x, σ_y, ϕ_1, ϕ_2, and ϕ_3 we replicated the simulation 500 times, and calculated the root mean-squared error (RMSE) and bias of $\hat{\beta}_x$, and coverage rate of nominal 95% confidence intervals for β_x. We additionally calculated the median df selected by each approach.

Although our primary focus was on semiparametric methods, we also compared the restricted spatial regression (RSR) method introduced by Hanks et al. (2015) described in Section B.1.3.

2.3.2 Results

We first consider the continuous outcome cases, starting with the scenario where there was a moderate amount of non-spatial variation in the outcome and a small amount of non-spatial variation in the exposure ($\sigma_x = 0.1, \sigma_y = 1$: Table 2.1). When $\phi_1 = 50$, we see that the KS-AIC and KS-BIC estimators outperformed the Spatial+, gSEM, E-PS, and KS-fixed estimators in terms of both RMSE and bias. Within the KS approaches, using AIC to select the df of the TPRS basis outperformed BIC, in terms of bias (0.101 compared to 0.114), but BIC outperformed AIC in terms of RMSE (0.155, compared to 0.162). Across the Spatial+ methods, Spatial+BIC outperformed all other approaches in terms of RMSE but Spatial+AIC and Spatial+fixed outperformed Spatial+BIC in terms of bias (Table 2.1). The E-PS estimator performed better in terms of RMSE than all Spatial+ methods except Spatial+BIC and performed better in terms of the bias than Spatial+BIC and Spatial+GCV. Overall, the gSEM-fixed produced equivalent RMSE and bias to the Spatial+fixed and using GCV to determine the df produced worse estimates that fixing the df for both Spatial+ and gSEM (Table 2.1). When we set $\phi_1 = 150$, conclusions were similar to when $\phi_1 = 50$ (see Table B.2).

Now, we consider the scenario where we expected the methods to not mitigate the bias induced from spatial confounding as well ($\phi_1 = 5$) when $\sigma_x = 0.1$ and $\sigma_y = 1$ (Table 2.1). Similar to when the confounded surface has a larger range, the KS-AIC and KS-BIC estimators outperformed the Spatial+, gSEM, E-PS, and KS-fixed estimators terms of both RMSE and bias. However,

Table 2.1: Root mean squared error (RMSE), estimated bias, and coverage rates for 95% confidence intervals for the estimates of β_x in the simulation where $\phi_2 = 10$, $\phi_3 = 100$, $\sigma_x = 0.1$, and $\sigma_y = 1$. The median degrees of freedom (df) for each method is given for both the exposure and outcome model.

ϕ_1	Method	df Selector	RMSE	Bias	Coverage (95% Nominal)	Median df Selected (Exposure)	(Outcome)
50	Unadjusted	–	0.268	0.120	0.222	–	–
	Spatial+	Fixed	0.248	0.127	0.892	300	300
	Spatial+	GCV	0.403	0.352	0.598	220.4	61.2
	Spatial+	AIC	0.247	0.128	0.886	300	300
	Spatial+	BIC	0.199	0.140	0.804	70	70
	gSEM	Fixed	0.248	0.127	0.822	300	300
	gSEM	GCV	0.396	0.343	0.590	220.4	62.1
	KS	AIC	0.162	0.101	0.870	–	41.5
	KS	BIC	0.155	0.114	0.756	–	11
	KS	Fixed	0.185	0.153	0.576	–	10
	E-PS	GCV	0.209	0.135	0.888	220.4	220.2
5	Unadjusted	–	0.642	0.621	0.002	–	–
	Spatial+	Fixed	0.590	0.571	0.026	300	300
	Spatial+	GCV	0.824	0.821	0.000	229.7	126.4
	Spatial+	AIC	0.590	0.571	0.028	300	300
	Spatial+	BIC	0.620	0.611	0.000	75	75
	gSEM	Fixed	0.589	0.570	0.016	300	300
	gSEM	GCV	0.794	0.781	0.000	229.7	128.7
	KS	AIC	0.571	0.560	0.008	–	100
	KS	BIC	0.614	0.608	0.000	–	27
	KS	Fixed	0.648	0.643	0.000	–	10
	E-PS	GCV	0.601	0.590	0.000	229.7	229.5

Table 2.2: Root mean squared error (RMSE), estimated bias, and coverage rates for 95% confidence intervals for the estimates of β_x in the simulation where $\phi_2 = 10$, $\phi_3 = 100$, $\sigma_x = 0.1$ and the outcome is binary. The median degrees of freedom (df) for each method is given for both the exposure and outcome model.

ϕ_1	Method	df Selector	RMSE	Bias	Coverage (95% Nominal)	Median df Selected (Exposure)	(Outcome)
50	Unadjusted	–	0.159	0.087	0.756	–	–
	Spatial+	Fixed	0.461	0.128	0.888	200	200
	Spatial+	GCV	0.383	0.089	0.918	180.7	10.9
	Spatial+	AIC	0.461	0.128	0.888	200	200
	Spatial+	BIC	0.375	0.102	0.930	125	125
	gSEM	Fixed	0.499	0.063	0.934	200	200
	gSEM	GCV	0.383	0.080	0.918	180.7	10.9
	KS	AIC	0.174	0.017	0.952	–	7
	KS	BIC	0.157	0.057	0.914	–	3
	KS	Fixed	0.194	0.075	0.922	–	10
	E-PS	GCV	0.318	0.094	0.946	180.7	119.7
5	Unadjusted	–	0.322	0.301	0.062	–	–
	Spatial+	Fixed	0.504	0.412	0.666	200	200
	Spatial+	GCV	0.410	0.333	0.718	183.6	30.9
	Spatial+	AIC	0.504	0.412	0.666	200	200
	Spatial+	BIC	0.448	0.373	0.658	125	125
	gSEM	Fixed	0.397	0.235	0.858	200	200
	gSEM	GCV	0.380	0.289	0.780	183.6	30.7
	KS	AIC	0.310	0.276	0.504	–	19
	KS	BIC	0.316	0.297	0.198	–	5
	KS	Fixed	0.346	0.326	0.206	–	10
	E-PS	GCV	0.403	0.354	0.622	183.6	123.6

performance of all methods was worse than when $\phi_1 = 50$. In the setting where $\phi_1 = 5$, using an AIC criterion outperformed using a BIC criterion for both KS and Spatial+ methods (RMSE of 0.571 and 0.614 and bias of 0.560 and 0.608, for KS-AIC and KS-BIC methods respectively) and Spatial+fixed outperformed Spatial+BIC in both RMSE and bias. In this setting, there was also a great reduction in the coverage rate due to the inability of the approaches to effectively account for the unmeasured confounder (Table 2.1).

We next view the case where the non-spatial variation was removed and keep a moderate amount of non-spatial variation in the outcome ($\sigma_x = 0, \sigma_y = 1$: Table B.3). Across all ranges for the confounded spatial field (ϕ_1), the KS-AIC and KS-BIC methods outperformed all of the other methods in terms of RMSE and bias (Table B.3). When $\phi_1 = 5$, KS-AIC slightly outperformed KS-BIC, but for $\phi_1 = 50, 150$, we saw similar trends of AIC outperforming BIC in terms of bias but not RMSE (Table B.3).

When increasing the non-spatial variation in the exposure and keeping a moderate amount of non-spatial variation in the outcome ($\sigma_x = 2, \sigma_y = 1$: Table B.4), it became difficult to distinguish the best method as the RMSE and bias values are equivalent across methods and outperformed the spatially unadjusted model when $\phi_1 = 50, 150$. When $\phi_1 = 5$, there existed slight differences between methods where Spatial+fixed and both gSEM methods produce slightly better RMSE and gSEM-GCV produced the smallest overall bias (Table B.4).

The final continuous outcome case we considered had a small amount of non-spatial variation in the exposure and a large amount of non-spatial variation in the outcome ($\sigma_x = 0.1, \sigma_y = 10$: Table B.5). Again, we saw that the KS methodology outperformed all other methods for both RMSE and bias, expect the bias for KS-fixed when $\phi_1 = 5$ (Table B.5). Across all three range values of ϕ_1, the KS-AIC outperformed KS-BIC for both RMSE and bias (Table B.5). We also noted a larger difference between RMSE and bias for the estimators in this setting due to the RMSE being influenced by the large amount of non-spatial variation added to the outcome. This large amount of non-spatial variation obscures any short range patterns in the spatial structure. Therefore, the spatial splines cannot effectively model the spatial variation and are only able to

model long range patterns. Because of the ineffectiveness of modeling the spatial variation with the spatial splines, the unadjusted estimator has the best RMSE out of all estimators compared (Table B.5).

We now consider the cases where the outcome is binary. Generally, the results show similar trends to the continuous outcome, where the KS methodology outperformed all other methods in terms of RMSE and bias. The exceptions to this are when the range parameter for the confounded surface is large ($\phi_1 = 150$: Table B.6) and the spatially unadjusted model has the smallest RMSE; and when the range parameter for the confounded surface is small ($\phi_1 = 5$: Table 2.2) and there is no spatial variation in the exposure and gSEM-fixed has the smallest bias. Among the KS methods, using KS-AIC had a smaller bias and a better coverage rate than KS-BIC except when $\phi_1 = 150$ and $\sigma_x = 0.1$ (Table B.6) and KS-BIC had a smaller RMSE than KS-AIC except when $\phi_1 = 5$ and $\sigma_x = 0.1$ (Table 2.2). Differing from the setting with a continuous outcome, among the Spatial+ methodologies, using GCV penalization had a better bias for all ϕ_1 and both values of σ_x (Tables 2.2 and B.6). Spatial+GCV also had a better RMSE and coverage rate except when $\phi_1 = 50, 150$ and $\sigma_x = 0.1$ where using the Spatial+BIC has a better RMSE (Tables 2.2 and B.6).

When we considered the RSR estimators (Table B.9), we noted that the unconditional estimators perform similar to the spatially unadjusted estimators and that the unconditional and conditional estimators never outperform the KS-AIC and KS-BIC estimators (Table B.9).

Overall from this simulation, KS-AIC generally has the best performance. Across all settings compared, KS-AIC consistently produced the smallest or one of the smallest biases and produced smaller RMSE relative to the other methods we compared.

2.4 Data Analysis

We apply these adjustment methods to an analysis of ambient fine particulate matter ($PM_{2.5}$) and temperature and birth weight in a cohort of Colorado births. We restrict the analysis to only include full-term (estimated gestational age greater than 36 weeks), singleton births where the estimated year of conception was between 2007 and 2017. Our outcome of interest is birth weight for

gestational age z-scores (BWGAZ). Lower birth weights are known to correlate with adverse health outcomes in early life, which may persist throughout the life course (Hack et al., 1995; Ogonowski et al., 2014). Measures of two ambient environmental factors were used as the exposures of interest: average $PM_{2.5}$ concentration and maximum daily temperature. We obtained predictions of $PM_{2.5}$ at census tract centriods through the down-scaled models published by the US Environmental Protection Agency (https://www.epa.gov/hesc/rsig-related-downloadable-data-files) and temperature on a 4km grid that was linked to census tracts (Abatzoglou, 2013). Both measures were obtained at the daily level and averaged over the third trimester for each individual to represent a single exposure value during late pregnancy. Our final analysis data set contains 611,096 live births across 1,240 census tracts in all 64 counties across Colorado.

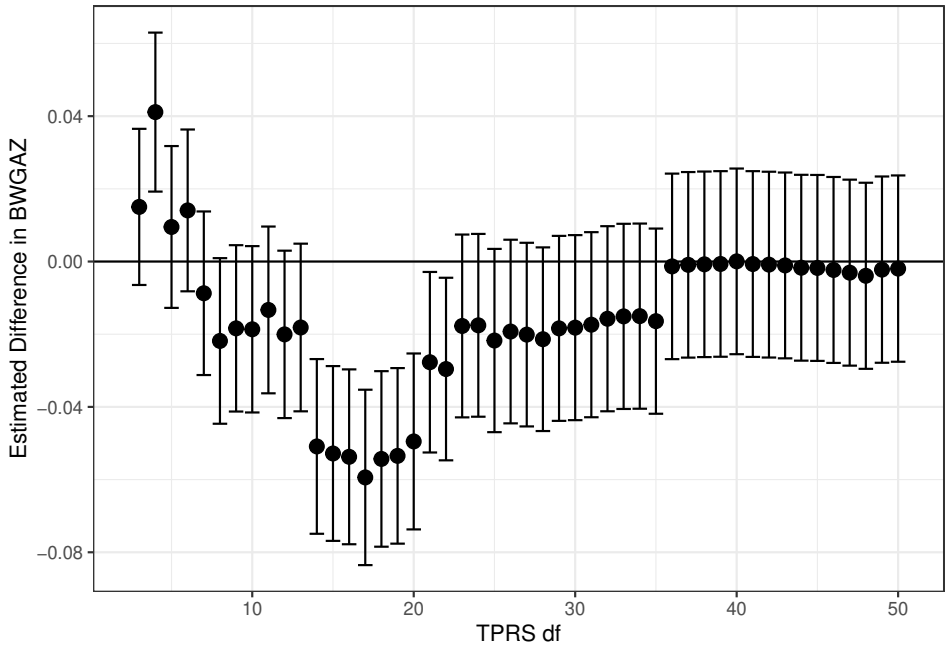
We fit ten models to the data: without spatial adjustment, Spatial+ (Fixed, GCV, AIC, BIC), gSEM (Fixed, GCV), KS (AIC, BIC), and E-PS. We computed the TPRS based on the census tract centroids using the mgcv and spconf packages (Wood, 2011; Keller and Rainey, 2024). For all models, 50df was the maximum TPRS df. We further adjusted for the sex of the child; the pregnant individual's age (as natural splines with 3 df), BMI, race and ethnicity (and their interaction), marital status, household income, and highest amount of education; and indicators for whether prenatal care was obtained and whether the pregnant individual smoked previously. We also adjusted for the month and year of conception to account for seasonal and long-term temporal trends of environmental exposures. For models with third trimester $PM_{2.5}$ as the exposure of interest, we adjusted for average $PM_{2.5}$ in the first and second trimester (similar for temperature). Due to the high correlation between the exposures in the first and second trimester and the third trimester, we investigated the first step in the KS method both adjusting for and not adjusting for the first and second trimester exposures. In both scenarios, the same df was selected and therefore produced the same point estimates.

2.4.1 Results

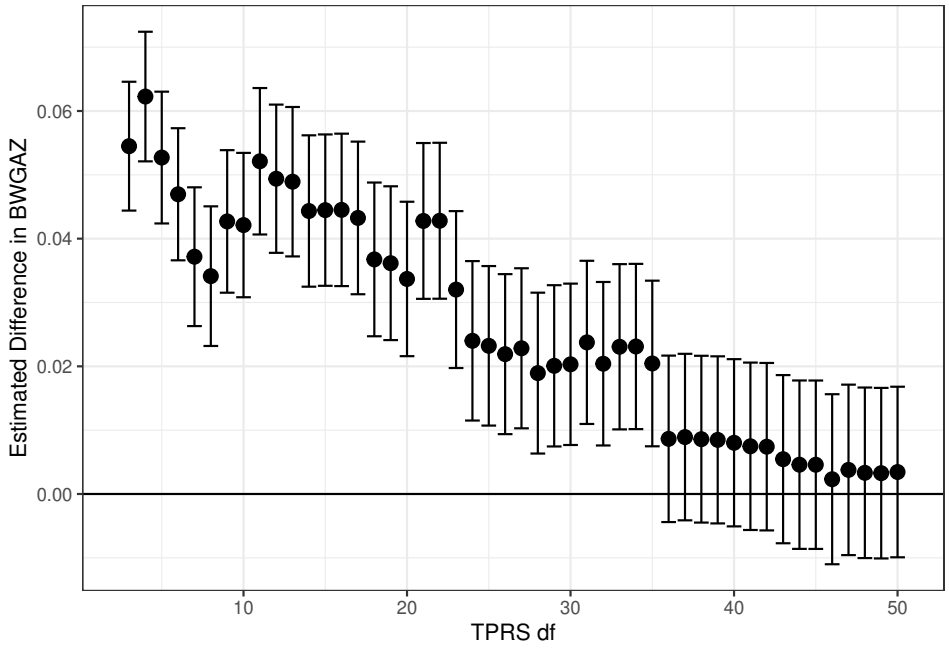
The third trimester $PM_{2.5}$ concentrations ranged from 2.3μg/m^3 to 17μg/m^3 with mean (SD) of 7 (1.5) μg/m^3 and the maxmimum temperatures ranged from -5.6°C to 37°C with mean (SD) of 18 (8.3) °C. BWGAZ ranged from -7.8 to 6.2 with a mean (SD) of -0.074 (0.87). The models that were not adjusted for space estimated a positive estimated difference of 0.031 (95% CI: 0.010, 0.052) in BWGAZ for each 10μg/m^3 difference in $PM_{2.5}$ and an estimated difference of 0.068 (95% CI: 0.058, 0.078) in BWGAZ for each 10°C difference in maximum temperature (Table 2.3).

When using a Spatial+, gSEM, KS, or E-PS model that adjusts for space via the inclusion of TPRS, the estimated association between the environmental exposure, either $PM_{2.5}$ or temperature, and BWGAZ was reduced and no longer statistically significant (Table 2.3). The estimated difference of BWGAZ for each 10μg/m^3 increment of $PM_{2.5}$ ranged from -0.0038 (95% CI: -0.0292, 0.0217) for gSEM using GCV, to -0.0013 (95% CI: -0.0276, 0.0237) for KS-BIC. The estimated difference of BWGAZ for each 10°C increment of maximum temperature ranged from 0.0033 (95% CI: -0.0100, 0.0170) for df-Spatial+ selecting via BIC, to 0.0086 (95% CI: -0.0044, 0.0217) for KS selecting via BIC. For the models selecting df using AIC or BIC, the maximum or close to the maximum df (50 df) was selected, except for KS BIC which selected 36 df (Table 2.3). The estimated associations of the KS-fixed method across all choices of df are given in Figure 2.1a for the $PM_{2.5}$-exposure and Figure 2.1b for the temperature-exposure. The models that used GCV penalization had effective df close to the maximum df as well.

Inspection of Figure 2.1 shows discontinuities as more splines are added to the model, e.g. between 13 and 14 df, 20 and 21 df, and 36 and 37 df. These jumps are due to spatial patterning of the additional spline, e.g. spline 21. There is a high correlation between the additional spline (21st spline) and the residual spatial information in the exposure and outcome after adjusting with the first 20 splines in the TPRS basis (Supplemental Table B.10). This correlation is much higher for some specific spline basis functions compared to other basis functions (see Supplemental Section B.2).



(a)



(b)

Figure 2.1: The estimated difference in BWGAZ by number of df included in the TPRS basis ($\hat{k}_3$ in Eq. 2.4) for each (a) difference of $10\mu/\text{m}^3$ in average $PM_{2.5}$ at the census tract level during the third trimester and (b) difference of 10°C in average maximum daily temperature during the third trimester.

Table 2.3: Point estimates and 95% confidence intervals of the difference in BWGAZ associated with a 10 unit difference (μg/m^3 for $PM_{2.5}$ and °C for maximum temperature) in exposure in the Colorado birth cohort study, and selected degrees of freedom (df) for the thin-plate regression spline basis in the exposure and outcome models

Exposure	Method	df Selector	Estimate	95% Confidence Interval	df Selected (Exposure)	df Selected (Outcome)
$PM_{2.5}$	Unadjusted	–	0.0312	(0.0102, 0.0522)	–	–
	Spatial+	Fixed	-0.0019	(-0.0276, 0.0237)	50	50
	Spatial+	GCV	-0.0023	(-0.0279, 0.0234)	49.8	47.6
	Spatial+	AIC	-0.0019	(-0.0276, 0.0237)	50	50
	Spatial+	BIC	-0.0019	(-0.0276, 0.0237)	50	50
	gSEM	Fixed	-0.0037	(-0.0291, 0.0218)	50	50
	gSEM	GCV	-0.0038	(-0.0292, 0.0217)	49.8	48.5
	KS	AIC	-0.0019	(-0.0276, 0.0237)	–	50
	KS	BIC	-0.0013	(-0.0269, 0.0242)	–	36
	E-PS	GCV	-0.0022	(-0.0278, 0.0235)	49.8	49.8
Maximum Temperature	Unadjusted	–	0.0678	(0.0578, 0.0777)	–	–
	Spatial+	Fixed	0.0034	(-0.0099, 0.0168)	50	50
	Spatial+	GCV	0.0041	(-0.0092, 0.0175)	49.6	47.7
	Spatial+	AIC	0.0034	(-0.0099, 0.0168)	50	50
	Spatial+	BIC	0.0033	(-0.0100, 0.0170)	48	48
	gSEM	Fixed	0.0037	(-0.0093, 0.0167)	50	50
	gSEM	GCV	0.0038	(-0.0092, 0.0169)	49.6	48.5
	KS	AIC	0.0034	(-0.0099, 0.0168)	–	50
	KS	BIC	0.0086	(-0.0044, 0.0217)	–	36
	E-PS	GCV	0.0039	(-0.0094, 0.0173)	49.6	49.6

The results from the spatially-unadjusted models suggest that higher ambient pollution levels and temperatures are associated with higher birth weights, on average. These results are in the opposite direction than would be expected based on substantial literature (Bosetti et al., 2010; Lakshmanan et al., 2015; Lamichhane et al., 2015; Šrám et al., 2005). This may be due in part to confounding by elevation across Colorado. $PM_{2.5}$ and maximum temperatures are lower in the Rocky Mountain region and higher in the Eastern Plains with $PM_{2.5}$ being highest along the Front Range corridor. If these spatial patterns, along with the spatial variation in BWGAZ, are not accounted for, bias may be induced due to spatial confounding. By adjusting for space to model or remove the spatial confounding in the data, we were able to produce estimates for $PM_{2.5}$ that are no longer in the opposite expected direction. The estimates that do adjust for space, however, suggest a lack of evidence in this cohort to determine an association between ambient pollution levels and maximum temperatures in the third trimester and birth weights, on average.

2.5 Discussion

We have examined multiple methods for mitigating spatial confounding bias and have made recommendations based on the amount of non-spatial variation in the exposure and outcome. All methods use a TPRS basis to model the confounded spatial variation. We compared methods from Thaden and Kneib (2018), Keller and Szpiro (2020), Dupont et al. (2022), and Bobb et al. (2022). We then introduced a method that uses the structure from Dupont et al. (2022) but instead of arbitrarily selecting a large number of degrees of freedom, use the degrees of freedom selection ideas from Keller and Szpiro (2020). While the simulation comparisons elucidated differences, it is important to note for most analyses, the values of ϕ_1, ϕ_2, and ϕ_3 will be unknown. Thus, it is difficult to select the method that will always produce the best overall results. However, in many settings it is possible to estimate the amount of non-spatial variation that is present in the exposure and outcome. Thus, decisions can be made for specific scenarios based on estimated values of σ_x and σ_y. For setting with large environmental epidemiology cohorts using predicted exposures, there is limited or no non-spatial variation in the exposure, and in these cases we recommend using

Keller and Szpiro's method using the AIC criterion since KS-AIC produced RMSE and bias values in this setting.

We also investigated scenarios where the outcome was continuous and had a small amount of non-spatial error ($\sigma_y = 0.1$). In these scenarios, we saw very low coverage rates across all of the models that we tested. However, in environmental epidemiology, the non-spatial error in the outcome is usually larger; and therefore, these scenarios are less likely and poor coverage in this scenario is less of a concern.

All the methods compared focus on modeling the spatial structure using a TPRS structure in a semiparametric framework. Further work extending the comparison past methods beyond those using TPRS could be beneficial, but computationally intensive approaches may be challenging to implement for large cohorts or complex survey designs. Large scale environmental epidemiological studies often contain multiple observations of an individual over a period of time. In the Colorado birth cohort study, we had daily measurements of air pollution and temperature across the entire pregnancy but had to use an average over the third trimester as our exposures of interest. Future work on extending the cross-sectional methods to described in this paper to include a time-varying component would be beneficial to many epidemiological studies.

In summary, we presented and compared several approaches that use TPRS to mitigate bias due to the presence of spatial confounding. While all methods improved the estimates obtained from the spatially-unadjusted model, we provided the recommendation to use the method developed by Keller and Szpiro (2020) using the AIC selection criterion for settings most common to large environmental epidemiology cohort studies.

Chapter 3

spconfShiny: An R Shiny application for calculating the spatial scale of smoothing splines for point data

3.1 Introduction

In studies that use regression models to estimate relationships between spatially-varying variables, such as air pollution concentrations or temperature and health outcomes, spatial confounding should be accounted for in the model (Paciorek, 2010). Spatial confounding is defined as the presence of any unmeasured, spatially-varying factor that impacts a spatially-varying response variable when the main predictor is also spatially-varying. In epidemiological contexts, a common way to account for this confounding is to include adjustment for space via spatial splines (Chan et al., 2015; Keller et al., 2022; Zhang et al., 2023). Several two-step approaches have been introduced that incorporate splines in differing models and also have different approaches for choosing the number of splines to include in the models (Bobb et al., 2022; Dupont et al., 2022; Keller and Szpiro, 2020; Thaden and Kneib, 2018). However, the relationship between the amount of spatial smoothing with a particular number of splines and the corresponding geographic distance is context-dependent. Generally, as additional splines are added to a model, finer spatial details can be modeled. But the size and shape of the geographic region can also impact the magnitude of the corresponding smoothing. A practical procedure is needed for interpreting the number of splines included in a spatial model in terms of spatial distances across different geographic regions.

R Shiny applications have become a beneficial tool to help researchers visualize and implement different spatial methodologies in their research (Salehi et al., 2021; Adin et al., 2019; Figueira et al., 2024; Aparicio et al., 2024; Silva et al., 2023; Johnson et al., 2021). For example, Salehi et al. (2021) created an application for the spatial visualization of COVID-19 data and Adin et al. (2019) developed one for spatiotemporal disease mapping. Figueira et al. (2024) developed an

application, BAYSPINS, that implements a Bayesian approach for species distribution models, creating a tool for researchers who are less experienced with those types of models or researchers who want a quick way to implement them. In other contexts, Aparicio et al. (2024) developed the Mr.Bean app to visualize spatial information from agricultural field trials, Silva et al. (2023) developed the movedesign app for animal movement studies, and Johnson et al. (2021) developed an application, MBGapp, aimed at teaching geostatistical analyses to researchers that do not have much statistical training.

To aid in the interpretation of spatial smoothing for point-level data, we present an R Shiny application called `spconfShiny`, that calculates the spatial distance corresponding to a chosen number of splines for a particular set of spatial locations. spconfShiny implements a modification, described below, of a procedure first developed by Keller and Szpiro (2020) for an effective bandwidth statistic. The core method is implemented in an accompanying R package, `spconf` (Keller and Rainey, 2024). Together, the package and application provide a user-friendly platform for researchers to calculate spatial scales for smoothing data from a custom set of geographic locations.

3.2 Methods

3.2.1 Statistical Framework

The motivating context for this work is an epidemiological analysis of the association between a health outcome, $y_i(\boldsymbol{s}_i)$, and a spatially-varying exposure, $x_i(\boldsymbol{s}_i)$, for each individual with corresponding location, $\boldsymbol{s}_i$. We assume that there are other measured covariates, $\boldsymbol{w}_i \in \mathbb{R}^p$. Unmeasured spatial confounding is a concern, so J spatial splines, which we denote $h_j(\boldsymbol{s}_i)$ for $j = 1, \ldots, J$, are included in the model (Bobb et al., 2022; Keller and Szpiro, 2020). A generalized linear model for estimating health effect associations in this context is:

$$g(\mathrm{E}[y_i(\boldsymbol{s}_i)]) = f(x_i(\boldsymbol{s}_i), \boldsymbol{\beta}) + \boldsymbol{w}_i^{\mathsf{T}}\boldsymbol{\gamma} + \sum_{j=1}^{J} \alpha_j h_j(\boldsymbol{s}_i) \tag{3.1}$$

where $\mathrm{E}[y_i(\boldsymbol{s}_i)]$ is the mean of the response, $g(\cdot)$ is a link function, $f(x_i(\boldsymbol{s}_i), \boldsymbol{\beta})$ is an exposure-response function, $\boldsymbol{\gamma} \in \mathbb{R}^p$ are regression coefficients for the measured covariates, and $\alpha_j \in \mathbb{R}$ are the regression coefficients of the splines. We choose $g(\cdot)$ to be the identity link, assuming that our response is continuous; however, other links may be assumed if the response is discrete.

Increasing the number of splines, J, included in (3.1) allows for finer scale spatial adjustments in the model. However, larger values of J do not necessarily equate to lower bias in the exposure-response association estimate (Keller and Szpiro, 2020). Decisions of how many splines to include should consider whether exposures are predicted or measured and the magnitude of non-spatial variation in exposure due to the possibility of over adjusting and nullifying the estimated association by adding too many splines (Keller and Szpiro, 2020; Bobb et al., 2022). The choice of the number of splines to include in the model is beyond the scope of this work but is an active area of research.

The target of the inferential analysis in (3.1) is to estimate the exposure-response relationships summarized by the parameter $\boldsymbol{\beta}$; however, the goal of this work is to provide an interpretation of the scale of the spatial splines $h_1(\boldsymbol{s}_i), \ldots, h_J(\boldsymbol{s}_i)$ so that the estimate for $\boldsymbol{\beta}$ can be interpreted more precisely.

3.2.2 Effective bandwidth of spatial splines

The common choice of basis to create spatial splines is the thin-plate regression spline (TPRS) basis (Wood, 2003), which can be calculated in R via the `mgcv` package (Wood, 2011). For unpenalized splines, the degrees of freedom (df) of a basis is equal to the number of splines, represented by J in (3.1). To interpret the choice of df for the TPRS basis, we propose an effective bandwidth, which we denote $\hat{k}$, using an approach adapted from a procedure developed by Keller and Szpiro (2020). We interpret the effective bandwidth as the approximate minimum radius of the area over which points are smoothed. In the context of (3.1), we can also think of the effective bandwidth as, given a specific location, the minimum distance at which confounding is being adjusted. In an epidemiological context, the inclusion of spatial splines can be interpreted as a

means for adjusting for confounding by location over a range given by the effective bandwidth. Smaller values of $\hat{k}$ mean that fewer locations are averaged across and thus finer-scale spatial details are adjusted for in (3.1).

The process of computing the effective bandwidth is illustrated in Figure 3.1 and in Algorithm 1. To determine $\hat{k}$, we first calculate a value $\hat{k}_i$ for each location i (or a random subset of locations). For a set of points $\mathcal{S} = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\}$, we first obtain the Euclidean distance matrix, $\boldsymbol{D} \in \mathbb{R}^{n \times n}$ between the given coordinates (step A in Figure 3.1). A TPRS basis, $\boldsymbol{H} \in \mathbb{R}^{n \times (df+1)}$, is computed based on $\boldsymbol{D}$ and is used to compute the smoothing matrix, $\boldsymbol{S} = \boldsymbol{H}(\boldsymbol{H}^\mathsf{T}\boldsymbol{H})^{-1}\boldsymbol{H}^\mathsf{T} \in \mathbb{R}^{n \times n}$ (step B in Figure 3.1). For each column $\boldsymbol{S}[, i]$, we order the values by the corresponding distances to all other points and find the distance at which the values from $\boldsymbol{S}[, i]$ first cross zero (step C in Figure 3.1). The median of these distances, $\hat{k}_i$, is what determines $\hat{k}$.

Algorithm 1 Computational algorithm for computing the effective bandwidth

Require: $\mathcal{S} \in \mathbb{R}^{n \times 2}$
  Initialize $k_{vec}[n]$
  Compute TPRS basis, $\boldsymbol{H} \in \mathbb{R}^{n \times (df+1)}$
  $\boldsymbol{D} = \text{distance}(\mathcal{S}, \mathcal{S})$
  $\boldsymbol{S} = \boldsymbol{H}(\boldsymbol{H}^\mathsf{T}\boldsymbol{H})^{-1}\boldsymbol{H}^\mathsf{T}$
  for $i \in 1 : n$ **do**
    order $\boldsymbol{S}[, i]$ by increasing $\boldsymbol{D}[, i]$
    $k_{vec}[i] = min(\boldsymbol{D}[, i])$ where $\boldsymbol{S}[, i] < 0$
  end for
  $k = \text{median}(k_{vec})$
  return k

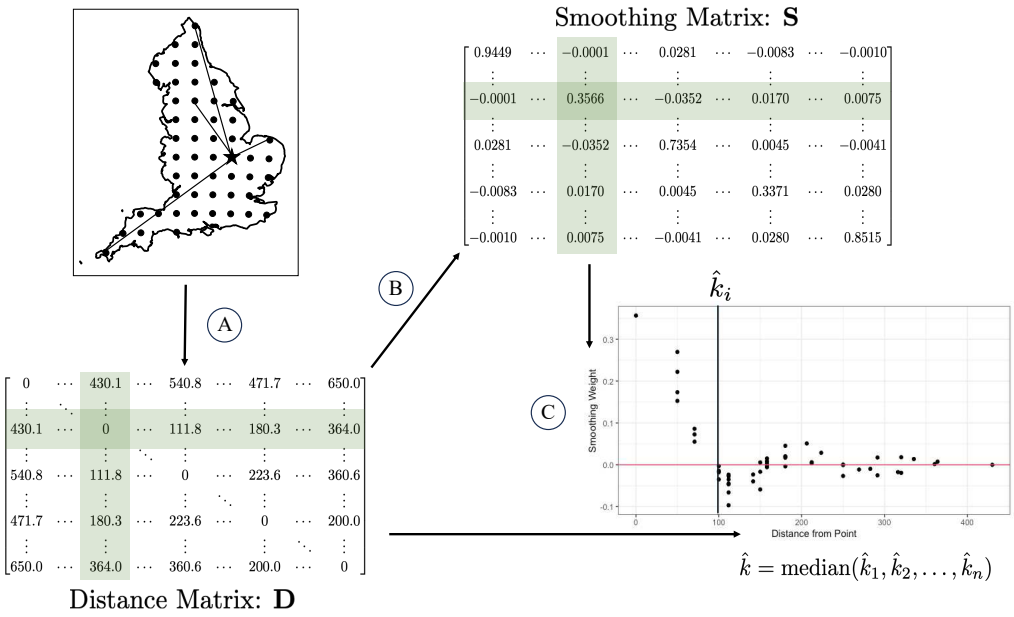


Figure 3.1: Visual representation of the process of computing the effective bandwidth. Step A computes the distance matrix, **D**, of the 50km grid across England. A smoothing matrix, **S**, is computed from the information in **D** via a TPRS basis in step B. The highlighted rows and columns correspond to the distance and smoothing values for the starred point. For each point in the grid, the smoothing weights are ordered by distance and the smallest distance with a negative smoothing weight, $\hat{k}_i$, is obtained and visually represented in step C. Finally, the effective bandwidth is computed by taking the median of the $\hat{k}_i$.

Keller and Szpiro's effective bandwidth

The procedure we developed to compute the effective bandwidth contrasts the methodology developed by Keller and Szpiro (2020) in how the relationships between the distances and the smoothing weights are used. In place of our Step C (Figure 3.1), Keller and Szpiro (2020) fit a loess curve to the smoothing weights $(\boldsymbol{S}[,i])$ as a function of distance, which also requires selecting a span value that controls the proportion of points included in the smoothing. They then predict smoothing weights for a set sequence of new distances and define the effective bandwidth as the distance at which the median predicted smoothing weights first cross zero. Our proposed methodology orders the empirical smoothing weights by distance and finds the smallest distance that has a negative smoothing weight, effectively finding where the points first cross the x-axis when plotting the smoothing weights by distance. The median of the selected distances determines the effective bandwidth. Compared to the original approach of Keller and Szpiro (2020), our approach does

not require the user to input the span for the smoothing calculations. This makes our approach faster and more user-friendly for differing geographic regions. However, we expect there to be a difference between the two computations. Using a loess curve to compute the effective bandwidth averages over a neighborhood of distances, creating an average radius for the area that is smoothed. In comparison, our proposed method takes the first point that below zero, not considering any other points, creating a minimum radius for the same area.

3.3 Computing the Effective Bandwidth in `spconfShiny`

`spconfShiny` is an interactive Shiny web application based on the `spconf` package in the R language, updated with our adaptation of the effective bandwidth (Keller and Rainey, 2024). We have integrated the modified effective bandwidth into `spconf`, which also retains functions for computing the version of the bandwidth measure proposed by Keller and Szpiro (2020).

3.3.1 Coordinate Input Options

In `spconfShiny`, we provide three different options to obtain spatial coordinates to compute the effective bandwidth:

- Create gridded coordinates in the application
- Select a set of preloaded coordinates
- Upload coordinates from a user file.

To create gridded coordinates, the length and the width of the grid must be entered and the user must select the distance between points (grid increment size). The preloaded coordinates in the application currently include the countries of England, India, Ireland, Northern Ireland, and the contiguous United States with grid sizes of 10km, 50km, 10km, 1km, and 50km, respectively. The user uploaded coordinates should be in .csv format, and the user must indicate the names of the columns that include the spatial coordinates.

3.3.2 Effective Bandwidth Options

The maximum number of splines must be selected in order to compute the effective bandwidth. The application offers the choice of 10, 25, 100, 300, or 500 splines. However, the number of splines may not exceed the number of coordinates included in the computations. The calculations slow as the number of coordinates increases; therefore, the application offers the option to subsample the coordinates to 1000, 2000, or 5000 locations to reduce computation time. If the number of coordinates in the computation is smaller than the selected number of points to subsample, all coordinates will be used.

3.3.3 Computing the Effective Bandwidth

To compute the effective bandwidth, the application first computes unpenalized TPRS on the coordinates via the computeTPRS() function from the `spconf` package (Keller and Rainey, 2024) with the chosen maximum number of splines. The computeTPRS() function relies on the `mgcv` and `stats` packages (Wood, 2011; R Core Team, 2023). Then, for each df between 3 and the maximum df, the effective bandwidth is computed using the compute_effective_range() function from the `spconf` package (Keller and Rainey, 2024), which implements Algorithm 1 and relies on functions from the `stats` and `flexclust` packages (R Core Team, 2023; Leisch, 2006).The application provides the output in both a tabular and graphical form, selected by switching tabs. A plot of the coordinates is also displayed in a third tab. The tabular results are available to download in .csv format.

3.3.4 Shiny Implementation

The `spconfShiny` application (deployed at https://g2aging.shinyapps.io/spconfShiny/) is implemented by the `Shiny` package (Chang et al., 2023) and the Shiny implementation also uses the `shinyjs`, `shinyWidgets`, and `bslib` packages (Attali, 2021; Perrier et al., 2023; Sievert et al., 2023) with plots created by `ggplot2` (Wickham, 2016). Additional parallelization of the smoothing curve estimation is done by the `parallel` package (R Core Team, 2023).

3.4 Demonstration of spconfShiny across different geographic regions

To demonstrate the utility of the application, we compared spatial bases created across England, India, Ireland, Northern Ireland, and the contiguous United States, which represent a range of different geographic sizes and are locations of current studies investigating the impacts of aging on cognition (Lee et al., 2021). We obtained shapefiles for these countries from Natural Earth (Earth, 2009). For each country, we created grids using Transverse Mercator projected coordinate system for England (1km, 10km, and 25km), Ireland (1km, 10km, and 25km), and Northern Ireland (1km and 10km) and Lambert Conformal Conic projected coordinate system for India (10km, 25km, and 50km) and the United States (10km, 25km, and 50km).

Using England with a 25km grid as an example, we uploaded the coordinates in the 'File Input:' section of the application. We then selected to compute the effective bandwidth for 100 splines and used either all points in the dataset or sampled 5000, whichever was smaller. After clicking the compute button, we downloaded the table of effective bandwidths and summarized the results for 5, 10, 25, 100 df in Table Table 3.1. An image of the application is shown in Figure 3.2. We proceeded with the other countries and grid sizes similarly. For countries with grids that have more than 300 points, 300 df was also summarized in the table.

Table 3.1: Effective bandwidth estimates, interpretable in kilometer distances, for thin-plate regression splines evaluated on different grid sizes across five countries. df indicates degrees of freedom.

Country	Grid Size	5 df	10 df	25 df	100 df	300 df
England	1km	156.4	124.3	77.2	37.6	21.0
	10km	160.3	125.3	80.0	40.0	28.3
	25km	167.7	127.5	79.1	50.0	–
India	10km	869.8	628.1	386.4	190.0	106.3
	25km	822.3	636.4	391.3	195.3	111.8
	50km	838.2	650.0	400.0	200.0	141.4
Ireland	1km	125.0	88.1	56.3	27.9	15.6
	10km	130.0	92.2	58.3	30.0	20.0
	25km	127.5	100.0	70.7	35.4	–
Northern Ireland	1km	53.5	39.5	25.0	12.6	7.1
	10km	56.6	41.2	28.3	14.1	–
United States	10km	1178.3.0	930.0	593.0	294.1	162.8
	25km	1153.9	927.7	594.2	300.0	167.7
	50km	1192.7	948.7	602.1	304.1	180.3

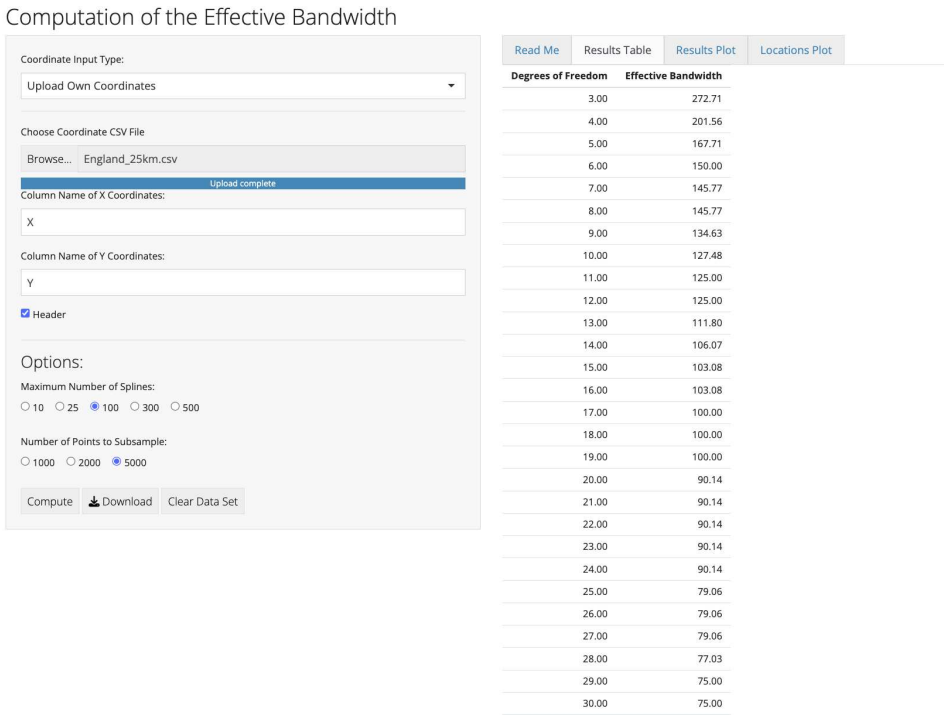


Figure 3.2: spconfShiny output for user inputted 25km grid across England.

3.4.1 Comparison of the effective bandwidth

Among the five countries that we compared, the least number of points that was considered was 115 for the 25km grid across Ireland and the most points considered was 130,382 for the 1km grid across England (Table 3.2). The smallest area that we compared was Northern Ireland, and the largest area that we compared was The United States (boundary height and width of 436km and 117km, and 2890km and 4610km, respectively:Table 3.2). Comparing the same df for the different countries, on the same grid size, $\hat{k}$ is smaller for smaller countries compared to larger countries: $\hat{k}$ of 41.2km, 92.2km, 125.3km, 628.1km, and 930.0km, for Northern Ireland, Ireland, England, India and the United States, respectively, for a TPRS basis with 10df on a 10km grid (Table 3.1, Figure 3.3).

Table 3.2: Characteristics of grids used to compute the effective bandwidths

Country	Boundary Width (km)	Boundary Height (km)	Points in Grid 1km	10km	25km	50km
England	567	646	130,382	1,302	210	–
India	2,840	3,090	–	32,558	5,217	1,300
Ireland	303	436	69,431	701	115	–
Northern Ireland	117	141	14,250	141	–	–
The United States	4,610	2,890	–	79,230	12,665	3,173

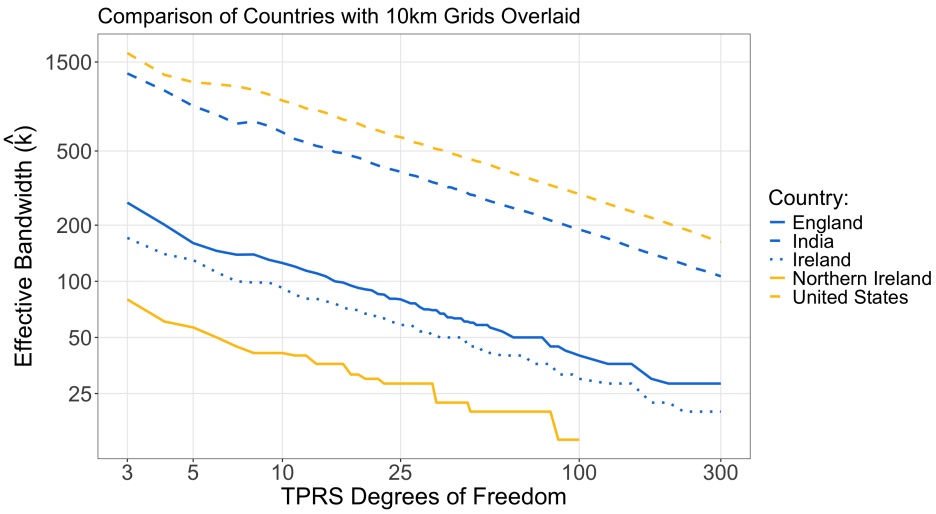


Figure 3.3: Comparison of the effective bandwidth computed for TPRS created on the 10km grid across England, India, Ireland, Northern Ireland, and the United States

Comparing different grid sizes for the same country, for the same df, the grid size does not have a meaningful influence on $\hat{k}$ ($\hat{k}$ of 628.1km, 636.4km, and 650.0km, for a TPRS basis of India with 10df with grid sizes of 10km, 25km, and 50km, respectively: Table 3.1). However, a user must still have reasonably fine resolution across the area as there must be more points than df included in the model.

3.4.2 Using the effective bandwidth in epidemiological studies

Ideally, the choice of the effective bandwidth, or number of splines included in (3.1), should be made before completing an analysis. When selecting an effective bandwidth, researchers should consider the relationship between the effective bandwidth, the complexity of the model, and the amount of spatial smoothing induced. Smaller effective bandwidths require more spatial splines to be included in the model, increasing the model complexity due to increasing the number of coefficients needed to be estimated. However, as stated previously, including more splines does not always equate to more accurate exposure-response association estimates (Keller and Szpiro, 2020). The number of locations also affects the effective bandwidth since the maximum number of splines that can be created is equivalent to the number of locations. Thus, some effective bandwidths may not be attainable due to the lack of spatial information in the data. `spconfShiny` can facilitate comparisons between similar sized countries for researchers who want to ensure the same amount of spatial smoothing. This can either be done by selecting an effective bandwidth, and determining the df needed for each country to spatially smooth at that range; or selecting the proportion of area of each country to smooth over, determining the effective bandwidth necessary for each country to achieve that proportion, and then determining the df needed for that effective bandwidth.

Suppose we wanted to compare a minimum smoothing radius of 100km in England and Ireland with a 10km grid. Using the Shiny application, we determine that we will need to include 7 df in the analysis for Ireland and 15 df in the analysis for England. However, if we want to smooth over the same proportion of area, for example 0.1 (i.e., 10% of the region), we need effective bandwidths of approximately 64km for England and 52km for Ireland, corresponding to including 36 df and 32 df in the analysis, respectively. Similarly, suppose we wanted to compare India and the United States with a 50km grid and want a minimum smoothing radius of 500km, we need 16 df included in the analysis for India and 36 df included in the analysis for the United States.

3.4.3 Comparison with alternative approaches

Finally, we provide two sensitivity analyses: a comparison of our proposed variant of the effective bandwidth with the original method of Keller and Szpiro (2020), and a comparison of our proposed approach using TPRS splines and using Duchon splines Duchon (1977). First, we compare the variant of the effective bandwidth with the original method for England with a 10km grid and the United States with a 50km grid using bases containing 5, 10, 25, and 100 df. For comparison with our proposed approach, we applied Keller and Szpiro (2020) method with spans of 0.1 and 0.5 (representing 10% and 50% of the data included in the loess curve smoothing). The original effective bandwidth produces larger values for the bandwidth than our variant (Table 3.3). This difference is to be expected since the proposed approach calculates minimum smoothing radius while the Keller and Szpiro (2020) approach calculates an average smoothing radius. The difference between the two methods decreases as the number of df in the basis increases (Table 3.3). Although both methods provide an effective bandwidth estimate, it is important that researchers use the same method when comparing across contexts.

Table 3.3: Effective bandwidth estimates for England with a 10km grid and the United States with a 50km grid comparing the original method of computing the effective bandwidth and our proposed computation. span = 0.1 and span = 0.5 indicate the original method with stated spans for the loess computation. df indicates degrees of freedom.

Country	Method	5 df	10 df	25 df	100 df
England	span = 0.1	268.6	156.9	98.2	52.9
	span = 0.5	268.2	157.9	102.4	87.6
	new	160.3	125.3	80.0	40.0
United States	span = 0.1	1660.4	1120.0	682.0	334.9
	span = 0.5	1658.7	1126.8	708.8	581.3
	new	1192.7	948.7	602.1	304.1

We use the same countries and degrees of freedom to compare the choice of spatial basis. We compare the TPRS basis with low rank Duchon splines. Duchon splines are a broader class

of spatial splines that encompasses thin plate splines (Duchon, 1977). For our comparison, we used the low-rank form of Duchon splines, implemented in the `mgcv` package (Wood, 2011) The Duchon splines are used as an input for the compute_effective_range() function from the `spconf` package, which provided the calculation of the effective bandwidth underlying the `spconfShiny` package (Keller and Rainey, 2024). The Duchon splines produce smaller effective bandwidths than the TPRS for smaller degrees of freedom, eventually converging to the same bandwidth as the df increases (Table 3.4.) While either set of splines could be used in practice, we implement only TPRS in `spconfShiny` due to their widepspread use in spatial analyses.

Table 3.4: Effective bandwidth estimates for England with a 10km grid comparing using TPRS or low rank Duchon splines to compute the spatial basis. df indicates degrees of freedom.

Country	Basis	5 df	10 df	25 df	100 df
England	TPRS	160.3	125.3	80.0	40.0
	Duchon	122.1	100.0	70.0	40.0
United States	TPRS	1192.7	948.7	602.1	304.1
	Duchon	930.1	743.3	522.0	304.1

3.5 Conclusion

`spconfShiny` is an R Shiny application that creates a user-friendly interface for the computation of the effective bandwidth for spatial splines. The effective bandwidth quantifies the amount of spatial smoothing induced in a model by including a given number of spatial splines in a model. Using the effective bandwidth, we can compare the impact of spatial smoothing across different geographic regions for differences in size and shape. As seen in our demonstration of `spconfShiny`, when creating models that will be applied to studies in different sized regions, different degrees of freedom should be used to model the same level of spatial detail and the smaller region will require including fewer splines compared to the larger region.

Chapter 4

Mitigating Spatial Confounding in Longitudinal and Time-to-Event Models

4.1 Introduction

Environmental epidemiology studies commonly examine the health outcomes of individuals over both a large geographic area and a given time period (Nawrot et al., 2006; Fischer et al., 2015; the CHILD study investigators et al., 2015; Di et al., 2017; Zhang et al., 2023). Due to the focus on the health effects of long-term environmental exposures and the spatial nature of both these exposures and the underlying trends in many health outcomes, any unmeasured spatially varying factor may induce bias in the association between the health outcome and environmental exposure. This phenomenon has been heavily studied for cross-sectional studies (Bobb et al., 2022; Dupont et al., 2022; Guan et al., 2023; Keller and Szpiro, 2020; Marques et al., 2022; Paciorek, 2010; Schnell and Papadogeorgou, 2020; Thaden and Kneib, 2018).

Many approaches take advantage of the ease of modeling mutli-dimensional space using thin-plate regression splines (TPRS) to either remove or model the confounded spatial relationships in the exposure and/or outcome. To incorporate TPRS into a regression model, researchers must select the number of splines to include in the model and the amount of penalization, and this has led to different strategies. Dupont et al. (2022)'s Spatial+ methodology, regresses the exposure on the TPRS to remove the spatial variation in the exposure, which in turn removes the spatial confounding from the outcome model. Keller and Szpiro (2020)'s, denoted here as KS, methodology, chose the amount of adjustment through an outcome model with everything except the exposure and using an information criterion to select the number of splines to include in the basis. Thus, the spatial variation present in the outcome that is not accounted for by the additional covariates is accounted for by the TPRS. Both of these methods, along with Bobb et al. (2022)'s exposure-penalized spline

method and Thaden and Kneib (2018)'s geoadditive structural equation model were compared in Chapter 2.

However, many environmental epidemiology studies use longitudinal or time-to-event data in which the spatial relationships may change over time (the CHILD study investigators et al., 2015; Di et al., 2017; Zhang et al., 2023). For longitudinal data where each individual has several observations over a given period, often a mixed model or a generalized estimating equation (GEE) model is used. Time-to-event data (e.g. onset of dementia or death) is commonly studied using a model for the hazard ratio. Utility of existing spatial confounding methods for these analyses is just beginning to be explored.

While spatial unmeasured confounding has yet to be investigated in longitudinal analyses, non-spatial unmeasured confounding has been studied extensively in this context (Palta and Yao, 1991; Gunasekara et al., 2014; Lin and Henley, 2016; Streeter et al., 2017; Keogh et al., 2018; Lee and Ma, 2024). Streeter et al. (2017)'s literature review of unmeasured confounding in longitudinal studies returned 121 studies, of which, most implemented either an instrumental variable analyses or a difference-in-difference analyses to mitigate the bias due to unmeasured confounding. However, in environmental epidemiology contexts, determining an instrument for an instrumental variable analysis or distinguishing between a 'treatment' and 'control' group for a difference-in-difference analysis is challenging due to the spatial dependence of environmental exposures. Time-dependent confounding has also been investigated in the GEE framework by Keogh et al. (2018) who introduced a sequential conditional means model that estimates the short-term temporal effects that the exposure has on the outcome by including the previous exposure measurement in the model. But, with environmental epidemiology studies' long-term exposure focus, extending these time-dependent methods would not be applicable.

Mitigating unmeasured confounding effects in survival studies has also primarily been investigated under causal frameworks (Klungsøyr et al., 2009; Bosco et al., 2010; MacKenzie et al., 2014); however, the exploration of the effect of unmeasured spatial confounding is starting for these studies (Banerjee, 2003; Xue et al., 2020; Azevedo et al., 2023). Similar to the longitu-

dinal contexts, instrumental variables methods have been proposed by MacKenzie et al. (2014) to mitigate the biases due to general additive confounding in proportional hazards (PH) models and showed a reduction in the bias; however they raised concerns about the generalization of their method. Klungsøyr et al. (2009) introduced using an extension of marginal structure for PH regression as a sensitivity analysis for causal inference to determine the impact of unmeasured confounding. Beginning the investigation on mitigating spatial confounding in survival frameworks, Azevedo et al. (2023) has extended the restricted spatial regression framework, originally introduced by Hodges and Reich (2010), to frailty models and found a reduction in the variation in the estimates using the restricted spatial frailty model compared to the spatial frailty model in spatially confounded scenarios. Similar to the original restricted spatial regression cross-sectional method, Azevedo et al. (2023)'s restricted spatial frailty model estimates the unconditional estimate, but conditioning upon the unmeasured spatial factor is of primary concern to researchers in environmental epidemiology studies.

In this chapter, we extend the cross-sectional methods that use TPRS to mitigate the bias induced by spatial confounding to longitudinal and survival frameworks. By adding a temporal component to the data, more than just a spatial TPRS basis must be considered and the effectiveness of the TPRS basis to completely model the spatiotemporal structure of the unmeasured confounder is uncertain. We investigate extensions of the Spatial+ and KS methodologies into both longitudinal and survival frameworks and provide an assessment of using TPRS to mitigate the bias in the estimated association between an environmental exposure and a health outcome induced by spatial confounding in spatiotemporal settings.

4.1.1 Wind Speeds and Preterm Birth in North Carolina from 1996 to 2017

We are motivated in part by an analysis investigating the association between preterm birth (birth occurring before 37 weeks of gestation) and exposure to a tropical cyclone during pregnancy in a North Carolina cohort between 1996 and 2017. Accounting for about 10% of births in the US, preterm birth is strongly linked to neonatal death and significantly raises the risk of both short-term

and long-term health complications for the child (Blencowe et al., 2012; Liu et al., 2015; Purisch and Gyamfi-Bannerman, 2017). In recent years, negative associations have been found between meteorological events and adverse health outcomes related to pregnancy, including preterm birth (Beltran et al., 2013; Sun et al., 2020). In Section 4.4, we analyze a random sample of the North Carolina cohort and evaluate the hazard ratio of preterm birth associated with tropical cyclone force winds while accounting for spatial confounding.

4.1.2 Outline of Chapter

The structure of the rest of this chapter is as follows. Section 4.2 defines the assumed longitudinal and survival structures of the data and introduces our proposed method, an outcome-model driven adjustment methodology, based on the cross-sectional methodology introduced by Keller and Szpiro (2020). In Section 4.3, we evaluate its performance and compare to alternative approaches in a simulation study. Section 4.4 applies the methods to the motivating tropical cyclone study and we conclude with Section 4.5 providing an overall discussion.

4.2 Methods for Mitigating Spatial Confounding in a Spatiotemporal Setting

4.2.1 Spatiotemporal Framework

Consider a sample of n individuals, each having a continuous outcome measured at multiple time points. For example, an outcome may be the resulting score of a memory test, weight of a baby through the first months of their life, or an individual's blood pressure (Zhang et al., 2023; the CHILD study investigators et al., 2015; AlGhatrif et al., 2013). Each individual, i, has a corresponding residential location, $\boldsymbol{s}_i \in \mathbb{R}^2$, and is measured at times $t_{ij} \in \mathbb{R}, j = 1, \ldots, J_i$ for some J_i. We denote the outcome of interest as $y_{ij}(\boldsymbol{s}_i, t_{ij})$ and model the mean of the outcome, $\mu_{ij}(\boldsymbol{s}_i, t_{ij})$, as

$$\mu_{ij}(\boldsymbol{s}_i, t_{ij}) = \beta_x x_{ij}(\boldsymbol{s}_i, t_{ij}) + \beta_f f_{ij}(\boldsymbol{s}_i, t_{ij}) + \boldsymbol{w}_{ij}^{\mathsf{T}}\boldsymbol{\gamma} \tag{4.1}$$

where $x_{ij}(\boldsymbol{s}_i, t_{ij})$ is the measured covariate of interest, also referred to as the exposure; $\boldsymbol{w}_{ij} \in \mathbb{R}^p$ are additional measured covariates that are possibly time-dependent or dependent on location; $f_{ij}(\boldsymbol{s}_i, t_{ij})$ is an unmeasured confounder; and $\beta_x, \beta_f \in \mathbb{R}$ and $\boldsymbol{\gamma} \in \mathbb{R}^p$ are regression coefficients, with $\boldsymbol{\gamma}$ including the intercept. For simplicity, we denote $f_{ij}(\boldsymbol{s}_i, t_{ij})$ as a single unmeasured confounder; however, in practice, it can be a composition of multiple unmeasured confounders that vary over space and time. Following an approach similar to the spatial settings in Chapter 2, we assume spatial and temporal correlation between $x_{ij}(\boldsymbol{s}_i, t_{ij})$ and $f_{ij}(\boldsymbol{s}_i, t_{ij})$ is induced through sharing at least one source of spatiotemporal information.

Two common approaches for estimating β_x in this longitudinal context are a mixed model and a GEE model. The mixed model is

$$y_{ij}(\boldsymbol{s}_i, t_{ij}) = \mu_{ij}(\boldsymbol{s}_i, t_{ij}) + a_i + \varepsilon_{y_{ij}} \tag{4.2}$$

where $a_i \sim N(0, \sigma_a^2)$ is a random effect for each individual and $\varepsilon_{y_{ij}} \sim N(0, \sigma^2)$ is additional non-spatial variation in the outcome. For the GEE model, we model the outcome as

$$y_{ij}(\boldsymbol{s}_i, t_{ij}) = \mu_{ij}(\boldsymbol{s}_i, t_{ij}) + \tilde{\varepsilon}_{y_{ij}} \tag{4.3}$$

where $\tilde{\varepsilon}_{y_{ij}} \sim N(0, \tilde{\sigma}^2 \boldsymbol{V})$ is non-spatial variation and $\boldsymbol{V}$ is assumed to have an exchangeable correlation structure such that all observations for an individual are equally correlated (Zeger et al., 1988).

For a second context, we consider a time-to-event outcome. Examples of time-to-event outcomes include the onset of dementia or the death of an individual. We define $y_i(\boldsymbol{s}_i)$ as the recorded event time and δ_i as the censoring indicator, such that $y_i(\boldsymbol{s}_i) = \min(E_i, C_i)$ and $\delta_i = I(E_i < C_i)$ where E_i is the actual event time and C_i is the censoring time. Thus, δ_i takes the values of 1 when an event is observed and 0 otherwise. We consider a Proportional Hazard (PH) regression model for these data and estimate the hazard ratio. In a PH model, $\mu_{ij}(\boldsymbol{s}_i, t_{ij})$ no longer represents the mean of the outcome but is still the linear predictor component. We denote the hazard function as

$h_{ij}(\boldsymbol{s}_i, e_{ij})$, where e_{ij} is any given time within the study, and model the time-varying hazard as

$$h_{ij}(\boldsymbol{s}_i, e_{ij}) = h_0(e_{ij})\text{exp}\{\mu_{ij}(\boldsymbol{s}_i, e_{ij})\}, \tag{4.4}$$

where $h_0(e_{ij})$ is the baseline hazard function.

In (4.2), (4.3), and (4.4), since $f_{ij}(\boldsymbol{s}_i, t_{ij})$ is unmeasured, a model without adjustment for space and/or time may lead to a biased estimate of β_x. To remove this bias, we investigate adding thin-plate regression splines (TPRS) to model the underlying confounded spatiotemporal structure in either the exposure or the outcome. For notation simplicity, we continue with suppressing the indexes i and j.

4.2.2 TPRS

Over the past decade, the use of thin-plate regressions splines (TPRS) to mitigate spatial confounding bias in cross-sectional semiparametric models has been studied extensively (Thaden and Kneib, 2018; Dupont et al., 2022; Keller and Szpiro, 2020; Bobb et al., 2022). With data varying across space and time, there are three potential domains on which to define the TPRS: the set of unique spatial locations, $\{\tilde{\boldsymbol{s}}\}$; the set of unique time points, $\{\tilde{\boldsymbol{t}}\}$; and the combination of unique spatial locations and time points $\{(\tilde{\boldsymbol{s}}, \tilde{\boldsymbol{t}})\}$. For each domain, we compute a set of L TPRS whose linear combination will represent the spatiotemporal confounded surface present in the exposure or outcome. The L vectors provide a computationally efficient basis for modeling multidimensional data (Wood, 2003). We denote a TPRS vector as $\boldsymbol{q}_{s,\ell}$, $\boldsymbol{q}_{t,\ell}$, or $\boldsymbol{q}_{st,\ell}$, indicating splines on the spatial, temporal, or spatiotemporal domain, respectively; and $\ell = 1, \ldots, L$ indexes the spline within the basis. We will represent the set of L vectors, as matrices such that, for the spatial matrix for example, $\boldsymbol{Q}_s = [\boldsymbol{q}_{s,1} \cdots \boldsymbol{q}_{s,L}] \in \mathbb{R}^{n_s \times L}$, and n_s represent the number of observations used to create the spatial basis. The matrices representing the temporal and spatiotemporal bases are constructed similarly and notated $\boldsymbol{Q}_t$ and $\boldsymbol{Q}_{st}$, respectfully. When evaluating the methods below, we also considered a linear combination of $\boldsymbol{Q}_s$ and $\boldsymbol{Q}_t$ for some contexts. Commonly, the number of splines, called the degrees of freedom (df), included in a regression model is less than L.

4.2.3 Adjustment via Outcome Model Selection

We propose an approach to spatiotemporal confounding adjustment based on addition of TPRS to the outcome model. This is an extension of the methodology introduced by Keller and Szpiro (2020). Our proposed method reduces the spatial confounding bias by using unpenalized TPRS to directly model the confounded spatial structure in the outcome model. We propose using a two-step approach that first fits the outcome to everything except the exposure and selecting the df by minimizing an information criterion. The second step fit the full outcome model, including the exposure and the TPRS basis with the selected df.

When modeling longitudinal data, we propose using $\boldsymbol{Q}_{st}$ as the TPRS basis to account for the interaction between space and time in the underlying confounded surface. Especially in scenarios where the exposure varies over time, not accounting for the spatio-temporal interaction in the underlying confounded surface will still leave the model underfit and not mitigate all spatial confounding bias. Using a mixed model, the procedure is

1. Fit $y = \delta_0 + a_1 + \sum_{\ell=1}^{k_1} \delta_\ell q_\ell + \boldsymbol{w}^\mathsf{T}\boldsymbol{\gamma} + \varepsilon_{1y_m}$ for varying k_1. Select $\hat{k}_1$ by minimizing AIC or BIC.

2. Fit the main outcome model $y = \beta_0 + a_2 + \beta_x x + \boldsymbol{w}^\mathsf{T}\tilde{\boldsymbol{\gamma}} + \sum_{\ell=1}^{\hat{k}_1} \beta_\ell q_\ell + \varepsilon_{2y_m}$ and make inference on β_x.

Here $a_1 \sim N(0, (\sigma_{a_1})^2)$, $a_2 \sim N(0, (\sigma_{a_2})^2)$, $\varepsilon_{1y_m} \sim N(0, (\sigma_{1y_m})^2)$, and $\varepsilon_{2y_m} \sim N(0, (\sigma_{2y_m})^2)$.

When using a GEE model, we follow a similar procedure. Specifically, we

1. Fit $y = \delta_0 + \sum_{\ell=1}^{k_2} \delta_\ell q_\ell + \boldsymbol{w}^\mathsf{T}\boldsymbol{\gamma} + \varepsilon_{1y_e}$ for varying k_2. Select $\hat{k}_2$ by minimizing QIC (Pan, 2001).

2. Fit the model $y = \beta_0 + \beta_x x + \boldsymbol{w}^\mathsf{T}\tilde{\boldsymbol{\gamma}} + \sum_{\ell=1}^{\hat{k}_2} \beta_\ell q_\ell + \varepsilon_{2y_e}$ and make inference on β_x.

Here, $\boldsymbol{\varepsilon}_{1y_e} \sim N(0, (\Sigma_{1y_e}))$ and $\boldsymbol{\varepsilon}_{2y_e} \sim N(0, (\Sigma_{2y_e}))$, and Σ_{1y_e} and Σ_{2y_e} are block diagonal matrices, assuming an exchangable correlation structure.

For survival data, we use a PH model in both steps and similar to the longitudinal analyses, propose using $\boldsymbol{Q}_{st}$ as the TPRS basis. Thus, we

1. Fit $h(\cdot) = h_{01}\text{exp}\{\boldsymbol{w}^\mathsf{T}\boldsymbol{\gamma} + \sum_{\ell=1}^{k_3} \delta_\ell q_\ell\}$ for varying k_3. Select $\hat{k}_3$ by minimizing AIC or BIC.

2. Fit $h(\cdot) = h_{02}\text{exp}\{\beta_x x + \boldsymbol{w}^\mathsf{T}\boldsymbol{\gamma} + \sum_{\ell=1}^{\hat{k}_3} \beta_\ell q_\ell\}$ and make inference on β_x.

where h_{01} and h_{02} are baseline hazard functions.

By using an outcome-model driven adjustment approach, we are mitigating the spatiotemporal confounding by modeling the association between the outcome and the unmeasured confounder through the addition of TPRS with a selected df. This modeling reduces the chance of oversmoothing the exposure, which could remove too much signal. Other cross-sectional methods reduce the spatial confounding by removing or modeling the confounded spatial structure in the exposure (Dupont et al., 2022; Thaden and Kneib, 2018; Bobb et al., 2022). In these methods, there exist the chance that by using TPRS to remove the confounded spatial variation in the exposure, all signal from the exposure will be removed, leaving nothing to identify an association between the exposure and outcome. The chance of removing all signal from the exposure increases when the exposure values are computed using predictions from an environmental model rather than using raw measurements that have non-spatial variation.

4.3 Simulation Study

We conducted a set of simulations to compare the the proposed approach for both mixed and GEE models for longitudinal data and PH models for survival analyses.

4.3.1 Set-up

For each simulation setting, we replicated the simulation 200 times and reported the average bias and root mean squared error (RMSE) of the estimates, the coverage rate of nominal 95% confidence intervals, and the median TPRS df used in the models. The RMSE is calculated as $\sqrt{\sum_{m=1}^{200}(\hat{\beta}_m - \beta)^2/200}$. For the PH regression models, we also report the average relative bias.

Simulating the spatiotemporal exposure and confounder

We sampled $n = 500$ (2000) locations for the longitudinal (survival) study simulations, across a $[0, 10] \times [0, 10]$ grid with incremental step size of 0.2 (0.1). Following an approach similar to the spatial settings in Chapter 2, to construct (4.1) we assume $x(\boldsymbol{s}, t)$ is a function of two centered and scaled spatiotemporal fields, $\boldsymbol{Z}_1(\boldsymbol{s}, t)$ and $\boldsymbol{Z}_2(\boldsymbol{s}, t)$, with no non-spatial variation. We also assume $f(\boldsymbol{s}, t)$ is also function of two centered and scaled spatiotemporal fields; $\boldsymbol{Z}_1(\boldsymbol{s}, t)$, shared with $x(\boldsymbol{s}, t)$, and $\boldsymbol{Z}_3(\boldsymbol{s}, t)$; such that

$$x(\boldsymbol{s}, t) = \boldsymbol{Z}_1(\boldsymbol{s}, t) + \boldsymbol{Z}_2(\boldsymbol{s}, t) \tag{4.5}$$

$$f(\boldsymbol{s}, t) = \boldsymbol{Z}_1(\boldsymbol{s}, t) + \boldsymbol{Z}_3(\boldsymbol{s}, t). \tag{4.6}$$

To generate $\boldsymbol{Z}_1(\boldsymbol{s}, t)$, $\boldsymbol{Z}_2(\boldsymbol{s}, t)$, and $\boldsymbol{Z}_3(\boldsymbol{s}, t)$, we follow an approach similar to Lindström et al. (2014) and Keller et al. (2015). We create 6 spatial fields, $\boldsymbol{z}_1(\boldsymbol{s}), \ldots, \boldsymbol{z}_6(\boldsymbol{s})$ and four temporal trends, $\boldsymbol{b}_1(t), \ldots, \boldsymbol{b}_4(t)$. Each spatial field is a realization of a zero-mean Gaussian Process with Matèrn covariance structure with range parameter $\phi = 10$, variance parameter $\sigma_z^2 = 1$, and varying smoothness parameter $\nu_1 = 2$, $\nu_2 = \nu_6 = 1.5$, $\nu_3 = 0.5$, $\nu_4 = \nu_5 = 1$. The temporal trends are defined as follows: $b_1(t)$ is a linear trend with a negative slope and $b_2(t)$, $b_3(t)$, and $b_4(t)$ are sinusoidal trends with parameters described in Appendix Table C.1. We combine the spatial fields and temporal trends to create

$$\boldsymbol{Z}_1(\boldsymbol{s}, t) = \boldsymbol{z}_1(\boldsymbol{s}) + \boldsymbol{z}_2(\boldsymbol{s}) \times b_1(t) \tag{4.7}$$

$$\boldsymbol{Z}_2(\boldsymbol{s}, t) = \boldsymbol{z}_3(\boldsymbol{s}) + \boldsymbol{z}_4(\boldsymbol{s}) \times b_2(t) + \boldsymbol{z}_4(\boldsymbol{s}) \times b_3(t) \tag{4.8}$$

$$\boldsymbol{Z}_3(\boldsymbol{s}, t) = \boldsymbol{z}_6(\boldsymbol{s}) \times b_4(t). \tag{4.9}$$

For each modeling framework, we consider the comparison of two scenarios: (1) the exposure and confounder are constant over the entire study but still vary spatially (i.e. $\mu_i(\boldsymbol{s}_i, \boldsymbol{t}) = \mu_{i1}(\boldsymbol{s}_i, \boldsymbol{t}_{i1}) = \cdots = \mu_{iJ_i}(\boldsymbol{s}_i, \boldsymbol{t}_{iJ_i})$ defined from (4.1)) and (2) the exposure and confounder vary

both spatially and temporally over the study (i.e. $\mu_{i1}(\boldsymbol{s}_i, \boldsymbol{t}_{i1}) \neq \cdots \neq \mu_{iJ_i}(\boldsymbol{s}_i, \boldsymbol{t}_{iJ_i})$ defined from (4.1)).

Simulating the longitudinal outcome

For the longitudinal studies, we sample eight time points, one in each of the intervals: [0,1.2], [1.4, 2.4], [2.6, 3.6], [4.8, 5], [5.2, 6.2], [6.4, 7.4], [7.6, 8.6], [8.8, 10] to mimic an individual with baseline measurements and seven additional follow-up observations across a 10-year span. We considered two scenarios based on the amount of correlation within an individual. We set the non-spatial error variance to $\sigma^2 = 1$ and, for mixed model outcomes, the random effect variance to σ_a^2 = 0.25 and 2.25. To create the outcome in the GEE model studies, we set $\sigma^2 = 1$ and the within cluster correlation to be 0.2 and 0.7 for scenarios low and high within person correlation.

Simulating the time-to-event outcome

In the survival framework, we partitioned the interval 0 to 10, representing years 2000 through 2010, into quarterly intervals. Additionally, we simulated two time-constant measured covariates: a three-level categorical predictor, $w_1 \in$ (1, 2, 3) with probabilities of 0.45, 0.45, and 0.1; and baseline age from a uniform distribution with bound of 55 and 70. We set the simulated study to run for a 10-year period.

After the spatiotemporal fields are created, we sampled a start date within the first three months for each individual and then computed the event times using quarterly intervals. For each individual, i, once they enter the study, for each quarterly interval, j, we compute an event time and a censor time, e_{ij} and c_{ij}, respectively, using the exposure value for that interval. To compute the times, we follow ther procedure of Austin (2012). We first sample two values, $u_e \sim \text{Unif}(0, 1)$ and $u_c \sim \text{Unif}(0, 1)$. Then we compute

$$e_{ij} = \frac{-\log(u_e)}{\lambda \exp\{\beta_x x + \beta_f f + \beta_{w2}\mathbf{I}(w_1 = 2) + \beta_{w3}(w_1 = 3) + \beta_{age} age\}} \tag{4.10}$$

$$c_{ij} = \frac{-\log(u_c)}{\lambda * \lambda_2} \tag{4.11}$$

where age is the baseline age of the individual, centered; λ is the baseline hazard ratio; λ_2 is a constant to control the amount of censoring; and β_x, β_f, β_{w2}, β_{w3}, and $\beta_{age} \in \mathbb{R}$ are the corresponding true coefficients. We set $\lambda = 0.04$, $\lambda_2 = 0.5$, $\beta_{w2} = -0.1$, $\beta_{w3} = -0.15$, $\beta_{age} = 0.12$, $\beta_x = 0.15$, and $\beta_f = 0.15$.

If the minimum of e_{ij} and c_{ij} is greater than the interval length of 1/4 (due to obtaining new exposure and confounder values quarterly), then the individual does not experience the event in the interval and a new e_{ij} and c_{ij} is computed for the next quarter. If the minimum of e_{ij} and c_{ij} is less than 1/4, the individual is removed from the study. Once the individual is removed, we compute the event time, $y_i = \lceil 365.25\{1/4(j-1) + \min(e_{ij}, c_{ij})\}\rceil$, and the censoring indicator, $\delta_i = I(e_j < c_j)$. However, if an individual has not experienced the event by the end of the study, the event time is set to $y_i = 10$ and $\delta_i = 0$. Thus, $\delta_i = 0$ unless it is the last included interval for an individual and they had a recorded event, in which case $\delta_i = 1$.

4.3.2 Additional Methods

In each simulation study, we compare our proposed method to four other approaches (Table 4.1). We investigate a model with no spatial or temporal adjustments as a baseline reference; a model where the number of df included in the TPRS basis is pre-selected to be either 5, 10, or 50; a one-step selection model using an information criterion to select the df of the basis; and an extension of Spatial+. In the one-step selection model, the information criterion selects from a full outcome model. The same model is used for estimating β_x.

To extend the Spatial+ method (Dupont et al., 2022) for longitudinal mixed models, the first step regresses the exposure on the TPRS basis and a random intercept. For either a fixed df or GCV penalization, 100 splines are included in the basis and the residuals are computed on all exposures. For information criterion selection, introduced in Chapter 2, the exposure model is fit with df ranging from 3 to 50, incrementing by 1, and 55 to 100, incrementing by 5, and the model that minimizes the df is used to compute the residuals. The second step models the outcome as a function of the residuals, TPRS basis with either 100 splines, fixed or penalized, or the selected

number of splines, and additional covariates. The longitudinal GEE models extend the Spatial+ method in a similar manner except instead of including a random intercept in both steps, an exchangable correlation structure, grouped by the individuals, is assumed. Extending the Spatial+ method to survival data using a PH model uses the same first step as the longitudinal mixed models as the exposure is still continuous. In the second step, the outcome model is a PH model with the residuals of the exposure, TPRS basis, and additional covariates composing the linear component in the model. In the scenario where a spatial basis, $\boldsymbol{Q}_s$, is used as the TPRS basis, we consider three subsets of the exposure for the step one model: all observations, baseline observations for each individual, or averaging the exposure for each individual.

Table 4.1: Representation of the methods used in comparison in the simulation study. The first step in the Spatial+ method, however, regresses the exposure on the TPRS basis to obtain the residuals used in the second step.

Method	Step #1		Step #2	
	Model	Selector	Model	Selector
Unadjusted	$\mu = x + \boldsymbol{w}$	–	–	
Preselected	$\mu = x + q_\ell + \boldsymbol{w}$	Fixed	–	
One-Step Selection	$\mu = x + q_\ell + \boldsymbol{w}$	AIC/BIC/QIC	–	
Proposed	$\mu = q_\ell + \boldsymbol{w}$	AIC/BIC/QIC	$\mu = x + q_\ell + \boldsymbol{w}$	$\hat{\ell}$
Spatial+	$x = q_\ell$	Fixed/GCV/AIC/BIC/QIC	$\mu = \hat{r}_x + q_\ell + \boldsymbol{w}$	Fixed/GCV/$\hat{\ell}$

4.3.3 Evaluating over various TPRS bases

We evaluate the effectiveness of the various TPRS bases, $\boldsymbol{Q}_s$, $\boldsymbol{Q}_{st}$, and the linear combination of $\boldsymbol{Q}_s$ and $\boldsymbol{Q}_t$, for mitigating the spatial confounding bias in each of the approaches in the simulation study. When evaluating the methods for longitudinal data, we fit each approach three times, one for each of the bases. We do not consider adding just the temporal basis, $\boldsymbol{Q}_t$, since we are primarily concerned with spatial confounding, and therefore, always want to be adjusting for space with the TPRS basis. If we are implementing a survival model, we only consider adding the spatial and spatiotemporal basis; since any function of time is absorbed by the baseline hazard, resulting in the same estimates for adding $\boldsymbol{Q}_s$ and the linear combination of $\boldsymbol{Q}_s$ and $\boldsymbol{Q}_t$.

4.3.4 Results

The results of the simulation study with low within person correlations ($\rho = 0.2$) are reported in Tables 4.2, 4.3, 4.4, 4.5, 4.6, and 4.7 with additional results for the time-constant exposure and time-varying exposure with high within-person correlation ($\sigma_a^2 = 2.25$ or $\rho = 0.7$) settings reported in the Appendix C.

Starting with the longitudinal mixed models (Table 4.2 and Table 4.3), the unadjusted model had the highest bias and RMSE (0.673 and 0.683, respectively), except for using the preselection model with 5 df using the spatiotemporal basis, which had a higher bias and RMSE (0.712 and 0.731, respectively) and the Spatial+ model with the linear combination of the spatial and temporal basis using BIC selection which had a higher RMSE (0.690). Among the preselection models, selecting 50 df using the spatiotemporal basis yielded the lowest bias and RMSE (0.017 and 0.086, respectively). Using the spatiotemporal basis yielded the overall best reduction of bias and RMSE in our proposed approach using AIC selection and the Spatial+ model using GCV penalization (0.006 and 0.057, bias and RMSE, respectively, for both models). Both of those models had a large df selected.

Looking at the longitudinal GEE models (Table 4.4 and Table 4.5), we see similar trends as the longitudinal mixed models. The unadjusted model now has the overall highest bias and RMSE. Again, using the spatiotemporal basis in the proposed and Spatial+ models produced the lowest bias and RMSE. Our proposed method had the best RMSE while Spatial+ selecting the df from QIC produced the lowest bias.

However, since Spatial+ uses the residuals of the exposure, there exists a risk that the first stage removes all signal existing in the exposure, greatly increasing the variability of the outcome. We see this phenomenon happening in the time-constant exposure with the high correlation within individual (Appendix Table C.2) Of the settings considered, the method with the overall best performance is our proposed method with AIC using a spatiotemporal basis.

Finally, we consider the results of the survival analysis simulation studies (Table 4.6 and Table 4.7). Differing from the longitudinal setting, the unadjusted model no longer has the highest

RMSE. All but our proposed method using the spatiotemporal basis have higher RMSE than the unadjusted. For Spatial+ with $\boldsymbol{Q}_{st}$, we note that these methods have the overall smallest average bias. However, the RMSE of these methods, except for when using a GEE model in the first step, are large in comparison. Using a GEE model in the first step greatly reduces the RMSE; however, the RMSE is still larger than the unadjusted model. These results indication that, at least for the scenarios considered here, the best model is the proposed method with AIC, using a spatiotemporal basis.

4.4 Analysis of North Carolina Birth Cohort Data

We illustrate the use of the proposed method and Spatial+ method for a time-to-event analysis in an investigation of the risk of preterm birth due to the wind speeds of a tropical cyclone in North Carolina birth cohort. We used birth data from the North Carolina Department of Health and Human Services (NCDHHS) vital statistics, covering all live births to North Carolina residents between 1996 and 2017. We restricted the data to singleton births, with no missing data and for births at 20+ weeks gestation. The event of interest is whether a birth was a preterm birth (births occurring before 37 weeks gestation). For each birth, we created weekly intervals from 20 weeks through 36 weeks gestation or through the end of the pregnancy, which ever occurs first. For all births that occur before 37 weeks, the event indicator was set to 1 during the final week of the pregnancy.

The exposure data were the maximum winds speed as a result of a tropical cyclone (wind speeds $\geq$ 17.5 m/s) calculated at the county level. To quantify the severity of the tropical cyclone exposure, we obtained the maximum wind speed, recorded at the county level, from the `hurricaneexposuredata` package in `R` Anderson et al. (2020). We note that the pregnant person may not have experienced wind speeds greater than 17.5 m/s, but resided in a county that experienced such wind speeds during a tropical cyclone. Therefore, exposures may be 0 (no tropical cyclone exposure), less that 17.5 m/s (experience wind speeds due to a tropical cyclone, but not tropical cyclone force winds), or greater than 17.5 m/s. Thus, we further restricted the birth cohort

Table 4.2: Root mean squared error (RMSE), estimated bias, coverage rates for 95% confidence intervals for the estimates of β_x in the mixed model simulations with a time-varying exposure where the within person correlation is low ($\sigma_a = 0.5$). The median degrees of freedom (df) for each method is reported for the respective basis used in each model.

Method	df Selector	Basis	Subset	Bias	RMSE	Coverage (95% Nominal)	df Selected (Spatial)	(Temporal)
Unadjusted	–	–	–	0.6733	0.6831	0.000	–	–
Preselect	Fixed	S	–	0.3089	0.4394	0.135	5	–
Preselect	Fixed	S	–	0.3423	0.4869	0.105	10	–
Preselect	Fixed	S	–	0.4099	0.5862	0.080	50	–
Preselect	Fixed	ST	–	0.7115	0.7306	0.005		5
Preselect	Fixed	ST	–	0.0843	0.2692	0.230		10
Preselect	Fixed	ST	–	0.0174	0.0855	0.770		50
Preselect	Fixed	S&T	–	0.3812	0.4705	0.140	5	5
Preselect	Fixed	S&T	–	0.4185	0.5070	0.125	10	10
Preselect	Fixed	S&T	–	0.4926	0.5806	0.115	50	10
One Step	AIC	S	–	0.3903	0.5588	0.115	18	–
One Step	BIC	S	–	0.3313	0.4739	0.125	4	–
One Step	AIC	ST	–	0.0091	0.0570	0.930		95
One Step	BIC	ST	–	0.0228	0.0899	0.765		48
One Step	AIC	S&T	–	0.4612	0.5605	0.130	19.5	7
One Step	BIC	S&T	–	0.4035	0.5012	0.120	4	5
Proposed	AIC	S	–	0.4047	0.5896	0.085	52.5	–
Proposed	BIC	S	–	0.3533	0.5058	0.110	11	–
Proposed	AIC	ST	–	0.0059	0.0571	0.925		100
Proposed	BIC	ST	–	-0.0070	0.0665	0.850		75
Proposed	AIC	S&T	–	0.4883	0.5791	0.095	55	7
Proposed	BIC	S&T	–	0.4262	0.5180	0.135	12	5

Table 4.3: Root mean squared error (RMSE), estimated bias, coverage rates for 95% confidence intervals for the estimates of β_x in the mixed model simulations with a time-varying exposure where the within person correlation is low ($\sigma_a = 0.5$). The median degrees of freedom (df) for each method is reported for the respective basis used in each model. (continued)

Method	df Selector	Basis	Subset	Bias	RMSE	Coverage (95% Nominal)	df Selected (Spatial)	df Selected (Temporal)
Spatial+	Fixed	S	–	0.4481	0.6380	0.080	100	–
Spatial+	Fixed	S	Mean	0.4351	0.6259	0.085	100	–
Spatial+	Fixed	S	Base	0.4351	0.6259	0.085	100	–
Spatial+	GCV	S	–	0.4525	0.6302	0.075	48.7	–
Spatial+	AIC	S	–	0.4438	0.6309	0.075	90	–
Spatial+	AIC	S	Mean	0.4350	0.6251	0.085	100	–
Spatial+	AIC	S	Base	0.4327	0.6216	0.080	95	–
Spatial+	BIC	S	–	0.4636	0.6221	0.080	27	–
Spatial+	BIC	S	Mean	0.4083	0.5881	0.085	43	–
Spatial+	BIC	S	Base	0.3976	0.5680	0.095	32	–
Spatial+	Fixed	ST	–	0.0702	0.1431	0.825	100	
Spatial+	GCV	ST	–	0.0061	0.0574	0.925	90.7	
Spatial+	AIC	ST	–	0.0700	0.1430	0.825	100	
Spatial+	BIC	ST	–	0.0702	0.1499	0.780	80	
Spatial+	Fixed	S&T	–	0.5985	0.6785	0.055	100	10
Spatial+	GCV	S&T	–	0.5255	0.6119	0.090	51.3	9.2
Spatial+	AIC	S&T	–	0.5983	0.6781	0.055	100	9
Spatial+	BIC	S&T	–	0.6093	0.6895	0.055	32	4

Table 4.4: Root mean squared error (RMSE), estimated bias, coverage rates for 95% confidence intervals for the estimates of β_x in the GEE model simulations with a time-varying exposure where the within person correlation is low ($\rho = 0.2$). The median degrees of freedom (df) for each method is reported for the respective basis used in each model.

Method	df Selector	Basis	Subset	Bias	RMSE	Coverage (95% Nominal)	df Selected (Spatial)	(Temporal)
Unadjusted	–	–	–	0.6755	0.6887	0.000	–	–
Preselect	Fixed	S	–	0.3094	0.4441	0.095	5	–
Preselect	Fixed	S	–	0.3446	0.4888	0.095	10	–
Preselect	Fixed	S	–	0.4086	0.5795	0.080	50	–
Preselect	Fixed	ST	–	0.7067	0.7268	0.000		5
Preselect	Fixed	ST	–	0.1080	0.2908	0.205		10
Preselect	Fixed	ST	–	0.0248	0.0886	0.675		50
Preselect	Fixed	S&T	–	0.3939	0.4761	0.085	5	5
Preselect	Fixed	S&T	–	0.4339	0.5113	0.070	10	10
Preselect	Fixed	S&T	–	0.5109	0.5817	0.055	50	10
One Step	QIC	S	–	0.4304	0.6102	0.075	100	–
One Step	QIC	ST	–	0.0108	0.0527	0.930		95
One Step	QIC	S&T	–	0.5028	0.5860	0.075	70	8

Table 4.5: Root mean squared error (RMSE), estimated bias, coverage rates for 95% confidence intervals for the estimates of β_x in the GEE model simulations with a time-varying exposure where the within person correlation is low ($\rho = 0.2$). The median degrees of freedom (df) for each method is reported for the respective basis used in each model. (continued)

Method	df Selector	Basis	Subset	Bias	RMSE	Coverage (95% Nominal)	df Selected (Spatial)	df Selected (Temporal)
Proposed	QIC	S	–	0.4297	0.6103	0.070	100	–
Proposed	QIC	ST	–	0.0077	0.0523	0.935	100	
Proposed	QIC	S&T	–	0.5314	0.6028	0.075	100	9
Spatial+	Fixed	S	–	0.4309	0.6111	0.070	100	–
Spatial+	Fixed	S	Mean	0.4309	0.6111	0.070	100	–
Spatial+	Fixed	S	Base	0.4309	0.6111	0.070	100	–
Spatial+	QIC	S	–	0.4308	0.6112	0.070	100	–
Spatial+	QIC	S	Mean	0.4309	0.6111	0.070	100	–
Spatial+	QIC	S	Base	0.4289	0.6080	0.070	95	–
Spatial+	Fixed	ST	–	0.0098	0.0524	0.940	100	
Spatial+	QIC	ST	–	0.0054	0.0607	0.920	95	
Spatial+	Fixed	S&T	–	0.5373	0.6066	0.055	100	10
Spatial+	QIC	S&T	–	0.5155	0.5895	0.060	75	5

Table 4.6: Root mean squared error (RMSE), estimated bias, coverage rates for 95% confidence intervals for the estimates of β_x in the PH model simulations with a time-varying exposure. The median degrees of freedom (df) for each method is reported for the respective basis used in each model.

Method	df Selector	Basis	Subset	Bias	Relative Bias	RMSE	Coverage (95% Nominal)	df Selected (Spatial)
Unadjusted	–	–	–	0.0444	0.2963	0.0662	0.650	–
Preselect	Fixed	S	–	0.0418	0.2784	0.0730	0.835	5
Preselect	Fixed	S	–	0.0483	0.3219	0.0850	0.820	10
Preselect	Fixed	S	–	0.0634	0.4228	0.1128	0.810	50
Preselect	Fixed	ST	–	0.0418	0.2787	0.0722	0.830	5
Preselect	Fixed	ST	–	0.0140	0.0930	0.0607	0.940	10
Preselect	Fixed	ST	–	0.0043	0.0287	0.0885	0.965	50
One Step	AIC	S	–	0.0429	0.2862	0.0765	0.800	3
One Step	BIC	S	–	0.0401	0.2673	0.0684	0.840	3
One Step	AIC	ST	–	0.0298	0.1987	0.0708	0.865	3
One Step	BIC	ST	–	0.0397	0.2648	0.0671	0.820	3
Proposed	AIC	S	–	0.0332	0.2211	0.0707	0.865	5
Proposed	BIC	S	–	0.0362	0.2414	0.0660	0.855	3
Proposed	AIC	ST	–	0.0062	0.0416	0.0572	0.980	8
Proposed	BIC	ST	–	0.0253	0.1683	0.0578	0.910	3

Table 4.7: Root mean squared error (RMSE), estimated bias, coverage rates for 95% confidence intervals for the estimates of β_x in the PH model simulations with a time-varying exposure. The median degrees of freedom (df) for each method is reported for the respective basis used in each model. (continued

Method	df Selector	Basis	Subset	Bias	Relative Bias	RMSE	Coverage (95% Nominal)	df Selected (Spatial)
Spatial+	Fixed	S	Mixed	0.0944	0.6296	3.8954	0.020	100
Spatial+	Fixed	S	GEE	0.0727	0.4844	0.1256	0.760	100
Spatial+	Fixed	S	Mean	0.0727	0.4844	0.1256	0.760	100
Spatial+	Fixed	S	Base	0.0727	0.4844	0.1256	0.760	100
Spatial+	AIC	S	Mixed	0.0950	0.6336	3.8924	0.020	100
Spatial+	AIC	S	Mean	0.0727	0.4847	0.1259	0.760	100
Spatial+	AIC	S	Base	0.0726	0.4842	0.1256	0.760	100
Spatial+	BIC	S	Mixed	0.1029	0.6861	3.8916	0.015	34.5
Spatial+	BIC	S	Mean	0.0607	0.4049	0.1080	0.790	42
Spatial+	BIC	S	Base	0.0724	0.4824	0.1250	0.760	100
Spatial+	QIC	S	GEE	0.0728	0.4855	0.1247	0.765	100
Spatial+	Fixed	ST	Mixed	0.0009	0.0060	0.3891	0.955	100
Spatial+	Fixed	ST	GEE	0.0018	0.0122	0.0980	0.950	100
Spatial+	AIC	ST	Mixed	0.0009	0.0060	0.3891	0.955	100
Spatial+	BIC	ST	Mixed	0.0005	0.0033	0.3886	0.960	100
Spatial+	QIC	ST	GEE	0.0021	0.0137	0.0966	0.950	100

to births in counties that experienced tropical cyclone force winds at some point between 1996 to 2017. We considered two exposures of interest: a baseline maximum wind exposure during the first 20 weeks of the pregnancy (a time-constant exposure for each person) and a 4-week sliding maximum wind speed exposure (a time-varying exposure) for weeks 20 through 36.

Along with the exposure, we included an indicator variable for month of birth and for the year of birth as additional covariate. After restrictions, the cohort contains 1,511,388 births of which 339,321 are preterm. Finally, we subsampled 20% of the births for computational speed. Our sample included 302,227 births of which 27,137 of the births were preterm and 5,364 of the preterm births experienced tropical cyclone force wind speeds.

For the analysis, we compared four models: the unadjusted model, the preselection model, the proposed model, and the Spatial+ model. We used both $\boldsymbol{Q}_s$ and $\boldsymbol{Q}_{st}$ in the preselection model, choosing a df of 5 and 15. Following the results of the simulation study, we used $\boldsymbol{Q}_{st}$ for the proposed approach, using the AIC selection criterion; and we used both $\boldsymbol{Q}_s$ and $\boldsymbol{Q}_{st}$ for the Spatial+ approach. For $\boldsymbol{Q}_s$, we used the exposure at 20 weeks as the baseline exposure in the first step and AIC to select the df, and for the spatiotemporal basis, we used the GEE model in the first step and QIC to select the df since the spatiotemporal basis has multiple exposures per individual and the baseline exposure only has one exposure per individual. The models that used information criteria to select the number of splines to include in the model selected from 3 to 20 df incrementing by 1 and 25 to 50 df incrementing by 5. We reported the estimated hazard ratio (HR) for the wind speed along with corresponding 95% confidence interval (CI) and df used in the final model in Table 4.8.

4.4.1 Results

When using a time-constant exposure, 193,404 births (17,616 preterm births) were exposed to winds due to a tropical cyclone and the average [sd] maximum wind speed of births that experienced a tropical cyclone during the first 20 weeks was 11.357 [7.989]m/s (11.458 [7.943]m/s). For the models that used a time-varying exposure, 195,587 births (16,210 preterm births) were ex-

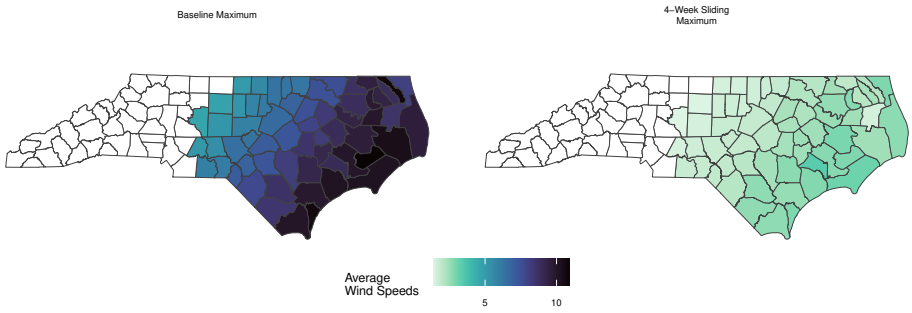


Figure 4.1: Shown here are the average maximum wind speeds experienced by a pregnant person when using a model that uses the maximum wind speed experienced during the first 20 weeks (Baseline Maximum) and a sliding window of maximum wind speeds of 4 weeks (4-Week Sliding Maximum).

posed to winds with the average [sd] maximum wind speed for births that experience a TC during a 4-week interval was 6.677 [7.068]m/s (6.644 [7.025]m/s) (Figure 4.1).

We saw that for the time-constant models, all models that included splines to adjust for space reported a hazard ratio (HR) of approximately twice as large as the HR computed from the model that did not include splines (Table 4.8). All time-constant models that added TPRS reported HR between 0.0014 and 0.0017 for a 1 m/s increase in wind speed experienced within the first 20 weeks of pregnancy compared to a HR of 0.0008 for the unadjusted model (Table 4.8). While all of the 95% confidence intervals contained 0, the doubling of the estimated hazard ratios suggests that the unadjusted model may be attenuated by an unmeasured spatial factor. However, across the models that do adjust for space, there exist variation in the df used but not in the estimated hazard ratios suggesting only slight adjustments were needed to account for the unmeasured spatial factor.

In contrast, the time-varying models all reported approximately the same HR and 95% confidence intervals. Also of note, all models with a time-varying exposure produced negative HR with 95% CIs that did not contain zero (Table 4.8). These results are counter-intuitive as they suggest that increasing the wind speeds experienced during pregnancy during a tropical cyclone

Table 4.8: Point estimates and 95% confidence intervals of the hazard ratio of preterm birth associated with a 1 m/s increase in maximum wind speed experienced during the first 20 weeks of a pregnancy (time-constant) or during a 4 week sliding window (time-varying) during a tropical cyclone event in the North Carolina birth cohort study, and selected degrees of freedom (df) for the thin-plate regression spline basis

Exposure Type	Basis	df Selector	Estimate	95% Confidence Interval	df Selected
Time-Constant	Unadjusted	–	0.0008	(-0.0011, 0.0028)	–
	Preselect	S	0.0015	(-0.0005, 0.0035)	5
	Preselect	S	0.0014	(-0.0006, 0.0034)	15
	Preselect	ST	0.0017	(-0.0003, 0.0037)	5
	Preselect	ST	0.0017	(-0.0004, 0.0037)	15
	Proposed	ST	0.0014	(-0.0006, 0.0035)	50
	Spatial+ (Base)	S	0.0014	(-0.0006, 0.0034)	18
Time-Varying	Unadjusted	–	-0.0103	(-0.0131, -0.0075)	–
	Preselect	S	-0.0102	(-0.0129, -0.0074)	5
	Preselect	S	-0.0102	(-0.0130, -0.0074)	15
	Preselect	ST	-0.0101	(-0.0129, -0.0073)	5
	Preselect	ST	-0.0101	(-0.0128, -0.0073)	15
	Proposed	ST	-0.0102	(-0.0130, -0.0074)	50
	Spatial+ (Base)	S	-0.0102	(-0.0130, -0.0075)	7
	Spatial+ (GEE)	ST	-0.0102	(-0.0130, -0.0074)	50

event decreases the risk of preterm birth (Beltran et al. (2013), Sun et al. (2020)). Possible explanations of these results include delayed care towards the end of a pregnancy, such as pushing back a scheduled preterm induction to then be a full term birth, but further investigation is warranted.

4.5 Discussion

In this chapter, we have introduced an extension of two cross-sectional semiparametric methods for longitudinal and time-to-event data containing either a time-constant or time-varying exposure to mitigate the bias due to spatial confounding using TPRS. Our extensions, along with a preselection and one-step method, allow both the proposed method and Spatial+ method to be implemented for mixed models, GEE models, and PH models. We showed that in the the presence of spatial confounding, adjustment for the confounded spatial variation is necessary to produce estimates with less bias for longitudinal and survival data.

In our simulation studies, we identified that using the proposed method with a spatiotemporal TPRS basis and the AIC criterion for mixed and survival model and QIC for GEE models pro-

duced estimates with a reduction in bias while also reducing the variability of the selected point estimates. Using just a spatial TPRS basis tended to increase the bias substantially compared to using a spatiotemporal basis. The Spatial+ extensions produced estimates with a lower bias in some scenarios but inflated the variation of the estimates.

In all of the studies, we did not include non-spatial variation when simulating the exposure, since we are motivated by environmental epidemiology studies where exposure values are prediction outputs from a model. Investigation of the extension of these methods when non-spatial variation is present in the exposure would benefit other contexts. We also only investigated the scenarios where the spatial variation in the exposure is greater than the spatial variation in the confounder and it is unclear if the methods would still mitigate bias when the spatial variation in the confounder is greater than in the exposure – a scenario known to challenge all approaches in cross-sectional studies. There also exist cross-sectional methods that mitigate bias due to spatial confounding that do not use TPRS (e.g. spectral filtering or Gaussian Process) and extension of these methods for longitudinal and time-to-event data are needed to continue to investigate the best methods for mitigating the bias for these data.

In summary, we investigated using TPRS to mitigate the bias due to spatial confounding for both longitudinal and time-to-event data by extending two existing cross-sectional methods for exposure that are either constant or vary over time. Using these methods allows for large-scale environmental epidemiologic studies that collect data over a period of time to account for spatial confounding in their analyses.

Chapter 5

Conclusions

In this dissertation, we compared and developed methods using thin-plate regression splines to mitigate the bias due to spatial confounding for large cohort environmental epidemiological studies and developed tools to assist researchers to apply these methods. In Chapter 2, we compared current methodologies and introduced a hybrid approach to mitigate the bias induced by the presence of spatial confounding in cross-sectional studies. In Chapter 3, we developed an R Shiny application, `spconfShiny`, based off of the R package `spconf`, to create a point-and-click dashboard that computes an effective bandwidth statistic to quantify the amount of spatial smoothing induced by adding a given number of splines into a regression model. In Chapter 4, we introduced new methods, which extended ideas from Chapter 2, for longitudinal and survival modeling frameworks, which include temporal variation.

5.1 Cross-Sectional vs. Spatiotemporal Models

For large cohort analyses, we investigated an outcome-model driven adjustment, originally introduced for cross-sectional data by Keller and Szpiro (2020) and compared in Chapter 2. We proposed an extension of this methodology for longitudinal and time-to-event data in Chapter 4 to reduce bias due to spatial confound while also controlling the variance of the point estimates. In both Chapter 2 and Chapter 4, we compared the outcome-model driven adjustment to an exposure-model driven adjustment. Originally introduced by Dupont et al. (2022) for cross-sectional data, we extended the Spatial+ approach, an exposure-model driven adjustment approach, and introduced new methodology for longitudinal and time-to-event data. More prominently seen in Chapter 4, our proposed outcome-model driven adjustment balances the bias-variance trade-off compared to the exposure-model driven adjustment, introducing a little bias for a great reduction of the variation. In the exposure-model driven adjustment, the removal of the spatial variation from the exposure risks removing all signal, especially when the exposure is computed from an exposure

model based on predictions at the residential locations of the individuals in the study. This can lead to either a null association between the exposure and outcome, or a highly variable estimate. However, in our proposed outcome-model driven adjustment methodology, the spatial variation in the outcome, not accounted for by any additional covariates in the model, is modeled and therefore still accounted for, not risking nullifying the association between the exposure and outcome and producing less variables estimates.

In Chapter 2, multiple simulations studies were completed, comparing both low and high frequency spatial confounding, and the smallest bias and RMSE were from approaches with the number of TPRS selected by AIC. In some simulation settings, either KS BIC produced a slightly smaller bias or our proposed Spatial+ BIC producing equivalent results. However, in Chapter 4, the results were less clear with our proposed outcome-model driven adjustment using AIC having the smallest RMSE, but not the smallest bias. The smaller biases were produced by various exposure-model driven adjustment methods we introduced; however, the RMSE were large in comparison to the outcome-model driven adjustment methods. Thus, by weighing the bias-variance trade-off, our proposed outcome-model driven adjustment methodology using AIC was recommended. Also, in Chapter 4, we compared a time-constant exposure to a time-varying exposure, but did not investigate the difference between low and high frequency spatial confounding. This difference would be of interest for researchers looking to mitigate spatial confounding in spatiotemporal data.

5.2 Connection to Other Approaches

In the extension to spatiotemporal data, we solely focused on the extension of methods that use thin-plate regression splines to reduce the bias due to spatial confounding through either the modeling of the confounded spatiotemporal surface or the removal of the confounded spatiotemporal variation. However, as discussed in Section 2.1, there exist other cross-sectional methods using various statistical techniques to mitigate spatial confounding bias. These techniques could include using the spectral domain or using Gaussian Processes, to directly model spatiotemporal variation (Guan et al., 2023; Wiecha et al., 2024; Schnell and Papadogeorgou, 2020; Marques et al.,

2022). Since Gaussian Processes are implemented using a Bayesian procedure, the additional temporal component requires selecting more priors and estimating more parameters, introducing more choices for the researcher to make before running an analysis and increasing the overall computation time. However, it would be of interest to see out the extension introduced in Chapter 4 compare to extensions of methods beyond the implementation of TPRS.

All of the work discussed in this dissertation revolves around computing an association between an environmental exposure and a health outcome of interest while removing bias due to spatial confounding. A majority of work mitigating confounding bias has been done in a causal inference framework, as discussed in both Section 2.1 and Section 4.1. However, environmental exposures are recorded as continuous measures where the exposure experienced by one individual is related, via distance, to the exposure experienced by another individual in the study. This poses two problems for causal inference methodologies. The first being that, for the majority of causal procedures, the treatment, or exposure, is binary, making matching procedures such as computing the propensity score much more challenging. Papadogeorgou and Dominici (2020) introduced a causal procedure using an continuous exposure as the treatment, however restricts the individuals in the study to be within a certain radius of the exposure sources. This restriction greatly reduces the number of individuals that would be able to be considered in large cohort environmental epidemiology studies and would therefore not be applicable. The second problem that using environmental exposures as the treatment for a causal inference methodology poses is that it breaks the assumption of no interference, or that the outcome of an individual is unaffected by the treatments assigned to other individuals. Papadogeorgou et al. (2019a) and Papadogeorgou et al. (2019b) have proposed methods to combat the interference that spatial exposures pose to causal inference through a distance adjusted propensity score and through the clustering of individuals in a study, respectively; however, both of these methods are only applicable to cross-sectional studies and the extension to time-varying data is nontrivial.

5.3 Future Work

An important consideration when completing a survival analysis with a time-varying exposure is the underlying time-scale assumed by the model. In Chapter 4, we assumed a follow-up time using calendar months as the underlying time-scale for the survival models in the simulation study. When age is included as a linear predictor in the exponential term of a PH model, a log-linear relationship between age and the health outcome of interest is inherently assumed. This assumption is one that needs to be considered in epidemiological studies and may not hold in analyses (Oakes, 1995; Cologne et al., 2012; Vyas et al., 2021). Rather than using the follow-up time, age can be used as the underlying time-scale for survival models. This breaks the log-linear relationship; however, makes implementing TPRS much more challenging as the TPRS are constructed over calendar time, not the age of an individual. Investigating how to use TPRS to mitigate bias when using age as a time-scale of a survival model would be of interest.

All methods investigated in this work rely on the assumption that TPRS basis can completely model the underlying confounded structure. However, there is no simple way in an analysis to determine whether this assumption is valid as the confounded variables are unmeasured, and so the underlying confounded structure will always be unknown. It is unknown how the various confounded surfaces would perform with time-varying spatial data. Also, none of the methods are tested using a complex survey design (e.g. using survey weights). Using survey weights is common in environmental epidemiology studies (Zhang et al., 2023), however, none of the simulations used survey weights to test the mitigation effect of the methods proposed in this work. Further work is needed to make conclusions for various survey designs.

Bibliography

Abatzoglou, J. T. (2013). Development of gridded surface meteorological data for ecological applications and modelling. *International Journal of Climatology*, 33(1):121–131. _eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/joc.3413.

Adar, S. D., Chen, Y.-H., D'Souza, J. C., O'Neill, M. S., Szpiro, A. A., Auchincloss, A. H., Park, S. K., Daviglus, M. L., Diez Roux, A. V., and Kaufman, J. D. (2018). Longitudinal Analysis of Long-Term Air Pollution Levels and Blood Pressure: A Cautionary Tale from the Multi-Ethnic Study of Atherosclerosis. *Environmental Health Perspectives*, 126(10):107003.

Adin, A., Goicoa, T., and Ugarte, M. D. (2019). Online relative risks/rates estimation in spatial and spatio-temporal disease mapping. *Computer Methods and Programs in Biomedicine*, 172:103–116.

AlGhatrif, M., Strait, J. B., Morrell, C. H., Canepa, M., Wright, J., Elango, P., Scuteri, A., Najjar, S. S., Ferrucci, L., and Lakatta, E. G. (2013). Longitudinal Trajectories of Arterial Stiffness and the Role of Blood Pressure: The Baltimore Longitudinal Study of Aging. *Hypertension*, 62(5):934–941.

Anderson, G. B., Ferreri, J., Al-Hamdan, M., Crosson, W., Schumacher, A., Guikema, S., Quiring, S., Eddelbuettel, D., Yan, M., and Peng, R. D. (2020). Assessing United States County-Level Exposure for Research on Tropical Cyclones and Human Health. *Environmental Health Perspectives*, 128(10):107009.

Aparicio, J., Gezan, S. A., Ariza-Suarez, D., Raatz, B., Diaz, S., Heilman-Morales, A., and Lobaton, J. (2024). Mr.Bean: a comprehensive statistical and visualization application for modeling agricultural field trials data. *Frontiers in Plant Science*, 14:1290078.

Attali, D. (2021). shinyjs: Easily Improve the User Experience of Your Shiny Apps in Seconds.

Austin, P. C. (2012). Generating survival times to simulate Cox proportional hazards models with time-varying covariates. *Statistics in Medicine*, 31(29):3946–3958.

Azevedo, D. R. M., Prates, M. O., and Bandyopadhyay, D. (2023). Alleviating spatial confounding in frailty models. *Biostatistics*, 24(4):945–961.

Banerjee, S. (2003). Frailty modeling for spatially correlated survival data, with application to infant mortality in Minnesota. *Biostatistics*, 4(1):123–142.

Banerjee, S., Gelfand, A. E., Carlin, B. P., and Corporation, E. (2015). *Hierarchical modeling and analysis for spatial data*. Ebook Library (EBL). Chapman & Hall/CRC Press, Boca Raton, Florida ;, second edition. edition. Publication Title: Hierarchical modeling and analysis for spatial data.

Beltran, A., Wu, J., and Laurent, O. (2013). Associations of Meteorology with Adverse Pregnancy Outcomes: A Systematic Review of Preeclampsia, Preterm Birth and Birth Weight. *International Journal of Environmental Research and Public Health*, 11(1):91–172.

Blencowe, H., Cousens, S., Oestergaard, M. Z., Chou, D., Moller, A.-B., Narwal, R., Adler, A., Vera Garcia, C., Rohde, S., Say, L., and Lawn, J. E. (2012). National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends since 1990 for selected countries: a systematic analysis and implications. *The Lancet*, 379(9832):2162–2172.

Bobb, J. F., Cruz, M. F., Mooney, S. J., Drewnowski, A., Arterburn, D., and Cook, A. J. (2022). Accounting for Spatial Confounding in Epidemiological Studies with Individual-Level Exposures: An Exposure-Penalized Spline Approach. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 185(3):1271–1293.

Bosco, J. L., Silliman, R. A., Thwin, S. S., Geiger, A. M., Buist, D. S., Prout, M. N., Yood, M. U., Haque, R., Wei, F., and Lash, T. L. (2010). A most stubborn bias: no adjustment method fully resolves confounding by indication in observational studies. *Journal of Clinical Epidemiology*, 63(1):64–74.

Bosetti, C., Nieuwenhuijsen, M. J., Gallus, S., Cipriani, S., La Vecchia, C., and Parazzini, F. (2010). Ambient particulate matter and preterm birth or birth weight: a review of the literature. *Archives of Toxicology*, 84(6):447–460.

Chan, S. H., Van Hee, V. C., Bergen, S., Szpiro, A. A., DeRoo, L. A., London, S. J., Marshall, J. D., Kaufman, J. D., and Sandler, D. P. (2015). Long-Term Air Pollution Exposure and Blood Pressure in the Sister Study. *Environmental Health Perspectives*, 123(10):951–958.

Chang, W., Cheng, J., Allaire, J. J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., and Borges, B. (2023). shiny: Web Application Framework for R.

Cologne, J., Hsu, W.-L., Abbott, R. D., Ohishi, W., Grant, E. J., Fujiwara, S., and Cullings, H. M. (2012). Proportional Hazards Regression in Epidemiologic Follow-up Studies: An Intuitive Consideration of Primary Time Scale. *Epidemiology*, 23(4):565–573.

Demateis, D., Keller, K. P., Rojas-Rueda, D., Kioumourtzoglou, M., and Wilson, A. (2024). Penalized distributed lag interaction model: Air pollution, birth weight, and neighborhood vulnerability. *Environmetrics*, 35(4):e2843.

Di, Q., Amini, H., Shi, L., Kloog, I., Silvern, R., Kelly, J., Sabath, M. B., Choirat, C., Koutrakis, P., Lyapustin, A., Wang, Y., Mickley, L. J., and Schwartz, J. (2020). Assessing NO $_2$ Concentration and Model Uncertainty with High Spatiotemporal Resolution across the Contiguous United States Using Ensemble Model Averaging. *Environmental Science & Technology*, 54(3):1372–1384.

Di, Q., Wang, Y., Zanobetti, A., Wang, Y., Koutrakis, P., Choirat, C., Dominici, F., and Schwartz, J. D. (2017). Air Pollution and Mortality in the Medicare Population. *New England Journal of Medicine*, 376(26):2513–2522.

Diggle, P. J., Heagerty, P. J., Liang, K.-y., and Zeger, S. L. (2002). *Analysis of Longitudinal Data*. Oxford University Press.

Donato, G. and Belongie, S. (2002). Approximate Thin Plate Spline Mappings. In Goos, G., Hartmanis, J., Van Leeuwen, J., Heyden, A., Sparr, G., Nielsen, M., and Johansen, P., editors, *Computer Vision — ECCV 2002*, volume 2352, pages 21–31. Springer Berlin Heidelberg, Berlin, Heidelberg. Series Title: Lecture Notes in Computer Science.

Duchon, J. (1977). Splines minimizing rotation-invariant semi-norms in Sobolev spaces. In Schempp, W. and Zeller, K., editors, *Constructive Theory of Functions of Several Variables*, pages 85–100, Berlin, Heidelberg. Springer Berlin Heidelberg.

Dupont, E., Marques, I., and Kneib, T. (2023). Demystifying Spatial Confounding. arXiv:2309.16861 [math, stat].

Dupont, E., Wood, S. N., and Augustin, N. H. (2022). Spatial+: A novel approach to spatial confounding. *Biometrics*, 78(4):1279–1290.

Earth, N. (2009). Natural Earth, 1:10m Cultural Vectors - Free vector and raster map data at 1:10m, 1:50m, and 1:110m scales.

Figueira, M., Conesa, D., and López-Quílez, A. (2024). A shiny R app for spatial analysis of species distribution models. *Ecological Informatics*, 80:102542.

Fischer, P. H., Marra, M., Ameling, C. B., Hoek, G., Beelen, R., De Hoogh, K., Breugelmans, O., Kruize, H., Janssen, N. A., and Houthuijs, D. (2015). Air Pollution and Mortality in Seven Million Adults: The Dutch Environmental Longitudinal Study (DUELS). *Environmental Health Perspectives*, 123(7):697–704.

Friedman, J. H. (1991). Multivariate Adaptive Regression Splines. *The Annals of Statistics*, 19(1):1 – 67. Publisher: Institute of Mathematical Statistics.

Gilbert, B., Datta, A., Casey, J. A., and Ogburn, E. L. (2024). A causal inference framework for spatial confounding. arXiv:2112.14946 [stat].

Guan, Y., Page, G. L., Reich, B. J., Ventrucci, M., and Yang, S. (2023). Spectral adjustment for spatial confounding. *Biometrika*, 110(3):699–719.

Gunasekara, F. I., Richardson, K., Carter, K., and Blakely, T. (2014). Fixed effects analysis of repeated measures data. *International Journal of Epidemiology*, 43(1):264–269.

Hack, M., Klein, N. K., and Taylor, H. G. (1995). Long-Term Developmental Outcomes of Low Birth Weight Infants. *The Future of Children*, 5(1):176.

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(4):243–254.

Hastie, T., Friedman, J., and Tibshirani, R. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York, New York, NY.

Havard, S., Deguen, S., Zmirou-Navier, D., Schillinger, C., and Bard, D. (2009). Traffic-Related Air Pollution and Socioeconomic Status: A Spatial Autocorrelation Study to Assess Environmental Equity on a Small-Area Scale:. *Epidemiology*, 20:S33.

Hodges, J. S. and Reich, B. J. (2010). Adding Spatially-Correlated Errors Can Mess Up the Fixed Effect You Love. *The American Statistician*, 64(4):325–334.

Hughes, J. and Haran, M. (2013). Dimension reduction and alleviation of confounding for spatial generalized linear mixed models: *Dimension Reduction and Alleviation of Confounding*. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(1):139–159.

Jia, P., Cao, X., Yang, H., Dai, S., He, P., Huang, G., Wu, T., and Wang, Y. (2021). Green space access in the neighbourhood and childhood obesity. *Obesity Reviews*, 22(S1):e13100.

Johnson, O., Fronterre, C., Diggle, P. J., Amoah, B., and Giorgi, E. (2021). MBGapp: A Shiny application for teaching model-based geostatistics to population health scientists. *PLOS ONE*, 16(12):e0262145.

Kaufman, J. D., Adar, S. D., Barr, R. G., Budoff, M., Burke, G. L., Curl, C. L., Daviglus, M. L., Roux, A. V. D., Gassett, A. J., Jacobs, Jr, D. R., Kronmal, R., Larson, T. V., Navas-Acien, A., Olives, C., Sampson, P. D., Sheppard, L., Siscovick, D. S., Stein, J. H., Szpiro, A. A., and Watson, K. E. (2016). Association between air pollution and coronary artery calcification within six metropolitan areas in the USA (the Multi-Ethnic Study of Atherosclerosis and Air Pollution): a longitudinal cohort study. *The Lancet*, 388(10045):696–704. Publisher: Elsevier.

Keller, J. P., Dunlop, J. H., Ryder, N. A., Peng, R. D., and Keet, C. A. (2022). Long-Term Ambient Air Pollution and Childhood Eczema in the United States. *Environmental Health Perspectives*, 130(5):057702.

Keller, J. P., Olives, C., Kim, S.-Y., Sheppard, L., Sampson, P. D., Szpiro, A. A., Oron, A. P., Lindström, J., Vedal, S., and Kaufman, J. D. (2015). A Unified Spatiotemporal Modeling Approach for Predicting Concentrations of Multiple Air Pollutants in the Multi-Ethnic Study of Atherosclerosis and Air Pollution. *Environmental Health Perspectives*, 123(4):301–309.

Keller, J. P. and Szpiro, A. A. (2020). Selecting a scale for spatial confounding adjustment. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 183(3):1121–1143.

Keller, K. and Rainey, M. (2024). spconf: Computing Scales of Spatial Smoothing for Confounding Adjustment.

Keogh, R. H., Daniel, R. M., VanderWeele, T. J., and Vansteelandt, S. (2018). Analysis of Longitudinal Studies With Repeated Outcome Measures: Adjusting for Time-Dependent Confounding Using Conventional Methods. *American Journal of Epidemiology*, 187(5):1085–1092.

Khan, K. and Berrett, C. (2023). Re-thinking Spatial Confounding in Spatial Linear Mixed Models. arXiv:2301.05743 [stat].

Khan, K. and Calder, C. A. (2022). Restricted Spatial Regression Methods: Implications for Inference. *Journal of the American Statistical Association*, 117(537):482–494.

Klungsøyr, O., Sexton, J., Sandanger, I., and Nygård, J. F. (2009). Sensitivity analysis for unmeasured confounding in a marginal structural Cox proportional hazards model. *Lifetime Data Analysis*, 15(2):278–294.

Lakshmanan, A., Chiu, Y.-H. M., Coull, B. A., Just, A. C., Maxwell, S. L., Schwartz, J., Gryparis, A., Kloog, I., Wright, R. J., and Wright, R. O. (2015). Associations between prenatal traffic-related air pollution exposure and birth weight: Modification by sex and maternal pre-pregnancy body mass index. *Environmental Research*, 137:268–277.

Lamichhane, D. K., Leem, J.-H., Lee, J.-Y., and Kim, H.-C. (2015). A meta-analysis of exposure to particulate matter and adverse birth outcomes. *Environmental Health and Toxicology*, 30:e2015011.

Lee, J., Phillips, D., Wilkens, J., and Gateway to Global Aging Data Team (2021). Gateway to Global Aging Data: Resources for Cross-National Comparisons of Family, Social Environment, and Healthy Aging. *The Journals of Gerontology: Series B*, 76(Supplement_1):S5–S16.

Lee, N. and Ma, S. (2024). A joint modeling approach to treatment effects estimation with unmeasured confounders. arXiv:2411.10980 [stat].

Leisch, F. (2006). A Toolbox for K-Centroids Cluster Analysis. *Computational Statistics and Data Analysis*, 51(2):526–544.

Lin, N. X. and Henley, W. E. (2016). Prior event rate ratio adjustment for hidden confounding in observational studies of treatment effectiveness: a pairwise Cox likelihood approach. *Statistics in Medicine*, 35(28):5149–5169.

Lindström, J., Szpiro, A. A., Sampson, P. D., Oron, A. P., Richards, M., Larson, T. V., and Sheppard, L. (2014). A flexible spatio-temporal model for air pollution with spatial and spatio-temporal covariates. *Environmental and Ecological Statistics*, 21(3):411–433.

Liu, L., Oza, S., Hogan, D., Perin, J., Rudan, I., Lawn, J. E., Cousens, S., Mathers, C., and Black, R. E. (2015). Global, regional, and national causes of child mortality in 2000–13,

with projections to inform post-2015 priorities: an updated systematic analysis. *The Lancet*, 385(9966):430–440.

MacKenzie, T. A., Tosteson, T. D., Morden, N. E., Stukel, T. A., and O'Malley, A. J. (2014). Using instrumental variables to estimate a Cox's proportional hazards regression subject to additive confounding. *Health Services and Outcomes Research Methodology*, 14(1-2):54–68.

Marques, I. and Kneib, T. (2022). Discussion on "Spatial+: A novel approach to spatial confounding" by Emiko Dupont, Simon N. Wood, and Nicole H. Augustin. *Biometrics*, 78(4):1295–1299.

Marques, I., Kneib, T., and Klein, N. (2022). Mitigating spatial confounding by explicitly correlating Gaussian random fields. *Environmetrics*, 33(5):e2727.

McCormick, R. (2017). Does Access to Green Space Impact the Mental Well-being of Children: A Systematic Review. *Journal of Pediatric Nursing*, 37:3–7.

McDougall, C. W., Quilliam, R. S., Hanley, N., and Oliver, D. M. (2020). Freshwater blue space and population health: An emerging research agenda. *Science of The Total Environment*, 737:140196.

McDuffie, E. E., Martin, R. V., Spadaro, J. V., Burnett, R., Smith, S. J., O'Rourke, P., Hammer, M. S., Van Donkelaar, A., Bindle, L., Shah, V., Jaeglé, L., Luo, G., Yu, F., Adeniran, J. A., Lin, J., and Brauer, M. (2021). Source sector and fuel contributions to ambient PM2.5 and attributable mortality across multiple spatial scales. *Nature Communications*, 12(1):3594.

Mork, D., Kioumourtzoglou, M.-A., Weisskopf, M., Coull, B. A., and Wilson, A. (2024). Heterogeneous Distributed Lag Models to Estimate Personalized Effects of Maternal Exposures to Air Pollution. *Journal of the American Statistical Association*, 119(545):14–26.

Nawrot, T., Plusquin, M., Hogervorst, J., Roels, H. A., Celis, H., Thijs, L., Vangronsveld, J., Van Hecke, E., and Staessen, J. A. (2006). Environmental exposure to cadmium and risk of cancer: a prospective population-based study. *The Lancet Oncology*, 7(2):119–126.

NRC, N. R. C. (1991). Environmental Epidemiology, Volume 1: Public Health and Hazardous Wastes. Publisher: National Academies Press.

Oakes, D. (1995). Multiple time scales in survival analysis. *Lifetime Data Analysis*, 1(1):7–18.

Ogonowski, J., Miazgowski, T., Engel, K., and Celewicz, Z. (2014). Birth weight predicts the risk of gestational diabetes mellitus and pregravid obesity. *Nutrition*, 30(1):39–43.

Paciorek, C. J. (2010). The Importance of Scale for Spatial-Confounding Bias and Precision of Spatial Regression Estimators. *Statistical Science*, 25(1).

Palta, M. and Yao, T.-J. (1991). Analysis of Longitudinal Data with Unmeasured Confounders. *Biometrics*, 47(4):1355.

Pan, W. (2001). Akaike's Information Criterion in Generalized Estimating Equations. *Biometrics*, 57(1):120–125.

Papadogeorgou, G. (2022). Discussion on "Spatial+: a novel approach to spatial confounding" by Emiko Dupont, Simon N. Wood, and Nicole H. Augustin. *Biometrics*, 78(4):4.

Papadogeorgou, G., Choirat, C., and Zigler, C. M. (2019a). Adjusting for unmeasured spatial confounding with distance adjusted propensity score matching. *Biostatistics*, 20(2):256–272.

Papadogeorgou, G. and Dominici, F. (2020). A causal exposure response function with local adjustment for confounding: Estimating health effects of exposure to low levels of ambient fine particulate matter. *The Annals of Applied Statistics*, 14(2).

Papadogeorgou, G., Mealli, F., and Zigler, C. M. (2019b). Causal Inference With Interfering Units for Cluster and Population Level Treatment Allocation Programs. *Biometrics*, 75(3):778–787.

Pekkanen, J. and Pearce, N. (2001). Environmental epidemiology: challenges and opportunities. *Environmental Health Perspectives*, 109(1).

Peng, R. D., Chang, H. H., Bell, M. L., McDermott, A., Zeger, S. L., Samet, J. M., and Dominici, F. (2008). Coarse Particulate Matter Air Pollution and Hospital Admissions for Cardiovascular and Respiratory Diseases Among Medicare Patients. *JAMA*, 299(18):2172–2179. _eprint: https://jamanetwork.com/journals/jama/articlepdf/181898/joc80039_2172_2179.pdf.

Perrier, V., Meyer, F., and Granjon, D. (2023). shinyWidgets: Custom Inputs Widgets for Shiny.

Purisch, S. E. and Gyamfi-Bannerman, C. (2017). Epidemiology of preterm birth. *Seminars in Perinatology*, 41(7):387–391.

R Core Team (2023). R: A Language and Environment for Statistical Computing.

Rainey, M. J. and Keller, K. P. (2024). spconfShiny: An R Shiny application for calculating the spatial scale of smoothing splines for point data. *PLOS ONE*, 19(10):e0311440.

Rasmussen, C. E. and Williams, C. K. I. (2008). *Gaussian processes for machine learning*. Adaptive computation and machine learning. MIT Press, Cambridge, Mass., 3. print edition.

Reich, B. J., Yang, S., and Guan, Y. (2022). Discussion on "Spatial+: A novel approach to spatial confounding" by Dupont, Wood, and Augustin. *Biometrics*, 78(4):1291–1294.

Salehi, M., Arashi, M., Bekker, A., Ferreira, J., Chen, D.-G., Esmaeili, F., and Frances, M. (2021). A Synergetic R-Shiny Portal for Modeling and Tracking of COVID-19 Data. *Frontiers in Public Health*, 8:623624.

Sampson, P. D., Richards, M., Szpiro, A. A., Bergen, S., Sheppard, L., Larson, T. V., and Kaufman, J. D. (2013). A regionalized national universal kriging model using Partial Least Squares regression for estimating annual PM2.5 concentrations in epidemiology. *Atmospheric Environment*, 75:383–392.

Schnell, P. M. and Papadogeorgou, G. (2020). Mitigating unobserved spatial confounding when estimating the effect of supermarket access on cardiovascular disease deaths. *The Annals of Applied Statistics*, 14(4).

Sievert, C., Cheng, J., and Aden-Buie, G. (2023). bslib: Custom 'Bootstrap' 'Sass' Themes for 'shiny' and 'rmarkdown'.

Silva, I., Fleming, C. H., Noonan, M. J., Fagan, W. F., and Calabrese, J. M. (2023). movedesign: Shiny R app to evaluate sampling design for animal movement studies. *Methods in Ecology and Evolution*, 14(9):2216–2225.

Stieb, D. M., Chen, L., Eshoul, M., and Judek, S. (2012). Ambient air pollution, birth weight and preterm birth: A systematic review and meta-analysis. *Environmental Research*, 117:100–111.

Streeter, A. J., Lin, N. X., Crathorne, L., Haasova, M., Hyde, C., Melzer, D., and Henley, W. E. (2017). Adjusting for unmeasured confounding in nonrandomized longitudinal studies: a methodological review. *Journal of Clinical Epidemiology*, 87:23–34.

Sun, S., Weinberger, K. R., Yan, M., Brooke Anderson, G., and Wellenius, G. A. (2020). Tropical cyclones and risk of preterm birth: A retrospective analysis of 20 million births across 378 US counties. *Environment International*, 140:105825.

Svechkina, A., Portnov, B. A., and Trop, T. (2020). The impact of artificial light at night on human and ecosystem health: a systematic literature review. *Landscape Ecology*, 35(8):1725–1742.

Szpiro, A. A., Sheppard, L., and Lumley, T. (2011). Efficient measurement error correction with spatially misaligned data. *Biostatistics*, 12(4):610–623. _eprint: https://academic.oup.com/biostatistics/article-pdf/12/4/610/34305154/kxq083.pdf.

Thaden, H. and Kneib, T. (2018). Structural Equation Models for Dealing With Spatial Confounding. *The American Statistician*, 72(3):239–252.

the CHILD study investigators, Takaro, T. K., Scott, J. A., Allen, R. W., Anand, S. S., Becker, A. B., Befus, A. D., Brauer, M., Duncan, J., Lefebvre, D. L., Lou, W., Mandhane, P. J., McLean, K. E., Miller, G., Sbihi, H., Shu, H., Subbarao, P., Turvey, S. E., Wheeler, A. J., Zeng, L.,

Sears, M. R., and Brook, J. R. (2015). The Canadian Healthy Infant Longitudinal Development (CHILD) birth cohort study: assessment of environmental exposures. *Journal of Exposure Science & Environmental Epidemiology*, 25(6):580–592.

Tétreault, L.-F., Doucet, M., Gamache, P., Fournier, M., Brand, A., Kosatsky, T., and Smargiassi, A. (2016). Childhood Exposure to Ambient Air Pollutants and the Onset of Asthma: An Administrative Cohort Study in Québec. *Environmental Health Perspectives*, 124(8):1276–1282.

Van Donkelaar, A., Hammer, M. S., Bindle, L., Brauer, M., Brook, J. R., Garay, M. J., Hsu, N. C., Kalashnikova, O. V., Kahn, R. A., Lee, C., Levy, R. C., Lyapustin, A., Sayer, A. M., and Martin, R. V. (2021). Monthly Global Estimates of Fine Particulate Matter and Their Uncertainty. *Environmental Science & Technology*, 55(22):15287–15300.

Vaneckova, P., Beggs, P. J., and Jacobson, C. R. (2010). Spatial analysis of heat-related mortality among the elderly between 1993 and 2004 in Sydney, Australia. *Social Science & Medicine*, 70(2):293–304.

Vyas, M. V., Fang, J., Kapral, M. K., and Austin, P. C. (2021). Choice of time-scale in time-to-event analysis: evaluating age-dependent associations. *Annals of Epidemiology*, 62:69–76.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wiecha, N., Hoppin, J. A., and Reich, B. J. (2024). Two-stage Estimators for Spatial Confounding. arXiv:2404.09358 [stat].

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 65(1):95–114.

Wood, S. N. (2011). Fast Stable Restricted Maximum Likelihood and Marginal Likelihood Estimation of Semiparametric Generalized Linear Models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 73(1):3–36.

Xue, Y., Schifano, E. D., and Hu, G. (2020). Geographically Weighted Cox Regression for Prostate Cancer Survival Data in Louisiana. *Geographical Analysis*, 52(4):570–587.

Zeger, S. L., Liang, K.-Y., and Albert, P. S. (1988). Models for Longitudinal Data: A Generalized Estimating Equation Approach. *Biometrics*, 44(4):1049.

Zhang, B., Weuve, J., Langa, K. M., D'Souza, J., Szpiro, A., Faul, J., Mendes De Leon, C., Gao, J., Kaufman, J. D., Sheppard, L., Lee, J., Kobayashi, L. C., Hirth, R., and Adar, S. D. (2023). Comparison of Particulate Air Pollution From Different Emission Sources and Incident Dementia in the US. *JAMA Internal Medicine*, 183(10):1080.

Zimmerman, D. L. and Ver Hoef, J. M. (2022). On Deconfounding Spatial Confounding in Linear Models. *The American Statistician*, 76(2):159–167.

Šrám, R. J., Binková, B., Dejmek, J., and Bobak, M. (2005). Ambient Air Pollution and Pregnancy Outcomes: A Review of the Literature. *Environmental Health Perspectives*, 113(4):375–382.

Appendix A

Supplemental Material for the Introduction

A.1 Multivariate Confounding Calculations

Let's say we fit the model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}$ when, in actuality, the true model is $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{C}\boldsymbol{\delta}+\boldsymbol{\epsilon}$. Then, assuming $\mathrm{E}[\boldsymbol{\epsilon}] = 0$, we can compute the bias of $\hat{\boldsymbol{\beta}}$ as

$$\begin{aligned}
\mathrm{E}[\hat{\boldsymbol{\beta}}] &= \mathrm{E}[(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{Y}] \\
&= (\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}\mathrm{E}[\boldsymbol{Y}] \\
&= (\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{C}\boldsymbol{\delta}) \\
&= (\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{X}\boldsymbol{\beta} + (\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{C}\boldsymbol{\delta} \\
&= \boldsymbol{\beta} + (\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{C}\boldsymbol{\delta}. \qquad \text{(A.1)}
\end{aligned}$$

Therefore, $\hat{\boldsymbol{\beta}}$ has a bias of $(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{C}\boldsymbol{\delta}$. If $\boldsymbol{C}$ is a confounder, not including it in the model will impact the results.

A.2 Restricted Spatial Regression Methodology

Introduced by Hodges and Reich (2010) and extended by Hanks et al. (2015), we define $\boldsymbol{y}$ to be a spatially varying outcome conditional upon the mean, $\boldsymbol{\mu}$ and additional covariates $\boldsymbol{\gamma}$. Additionally, we assume, for a given link function, $g(\cdot)$, the mean of the outcome is a transformed function of a spatially-varying covariates, $\boldsymbol{X}$, and a zero-mean spatial random effect, $\boldsymbol{\eta}$, quantifying the

random spatial variation present in the model. That is, we define

$$\boldsymbol{y} \sim [\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\gamma}] \tag{A.2}$$

$$g(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\eta} \tag{A.3}$$

$$\boldsymbol{\eta} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_\eta) \tag{A.4}$$

$$\boldsymbol{\beta} \sim N(\mu_\beta + \boldsymbol{\Sigma}_\beta) \tag{A.5}$$

where $\boldsymbol{\Sigma}$ are covariance matrices; $\boldsymbol{\beta}$ are the fixed regression parameters on the covariates in $\boldsymbol{X}$. Under this framework, spatial confounding is defined as random effect collinearity between the spatial covariates of interest, $\boldsymbol{X}$, and the spatial random effect, $\boldsymbol{\eta}$.

To mitigate the multicollinearity, Hodges and Reich (2010) constrained the random effect to be orthogonal to the spatial covariates fitting

$$g(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\delta} + (\boldsymbol{I} - \boldsymbol{P_X})\boldsymbol{\eta} \tag{A.6}$$

instead of (A.3), where $\boldsymbol{I}$ is the identity matrix, $\boldsymbol{P_X} = \boldsymbol{X}(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}$ is the projection matrix onto the column space of $\boldsymbol{X}$, and $\boldsymbol{\delta}$ is the unconditional relationship between $\boldsymbol{X}$ and $\boldsymbol{y}$. By constraining the random effect, the random effect is now independent of the spatially-varying covariates, relieving the multicollinearity and therefore removing the spatial confounding. Hanks et al. (2015) extended this framework to model the conditional relationship of $\boldsymbol{X}$ and $\boldsymbol{y}$, conditional on the random effect, $\boldsymbol{\eta}$, by transforming the unconditional effect such that $\boldsymbol{\beta} = \boldsymbol{\delta} + (\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{\eta}$.

Appendix B

Supplemental Material for Chapter 2

B.1 Additional Simulation Details

B.1.1 Simulation Study Information

Table B.1: Description of the degrees of freedom (df) used by the various methods in simulation study.

Response Type	Method	df Selector	df Implemented
Continuous	Spatial+ and gSEM	Fixed	300
	KS	Fixed	10
	KS and Spatial+	AIC/BIC	3 to 50 incrementing by 1
			55 to 100 incrementing by 5
			125 to 300 incrementing by 25
Binary	Spatial+ and gSEM	Fixed	200
	KS	Fixed	10
	KS and Spatial+	AIC/BIC	3 to 50 incrementing by 1
			55 to 100 incrementing by 5
			125 to 200 incrementing by 25

B.1.2 Additional Simulation Results

Results from Continuous Outcomes

Table B.2: Root mean squared error (RMSE), estimated bias, and coverage rates for 95% confidence intervals for the estimates of β_x in the simulation where $\phi_1 = 150$, $\phi_2 = 10$, $\phi_3 = 100$, $\sigma_x = 0.1$, and $\sigma_y = 1$. The median degrees of freedom (df) for each method is given for both the exposure and outcome model.

Method	df Selector	RMSE	Bias	Coverage (95% Nominal)	Median df Selected (Exposure)	(Outcome)
Unadjusted	–	0.176	0.079	0.434	–	–
Spatial+	Fixed	0.228	0.079	0.946	300	300
Spatial+	GCV	0.334	0.265	0.756	218.6	50.3
Spatial+	AIC	0.225	0.055	0.940	300	300
Spatial+	BIC	0.155	0.054	0.938	65	65
gSEM	Fixed	0.228	0.055	0.876	300	300
gSEM	GCV	0.329	0.259	0.746	218.6	50.7
KS	AIC	0.127	0.011	0.960	–	33
KS	BIC	0.108	0.022	0.920	–	9
KS	Fixed	0.118	0.055	0.876	–	10
E-PS	GCV	0.175	0.056	0.964	218.6	218.4

Table B.3: Root mean squared error (RMSE), estimated bias, and coverage rates for 95% confidence intervals for the estimates of β_x in the simulation where $\phi_2 = 10$, $\phi_3 = 100$, $\sigma_x = 0$, and $\sigma_y = 1$. The median degrees of freedom (df) for each method is given for both the exposure and outcome model.

ϕ_1	Method	df Selector	RMSE	Bias	Coverage (95% Nominal)	Median df Selected (Exposure)	(Outcome)
5	Unadjusted	–	0.650	0.629	0.004	–	–
	Spatial+	Fixed	0.696	0.676	0.012	300	300
	Spatial+	GCV	0.949	0.935	0.000	237.8	127.8
	Spatial+	AIC	0.695	0.675	0.012	300	300
	Spatial+	BIC	0.673	0.664	0.000	80	80
	gSEM	Fixed	0.696	0.676	0.008	300	300
	gSEM	GCV	0.914	0.898	0.000	237.8	129.4
	KS	AIC	0.638	0.626	0.004	–	100
	KS	BIC	0.640	0.633	0.000	–	27
	KS	Fixed	0.665	0.660	0.000	–	10
	E-PS	GCV	0.682	0.670	0.000	237.8	237.6
50	Unadjusted	–	0.279	0.207	0.210	–	–
	Spatial+	Fixed	0.312	0.186	0.876	300	300
	Spatial+	GCV	0.499	0.434	0.592	238.2	60.0
	Spatial+	AIC	0.313	0.172	0.876	300	300
	Spatial+	BIC	0.241	0.168	0.816	80	80
	gSEM	Fixed	0.312	0.186	0.824	300	300
	gSEM	GCV	0.493	0.427	0.572	238.3	60.2
	KS	AIC	0.180	0.111	0.866	–	40
	KS	BIC	0.165	0.124	0.756	–	11.5
	KS	Fixed	0.197	0.164	0.562	–	10
	E-PS	GCV	0.264	0.177	0.886	238.3	238.1
150	Unadjusted	–	0.188	0.084	0.442	–	–
	Spatial+	Fixed	0.275	0.083	0.946	300	300
	Spatial+	GCV	0.410	0.318	0.784	238.2	49.3
	Spatial+	AIC	0.276	0.083	0.946	300	300
	Spatial+	BIC	0.190	0.069	0.928	80	80
	gSEM	Fixed	0.275	0.083	0.872	300	300
	gSEM	GCV	0.406	0.314	0.776	238.2	49.4
	KS	AIC	0.141	0.009	0.940	–	32
	KS	BIC	0.111	0.021	0.932	–	9
	KS	Fixed	0.123	0.091	0.874	–	10
	E-PS	GCV	0.219	0.075	0.956	238.2	238.0

Table B.4: Root mean squared error (RMSE), estimated bias, and coverage rates for 95% confidence intervals for the estimates of β_x in the simulation where $\phi_2 = 10$, $\phi_3 = 100$, $\sigma_x = 2$, and $\sigma_y = 1$. The median degrees of freedom (df) for each method is given for both the exposure and outcome model.

ϕ_1	Method	df Selector	RMSE	Bias	Coverage (95% Nominal)	Median df Selected (Exposure)	(Outcome)
5	Unadjusted	–	0.129	0.119	0.020	–	–
	Spatial+	Fixed	0.021	0.010	0.922	300	300
	Spatial+	GCV	0.034	0.029	0.586	34.5	81.0
	Spatial+	AIC	0.034	0.029	0.590	22	22
	Spatial+	BIC	0.055	0.051	0.190	6	6
	gSEM	Fixed	0.021	0.010	0.844	300	300
	gSEM	GCV	0.021	-0.004	0.844	34.5	58.3
	KS	AIC	0.028	0.020	0.748	–	39
	KS	BIC	0.042	0.038	0.388	–	11
	KS	Fixed	0.050	0.046	0.246	–	10
	E-PS	GCV	0.037	0.033	0.498	34.5	34.5
50	Unadjusted	–	0.034	0.022	0.636	–	–
	Spatial+	Fixed	0.019	0.002	0.960	300	300
	Spatial+	GCV	0.020	0.010	0.894	19.2	33.2
	Spatial+	AIC	0.017	0.004	0.938	13	13
	Spatial+	BIC	0.018	0.007	0.924	4	4
	gSEM	Fixed	0.019	0.002	0.888	300	300
	gSEM	GCV	0.018	0.000	0.914	19.2	21.7
	KS	AIC	0.016	0.002	0.946	–	15
	KS	BIC	0.017	0.005	0.946	–	4
	KS	Fixed	0.017	0.005	0.944	–	10
	E-PS	GCV	0.017	0.005	0.942	19.2	19.2
150	Unadjusted	–	0.023	0.008	0.834	–	–
	Spatial+	Fixed	0.019	0.001	0.958	300	300
	Spatial+	GCV	0.019	0.008	0.904	18.0	27.7
	Spatial+	AIC	0.016	0.002	0.944	12	12
	Spatial+	BIC	0.016	0.002	0.938	4	4
	gSEM	Fixed	0.019	0.001	0.890	300	300
	gSEM	GCV	0.017	0.001	0.918	18.0	18.0
	KS	AIC	0.016	0.000	0.944	–	12
	KS	BIC	0.016	0.001	0.948	–	4
	KS	Fixed	0.016	0.002	0.948	–	10
	E-PS	GCV	0.016	0.002	0.944	18.0	18.0

Table B.5: Root mean squared error (RMSE), estimated bias, and coverage rates for 95% confidence intervals for the estimates of β_x in the simulation where $\phi_2 = 10$, $\phi_3 = 100$, $\sigma_x = 0.1$, and $\sigma_y = 10$. The median degrees of freedom (df) for each method is given for both the exposure and outcome model.

ϕ_1	Method	df Selector	RMSE	Bias	Coverage (95% Nominal)	Median df Selected (Exposure)	(Outcome)
5	Unadjusted	–	0.718	0.620	0.526	–	–
	Spatial+	Fixed	1.621	0.632	0.922	300	300
	Spatial+	GCV	1.745	1.051	0.884	229.7	8.1
	Spatial+	AIC	1.618	0.640	0.924	300	300
	Spatial+	BIC	1.181	0.646	0.896	75	75
	gSEM	Fixed	1.621	0.632	0.872	300	300
	gSEM	GCV	1.745	1.051	0.880	229.7	8.1
	KS	AIC	0.703	0.486	0.860	–	5
	KS	BIC	0.767	0.615	0.742	–	3
	KS	Fixed	0.872	0.659	0.796	–	10
	E-PS	GCV	1.296	0.653	0.940	229.7	229.5
50	Unadjusted	–	0.537	0.189	0.930	–	–
	Spatial+	Fixed	2.144	0.247	0.954	300	300
	Spatial+	GCV	2.030	0.517	0.952	220.4	3.5
	Spatial+	AIC	2.121	0.256	0.952	300	300
	Spatial+	BIC	1.421	0.180	0.970	70	70
	gSEM	Fixed	2.144	0.247	0.892	300	300
	gSEM	GCV	2.030	0.517	0.950	220.4	3.5
	KS	AIC	0.691	0.020	0.960	–	3
	KS	BIC	0.679	0.150	0.948	–	3
	KS	Fixed	0.899	0.148	0.950	–	10
	E-PS	GCV	1.614	0.235	0.972	220.4	220.2
150	Unadjusted	–	0.533	0.068	0.950	–	–
	Spatial+	Fixed	2.227	0.182	0.952	300	300
	Spatial+	GCV	2.086	0.418	0.958	218.6	2.8
	Spatial+	AIC	2.194	0.185	0.948	300	300
	Spatial+	BIC	1.465	0.096	0.962	65	65
	gSEM	Fixed	2.227	0.182	0.890	300	300
	gSEM	GCV	2.086	0.418	0.952	218.6	2.8
	KS	AIC	0.741	-0.071	0.956	–	3
	KS	BIC	0.698	0.045	0.946	–	3
	KS	Fixed	0.941	0.042	0.950	–	10
	E-PS	GCV	1.673	0.156	0.980	218.6	218.4

Binary Outcome

Table B.6: Root mean squared error (RMSE), estimated bias, and coverage rates for 95% confidence intervals for the estimates of β_x in the simulation where $\phi_1 = 150$, $\phi_2 = 10$, $\phi_3 = 100$, $\sigma_x = 0.1$ and the outcome is binary. The median degrees of freedom (df) for each method is given for both the exposure and outcome model.

Method	df Selector	RMSE	Bias	Coverage (95% Nominal)	Median df Selected (Exposure)	(Outcome)
Unadjusted	–	0.131	0.031	0.886	–	–
Spatial+	Fixed	0.472	0.091	0.898	200	200
Spatial+	GCV	0.392	0.055	0.928	180.1	8.4
Spatial+	AIC	0.472	0.091	0.898	200	200
Spatial+	BIC	0.372	0.058	0.934	100	100
gSEM	Fixed	0.552	0.054	0.942	200	200
gSEM	GCV	0.388	0.049	0.928	180.1	8.4
KS	AIC	0.172	-0.026	0.948	–	6
KS	BIC	0.149	0.009	0.954	–	3
KS	Fixed	0.190	0.029	0.948	–	10
E-PS	GCV	0.320	0.055	0.940	180.1	118.8

Table B.7: Root mean squared error (RMSE), estimated bias, and coverage rates for 95% confidence intervals for the estimates of β_x in the simulation where $\phi_2 = 10$, $\phi_3 = 100$, $\sigma_x = 0$ and the outcome is binary. The median degrees of freedom (df) for each method is given for both the exposure and outcome model.

ϕ_1	Method	df Selector	RMSE	Bias	Coverage (95% Nominal)	Median df Selected (Exposure)	(Outcome)
5	Unadjusted	–	0.327	0.308	0.056	–	–
	Spatial+	Fixed	0.599	0.496	0.588	200	200
	Spatial+	GCV	0.478	0.394	0.674	185.7	30.7
	Spatial+	AIC	0.599	0.496	0.588	200	200
	Spatial+	BIC	0.533	0.445	0.600	150	150
	gSEM	Fixed	0.467	0.294	0.822	200	200
	gSEM	GCV	0.451	0.352	0.714	185.7	30.5
	KS	AIC	0.325	0.292	0.466	–	19
	KS	BIC	0.321	0.302	0.198	–	4
	KS	Fixed	0.355	0.334	0.198	–	10
	E-PS	GCV	0.454	0.398	0.592	185.7	128.1
50	Unadjusted	–	0.175	0.103	0.712	–	–
	Spatial+	Fixed	0.558	0.238	0.894	200	200
	Spatial+	GCV	0.452	0.171	0.936	185.8	10.6
	Spatial+	AIC	0.558	0.238	0.894	200	200
	Spatial+	BIC	0.497	0.194	0.920	150	150
	gSEM	Fixed	0.598	0.124	0.922	200	200
	gSEM	GCV	0.456	0.157	0.934	185.8	10.6
	KS	AIC	0.179	0.031	0.964	–	7
	KS	BIC	0.166	0.074	0.922	–	3
	KS	Fixed	0.211	0.095	0.912	–	10
	E-PS	GCV	0.368	0.154	0.956	185.8	130.4
150	Unadjusted	–	0.149	0.045	0.816	–	–
	Spatial+	Fixed	0.546	0.180	0.920	200	200
	Spatial+	GCV	0.453	0.120	0.948	185.8	8.6
	Spatial+	AIC	0.546	0.180	0.920	200	200
	Spatial+	BIC	0.491	0.147	0.940	150	150
	gSEM	Fixed	0.614	0.089	0.936	200	200
	gSEM	GCV	0.458	0.114	0.952	185.8	8.5
	KS	AIC	0.180	-0.011	0.962	–	5
	KS	BIC	0.153	0.028	0.956	–	3
	KS	Fixed	0.204	0.047	0.940	–	10
	E-PS	GCV	0.364	0.103	0.960	185.8	112.9

Table B.8: Root mean squared error (RMSE), estimated bias, and coverage rates for 95% confidence intervals for the estimates of β_x in the simulation where $\phi_2 = 1$, $\phi_3 = 100$, $\sigma_x = 0.1$ and the outcome is binary. The median degrees of freedom (df) for each method is given for both the exposure and outcome model.

ϕ_1	Method	df Selector	RMSE	Bias	Coverage (95% Nominal)	Median df Selected (Exposure)	(Outcome)
5	Unadjusted	–	0.195	0.179	0.210	–	–
	Spatial+	Fixed	0.242	0.185	0.738	200	200
	Spatial+	GCV	0.172	0.115	0.844	184.4	41.4
	Spatial+	AIC	0.242	0.185	0.738	200	200
	Spatial+	BIC	0.208	0.155	0.760	150	150
	gSEM	Fixed	0.292	0.046	0.928	200	200
	gSEM	GCV	0.148	0.071	0.990	184.4	41.4
	KS	AIC	0.113	0.079	0.864	–	26
	KS	BIC	0.123	0.105	0.662	–	5
	KS	Fixed	0.131	0.112	0.634	–	10
	E-PS	GCV	0.168	0.130	0.808	184.4	125.1
50	Unadjusted	–	0.083	0.041	1.000	–	–
	Spatial+	Fixed	0.206	0.102	0.856	200	200
	Spatial+	GCV	0.153	0.046	0.924	184.0	24.2
	Spatial+	AIC	0.206	0.102	0.856	200	200
	Spatial+	BIC	0.176	0.074	0.898	150	150
	gSEM	Fixed	0.200	0.020	0.952	200	200
	gSEM	GCV	0.145	0.024	0.928	184.0	23.8
	KS	AIC	0.081	-0.010	0.960	–	14
	KS	BIC	0.071	0.014	0.954	–	3
	KS	Fixed	0.078	0.018	0.952	–	10
	E-PS	GCV	0.130	0.050	0.938	184.0	126.2
150	Unadjusted	–	0.073	0.015	1.000	–	–
	Spatial+	Fixed	0.204	0.091	0.876	200	200
	Spatial+	GCV	0.153	0.037	0.926	184.0	22.0
	Spatial+	AIC	0.204	0.091	0.876	200	200
	Spatial+	BIC	0.171	0.064	0.902	150	150
	gSEM	Fixed	0.211	0.014	0.944	200	200
	gSEM	GCV	0.145	0.016	0.934	184.0	21.9
	KS	AIC	0.084	-0.018	0.948	–	12
	KS	BIC	0.073	0.003	0.950	–	3
	KS	Fixed	0.079	0.008	0.944	–	10
	E-PS	GCV	0.130	0.040	0.948	184.0	126.5

B.1.3 Restricted Spatial Regression

Restricted spatial regression (RSR) Hanks et al. (2015) is a Bayesian approach where the unmeasured confounder is modeled by a spatial random effect, $\boldsymbol{\eta} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$. Under RSR, $\boldsymbol{\eta}$ is restricted to be orthogonal to the exposure and define $\tilde{\boldsymbol{\eta}} = (\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T})\boldsymbol{\eta}$ where $\boldsymbol{X} = [\mathbf{1}, \boldsymbol{x}]$. We assume $\boldsymbol{\Sigma}$ has an exponential covariance structure and use a Metropolis-Gibbs sampler to obtain the posterior means of both the conditional, β_x, and unconditional $\delta = \beta_x + (\boldsymbol{X}^\mathsf{T}\boldsymbol{X})^{-1}\boldsymbol{X}^\mathsf{T}\boldsymbol{\eta}$ estimates.

Table B.9: Root mean squared error (RMSE), estimated bias, and coverage rates (95% nominal) for restricted spatial regression (RSR) conditional and unconditional simulation estimates of β_x when $\phi_2 = 10$ and $\phi_3 = 100$.

			Conditional Estimator			Unconditional Estimator		
ϕ_1	σ_x	σ_y	RMSE	Bias	Coverage	RMSE	Bias	Coverage
5	0	1	0.661	0.658	0.000	0.650	0.629	0.004
	0.1	1	0.637	0.634	0.000	0.642	0.621	0.002
	2	1	0.032	0.028	0.620	0.129	0.119	0.008
	0.1	10	0.732	0.634	0.648	0.718	0.619	0.528
50	0	1	0.185	0.162	0.560	0.268	0.120	0.210
	0.1	1	0.198	0.175	0.530	0.279	0.207	0.196
	2	1	0.017	0.005	0.944	0.034	0.022	0.604
	0.1	10	0.548	0.181	0.954	0.537	0.189	0.928
150	0	1	0.108	0.062	0.880	0.175	0.079	0.433
	0.1	1	0.114	0.068	0.874	0.188	0.084	0.430
	2	1	0.016	0.002	0.946	0.023	0.008	0.816
	0.1	10	0.552	0.064	0.974	0.533	0.068	0.952

B.2 Additional Data Analysis Results

To better understand jumps in point estimates seen in Figure 1 of the main text, we explored the spatial structure of specific splines. We fit the $PM_{2.5}$ concentrations for each census tract to a TPRS basis with 20 df. We obtained the residuals of the census tract averaged $PM_{2.5}$ values for each observation. Then, for each census tract, we averaged the residual $PM_{2.5}$ values for all births in the tract and created a chloropleth map of these values across Colorado (Figure B.1a). We repeated this procedure using 19 and 21 df. We followed the same procedure with BWGAZ, except we included individual-level covariates in the model (Figure B.1b). We also created a plot of the 21st spline in the TPRS basis across Colorado census tracts(Figure B.1c). In both plots, the darker shades represent more negative values and the lighter shades represent more positive values. We also computed the correlations between the mean residuals of $PM_{2.5}$ and BWGAZ modeled with TPRS basis with 19, 20, and 21 df and the 20th, 21st and 22nd spline in the TPRS basis, respectively (Table B.10).

Comparing Figures B.1a and B.1b with Figure B.1c, we see similar patterning, but in the inverse coloring, for Figures B.1a and B.1c. For example, in the northwest corner of Colorado, the shading for the residual $PM_{2.5}$ (Figure B.1a) indicate slightly positive values, and the corresponding census tracts shaded by the 21st spline (Figure B.1c) indicate negative values of greater magnitude. Thus, the 21st spline is negatively correlated with the residual $PM_{2.5}$ values (Table B.10). We also note that the residual variation for models that include 20 df in the TPRS for both BWGAZ and $PM_{2.5}$ is more correlated with the 21st spline than it is for other splines (Table B.10). This matching of spatial patterning and higher correlation creates a model that more accurately models the underlying spatial trends across the state with the addition of a single spline; thus, creating a jump in point estimates between the 20th and 21st spline.

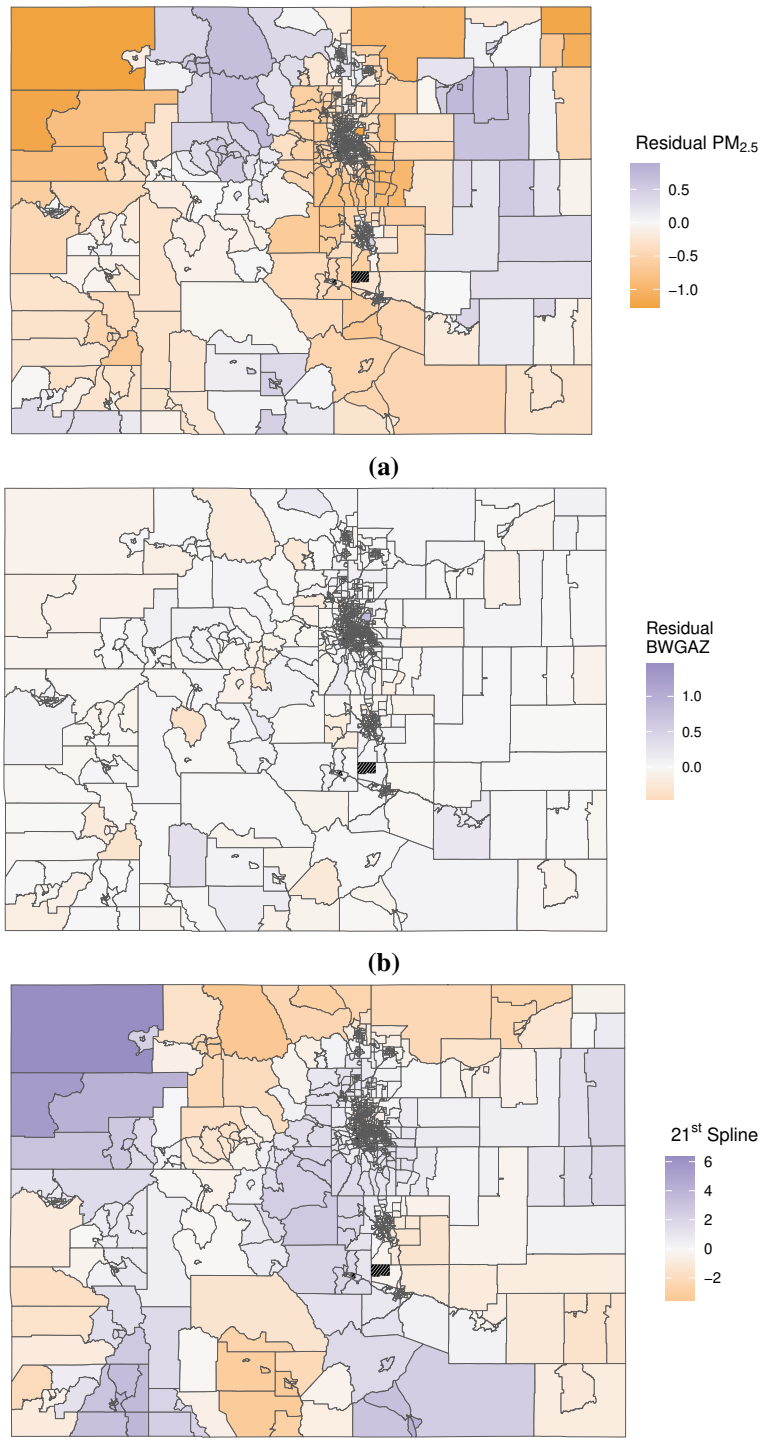


Figure B.1: Map of Colorado shaded by (a) residuals of modeling $PM_{2.5}$ as a function of a TPRS basis with 20 splines, averaged at the census tract, (b) residuals of modeling BWGAZ as a function of the additional individual-level covariates described in Section 2.4 and TPRS basis with 20 splines, averaged at the census tract, and (c) the 21st TPRS basis spline. Tracts not included in our study are hashed out.

Table B.10: Correlation between residuals and next spline in TPRS basis.

Model	Next Spline in Basis		
Outcome	20	21	22
$PM_{2.5}$	0.123	-0.637	0.237
BWGAZ	-0.072	0.100	0.006

Appendix C

Supplemental Material for Chapter 4

C.1 Temporal Trends Equations

Table C.1: Functions used to create temporal trends used in simulation study.

Trend	Shape	Study	Equation
$b_1(t)$	Linear	Longitudinal	$b_1(t) = -0.25(t-5)$
		Survival	$b_1(t) = -0.23(t-5.5)$
$b_2(t)$	Sinusoidal	Both	$b_2(t) = 0.25\text{sin}(\frac{2\pi}{7}t)$
$b_3(t)$	Sinusoidal	Both	$b_3(t) = 0.25\text{sin}(\frac{2\pi}{1.25}t)$
$b_4(t)$	Sinusoidal	Both	$b_4(t) = 0.25\text{sin}(\frac{2\pi}{7}t-1)$

C.2 Longitudinal Mixed Model Simulation Results

Table C.2: Root mean squared error (RMSE), estimated bias, coverage rates for 95% confidence intervals for the estimates of β_x in the mixed model simulations with a time-constant exposure where the within person correlation is high ($\sigma_a = 1.5$). The median degrees of freedom (df) for each method is reported for the respective basis used in each model.

Method	df Selector	Subset	Bias	RMSE	Coverage (95% Nominal)	df Selected (Spatial)
Unadjusted	–	–	0.4894	0.5939	0.180	–
Preselect	Fixed	–	0.0614	0.1824	0.780	5
Preselect	Fixed	–	0.0111	0.1380	0.950	10
Preselect	Fixed	–	-0.0111	0.2200	0.955	50
One Step	AIC	–	-0.0030	0.1492	0.930	9
One Step	BIC	–	0.0395	0.1663	0.795	4
Proposed	AIC	–	-0.0763	0.1843	0.935	24.5
Proposed	BIC	–	-0.0448	0.1432	0.895	7
Spatial+	Fixed	–	-0.0384	0.2839	0.945	100
Spatial+	Fixed	Mean	-0.0384	0.2839	0.945	100
Spatial+	Fixed	Base	-0.0384	0.2839	0.945	100
Spatial+	GCV	–	-0.0228	0.2818	0.945	31.8
Spatial+	AIC	–	-0.0383	0.2840	0.945	100
Spatial+	AIC	Mean	-0.0367	0.2801	0.950	100
Spatial+	AIC	Base	-0.0367	0.2801	0.950	100
Spatial+	BIC	–	-0.0384	0.2830	0.950	100
Spatial+	BIC	Mean	-0.0123	0.2227	0.955	46
Spatial+	BIC	Base	-0.0123	0.2227	0.955	46

Table C.3: Root mean squared error (RMSE), estimated bias, coverage rates for 95% confidence intervals for the estimates of β_x in the mixed model simulations with a time-constant exposure where the within person correlation is low ($\sigma_a = 0.5$). The median degrees of freedom (df) for each method is reported for the respective basis used in each model.

Method	df Selector	Subset	Bias	RMSE	Coverage (95% Nominal)	df Selected (Spatial)
Unadjusted	–	–	0.4943	0.5953	0.090	–
Preselect	Fixed	–	0.0717	0.1643	0.525	5
Preselect	Fixed	–	0.0202	0.0843	0.825	10
Preselect	Fixed	–	-0.0059	0.0849	0.960	50
One Step	AIC	–	-0.0008	0.0723	0.935	18
One Step	BIC	–	0.0165	0.0909	0.765	8
Proposed	AIC	–	-0.0411	0.1045	0.930	65
Proposed	BIC	–	-0.0269	0.0761	0.900	15.5
Spatial+	Fixed	–	-0.0140	0.1135	0.935	100
Spatial+	Fixed	Mean	-0.0140	0.1135	0.935	100
Spatial+	Fixed	Base	-0.0140	0.1135	0.935	100
Spatial+	GCV	–	-0.0051	0.1126	0.940	64.2
Spatial+	AIC	–	-0.0139	0.1135	0.935	100
Spatial+	AIC	Mean	-0.0133	0.1118	0.940	100
Spatial+	AIC	Base	-0.0133	0.1118	0.940	100
Spatial+	BIC	–	-0.0139	0.1129	0.940	100
Spatial+	BIC	Mean	-0.0068	0.0881	0.965	46
Spatial+	BIC	Base	-0.0068	0.0881	0.965	46

Table C.4: Root mean squared error (RMSE), estimated bias, coverage rates for 95% confidence intervals for the estimates of β_x in the mixed model simulations with a time-varying exposure where the within person correlation is high ($\sigma_a = 1.5$). The median degrees of freedom (df) for each method is reported for the respective basis used in each model.

Method	df Selector	Basis	Subset	Bias	RMSE	Coverage (95% Nominal)	df Selected (Spatial)	(Temporal)
Unadjusted	–	–	–	0.7942	0.8056	0.000	–	–
Preselect	Fixed	S	–	0.4385	0.6271	0.080	5	–
Preselect	Fixed	S	–	0.4547	0.6544	0.070	10	–
Preselect	Fixed	S	–	0.4774	0.6950	0.075	50	–
Preselect	Fixed	ST	–	0.8211	0.8360	0.005		5
Preselect	Fixed	ST	–	0.1539	0.5309	0.180		10
Preselect	Fixed	ST	–	0.0369	0.1651	0.615		50
Preselect	Fixed	S&T	–	0.5206	0.6078	0.100	5	5
Preselect	Fixed	S&T	–	0.5407	0.6275	0.095	10	10
Preselect	Fixed	S&T	–	0.5657	0.6543	0.090	50	10
One Step	AIC	S	–	0.4607	0.6666	0.075	8.5	–
One Step	BIC	S	–	0.4363	0.6285	0.075	3	–
One Step	AIC	ST	–	0.0218	0.0930	0.900		95
One Step	BIC	ST	–	0.0411	0.1786	0.665		48
One Step	AIC	S&T	–	0.5420	0.6334	0.085	8	7
One Step	BIC	S&T	–	0.5176	0.6064	0.090	3	5
Proposed	AIC	S	–	0.4607	0.6698	0.075	19	–
Proposed	BIC	S	–	0.4367	0.6313	0.075	5	–
Proposed	AIC	ST	–	0.0163	0.0920	0.905		95
Proposed	BIC	ST	–	-0.0162	0.1352	0.735		65
Proposed	AIC	S&T	–	0.5437	0.6343	0.100	19	7
Proposed	BIC	S&T	–	0.5171	0.6068	0.095	5	5

Table C.5: Root mean squared error (RMSE), estimated bias, coverage rates for 95% confidence intervals for the estimates of β_x in the mixed model simulations with a time-varying exposure where the within person correlation is high ($\sigma_a = 1.5$). The median degrees of freedom (df) for each method is reported for the respective basis used in each model (continued).

Method	df Selector	Basis	Subset	Bias	RMSE	Coverage (95% Nominal)	df Selected (Spatial)	df Selected (Temporal)
Spatial+	Fixed	S	–	0.4870	0.7109	0.075	100	–
Spatial+	Fixed	S	Mean	0.4842	0.7082	0.080	100	–
Spatial+	Fixed	S	Base	0.4842	0.7082	0.080	100	–
Spatial+	GCV	S	–	0.4905	0.7094	0.075	22.1	–
Spatial+	AIC	S	–	0.4860	0.7088	0.075	90	–
Spatial+	AIC	S	Mean	0.4843	0.7079	0.080	100	–
Spatial+	AIC	S	Base	0.4839	0.7071	0.080	95	–
Spatial+	BIC	S	–	0.4898	0.7018	0.070	27	–
Spatial+	BIC	S	Mean	0.4768	0.6968	0.075	43	–
Spatial+	BIC	S	Base	0.4740	0.6893	0.070	32	–
Spatial+	Fixed	ST	–	0.0425	0.1304	0.850		100
Spatial+	GCV	ST	–	-0.0053	0.0920	0.900		84.5
Spatial+	AIC	ST	–	0.0423	0.1304	0.850		100
Spatial+	BIC	ST	–	0.0429	0.1380	0.790		80
Spatial+	Fixed	S&T	–	0.5885	0.6759	0.075	100	10
Spatial+	GCV	S&T	–	0.5742	0.6628	0.080	21.7	9.2
Spatial+	AIC	S&T	–	0.5883	0.6757	0.075	100	9
Spatial+	BIC	S&T	–	0.5900	0.6772	0.075	32	4

C.3 Longitudinal GEE Model Simulation Results

Table C.6: Root mean squared error (RMSE), estimated bias, coverage rates for 95% confidence intervals for the estimates of β_x in the GEE model simulations with a time-constant exposure where the within person correlation is high ($\rho = 0.7$). The median degrees of freedom (df) for each method is reported for the respective basis used in each model.

Method	df Selector	Subset	Bias	RMSE	Coverage (95% Nominal)	df Selected (Spatial)
Unadjusted	–	–	0.5059	0.6099	0.110	–
Preselect	Fixed	–	0.0584	0.1740	0.600	5
Preselect	Fixed	–	0.0266	0.1114	0.840	10
Preselect	Fixed	–	0.0104	0.1243	0.925	50
One Step	QIC	–	0.0116	0.1032	0.870	15
Proposed	QIC	–	-0.0340	0.1196	0.930	49
Spatial+	Fixed	–	0.0107	0.1508	0.915	100
Spatial+	Fixed	Mean	0.0107	0.1508	0.915	100
Spatial+	Fixed	Base	0.0107	0.1508	0.915	100
Spatial+	QIC	–	0.0090	0.1496	0.895	95
Spatial+	QIC	Mean	0.0088	0.1502	0.920	100
Spatial+	QIC	Base	0.0088	0.1502	0.920	100

Table C.7: Root mean squared error (RMSE), estimated bias, coverage rates for 95% confidence intervals for the estimates of β_x in the GEE model simulations with a time-constant exposure where the within person correlation is low ($\rho = 0.2$). The median degrees of freedom (df) for each method is reported for the respective basis used in each model.

Method	df Selector	Subset	Bias	RMSE	Coverage (95% Nominal)	df Selected (Spatial)
Unadjusted	–	–	0.5043	0.6095	0.085	–
Preselect	Fixed	–	0.0584	0.1691	0.465	5
Preselect	Fixed	–	0.0256	0.0921	0.750	10
Preselect	Fixed	–	0.0041	0.0803	0.915	50
One Step	QIC	–	0.0064	0.0679	0.900	22
Proposed	QIC	–	-0.0182	0.0922	0.935	95
Spatial+	Fixed	–	0.0009	0.0973	0.935	100
Spatial+	Fixed	Mean	0.0009	0.0973	0.935	100
Spatial+	Fixed	Base	0.0009	0.0973	0.935	100
Spatial+	QIC	–	0.0004	0.0968	0.915	95
Spatial+	QIC	Mean	0.0003	0.0964	0.940	100
Spatial+	QIC	Base	0.0003	0.0964	0.940	100

Table C.8: Root mean squared error (RMSE), estimated bias, coverage rates for 95% confidence intervals for the estimates of β_x in the GEE model simulations with a time-varying exposure where the within person correlation is high ($\rho = 0.7$). The median degrees of freedom (df) for each method is reported for the respective basis used in each model.

Method	df Selector	Basis	Subset	Bias	RMSE	Coverage (95% Nominal)	df Selected (Spatial)	(Temporal)
Unadjusted	–	–	–	0.7631	0.7766	0.000	–	–
Preselect	Fixed	S	–	0.4135	0.5838	0.060	5	–
Preselect	Fixed	S	–	0.4386	0.6173	0.055	10	–
Preselect	Fixed	S	–	0.4742	0.6726	0.055	50	–
Preselect	Fixed	ST	–	0.7915	0.8085	0.005		5
Preselect	Fixed	ST	–	0.1715	0.4792	0.130		10
Preselect	Fixed	ST	–	0.0451	0.1644	0.410		50
Preselect	Fixed	S&T	–	0.5394	0.6071	0.040	5	5
Preselect	Fixed	S&T	–	0.5648	0.6295	0.030	10	10
Preselect	Fixed	S&T	–	0.5999	0.6635	0.030	50	10
One Step	QIC	S	–	0.4834	0.6880	0.050	100	–
One Step	QIC	ST	–	0.0176	0.0607	0.880		95
One Step	QIC	S&T	–	0.5899	0.6577	0.040	36.5	8

Table C.9: Root mean squared error (RMSE), estimated bias, coverage rates for 95% confidence intervals for the estimates of β_x in the GEE model simulations with a time-varying exposure where the within person correlation is high ($\rho = 0.7$). The median degrees of freedom (df) for each method is reported for the respective basis used in each model. (continued)

Method	df Selector	Basis	Subset	Bias	RMSE	Coverage (95% Nominal)	df Selected (Spatial)	df Selected (Temporal)
Proposed	QIC	S	–	0.4838	0.6885	0.050	100	–
Proposed	QIC	ST	–	0.0155	0.0606	0.870	100	
Proposed	QIC	S&T	–	0.6033	0.6686	0.040	100	9
Spatial+	Fixed	S	–	0.4850	0.6893	0.050	100	–
Spatial+	Fixed	S	Mean	0.4850	0.6893	0.050	100	–
Spatial+	Fixed	S	Base	0.4850	0.6893	0.050	100	–
Spatial+	QIC	S	–	0.4850	0.6893	0.050	100	–
Spatial+	QIC	S	Mean	0.4850	0.6893	0.050	100	–
Spatial+	QIC	S	Base	0.4843	0.6881	0.050	95	–
Spatial+	Fixed	ST	–	0.0167	0.0589	0.880	100	
Spatial+	QIC	ST	–	0.0089	0.0959	0.825	95	
Spatial+	Fixed	S&T	–	0.6098	0.6736	0.030	100	10
Spatial+	QIC	S&T	–	0.5986	0.6636	0.035	75	5

C.4 Survival Model Simulation Results

Table C.10: Root mean squared error (RMSE), estimated bias, coverage rates for 95% confidence intervals for the estimates of β_x in the PH model simulations with a time-constant exposure. The median degrees of freedom (df) for each method is reported for the respective basis used in each model.

Method	df Selector	Subset	Bias	Relative Bias	RMSE	Coverage (95% Nominal)	df Selected (Spatial)
Unadjusted	–	–	0.0797	0.5313	0.0999	0.37	–
Preselect	Fixed	–	0.0135	0.0897	0.0675	0.935	5
Preselect	Fixed	–	0.0038	0.0252	0.0732	0.950	10
Preselect	Fixed	–	0.0045	0.0303	0.1195	0.950	50
One Step	AIC	–	0.0112	0.0746	0.0706	0.890	4
One Step	BIC	–	0.0243	0.1622	0.0684	0.875	3
Proposed	AIC	–	-0.0090	-0.0601	0.0658	0.945	6
Proposed	BIC	–	0.0082	0.0549	0.0607	0.945	3
Spatial+	Fixed	–	0.0049	0.0329	0.1429	0.940	100
Spatial+	Fixed	Mean	0.0049	0.0329	0.1429	0.940	100
Spatial+	Fixed	Base	0.0049	0.0329	0.1429	0.940	100
Spatial+	AIC	–	0.0049	0.0329	0.1429	0.940	100
Spatial+	AIC	Mean	0.0050	0.0333	0.1430	0.940	100
Spatial+	AIC	Base	0.0050	0.0333	0.1430	0.940	100
Spatial+	BIC	–	0.0049	0.0329	0.1429	0.940	100
Spatial+	BIC	Mean	0.0054	0.0357	0.1443	0.920	100
Spatial+	BIC	Base	0.0054	0.0357	0.1443	0.920	100