UNIVERSITY OF CALIFORNIA,
IRVINE

Usage of Kernel Smoothing in Generalized Additive Models for Disease Mapping with Individual-level Point-referenced Data: Stratified Smoothers and Generalized Additive Mixed Models

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Statistics

by

Yannan Tang

Dissertation Committee:
Professor Daniel L. Gillen, Chair
Professor Michele Guindani
Professor Veronica M. Vieira
Professor Scott M. Bartell

2020

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ALGORITHMS

ACKNOWLEDGMENTS

I wish to thank my dissertation committee members, Professor Michele Guindani, Professor Veronica Vieira and Professor Scott Bartell for the dedicated time, efforts and thoughts on the research work and thesis preparation.

I would like to thank my advisor, Professor Daniel Gillen. I have learnt enormously from him about what a Ph.D. degree is, how research work should proceed and what a decent work should look like. He has been so supportive that I somehow believe, and still so, that I could turn to him for help anytime on anything, including but not limited to background knowledge, research ideas, funs to have, beers to enjoy and how to live a phenomenal life.

I am much grateful to our collaborators, Professor Veronica Vieira and Professor Scott Bartell, for their generous support, inspiring inputs and bright encouragements. Thanks to Professor Vieira, my dissertation work is supported by the NIH grant, NIEHS P42ES007381.

The department of statistics, UCI is a treasure to me, with no doubts. I always feel lucky to be a part of it. I got many advices and thoughts from the faculty members. The students here are the ones you call best friends. Sometimes I wonder how this magic department always manage to hire and admit these top-tier people.

Last but not least, I want to thank my parents, who have been supporting me for 29 years and surprisingly, there is no clear sign that they will stop doing that.

VITA

Yannan Tang

EDUCATION

University of California, Irvine — **Irvine, CA, USA**
Ph.D. in Statistics — *2015 - 2020*
Dissertation Advisor: Professor Daniel L. Gillen

The George Washington University — **Washington, DC, USA**
M.S. in Statistics — *2013 - 2015*

Tsinghua University — **Beijing, China**
B.E. in Civil Engineering — *2008 - 2012*

PUBLICATIONS

- **Tang Y.**, Vieira V., Bartell S. and Gillen D., "A Stratified Generalized Additive Model and Permutation Test for Temporal Heterogeneity of Smoothed Bivariate Spatial Effects" (revised and submitted to *Statistics in Medicine*)
- **Tang Y.**, Vieira V., Bartell S. and Gillen D., "Additive Mixed Models with Kernel Smoothers for Disease Mapping Using Individual-level Data" (Submitted)
- **Tang Y.**, Vieira V., Bartell S. and Gillen D., "Disease Mapping using Generalized Additive Mixed Models with Kernel Smoothers" (To submit)

COLLABORATIVE RESEARCH

- Spatio-temporal analysis of birth defects and infant morbidity in relation to air pollution using generalized additive models (GAM) in a geographic framework
  Grant Funding Number: P42ES007381, NIEHS Grant, NIH
  PI: Veronica Vieira, D.Sc., Professor of Public Health, UC Irvine
  Role: Research Assistant
- Leveraging external data for regulatory decision making using propensity scores, with application in label expansion for multiple medical devices
  Sponsor: Allergan plc
  Supervisor: Jingyuan Yang, Ph.D., Director, Biostatistics, Allergan plc

TEACHING EXPERIENCE

University of California, Irvine **Irvine, CA, USA**
Tutor for Ph.D. qualification exams *2017 - 2019*

University of California, Irvine **Irvine, CA, USA**
Teaching Assistant, Reader *2015 - 2019*

CONTRIBUTED PRESENTATIONS AT ACADEMIC MEETINGS

- 13th International Conference on Health Policy Statistics, "An Additive Linear Mixed-effects Model (ALMM) with Kernel Smoothers and a Permutation Test on Temporal Heterogeneity of Geospatial Risk Patterns" (San Diego, CA, USA; JAN 2020)
- Joint Statistical Meetings of the ASA, "Time-Stratified LOESS Smoothers for Estimating and Testing Temporal Heterogeneity in Spatial Risk Patterns" (Vancouver, Canada; JUL 2018)

PROFESSIONAL MEMBERSHIPS

- American Statistical Association (2017-present)
- International Chinese Statistical Association (2018-present)

DEPARTMENT SERVICE

Department of Statistics **UC Irvine**
Graduate Student Representative (elected) *2017 - 2018*

AWARDS

- Early Advancement Award, Department of Statistics, UC Irvine (2017)
- University Scholarship, The George Washington University (2014, 2015)

ABSTRACT OF THE DISSERTATION

Usage of Kernel Smoothing in Generalized Additive Models for Disease Mapping with Individual-level Point-referenced Data: Stratified Smoothers and Generalized Additive Mixed Models

By

Yannan Tang

Doctor of Philosophy in Statistics

University of California, Irvine, 2020

Professor Daniel L. Gillen, Chair

Epidemiologists frequently aim to quantify geospatial heterogeneity in disease occurrence to identify relevant hidden health disparities. With the growing prevalence of individual-level point-referenced data, generalized additive models (GAMs) are becoming increasingly popular to map geospatial disease risk patterns while adjusting for confounding effects when the study is a cross-sectional one with an exponential family response. In the meanwhile, local regression smoothers are frequently adopted for spatial effects estimation in GAM framework by researchers partially due to their intuitive ideas and adaptation to changing population density.

However, studies with records over a (potentially long) period of time, including those with repeated measurements on subjects, commonly come into play nowadays. For these studies, traditional GAMs could be problematic. Firstly, since data could be recorded over a period of time while spatial risk patterns should not be assumed to be invariant in many cases, statistical tools to access time-varying spatial effects are required. On the other hand, if the study is longitudinally designed, traditional GAMs could lead to incorrect inference due to their incapability of accomodating within-individual correlation.

This dissertation work sought to develop statistical methodologies to address these problems under the GAM framework with kernel smoothers, using local regression smoothers in particular. In Chapter 3, we proposed GAMs with stratified kernel smoothers that could be applied for time-specific spatial effects modeling. Based on the new class of GAMs, we further designed a hypothesis testing procedure to formally detect temporal heterogeneity of spatial effects. In Chapter 4 and 5, we incorporated random effects, as well as kernel smoothers, into GAM, resulting in a class of generalized additive mixed models (GAMMs) with kernel smoothers. We further elaborated the novel fitting and inference procedures for the proposed models.

Relevant empirical results showed the utility and advantages in model fitting under some fairly designed scenarios, with comparison to classic models. We further applied our proposed methods in a study on birth defects in Massachusetts in Chapter 3 and a study on residents' serum PFOA concentration in Lubeck, WV, and Little Hocking, OH region.

Chapter 1

Introduction

1.1 Disease mapping with individual-level point-referenced data in epidemiology studies

In epidemiology studies, geospatial disparities of certain disease risks are of common interest since heterogenous risks over geographic areas may indicate location-related health disparities and/or risk factors. These disparities or factors may be environmental, demographic, socioeconomic in nature. In plain language, when investigating a specific disease, epidemiologists frequently aim to identify areas where residents are more likely to develop the disease. Based on the identified areas with high risk, it would be more probable to investigate the underlying risk factors that are associated with occurrence rate of the disease by exploring the difference in potentially relevant factors between high and low risk areas. Once one or more factors are identified, corresponding actions, such as environmental treatment or policy modification, would be possible. For instance, Bristow et al. (2015) conducted a spatial analysis on advanced-stage ovarian cancer mortality in California and found significant geospatial disparity in mortality rates. Based on these findings they were able to further

identify the receipt of NCCN (National Comprehensive Cancer Network) guideline adherent care and treatment at an HVH (High Volume Hospital) as potential explanations for the observed spatial patterns. The end result of this investigation is the identification of risk factors that impact mortality in advanced-stage ovarian cancer that may be modified by broadening access to care and proposing that guidelines be instituted consistently across care facilities.

Partially due to the lack of high resolution data collection and insufficient computing power, traditional disease mapping commonly focuses on areal data where a specific area, such as a country, a state, or a county is treated as a single sampling unit. In this case, inference is made at the aggregate level rather than at the specific individual-level or at specific locations. While this type of analytic approach may yield meaningful inference it is subject to ecological bias and may not provide sufficient resolution for identifying spatially related health disparities and risk factors.

Individual-level datasets contain outcome and covariate information on specific individuals. In spatial analysis, point-referenced data, or point-level data indicates a data collection setting where items are observed at precise spots on a map. Along with advances in data storage and measurement techniques, point-referenced data are becoming increasingly prevalent. The work presented in this dissertation focuses on methodology for spatial analyses of individual-level point-referenced data. Since these data provide information on unique individuals and locations, inference with higher resolution is possible when compared to classic methods geared towards analyses of data at the aggregate level. In particular, with individual-level point-referenced data, spatial epidemiologist, along with statisticians, seek to provide model-based estimates of disease risk and inference at virtually all locations on a spatial map of interest rather than marginal inference over an entire area.

As such, statistical tools for individual-level point-referenced data are much needed. In general, nonparametric smoothing techniques, including freqentist and Bayesian approaches,

are used to estimate underlying spatial risk patterns in order to render inference on a given disease risk at virtually every location. Further, generalized additive models (GAMs) (Hastie and Tibshirani, 1990) with bivariate smoothers play an increasingly popular role when both geospatial and confounding effects exist and the distribution of the response variable is assumed to belong to the exponential family.

In the following sections we introduce two research studies that rely on the analysis of individual-level point-referenced data. We briefly present each, but note that further information can be found in Vieira et al. (2009) and Bristow et al. (2015).

1.2 Motivating examples

1.2.1 Birth defects study in Massachusetts

A fairly recent study of birth defects in the state of Massachusetts was conducted by Girguis et al. (2016). In the study, all recorded births in the Massachusetts Birth Defects Registry (MBDR) having cardiac, orofacial and neural tube defects from 2001 to 2009 were identified as cases and 1000 live births per year without defects were sampled as common controls. Among the recorded defects, one of the most common was patent ductus arteriosus (PDA). PDA is a cardiovascular birth defect in which abnormal blood flow occurs between two of the major arteries connected to the heart and is associated with high morbidity and mortality. Residential longitude and latitude were recorded for all observations as well as potential confounding variables including maternal age, adequacy of prenatal care, maternal race, maternal education level and number of siblings.

A primary goal of the MBDR study is to quantify geospatial risks for PDA with adjustment for known risk factors, thus allowing epidemiologists to further explore the underlying space-

related risk factors. Moreover, since data are collected over 9 years and the spatial risk pattern could possibly change over the years, statistical tools to estimate time-specific spatial risk pattern are in need, as well as a class of hypothesis tests that formally decide if the spatial risk patterns at each time significantly differ from each other.

1.2.2 Serum PFOA concentration study

Another recent spatial epidemiology study was conducted by Bartell et al. (2010) to investigate serum perfluorooctanoic acid (PFOA) concentration among residents in Lubeck, West Virginia and Little Hocking, Ohio. In this study, researchers aimed to understand the declining behavior of PFOA concentration after granular activated carbon filtration on the public water systems in 2007. By design, 200 residents were included and 6 blood samples were to collect from each resident from May 2007 to August 2008 so that a trend of PFOA concentration could be observed. Besides PFOA concentration, residents' information such as gender, age and recent water consumption type (public or bottled water) was recorded as well as precise residential location (recorded as longitude and latitude).

One of the objectives is to understand the geospatial distribution of residents' serum PFOA concentration in order to help identify potential latent space-confounded risk factors. However, the since this study is a longitudinal one where individuals get repeated measurements, the estimation of spatial effects should be achieved with adjustment of confounding variables as well as the within individual correlation.

1.3 Overview of this dissertation

In this chapter we briefly introduced the motivating examples that have led to the methodologic developments presented in the remainder of the dissertation. In Chapter 2, we present

a background of the statistical methodology on which our approaches are based. The covered statistical background includes frequentist and Bayesian smoothing techniques, generalized additive models (GAMs) and generalized linear mixed models. In Chapter 3, we propose stratified smoothers and incorporate these smoothers into a GAM framework. We further develop a class of permuted mean squared difference (PMSD) tests to detect temporal heterogeneity of geospatial effects. The methods are applied to the previously discussed data on birth defects in Massachusetts state. In Chapter 4, we generalize kernel smoothers using variance-covariance adjustment, describe a novel additive mixed models (AMMs) framework with kernel smoothers and further propose a new backfitting algorithm to fit AMMs that incorporate kernel smoothers. Chapter 5 provides an extension of Chapter 4, accommodating exponential family responses via a novel fitting algorithm that relies on a combination of penalized quasi-likelihood (PQL) and the fitting procedure introduced in Chapter 4. Both Chapter 4 and 5 present an application of the proposed methods to data on serum PFOA levels, identifying areas of high and low risk in Lubeck, WV and the Little Hocking, OH area. Chapter 6 covers relevant discussion and some insights on probable future research directions.

Chapter 2

Statistical background

2.1 Smoothers for disease mapping

Smoothing functions, also known as smoothers, are commonly used to explore the relationship between two or more variables when it is desirable to avoid assuming a parametric function for the relationship. Popular smoothing techniques include running-mean, running-line, K-nearest neighbors and kernel smoothers. Since flexible fitting and estimation are provided, smoothing functions are widely used when spatial risk patterns are to be investigated. Commonly used smoothers in the context of bivariate spatial smoothing include local regression (LOESS), regression splines and Gaussian process models.

In this section, we introduce and review methods to specify and estimate a smooth function f relating a univariate response, y_i to covariate(s) $\mathbf{x}_i$ (2.1). Note that $\mathbf{x}_i$ could be univariate or a multidimensional vector. When introducing smoothing techniques, we present the methods for univariate smoothers where $\mathbf{x}_i$ is univariate and further generalize them to bivariate

smoothing situations.

$$y_i = f(\mathbf{x}_i) + \epsilon_i, \quad i = 1, 2, \ldots, N, \tag{2.1}$$

where ϵ_i is a random error.

2.1.1 LOESS

Locally weighted regression smoothing (LOESS - LOcal regrESSion) follows the intuitive idea that the underlying curve can be assumed to be linear in a local region. Based on this idea, when estimation of a certain location is desired, one can consider a weighted average of the response among k neighboring observations. Using the k nearest neighbors, instead of a simple linear model, a weighted linear regression is adopted so that closer observations render more weight in the fitting process. The procedure for finding the LOESS estimate of the smoothing function $\hat{f}(x_0)$ at location x_0, when utilizing the tri-cube weight function, is shown in Algorithm 1.

Algorithm 1 LOESS fitting procedure

Identify k nearest neighbors of x_0, denoted by $N_b(x_0)$.
Find the greatest distance among the neighbors, $\Delta(x_0) = \max_{x_j \in N_b(x_0)} |x_j - x_0|$.
Assign weights to the neighbors using the tri-cube weight function $W(\frac{|x_j - x_0|}{\Delta(x_0)})$ where

$$W(z) = \begin{cases} (1 - z^3)^3, & \text{for} 0 \leq z < 1; \\ 0 & \text{otherwise.} \end{cases} \tag{2.2}$$

Achieve the fitted value $\hat{f}(x_0)$ from the weighted least-squares fit of y to x confined to $N(x_0)$ using the tri-cube weights $W(\frac{|x_j - x_0|}{\Delta(x_0)})$.

If bivariate smoothing is of interest, as in the case of spatial smoothing for disease risk, the above LOESS procedure can be easily generalized. In this case, when defining the nearest k observations, Euclidean distances between every pair of points is commonly used. Analogous

to the univariate case, local weighted linear regressions using the two covariates where weights are decided by tri-cube weight function are generally employed.

For a simple example of univariate smoothing problem, we aim to estimate y as a function of x, where a scatter plot is shown in Figure 2.1. Relationship between the 2 variables does not appear to be linear and a more flexible fitting would be required to achieve a reasonable estimation of the function $\mathbb{E}(Y) = f(x)$.

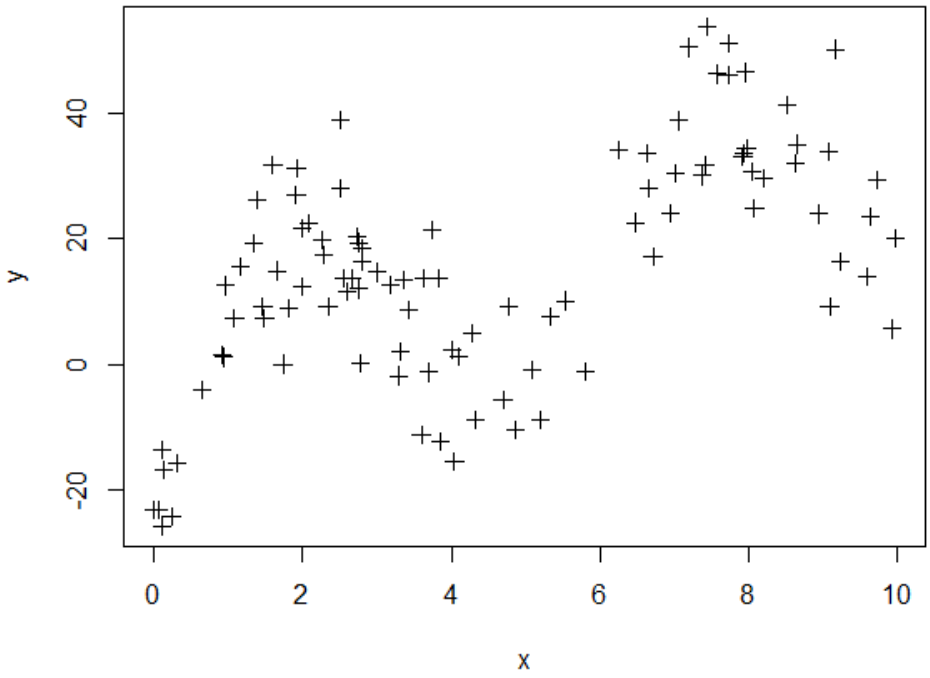

Figure 2.1: Simulated data for smoother illustration.

If LOESS approach is adopted, nearest neighbors of every single estimand value of x. The number of neighbor observations depends on a smoothing parameter span, which is defined as the proportion of total data that are used for the local weighted linear models. In Figure 2.2, 2 examples of span sizes were illustrated while the fitted curves of these 2 span sizes, along with an AIC optimally selected span fitting are shown in

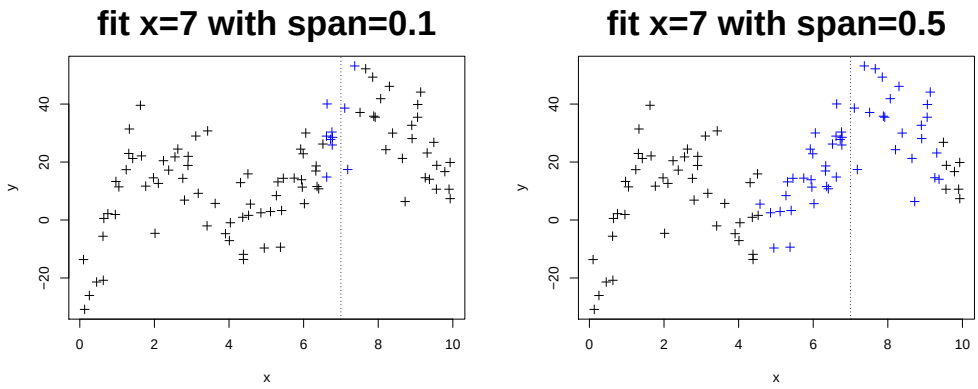

Figure 2.2: Neighborhood selection at $x = 7$ with span size 0.1 and 0.5. Blue are the neighbor observations that are used in the local weighted linear model for fitting of the extimand $x = 7$.

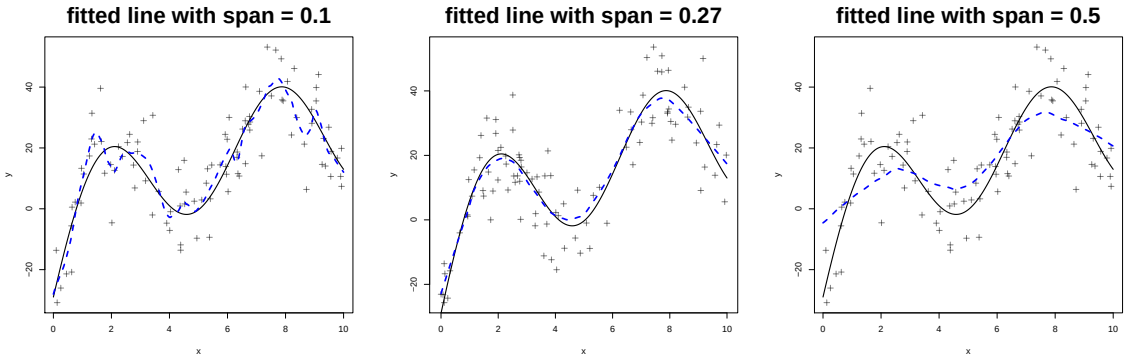

Figure 2.3: Fitted lines (blue) with span size 0.1 (left), 0.27 (middle) and 0.5 (right). Black lines are the true curve that is used in simulation. Span=0.27 results in minimal AIC value hence 0.27 provides the "best" fit of the data, which agrees with visual judgment for most.

2.1.2 Basis expansion methods

Basis expansion smoothers assume that the relationship f associating $\mathbf{x}$ to response y can be expressed as

$$f(\mathbf{x}) = \sum_{m=1}^{M} \beta_m h_m(\mathbf{x}), \tag{2.3}$$

where $h_m(\mathbf{x}), m = 1, \ldots, M$ are transformations defined on $\mathbb{R}^p \to \mathbb{R}$ and p stands for the dimension of $\mathbf{x}$. Hence $p = 1$ results in univariate smoothing while $p = 2$ will render bivariate

smoothing.

As a simple example, polynomial expansions for univariate $\mathbf{x}$ use basis functions defined as

$$h_m(x) = x^m, m = 1, 2, \ldots, d, \tag{2.4}$$

where d is the chosen degree of polynomial.

Another class of basis expansion smoothers follows the spirit of piecewise regressions. While the whole curve is flexible where no parametric form is proper, this class of smoothers assumes a parametric relationship between response and explanatory variables such as linear or polynomial patterns. This sounds similar to LOESS smoothers, however the difference is that piecewise regressions place knots on the support of the explanatory variables so that local regressions are restricted to the generated intervals. In contrast LOESS smoothers use the nearest k observations and hence the local intervals used for regressions depend on where the nearest neighbors are.

A widely used type of univariate piecewise regression is the natural cubic regression spline smoother. Assume knots are placed at $\xi_1 < \xi_2 < \cdots < \xi_K$. Natural cubic regression splines use cubic regressions between ξ_i and ξ_{i+1}, $i = 1, 2, \ldots, K-1$. In addition, to achieve smooth fitting, continuity of the underlying function, as well as continuous first and second derivatives, is required. With this specification, it can be shown that a the induced basis expansion can be expressed as

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 h(x, \xi_1) + \cdots + \beta_{K+3} h(x, \xi_K) \tag{2.5}$$

where

$$h(x,\xi) = \begin{cases} (x-\xi)^3, & \text{for } x > \xi; \\ 0 & \text{otherwise.} \end{cases} \tag{2.6}$$

This smoother is called a cubic regression spline. One arising issue is that since the edge of the support has data on only one side and observations are commonly sparse at edges, cubic regressions can be vulnerable to overfitting. To address this issue, one can force the function to be linear at the boundaries, i.e. for $x < \xi_1$ and $x > \xi_K$. This constraints transforms cubic regression splines to natural cubic regression splines and renders a model shown in (2.7).

$$f(x) = \beta_0 + \beta_1 x + \beta_2 h(x,\xi_1) + \cdots + \beta_K h(x,\xi_{K-1}) + \beta_{K+1}(x-\xi_K)^+ \tag{2.7}$$

Natural cubic regression splines provide an elegant way to model an unknown relationship when researchers have a good sense of where the knots should be placed. However, the location of knots is not trivial to decide in most cases. Model selection with respect to knots is subjective since both number and locations of knots need to be selected, resulting in a generally large search space for optimal not selection. To simplify the search space, evenly spaced knots are generally adopted. Obviously if few knots are used, one cannot achieve enough smoothing for the underlying functions. Conversely, if more knots than needed are specified, overfitting is a problem.

To avoid issues that arise with specification of the number and placement of knots, penalized regression splines were proposed by Wahba (1980). In this case, a knot is specified at every observed location in the support of the predictor covariate. Obviously this would lead to overfitting, so a term to penalize the wiggliness of the resulting function is introduced. More concretely, penalized regressions splines seek to find the smoothing function $f(x)$ that

minimizes

$$\sum_{i=1}^{N}\{y_i - f(x_i)\}^2 + \lambda \int f''(x)^2 dx \tag{2.8}$$

where λ is tunable parameter for smoothness controlling. This renders a currently popular method for univariate smoothing with smoothness controlling.

Further, when we want to smooth with respect to two or more input. For instance, in many cases, we want to find a smooth function $f(x_1, x_2)$ instead of $f(x)$. Knots placing could be less attractive since sparsity of data could become a serious problem when dimension grows. Especially for geographical smoothing, observations on map are generally not uniformly distributed hence if uniform grid is placed, there could be many small areas with no data in there. Observed subjects in epidemiological studies are frequently people and people do not live uniformly on map. Consequently, using knots-based basis expansion could suffer from over parametrization. Thin plate regression splines address this problem in a basis expansion framework (Wood, 2003).

Similar to the idea of natural cubic regression splines that minimize Eq. (2.8), bivariate thin plate regression spline smoothers seek to minimize

$$\sum_{i=1}^{N}\{y_i - f(\mathbf{x}_i)\}^2 + \lambda \int\int (\frac{\partial^2 f}{\partial x_1^2}) + 2(\frac{\partial^2 f}{\partial x_1 \partial x_2})^2 + (\frac{\partial^2 f}{\partial x_2^2}) dx_1 dx_2. \tag{2.9}$$

Similarly, λ acts as a tuning parameter that controls the smoothness of fitting and out-of-sample predictive criteria such as AIC, BIC, or GCV (Golub et al. (1979)) can be applied for selection of λ. An advantage of thin plate regression splines is that they offer a smoothing option without specification of a fixed set of knots. In addition, Wood (2017, Chapter 5.5.1) discussed methods to truncate the dimension of basis functions as an effort to reduce then computational burden when data grows big.

In spatial epidemiologic studies, if exact locations of the observations are known, thin plate regression splines are attractive options given their advantage over knot-based methods. Moreover, if the map of interest is not regularly shaped or there is a strong belief that edge effects may exist, soap film regression splines are potentially preferable due to their advantage in edge effects controlling. (Wood et al., 2008).

Here we show the fitting of natural cubic splines to the data presented in Figure 2.1 as an example. By putting knots at $x = 2, 5, 8$, the fitted curve is shown in Figure 2.4.

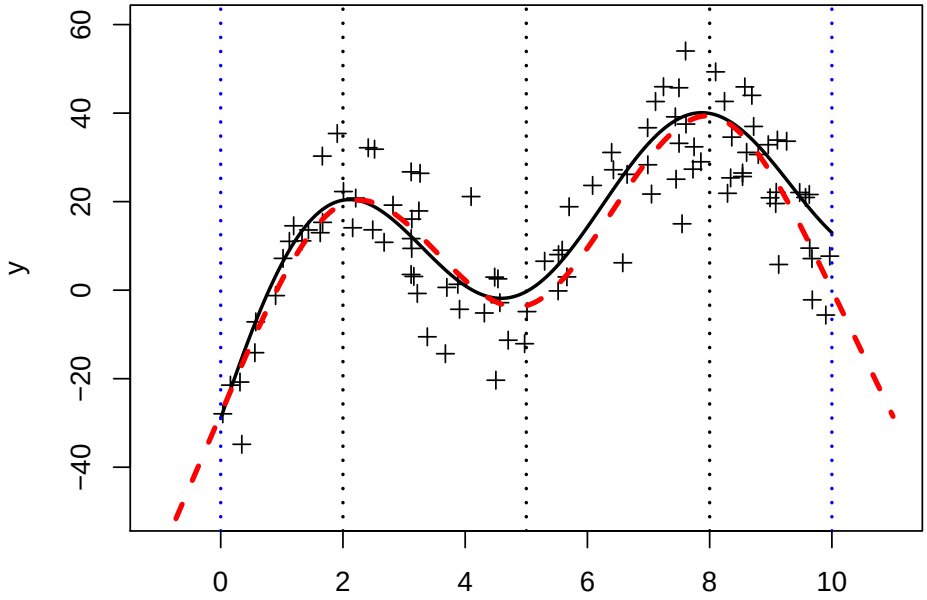

Figure 2.4: Fitted curve using natural cubic splines (red) with knots $x = 2, 5, 8$, with $x = 0, 10$ as the boundaries. Black solid line indicates the truth for simulation.

2.1.3 Smoothers using Gaussian processes

A Gaussian process (GP) is a stochastic process indexed by time or space. Consider a GP on domain $\mathbf{x}$. Then the GP is a distribution of function $f(\mathbf{x})$. That is to say, every realization of this GP will be a function $f(\mathbf{x})$. The process is called Gaussian since given any finite collection of $\{\mathbf{x}_i, i = 1, \ldots, n\}$, the distribution of $\{f(\mathbf{x}_i), i = 1, \ldots, n\}$ is joint Gaussian where the parameters, mean and variance of the Gaussian distribution, are defined by the

parameters of the GP.

To fully specify a Gaussian process, it is necessary to specify the mean and variance function in order to induce the parameters for the joint Gaussian distribution. Consider a Gaussian process defined on the vector space of $\{\mathbf{x}\}$. In formula 2.11 and 2.12, $m(\mathbf{x})$ and $k(\mathbf{x}, \mathbf{x}')$ stand for mean and covariance functions, respectively. Every realization of a Gaussian process is continuous since the correlation coefficient between $f(\mathbf{x})$ and $f(\mathbf{x}')$ equals to 1 if $\mathbf{x} = \mathbf{x}'$. Since shape of the generated curves is not restricted to any specific patterns, Gaussian processes could be promising smoothers. In practice, $m(\mathbf{x})$ is usually set to be 0 and squared-exponential covariance function shown in Formula 2.13 is a commonly used.

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')) \tag{2.10}$$

$$m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})] \tag{2.11}$$

$$k(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(f(\mathbf{x}) - f(\mathbf{x}))(f(\mathbf{x}') - f(\mathbf{x}'))] \tag{2.12}$$

$$k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp(-l^2 |\, \mathbf{x} - \mathbf{x}' \,|^2) \tag{2.13}$$

The definition of GP leads to a basic understanding of Gaussian processes as following:

1. A Gaussian process has infinitely many dimensions.
2. Only mean and covariance rule are specified. In most cases, the closer two inputs are, the higher correlation they have.
3. Every realization of a Gaussian process is a curve (surface). This result is drawn directly from the fact that correlation coefficient of two inputs $\mathbf{x}$ and $\mathbf{x}'$ goes to 1 as the distance between them goes to 0.

For estimation, the GP model can be specified as

$$\begin{bmatrix} f \\ f^\star \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K(X,X) & K(X,X^\star) \\ K(X^\star,X) & K(X^\star,X^\star) \end{bmatrix}\right), \tag{2.14}$$

where f are training outputs while $f^\star$ stand for test outputs. Also, X stands for the observed inputs and $X^\star$ is the collection of test inputs.

The GP model is attractive in terms of the fact that on any subset of the whole support, f follows a multivariate Gaussian distribution, allowing users to lean on vast experience with multivariate normal random variables. This is a primary reason why GP is preferred over other stochastic process models.

In practice, it is frequently assumed that instead of an exact GP, the observable outcome is an additive combination of random errors and a GP.

$$\begin{bmatrix} \mathbf{y} \\ f^\star \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K(X,X)+\sigma_n^2 I & K(X,X^\star) \\ K(X^\star,X) & K(X^\star,X^\star) \end{bmatrix}\right). \tag{2.15}$$

From this setting, we draw inference on $f^\star$ by noting that

$$f^\star|X,\mathbf{y},X^\star \sim \mathcal{N}(\bar{f^\star}, \mathrm{cov}(f^\star)), \tag{2.16}$$

where

$$\bar{f^\star} = E[f^\star|X,Y,X^\star] = K(X^\star,X)[K(X,X)+\sigma_n^2 I]^{-1}\mathbf{y} \tag{2.17}$$

$$\mathrm{cov}(f^\star) = K(X^\star,X^\star) - K(X^\star,X)[K(X,X)+\sigma_n^2 I]^{-1}K(X,X^\star). \tag{2.18}$$

Since properties of multivariate Gaussian distributions are well known, with properly specified priors for the hyperparameters in GP models, posterior distributions can be easily

derived. Markov chain Monte Carlo sampling could be further adopted to achieve posterior distributions of the hyperparameters as well as point estimation at explanatory variable locations of interest.

This idea is applicable in geospatial smoothing problems by applying two-dimensional vectors as inputs $\mathbf{x}_i$ to the GP models. Since we are interested in spatial epidemiologic studies with exact longitude and latitude of each location, a spatial GP could be defined as one with mean 0 and covariance rules defined in Formula 2.13 using $\mathbf{x} = (u, v)$, where (u, v) stands for longitude and latitude, respectively. Similarly, sampling strategies could be used to approximate the posterior distribution of model parameters as well as underlying spatial risk patterns.

2.2 Generalized additive models

In various studies, merely smoothing over space is generally not sufficient since factors other than space could also have effects on the response of interest or could confound the association between space and the response. Thus, to explore the spatial risk patterns researchers would appreciate a method that estimates the spatial effect with adjustment for those factors. For instance, if we want to explore the spatial pattern of survival rates as in the California ovarian cancer study, social-economic status should be considered in the analysis since it could be a potential confounding variable in the sense that social-economic status could be related to both spatial location and survival.

Generalized additive models, originally developed by Hastie and Tibshirani (1990), are designed to achieve this goal. Based on the linear terms in the mean model defined in Eq. (2.19), flexible functions are added, as is shown in Eq. (2.20), rendering a generalized additive models. In each case, $\mu_i = \mathbb{E}(y_i|\,\mathbf{x}_i)$, $g()$ denotes the link function, p is the number of

additive components and $s_k()$, $k = 1, \ldots, p$ is an arbitrary curve and is commonly defined to be smooth.

$$g(\mu_i) = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ik} \tag{2.19}$$

$$g(\mu_i) = \beta_0 + \sum_{k=1}^{p} s_k(\mathbf{x}_{ik}) \tag{2.20}$$

Maximum likelihood estimation in generalized linear models can be carried out via an iteratively reweighted least squares (IRLS) algorithm, which could be naturally extended to model fitting for generalized additive models when the flexible functions $s_k()$, $k = 1, \ldots, p$ are fully parametrized. Spline smoothers, such as cubic regression splines and thin plate regression splines, can be expressed using basis expansions and hence IRLS can be directly applied to GAMs with spline smoothers.

However, kernel smoothers, such as LOESS, could not be expressed by basis expansion hence the IRLS algorithm does not apply directly in this case. To carry out estimation, a backfitting algorithm (Breiman and Friedman, 1985) can be used instead. The idea of backfitting algorithm is to fit partial residuals iteratively on each additive component of the mean model until convergence. Using a GAM with continuous outcome specified in (2.21), Algorithm 2 is present as an example.

$$y_i = \beta_0 + \sum_{k=1}^{p} s_k(\mathbf{x}_{ik}) + \epsilon_i, \quad i = 1, 2, \ldots, N. \tag{2.21}$$

Algorithm 2 Backfitting Algorithm

Initialize $\hat{\beta}_0 = \frac{1}{N}\sum_{i=1}^{N} y_i$ and $\hat{s}_k = 0$ for all k.
while At least one of functions $\hat{s}_k$, $k = 1, \ldots, p$, does not converge **do**
    for k from 1 to p **do**
        Fit $\hat{s}_k$ using $\{y_i - \hat{\beta}_0 - \sum_{j \neq k} \hat{f}_j(\mathbf{x}_{ik}), i = 1, \ldots, N\}$ as response.
        Center $\hat{s}_k$ using $\hat{s}_k = \hat{s}_k - \frac{1}{N}\sum_{i=1}^{N} \hat{s}_k(\mathbf{x}_{ik})$.
    end for
end while

2.3 Generalized linear mixed models

2.3.1 linear mixed models

Linear models provide a fundamental approach to model relationship between a Gaussian distributed outcome variable and several explanatory variables via linear terms under the assumption that observations are independent within the dataset in use. A typical linear model is expressed as

$$y_i = x_i\beta + \epsilon_i, \tag{2.22}$$

where $\epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$.

However, when the dataset of interest includes multiple measurements on some individuals or the dataset is composed of multiple clusters, the independence assumption does not inherently hold since measurements on one particular individual or within one certain cluster should not be considered independent in many cases. Model (2.22) does not take the probable correlation into account, incorrect inference could be yielded when it is adopted to fit longitudinal or cluster data.

To account for within-individual correlation arising from longitudinal sampling of individuals

over time, Laird and Ware (1982) proposed a class linear mixed models (LMMs) given by

$$y_{ij} = x_{ij}\beta + z_{ij}b_i + \epsilon_{ij}, \tag{2.23}$$

where $b_i \overset{i.i.d.}{\sim} N(0, D)$ and $\epsilon_{\mathbf{i}} = (\epsilon_{i1}, \epsilon_{i2}, \ldots, \epsilon_{iJ_i})' \sim N(0, R)$, with b_i and ϵ_i independent. Typically, the fixed effects $x_{ij}\beta$ component of the linear predictor is used to model the scientific association of interest and adjust for potential confounding covariates, the random effects $z_{ij}b_i$ component is used to model individual-specific effects and the ϵ_{ij}'s are assumed to be i.i.d. conditional upon the random effects. LMMs are then fitted by a maximum likelihood (ML) or a restricted maximum likelihood (REML) procedure.

2.3.2 Generalized linear mixed models

LMMs concentrate on scenarios where the outcome variable follows a Gaussian distribution. However, outcomes in longitudinal studies could be more generally distributed. For example When the response follows a Bernoulli or Poisson distribution, LMMs do not apply instantly.

Analogous to the generalization from a linear model to a generalized linear model, Breslow and Clayton (1993) described a class of generalized linear mixed models (GLMMs) that accommodate correlation in a linear model by incorporating random effects to model non-Gaussian response.

Under a GLMM framework, the outcome y_i follows a distribution of exponential family with $E(y_i) = \mu_i$ and $\text{var}(y_i) = v(\mu_i)$, where function $v()$ depends on the specific distribution of y_i. The mean μ_i is linked to the linear predictor $x_i^T\beta$ by the link function $g()$. Hence the systematic component of a GLMM is commonly written as

$$g(\mu_{ij}) = x_{ij}\beta + z_{ij}b_i. \tag{2.24}$$

Breslow and Clayton (1993) raised 2 potential approximate inference approaches to fit a GLMM, penalized quasi-likelihood (PQL) and marginal quasi-likelihood (MQL). Both of the approaches use iterative procedures with iteratively updated working response and weights. Since our work in Chapter 5 would be based on PQL procedure, an introduction to PQL is provided here. To perform a PQL fitting procedure, Working response y_{ij}^w is defined as

$$y_{ij}^w = g(\hat{\mu}_{ij}) + (y_{ij} - \hat{\mu}_{ij})g'(\hat{\mu}_{ij}), \tag{2.25}$$

with reasonably initialized $\hat{\mu}_{ij}$. The working response vector $\mathbf{y}_{\hat{\mu}}^w$ could then be approximated by a Gaussian distribution

$$N[X\beta + Zb,\ g'(\hat{\mu})R_{\hat{\mu}}g'(\hat{\mu})], \tag{2.26}$$

where $R_{\hat{\mu}}$ is the variance-covariance matrix defined by the assumed outcome distribution given the estimated mean vector $\hat{\mu}$. It follows that a weighted linear mixed model

$$\mathbf{y}_{\hat{\mu},ij}^w = x_{ij}\beta + z_{ij}b_i + \epsilon_{ij} \tag{2.27}$$

with working diagonal weight matrix

$$\hat{W}_{\hat{\mu}} = R_{\hat{\mu}}^{-1}[g'(\hat{\mu})]^{-2} \tag{2.28}$$

could be used to model the working response $\mathbf{y}_{\hat{\mu}}^w$. The PQL estimating procedure iteratively fits weighted linear mixed model with updated working response $\mathbf{y}_{\hat{\mu}}^w$ and working weight matrix $\hat{W}_{\hat{\mu}}$ based on the updated $\hat{\mu}$ at each iteration until the difference in parameter estimations are sufficiently small.

Chapter 3

GAMs with stratified smoothers and PMSD tests

3.1 Introduction

Spatial differences in disease risk are potential indicators of space-related disease factors such as environmental exposures or availability of sufficient health care in certain areas. As such, quantification of heterogeneity in disease risk patterns over geographical space is of common interest in epidemiology studies.

While traditional geographic modeling methods focus on analyzing aggregated area-level data that treat area-defined partitions as one unit, more recent spatial epidemiology studies avoid aggregation bias and ecological fallacy by modeling individual-level data. With accurate records of geospatial information over a period of time, researchers seek to draw inference on both the existence of spatial effects on risks of disease as well as potential changes in spatial patterns of disease over time. As one example, Girguis et al. (2016) conducted a fairly recent study of birth defects in the state of Massachusetts. In the study, all recorded

births in the Massachusetts Birth Defects Registry (MBDR) having cardiac, orofacial and neural tube defects from 2001 to 2009 were identified as cases and 1000 live births per year without defects were sampled as common controls. Among the birth defects considered in the case definition, one of the most common was patent ductus arteriosus (PDA). PDA is a cardiovascular birth defect in which abnormal blood flow occurs between two of the major arteries connected to the heart and is associated with high morbidity and mortality. Residential longitude and latitude were recorded for all observations as well as potential confounding variables including maternal age, adequacy of prenatal care (measured by the Adequacy of Prenatal Care Utilization Index), maternal race, maternal education level and number of siblings.

A primary goal of the MBDR study is to quantify geospatial risks for PDA, after adjusting for known risk factors. It is not generally reasonable to assume an *a priori* parametric form on spatial effects, as spatial disease patterns are often complex and require flexible modeling techniques. Because of this, smoothers are commonly used in such settings. In spatial analyses, smoothers consider the underlying spatial risk pattern as a flexible bivariate function of (u, v), the longitude and latitude associated with a given response. Popular smoothers for spatial risk pattern estimation generally belong to two broad categories: kernel smoothers and spline smoothers. Hastie and Tibshirani (1990) introduced both categories and consider the incorporation of the smoothing techniques into regression-based methods otherwise known as generalized additive models (GAMs). Wood (2017) focused on spline smoothers and offered a full introduction to knot-based splines, smoothing splines and regression splines.

Popular smoothing functions for geographical analysis include local weighted scatterplot smoothing (LOESS) and thin-plate regression splines. LOESS was first proposed by Cleveland (1979) for flexible smoothing with moving weighted linear regressions. LOESS assumes local (weighted) linearity, resulting in flexible functional estimation for the whole domain. The method was further applied to geospatial analysis. (Brunsdon et al., 1996) One ad-

vantage of LOESS for spatial analyses is that it intuitively adapts to changing population density by varying the size of the smoothing neighborhood based on the local data density given a fixed span size, which is defined as the proportion of observations used for local regressions. On the other hand, for geospatial risk pattern estimation, thin-plate splines have been used by Duchon (1977) among others. Development of thin-plate regression splines (Wood, 2003) was further provided and well illustrated by Wood (2017).

Another widely used class of smoothing strategies considers the spatial effects as a realization of an underlying spatial stochastic process (or a random field). Among stochastic process strategies, spatial Kriging is one of the most popular in geospatial risk estimation. The terminology, history and general ideas of Kriging were illustrated by Cressie (1990) in a concise while comprehensive manner. Briefly, Kriging aims to seek the best linear unbiased prediction of an underlying function given observed data and a known covariance structure. In general, Kriging models do not specify the full distribution of the underlying process.(Stein, 2012; Cressie, 1992) Bayesian Kriging models have further been developed by merging prior information into Kriging models. Since a full likelihood specification is required for Bayesian parametric models, the distribution of the spatial stochastic process needs to be specified although uncertainty is incorporated through hyperparameters. Prior information on mean of the underlying process were discussed by Omre and others.(Omre, 1987; Omre and Halvorsen, 1989) Handcock and Stein (1993) further incorporated uncertainty into the covariance structure. Diggle and Ribeiro Jr (2002) formalized a comprehensive "model-based" geostatistics framework which explicitly specified a stochastic model along with corresponding model fitting strategies.(Diggle et al., 2003) A more recent and comprehensive work on Hierarchical Bayesian spatial modeling is provided by Banerjee et al. (2014) while a recent review by Gelfand and Banerjee (2017) covered a variety of topics on Bayesian geospatial data modeling in a succinct fashion.

In this manuscript, our methods are developed based on smoothers and generalized additive

models. Using any of the available flexible smoothers, spatial epidemiologists are able to fit flexible cross-sectional models that incorporate kernel-based or spline-based methods into the mean model of a generalized linear model. However, in epidemiology studies, data collected over a period of time are becoming increasingly available. Cross-sectional models do not suffice when geospatial risk patterns are heterogeneous over time. Epidemiologists have interest in estimating and comparing time-specific spatial risk patterns in order to better understand diseases and related factors. A formal test to determine if the spatial risk pattern changes over a period of time and when the change occurs would be attractive, as the test result would offer valuable clues to identifying factors that elevate or reduce the risk of adverse health outcomes. Although R package *mgcv* (Wood, 2017, 2003, 2011; Wood et al., 2016; Wood, 2004) offers stratified regression splines, which can be used to estimate time-specific geospatial risk patterns and a corresponding ANOVA F test, there is a dearth of easily implementable tools to estimate time-specific spatial risks in the GAM framework with kernel smoothers such as LOESS. Further, since the approximate ANOVA F test for significance of LOESS smoothers renders inflated type I errors, (Young et al., 2011) there is a lack of intuitive formal tests for temporal homogeneity of spatial risk patterns. Due to the popular usage of LOESS in epidemiology studies, in this paper, we aim to provide intuitive and effective methods to solve both of these problems using kernel smoothers with a focus on LOESS.

The remainder of the current manuscript is devoted to developing an extension of GAMs that incorporates time-stratified smoothers and an accompanying permutation-based testing procedure for assessing geospatial risk pattern changes over time. In Section 3.2, we introduce notation and describe our proposed time-stratified GAM and permutation testing procedure. In Section 3.3, we perform simulation studies to assess the operating characteristics of our proposed methods. In Section 3.4, we use our proposed procedure to test for temporal variation in the estimated spatial risk patterns of PDA using the MBDR data. Section 3.5 provides further discussion of the proposed work and considers avenues of future research.

3.2 Methods

3.2.1 Notation

We begin by introducing notation used throughout the remainder of the manuscript. Let $j = 1, \ldots, J$, denote a discrete time index and $i = 1, \ldots, n_j$, denote the observation index, indicating that there are n_j independent observations at time j. Specifically, in this study, we focus on studies where no repeated measurements are taken on the same subject so that independence among observations holds for the entire dataset. Thus, identical values of i at distinct time points j's do not refer to the same subject but simply an index of a unique subject. Let (u_{ij}, v_{ij}) denote the geographic location (longitude and latitude) of observation i at time j, and the function $s()$ denotes a general (bivariate) smoothing function to be applied over spatial location. Finally, X_{ij} denotes a q-vector of potentially time-dependent adjustment variables corresponding to observation i at time j.

3.2.2 GAMs with time-stratified smoothers

As it is increasingly prevalent that spatio-temporal occurrence of disease is routinely collected at the individual level, a common scientific goal is recently to determine if spatial patterns in disease vary over time, i.e., to determine if an interaction effect between time and space on disease risks exists. To address this problem, we consider the use of time-stratified smoothers.

We consider the case where time is discretized into multiple time points. At each time point, data including disease outcome, confounding variables and geographical locations of subjects are recorded. If spatial effects are homogeneous across time, one could reasonably employ a single smoother, $s(u, v)$, over space in order to model the mean of response y_{ij}, denoted as

μ_{ij}. In this case, researchers could use a model of the form

$$g(\mu_{ij}) = \beta_0 + s(u_{ij}, v_{ij}) + X_{ij}\boldsymbol{\beta}, \tag{3.1}$$

where data over all time points are pooled to estimate a single spatial risk pattern with adjustment for confounding factors. However, if spatial effects vary from one time point to another, one smoother for all observations is not sufficient to capture time-specific geospatial risk patterns. Instead, one smoother at every time point would be preferable. Thus we consider a class of GAMs with stratified bivariate spatial smoothers given by

$$g(\mu_{ij}) = \beta_0 + s_j(u_{ij}, v_{ij}) + X_{ij}\boldsymbol{\beta}, \tag{3.2}$$

where the function $s(u_{ij}, v_{ij})$ is now indexed by j. Importantly, the model given by (3.2) uses a common effect of X_{ij} over time, thereby borrowing information of confounding effects across all observed time points. A major difference between our modeling strategy and many other commonly seen space-time models, such as Gaussian processes with separable time-space correlation structures, is that by stratifying the smoothing function, no assumption is made on the temporal correlation of the geospatial risk. Our strategy may sacrifice precision when the varying mechanism of geospatial risk is well understood, however, as the environmental risk factors are frequently believed to be uncertain models that do not assume a specific form of temporal correlation will be able to estimate geospatial risk patterns at each time point without restricting how the pattern changes over time. Also, given sufficient data at each time point, the loss of efficiency due to stratification is somewhat negligible since data within each strata can render reasonably precise estimation of geospatial risk patterns.

To the best of our knowledge, no estimation procedures are currently available for fitting of stratified kernel smoothers given in (3.2). In this work, we generalize backfitting algorithm to fill this gap. For reference, we begin with the standard backfitting algorithm (Algorithm

3) utilized in the GAM framework, using continuous response with identity link function as an example. We modify the classic backfitting algorithm and propose Algorithm 4 which incorporates time-stratified LOESS smoothers. Specifically, instead of regressing on the partial residuals with a marginal bivariate smoother, we stratify the working data and fit time-specific smoothers and then combine the fitted values. For GAMs with kernel smoothers other than LOESS, the same procedure could be applied by replacing LOESS with the smoother of interest. For GAMs with spline smoothers, the backfitting algorithm is also valid but not as necessary since splines can be expressed as a basis expansion of the covariates. Thus, for splines, classic fitting procedures for parametric models are more computationally attractive in general.

Algorithm 3 Backfitting algorithm (continuous response)

1: Initialize $\hat{\beta}_0 = (\sum_{j=1}^{J} n_j)^{-1} \sum_{i,j} y_{ij}$, $\hat{lo}_{ij} = \hat{f}_{ij} = 0$ for all i, j.
2: ($\hat{lo}_{ij}$ will denote the fitted values of the bivariate spatial LOESS smoothers and $\hat{f}_{ij}$ will denote the fitted values of the parametric component $X_{ij}\beta$.)
3: **while** $|\text{SSR}_0 - \text{SSR}_1| > 10^{-8}\text{SSR}_0$, **do**
4: Set $\text{SSR}_0 = \text{SSR}_1$
5: Fit linear model: $(y_{ij} - \hat{lo}_{ij}) = \beta_0 + X_{ij}\beta + \epsilon_{ij}$ and get fitted values $\hat{f}_{ij}$ for all i, j.
6: Centralize the fitted values using $\hat{f}_{ij} = \hat{f}_{ij} - (\sum_{j=1}^{J} n_j)^{-1} \sum_{i,j} \hat{f}_{ij}$.
7: Fit LOESS smoother $(y_{ij} - \hat{f}_{ij}) = lo(u_{ij}, v_{ij}) + \epsilon_{ij}$ and get fitted values $\hat{lo}_{ij}$ for all i, j.
8: Centralize the fitted values using $\hat{lo}_{ij} = \hat{lo}_{ij} - (\sum_{j=1}^{J} n_j)^{-1} \sum_{i,j} \hat{lo}_{ij}$.
9: Calculate residuals $e_{ij} = y_{ij} - \hat{\beta}_0 - \hat{f}_{ij} - \hat{lo}_{ij}$.
10: Calculate sum of squared residuals $\text{SSR}_1 = \sum_{i,j} e_{ij}^2$.
11: **end while**

More generally, in the case where the distribution of the response is a member of the exponential family where the variance of response $V_{ij} = Var(y_{ij})$ may depend upon μ_{ij} and the assumed link function $g(\cdot)$, an iteratively reweighted least squares algorithm can be incorporated into backfitting Algorithm 3. (Hastie and Tibshirani, 1990) In a similar fashion, we generalize Algorithm 4 to accommodate exponential family outcomes by iteratively reweighting the proposed stratified smoother to obtain Algorithm 5. Note that the partial derivatives $\frac{\partial \eta_{ij}}{\partial \mu_{ij}}$ and the working variance V_{ij}^0 depend on the corresponding link function, $g(\cdot)$.

Algorithm 4 Backfitting algorithm for GAMs with a time-stratified LOESS smoother (continuous response)

Initialize $\hat{\beta}_0 = (\sum_{j=1}^{J} n_j)^{-1} \sum_{i,j} y_{ij}$, $\hat{lo}_{ij} = \hat{f}_{ij} = 0$ for all i, j.
Initialize $\text{SSR}_0 = 1$, $\text{SSR}_1 = 2$.
while $|\text{SSR}_0 - \text{SSR}_1| > 10^{-8}\text{SSR}_0$, **do**
    Set $\text{SSR}_0 = \text{SSR}_1$
    Fit linear model $(y_{ij} - \hat{lo}_{ij}) = X_{ij}\boldsymbol{\beta} + \epsilon_{ij}$ and get fitted values $\hat{f}_{ij}$ for all i, j.
    Centralize the fitted values using
    for j from 1 to J **do** $\hat{f}_{ij} = \hat{f}_{ij} - (\sum_{j=1}^{J} n_j)^{-1} \sum_{i,j} \hat{f}_{ij}$.
        Fit LOESS smoother $(y_{ij} - \hat{f}_{ij}) = lo(u_{ij}, v_{ij}) + \epsilon_{ij}$ at time j.
        Get fitted values $\hat{lo}_{ij}$ at time j for all i.
        Centralize the fitted values using $\hat{lo}_{ij} = \hat{lo}_{ij} - (\sum_{j=1}^{J} n_j)^{-1} \sum_{i,j} \hat{lo}_{ij}$.
        Calculate residuals $e_{ij} = y_{ij} - \hat{\beta}_0 - \hat{f}_{ij} - \hat{lo}_{ij}$.
        Calculate sum of squared residuals $\text{SSR}_1 = \sum_{i,j} e_{ij}^2$.
    end for
end while

3.2.3 The permuted mean squared difference (PMSD) test

In order to determine if geospatial risk patterns change over time, we consider a global test of temporal heterogeneity of spatial effects in a GAM scheme. More formally, in a simple case where there are two time points under investigation, $j = 1, 2$, we wish to test the null hypothesis

$$H_0 : s_1(u, v) = s_2(u, v), \text{ for all locations } (u, v) \text{ on the map,} \tag{3.3}$$

where $s_j(u, v)$ stands for the geospatial risk effect at time j and location (u, v).

To construct a measure of temporal heterogeneity in geospatial patterns, we consider a mean squared difference (MSD) statistic given by

$$MSD = \frac{1}{N_g} \sum_{g=1}^{N_g} (\hat{s}_1(u^{(g)}, v^{(g)}) - \hat{s}_2(u^{(g)}, v^{(g)}))^2, \tag{3.4}$$

where $\hat{s}_j(u^{(g)}, v^{(g)})$ stands for the estimated spatial effect at time j and location $(u^{(g)}, v^{(g)})$.

Algorithm 5 Backfitting algorithm for GAMs with a time-stratified LOESS smoother for exponential family responses (e.g. binary and counting responses)

1: Initialize: $\hat{\beta}_0 = g[(\sum_{j=1}^{J} n_j)^{-1} \sum_{i,j} y_{ij}]$; $\hat{lo}_{ij}^{0} = \hat{f}_{ij}^{0} = 0$.

2: Update: Construct an adjusted dependent variable

$$z_{ij} = \eta_{ij}^{0} + (y_{ij} - \mu_{ij}^{0})\Big(\frac{\partial \eta_{ij}}{\partial \mu_{ij}}\Big)_0$$

with $\eta_{ij}^{0} = \hat{\beta}_0 + \hat{lo}_{ij}^{0} + \hat{f}_{ij}^{0}$ and $\mu_{ij}^{0} = g^{-1}(\eta_{ij}^{0})$. Construct weights

$$w_{ij} = \Big(\frac{\partial \eta_{ij}}{\partial \mu_{ij}}\Big)_0^2 (V_{ij}^{0})^{-1}$$

3: Fit a weighted additive model with stratified smoothers

$$z_{ij} = \beta_0 + X_{ij}\beta + lo_j(u_{ij}, v_{ij}) + \epsilon_{ij}$$

with Algorithm 4 using weights w_{ij}, to get estimated functions $\hat{lo}_{ij}^{1}$ and $\hat{f}_{ij}^{1}$, additive predictor η^1, and fitted values μ_{ij}^{1}. Compute the convergence criterion

$$\Delta(\eta^1, \eta^0) = \frac{||\hat{lo}_{ij}^{1} - \hat{lo}_{ij}^{0}|| + ||\hat{f}_{ij}^{1} - \hat{f}_{ij}^{0}||}{||\hat{lo}_{ij}^{0}|| + ||\hat{f}_{ij}^{0}||}$$

A natural candidate for $||f||$ is $||\mathbf{f}||$, the length of the vector of evaluations of f at the n sample points.

4: Repeat step 2 and 3.

5: Replace η^0 by η^1 until $\Delta(\eta^1, \eta^0) < 10^{-8}\eta^0$.

The set of location points $\{(u^{(g)}, v^{(g)}), g = 1, \ldots, N_g\}$ define the points of interest for determining heterogeneity. For example, a dense and uniformly distributed grid on the entire map could be chosen as the set of evaluation points if no specific regions are believed to be time-varying a priori.

If spatial effects are homogeneous over time, we would expect MSD to be low. Conversely, since MSD is a measure of disparity in spatial patterns by definition, large MSD values would indicate temporal heterogeneity of spatial patterns.

To construct a reference distribution for the MSD statistic, we propose a permutation strategy. The reference distribution will be developed based on the assumed exchangeability of

time labels under H_0. In other words, if H_0 holds, i.e. spatial patterns do not change over time, the sampling distribution of spatial effects at each time point given by the model in (3.2) can be approximated by permuting time labels randomly. Consequently, we randomly permute time labels among the dataset for N_{perm} times to obtain a set of permuted MSD statistics (referred to henceforth as PMSD or permuted mean squared difference statistics) $\{\text{PMSD}_p, \quad p = 1, 2, \ldots, N_{perm}\}$. Then under H_0, the observed MSD (named OMSD) would be a regular member of PMSD_p's. In other words, *OMSD* and PMSD_p's would come from the same distribution if H_0 holds. On the other hand, when H_0 is violated, MSD will likely be greater than most PMSD_p values, leading to a permutation test that rejects H_0 when

$$\frac{1}{N_{perm}} \sum_{p=1}^{N_{perm}} I\{\text{PMSD}_p > \mathit{OMSD}\} < \alpha, \tag{3.5}$$

where α is the desired level of significance for testing H_0 and $I\{\}$ is an indicator function.

Here we summarize the permutation test for the stratified GAM model given in (3.2) with pseudo-code given by Algorithm 6. For context, the procedure we present here takes the smoothing function to be a LOESS smoother hence replace $s(u, v)$ with $lo(u, v)$ correspondingly.

3.2.4 Extension to Greater Than 2 Time Points

In the above, we have considered the PMSD test for comparing spatial effects over 2 time points. When more than 2 time points are available, we propose an extension of (3.4) for $J > 2$, where each stratified smoother is compared to a "grand mean" at each location over all time points. Specifically, we propose the statistic

$$\text{MSD} = \frac{1}{N_t N_g}(\sum_{j=1}^{N_t} \sum_{g=1}^{N_g} (\hat{s}_j(u^{(g)}, v^{(g)}) - \hat{s}_0(u^{(g)}, v^{(g)}))^2) \tag{3.6}$$

Algorithm 6 Permutation test for spatial heterogeneity with 2 time points

1: Fit (generalized) linear model $M_0 : Y_{ij} = \beta_0 + X_{ij}\boldsymbol{\beta}$.
2: Extract (working) residuals R_{ij} from M_0. (remove effects other than spatial effects from the response)
3: Split data by time and get time-stratified datasets D_1 and D_2. (Stratify data)
4: **for** j from 1 to 2, **do**
5: Fit $R_{ij} = lo(u_{ij}, v_{ij}) + \epsilon_{ij}$ using D_j.
6: Get prediction $\hat{lo}_j(u^{(g)}, v^{(g)})$. (grid prediction)
7: **end for**
8: Calculate *OMSD* using Eq. (3.4).
9: **for** *ip* from 1 to N_{perm} **do**
10: Randomly permute time labels t in dataset after Step 2. (Permute under 2 time points)
11: Repeat Step 3-7.
12: Calculate PMSD_{ip} using Eq. (3.4).
13: **end for**
14: Calculate p-value using Eq. (3.5).

where $\hat{s}_0(u^{(g)}, v^{(g)})$ represents a marginal smoother utilizing data pooled over all available sampling times. The permutation procedure in this case is a natural extension to the 2 time-point setting. Specifically, the time label t within the dataset is randomly shuffled and hence every observation within the dataset could potentially end up having any t value while the sample size at each time point is held unchanged.

3.2.5 Selection of locations for the MSD statistic

The MSD statistic defined in (3.6) is computed by taking the mean of the squared differences of the estimated spatial effects at a specified set of locations $\{(u^{(g)}, v^{(g)}), g = 1, \ldots, N_g\}$, which may or may not coincide with the observed locations $\{(u_{ij}, v_{ij})\}$. We previously claimed that if the temporal heterogeneity of the spatial pattern on the whole map is of interest, an intuitive and reasonable choice is to use a dense uniform grid on the entire map. Alternatively, evaluation points could be chosen to match the observed density of observations sampled over the map. In general, we recommend that users choose evaluation

points according to the scientific question of interest while accounting for the sampling scheme of the study. Specifically, a uniform grid provides equal weight over the entire map while using the set of observed locations will give higher weight to areas with denser observations, resulting in an estimand that is skewed towards more densely populated or more densely sampled areas. One could reasonably argue that either choice would be more or less scientifically important in specific settings.

Of course, another factor in the choice of grid type is the impact on the statistical properties (type I error and power) of the proposed PMSD test. In the Section 3.3, we will empirically investigate the impacts of grid density and location on the statistical properties of the PMSD test using simulation studies.

3.3 Monte Carlo Studies

3.3.1 Simulation study with underlying nonlinear risk patterns

To explore the operating characteristics of our proposed methods, we consider a variety of simulation studies based on a 2×2 square map. A nonlinear geospatial risk pattern, denoted by "truth: shift=0", is shown in Panel (a) of Figure 3.1. The pattern is created by Equation (3.7). We design temporal heterogeneous geospatial risk patterns by shifting the "shift=0" pattern to left by an amount up to 0.2 in order to achieve a metric of extremity of heterogeneity. Two examples of shifted patterns are shown in Panel (b) and (c) of Figure 3.1.

$$y = -u + 0.1\log(1.3)v + 1.2\sin(3(u + 0.1)) + 2uv + 6\log(0.6)v^2 \tag{3.7}$$

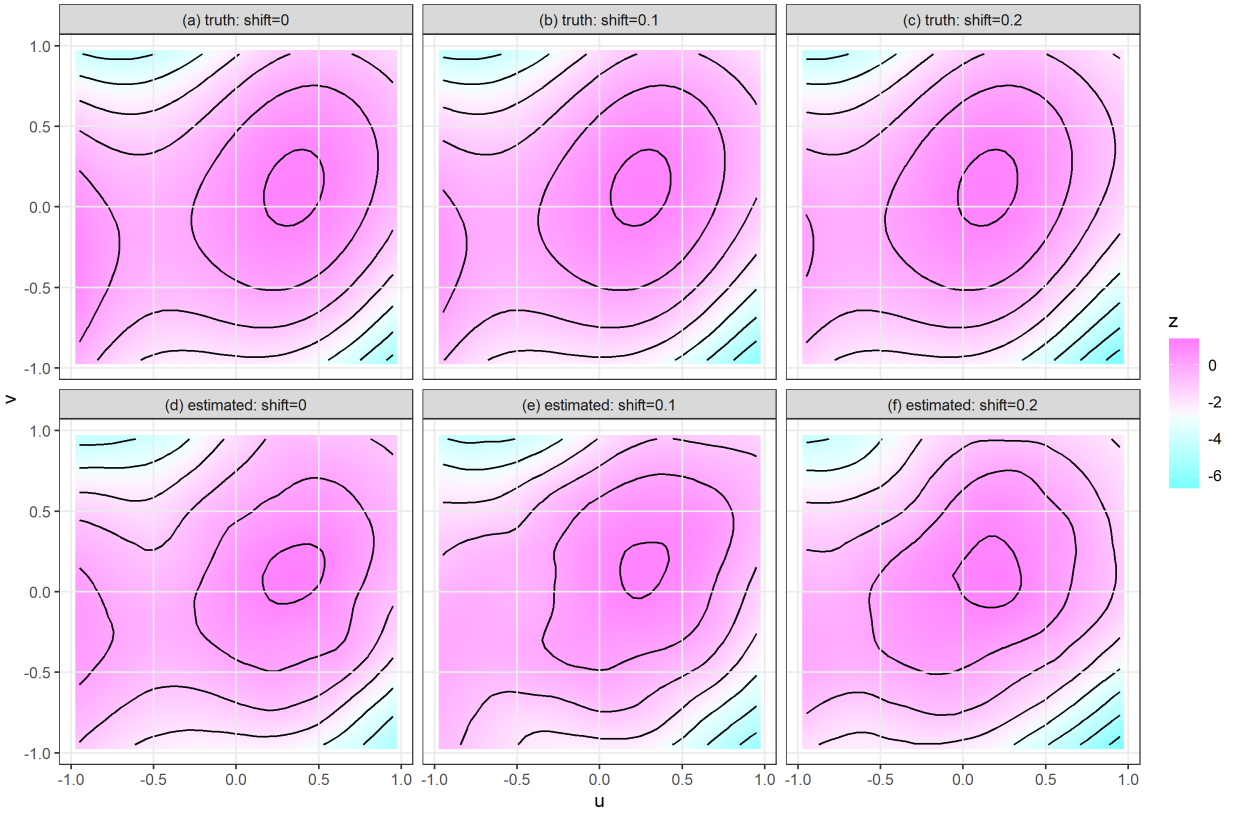

Figure 3.1: **Top**: Patterns used in nonlinear risk pattern simulations. "shift" stands for the shifting amount of the whole pattern to left; **Bottom**: Estimated spatial risk patterns using our proposed model (in (3.2)).

Using our proposed generalized additive model in (3.2), we are able to estimate the risk patterns at each time point. One simulation study is used to assess the performance of the stratified LOESS smoothers. 3 time points are set up with time-varying geospatial risk patterns where shift amounts are 0, 0.1 and 0.2, respectively. The 3 plots in the bottom of Figure 3.1 show the estimated risk patterns at each time point. From the plots, it is shown that the proposed time-specific LOESS smoothers are capable of recreating each of the time-specific geospatial risk patterns.

Several simulations are conducted with 2 time points in total. Under H_0, we assume homogeneous underlying spatial effects ("shift=0") at both time points. For simulations under H_1, "shift=0" is used for Time 1 and a shifted pattern is used for Time 2. The amount of shift indicates extremity of deviation from H_0, hence greater shift amounts should be detected

with greater statistical powers.

Empirical powers are defined as the proportion of simulations in which H_0 is rejected. A level of significance of $\alpha = 0.05$ is used. In each simulation scenario, simulation is repeated 500 times and $N_{perm} = 1000$. As a comparison, we also consider ANOVA tests for temporal homogeneity in the above simulation settings when a naive parametric form of spatial effects is assumed. Specifically, consider Model (3.8) and and test $H_0 : \beta_5 = \beta_6 = \beta_7 = 0$ using a permutation version of the F test. A permutation strategy is used in order to maintain a correct type I error. Specifically, for each simulated dataset, the permuted ANOVA procedure simply compares the observed F statistic and permuted F statistics, where the permuted F statistics are calculated using dataset with randomly permuted time labels. Also, the *mgcv* package offers a class of stratified thin-plate regression splines and corresponding ANOVA tests. We applied this test as well for comparison. The resulting powers are shown in Panel (a) of Figure 3.2.

$$E(Y_{ij}) = \beta_0 + \beta_1 u_{ij} + \beta_2 v_{ij} + \beta_3 t_{ij} + \beta_4 u_{ij} v_{ij} + t_{ij}(\beta_5 u_{ij} + \beta_6 v_{ij} + \beta_7 u_{ij} v_{ij}) \tag{3.8}$$

To assess the performance of the PMSD test on datasets with more than 2 time points, we create datasets with 4 time points. No shift is used at time point 1 and a shift amount S is applied at time point 4. For time 2 and 3, we used uniformly spaced shift amounts $S/3$ and $2S/3$. The resulting powers under these settings are shown in Panel (c) of Figure 3.2.

The Monte Carlo results presented in Figure 2 for the PMSD test utilize a uniformly distributed 20×20 grid on the entire 2×2 map. Although a 20×20 grid is seemingly sufficient for the risk patterns in the simulation studies we performed, we suggest a grid that is dense enough in order to sufficiently evaluate local behaviors.

In addition to the grid density, another decision to make when performing PMSD tests is grid selection strategy. Intuitively, one might choose uniformly distributed locations on the entire map or the observed locations within the dataset. Since the MSD statistic is defined as the mean of squared differences at a chosen set of locations, the pattern of the grid intrinsically defines the weights assigned over the map. For temporal heterogeneity detection, a uniform grid will place equal weights on all areas of the map while observed locations put higher weight on areas with denser observations. As a simple example, given a risk pattern where temporal variation exists in one area of the map, using more points in the varying area in the grid for PMSD tests would result in higher power since the test would be weighted more heavily in the varying area due to the grid design.

To assess the impact of grid selection strategy on power, we performed simulation studies in 2 more scenarios. In Scenario 1 (shown in the plots in the top row of Figure 3.3), we use the same pattern as the previous simulations with 2 time points where we use "shift=0" for Time 1 and "shift=0.1" at Time 2. The observations are designed to be uniformly distributed on the map. We vary N_g, the number of locations used in the grid, to explore the impact of grid density on power. For each N_g value, we simultaneously chose N_g random locations from a uniform 20×20 grid on the map and another random collection of N_g locations from the observed location set. Using the 2 sets of locations, PMSD tests were performed in order to assess the impact of grid selection strategy on power. In Scenario 2 (bottom row of Figure 3.3), we use another design of spatial risk surface which has 2 high risk areas. We use the pattern at Time 1 and mitigate the risk by 30% in one of the high risk areas at Time 2 to create temporal heterogeneity. In this setting, observations are designed to be denser in high risk areas. We performed similar grid selection and testing procedures to those of Scenario 1 and compared power with respect to grid density and selection strategy. From the empirical power results, the denser grids tend to render higher power, which is reasonable since local variations are more likely to be detected with more evaluated locations. In Scenario 1, tests using observed locations yield slightly higher power than those using uniform grids.

This result could be explained by higher precision in spatial effects estimation at observed locations. The performance of tests using observed locations is much better than that of tests using a uniform grid in Scenario 2, which is as expected since more locations in time-varying areas are evaluated by MSD statistic, resulting in greater weights in the areas when the test is performed.

3.3.2 Simulation studies with linear underlying patterns

In the previous Monte Carlo studies, we assume nonlinear underlying spatial patterns, as is shown in Figure 3.1. In these situations, our proposed PMSD test outperformed a permutation version of the F test, as is expected. In this section, we conduct a simulation study where the true underlying spatial risk patterns are created using a linear function of longitude and latitude and compare the finite sample behavior of the PMSD tests and the optimal F tests. Specifically, data were generated using the following model:

$$y_{ij} = -2 + 3u_{ij} - 3v_{ij} + u_{ij}v_{ij} + \tau t_{ij}(2u_{ij} + 2v_{ij} + u_{ij}v_{ij} + 1) + \epsilon_{ij}. \tag{3.9}$$

In the above model, values from 0 to 0.16 are chosen to be τ. ϵ_{ij}'s are $i.i.d.$ random errors from a standard normal distribution. We apply both a classical ANOVA F test of the space-time interaction terms (a correctly specified model) and the PMSD test. Using a similar strategy as previous sections, we compare the performance of the 2 classes of tests and plot the results in Panel (d) of Figure 3.2. Both tests yield correct type I errors while the ANOVA test yields slightly higher power. However, the increased power comes at a high cost, as the inflexible parametric model does not perform as well if the underlying risk pattern is more realistically nonlinear.

3.4 Application to birth defects study in Massachusetts

In this section we use our proposed methods to analyze the data achieved by the previously introduced MBDR study. Given the severity of PDA, it is of interest to determine if space-related risk factors place infants at increased PDA risk. Generalized additive models with a bivariate smoother incorporated, such as the model in (3.1), estimate cross-sectional geospatial risk patterns with adjustment for other potentially relevant factors including maternal age, adequacy of prenatal care (measured by the Adequacy of Prenatal Care Utilization Index), maternal race, maternal education level and number of siblings, as is applied in Girguis et al. (2016).

Beyond the analysis of cross-sectional data in each year, we further aim to investigate the existence of variation in spatial risk patterns over time. We applied our methods on data collected in 2003, 2006, and 2009. Available sample sizes for the three years are 1082, 969, and 877, respectively. Corresponding numbers of PDA cases are 111(10.3%), 90(9.3%) and 60(6.8%). The geographic distribution of the observations are shown in the top row of Figure 3.4. With adjustment for maternal age, adequacy of prenatal care, maternal race, maternal education level and number of siblings, the estimated geospatial risk patterns are shown in the bottom row of Figure 3.4. PMSD tests were performed to assess heterogeneity over the 3 years. The results are shown in Figure 3.5.

From the estimation in Figure 3.4, potential temporal heterogeneity of geospatial PDA risk is observed by visual inspection. To formally determine if the spatial risk changes over the period, we apply our proposed PMSD test and the resulting p-value is 0.028, indicating potential time-varying and space-related factors for PDA risk other than the adjusted ones in the model over the 7 years from 2003 to 2009. Note that since the PMSD test aims to find any temporal change in the geospatial risk patterns, the detected heterogeneity could be a result of either a change in the overall level of PDA rate at each time point or a change

in the spatial disparity patterns of PDA risk.

The presented analysis utilizes all observed locations as evaluation points. The PMSD test utilizing a uniform grid on the Massachusetts map renders a greater p-value (= 0.069) than the one using the observed locations (consistent with our previously presented simulation results). Corresponding PMSD and *OMSD* values are plotted in Figure 3.5. As discussed previously, the choice between the two grid choice strategies should also take into account the scientific goals of the study. The PMSD test using observed locations places weights that are roughly proportional to the population (under a simple random sample) while the PMSD test using a uniform grid equally weights all areas of Massachusetts.

3.5 Discussion

In this study, we brought in time-stratified kernel smoothers to the generalized additive model framework for estimation of time-specific geospatial disease risk patterns. Based on the proposed GAMs, we further formalized a permutation test for temporal heterogeneity of smoothed spatial risk effects.

Using simulation studies, we showed that our proposed PMSD test performed substantially better than the other 2 competitors in detecting underlying temporal heterogeneity given a specific nonlinear spatial pattern. Even when the difference between spatial risk patterns was not clearly noticeable by visual inspection, the proposed procedure yielded acceptable power. In situations with parametric spatial risk patterns, a correctly specified parsimonious parametric ANOVA F test only slightly outperformed the PMSD test. We did assume smoothness of the surface in the presented simulations as well as smooth shifts over time as there would likely not be abrupt shifts in the spatial or temporal patterns for the birth defect data that motivates our methodology. We do, however, acknowledge that non-smooth

patterns can exist and differential changes in hot spots may arise. In additional simulation studies, not presented here, we considered the performance of our methods in the setting of abrupt changes in the response surface. We found that our model managed to render reasonably good estimation although the estimated pattern is not as accurate at the areas where the risk changes dramatically, which is not surprising since the smoothness assumption is violated. We also found that the type I error rate of our proposed PMSD test was maintained and that relative power benefits compared to the F test were also maintained in this setting.

Straightforward extensions of the proposed methods may be of interest in particular settings. For instance, if multiple contiguous time points are assumed to share one common spatial risk pattern, these time points could be grouped as one strata. Thus the stratified smoothers and corresponding PMSD tests are naturally applied to multiple time stratas, rather than to all time points. A similar strategy could be adopted for cases with continuous time, where smoothers could be stratified at separate time intervals. Also, when researchers wish to investigate heterogeneity in geospatial risk patterns over a categorical factor other than time, such as sex or discretized age, the proposed methods remain applicable.

The proposed PMSD test considers a global test for temporal homogeneity of spatial effects, as opposed to identification of specific local area differences. This is a natural first step in the identification of changing spatial patterns over time. Note that when temporal variation of geospatial risk pattern exists merely localized while the grid for MSD statistics is designed on the entire map, the power may suffer from the inclusion of geographic locations where the risk does not vary over the time period in MSD statistic. One next step, as we have done in the MBDR analysis for cross-sectional data analysis, is to highlight areas with differential risks. This is akin to first establishing the existence of main effects then further investigating effect modification.

Other than what is proposed in this work, smoothing with respect to time seems to be

an intuitive way to model space-time data, but naively including time in smoothing terms is questionable. Using two separate smoothing terms in an additive way, one for spatial effects and another for time effects, fails to model time-space interaction while using a 3-way smoother including longitude, latitude and time offers flexible smoothing but suffers from anisotropy and potentially sparseness. The proposed procedures make no assumption on temporal correlation and put no restriction on how geospatial effects pattern varies. Some might be concerned about loss of efficiency due to the flexibility. We would argue that for large data sets, which is frequently the case in epidemiology studies, efficiency is maintained at each time point. As supporting evidence, Panel (d) of Figure 3.2 shows close performance between PMSD tests and the optimal ANOVA F-tests under parametric spatial risk patterns, indicating little efficiency loss.

For statistical inference on smoothing components in GAMs, an approximate F-test was introduced by Hastie and Tibshirani (1990). However, simulation studies showed that this approximate F-test renders inflated type I error in Young et al. (2011). According to simulation results of the approximate F-test (not shown in this manuscript) on our simulation problem, inflated type I errors are observed as well. However, due to its efficiency potential, we will explore calibrating methods for this class of approximate F-tests in our future work.

An intuitive Bayesian counterpart of the stratified smoother could be a time-stratified Gaussian process where one process is set up at each time point with or without shared hyperparameters. A stratified Gaussian process only requires modification of the likelihood. This stratification does not require a specific time-space covariance structure since no specific temporal correlation is assumed. Based on this class of models, temporal heterogeneity in spatial patterns can potentially be tested using the Bayes factor.(Kass and Raftery, 1995) While we appreciate the Bayesian framework in geospatial modeling, many health researchers are not closely familiar with Bayesian methods. In contrast, local linear regressions may be more intuitive and accessible to nonstatisticians. In particular, tuning a span size that is

understandable as the proportion of data utilized in the local window of smoothing, is more straightforward than choosing hyperparameters in covariance structures of a Gaussian process. In addition, GP models tend not to scale well with respect to computation. In contrast, the time-stratified generalized additive models proposed here are computationally cheap and easily scale to large data sets. We compared model fitting time between GP models and GAMs using the R package *spBayes* (Finley et al., 2007, 2015) and R package *gam* (Hastie and Tibshirani, 1990) respectively. GP models were roughly 20 times more computationally expensive when compared to GAMs for $N = 100$. Computation times were roughly 200 times greater for $N = 300$, 500 time greater for $N = 500$ and 1000 greater for $N = 700$, indicating a disadvantage of GP models with respect to scalability. Given these potential advantages, we view the proposed procedure as providing another useful tool to spatial epidemiologists.

One key assumption we made in this paper is mutual independence of all observations. That is to say, cross-sectional studies over multiple time points, rather than longitudinal studies, are discussed. Since longitudinal studies where individuals have multiple measurements over time are increasingly prevalent, in the future, we plan to further generalize stratified smoothers and investigate the performance of the PMSD tests on longitudinal data.

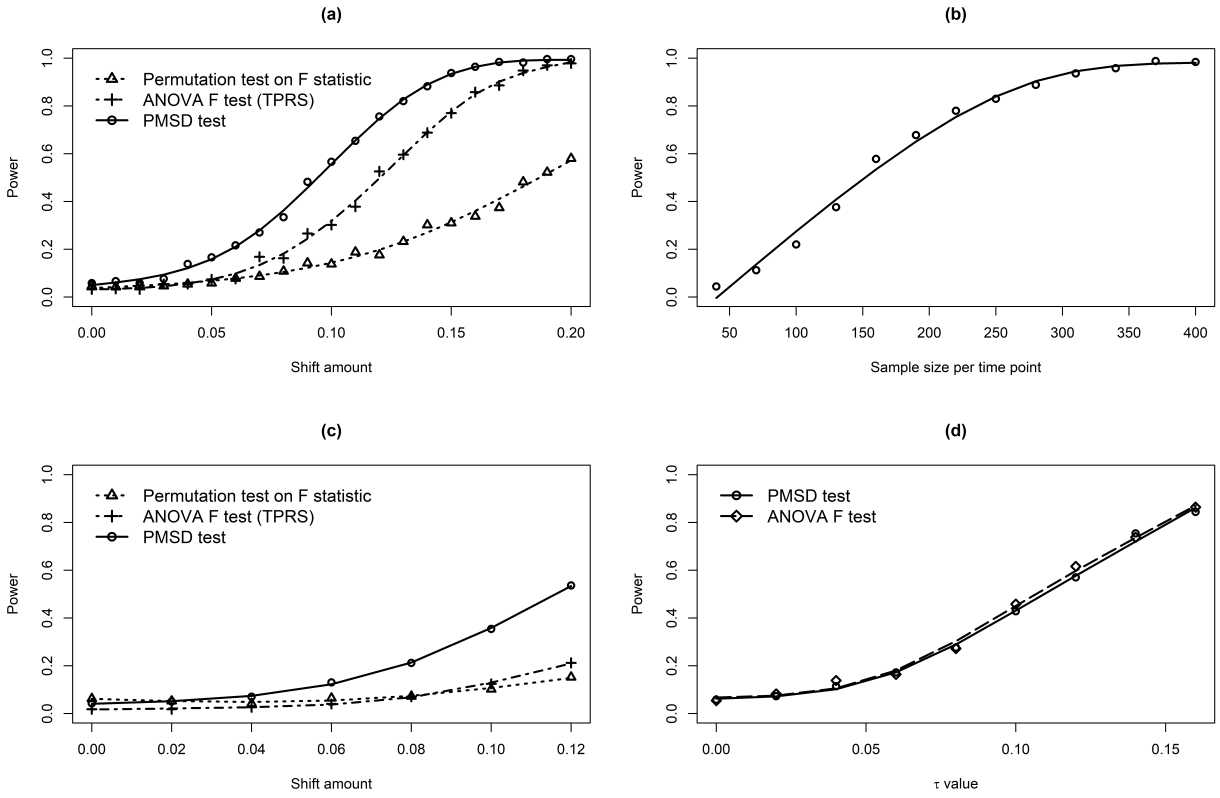

Figure 3.2: **Panel (a)**: Power vs. shift amount based on 500 simulations at each shift value. Increasing rejection proportion could be observed. Type I error (rejection proportion at shift=0) is 0.046 for permutation test on F statistic, 0.032 for ANOVA F test based on thin-plate regression splines and 0.056 for PMSD test. The proposed PMSD test has the highest power in detecting temporal heterogeneity and permutation test based on parametric models renders the lowest power; **Panel (b)**: Power vs. sample size based on 500 simulations at each sample size, given shift=0.15. Power increases along with sample size and approaches 1 near a sample size of 350 observations per time point; **Panel (c)**: Power vs. shift amount at time point 4 based on 500 simulations at each shift value. Increasing rejection proportion (or greater power) could be observed. Type I error (rejection proportion at shift=0) is 0.062 for permutation test on F statistic, 0.018 for ANOVA F test based on thin-plate regression splines and 0.042 for PMSD test. Similarly, our proposed PMSD tests have better performance in power in detecting temporal heterogeneity; **Panel (d)**: Power vs. multiplier τ in (3.9) based on 500 simulations at each value of τ. As expected, classical ANOVA F test performs better in terms of power since it is the "correct" test hence most powerful. In the meanwhile, the performance of PMSD is close to the F test.

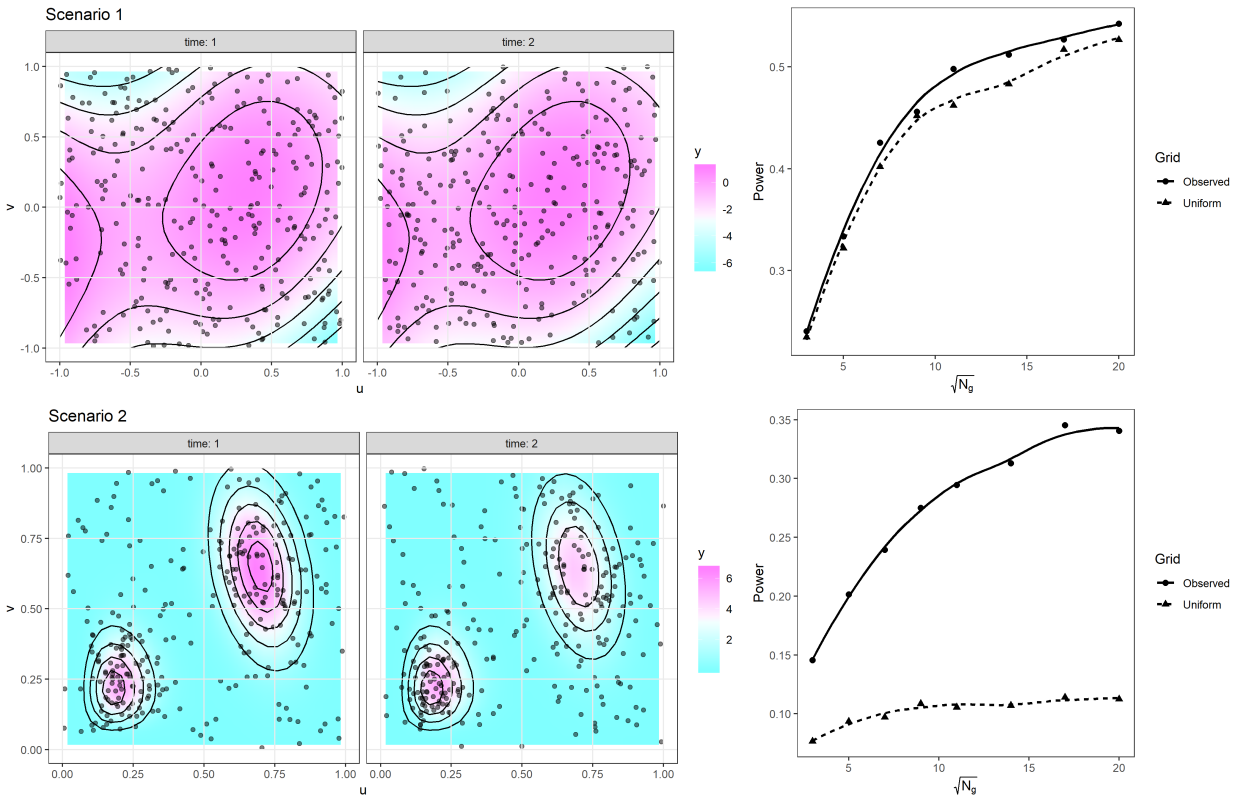

Figure 3.3: **Top**: Geospatial risk patterns at Time 1 and 2 with observed locations in Scenario 1 and corresponding power v.s. density curves for PMSD tests using uniform and observed grids. **Bottom**: Geospatial risk patterns at Time 1 and 2 with observed locations in Scenario 2 and corresponding power v.s. density curves for PMSD tests using uniform and observed grids.

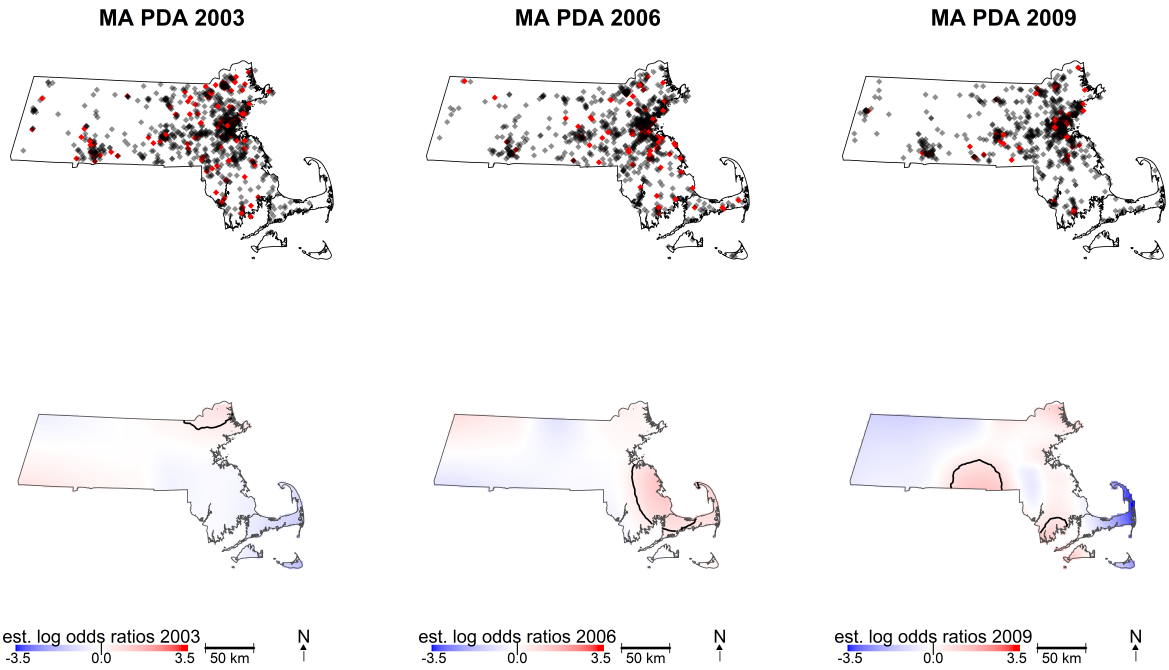

Figure 3.4: **Top**: Distribution of PDA cases (red) and controls (black) over selected years. Sparsity in western Massachusetts reflects lower population density; **Bottom**: Estimated geospatial risks for each year with adjustment for relevant variables. Values presented are estimated log odds ratios. The solid lines on the estimated patterns indicate areas with significant nonzero log-odds using $\alpha = 0.05$. The significance is determined by a permutation test described in Webster et al. (2006) The idea of the test is to randomly permute the locations of the observations and recalculate the log-odds for M times in order to achieve a point-wise reference distribution of log-odds at each point on map. The significant areas contain points where the estimated log-odds is outside of the 95% confidence interval constructed by the reference distribution. The test is applied using R package *MapGAM*. (Bai et al., 2019) According to the estimated surfaces, southeast Massachusetts has potentially significant high PDA risk in 2006 but low risk in both 2003 an 2009. In addition, a high PDA risk appears at central-southern Massachusetts in 2009.

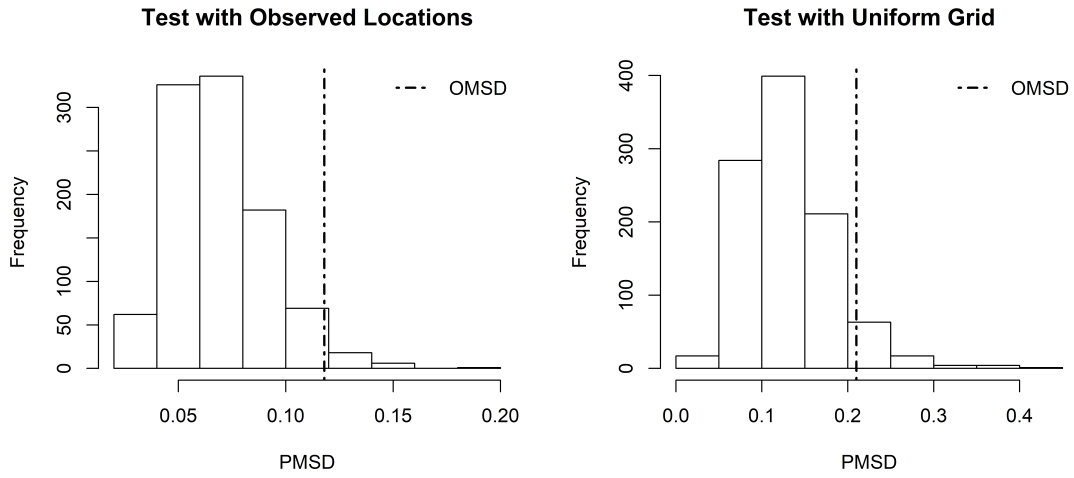

Figure 3.5: Histograms of PMSD values using 2 test location selection strategies with vertical lines indicating the value of OMSD. The p-values are 0.029 using observed locations and 0.069 using a uniform grid on the Massachusetts map.

Chapter 4

Additive mixed models with kernel smoothers

4.1 Introduction

Geographically heterogeneous disease rates are of common interest in epidemiology studies since local high or low rates may serve as a surrogate for space-related risk factors such as environmental exposures and local healthcare access or quality. Traditional geographic modeling methods focus on analyzing aggregated area-level data that treat area-defined partitions as one unit. More recent spatial epidemiology studies avoid aggregation bias and the ecological fallacy by modeling individual-level data that may be collected longitudinally over time. With accurate records of geospatial information (frequently longitude and latitude), researchers generally assume an underlying smooth surface for modeling the heterogeneity of disease risk over a given geospatial region, regardless of borders of inner areas. Based on this assumption, the estimation of the surface is an essential component in spatial data analysis.

Due to the complex nature of geospatial disease risk, it is not feasible to assume a parametric form to model the underlying risk surface in most cases. As such, nonparametric methods are popular in spatial effects modeling. Under a frequentist estimation framework, popular nonparametric methods include kernel and spline smoothers. Kernel smoothers utilize locally weighted models while spline smoothers are defined by a basis expansion of the design matrix over the full predictor support. In this work we consider locally weighted regression (LOESS) as proposed by Cleveland (1979). LOESS assumes local (weighted) polynomial relationship between response and explanatory variables and was applied to geospatial analysis by Brunsdon et al. (1996) among others. One advantage of LOESS for spatial analyses is that it intuitively adapts to changing population densities by varying the size of the smoothing neighborhood based on the local data density given a fixed span size (typically defined as the proportion of observations used for local regressions).

In addition to spatial risk pattern estimation, confounding variables, such as biomarkers in studies on public health, should be included in spatial models. Generalized additive models (GAMs) (Hastie and Tibshirani, 1990) offer a framework for incorporating frequentist smoothers and confounding variables in an additive fashion by assuming a linear predictor consisting of the sum of nonparametric smoothers and parametric adjustment covariates. Wood (2017) further developed GAMs by incorporating various types of splines and covered relevant topics such as computation, properties and applications.

Recently, data collection procedures such as patients revisits and health trackers have become increasingly prevalent. These procedures frequently result in multiple longitudinal measurements on individuals over time. This sampling framework results in within-subject correlation that must be accounted for in analytic methods in order to provide valid inferential results. Linear mixed models (LME) provide one framework for modeling within-subject correlation by adding random effects to linear models. Increased flexibility in the LME can be accomplished by incorporating random effects into a GAM framework. The resulting models

are commonly known as generalized additive mixed models (GAMMs). Lin and Zhang (1999) considered one strategy for the incorporation spline smoothers into the LME framework. To the best of our knowledge, however, no existing literature covers additive mixed models with kernel smoothers. The current manuscript seeks to fill this gap by propoposing a class of additive mixed models (AMM) that incorporate kernel smoothers and random effects into a linear model for continuous outcomes.

As one example, a fairly recent study was conducted to investigate serum perfluorooctanoic acid (PFOA) concentration among residents in Lubeck, West Virginia and Little Hocking, Ohio. (Bartell et al., 2010) In this study, researchers aimed to understand the declining behavior of PFOA concentration after granular activated carbon filtration on the public water systems in 2007. By design, 200 residents were included and 6 blood samples were to collect from each resident from May 2007 to August 2008 so that a trend of PFOA concentration could be observed. Besides PFOA concentration, residents' information such as gender, age and recent water consumption type (public water or bottled water) was recorded as well as precise residential location (recorded as longitude and latitude). One of the objectives is to understand the geospatial distribution of residents' serum PFOA concentration in order to help identify potential latent space-confounded risk factors.

The remainder of this chapter is organized as follows: In Section 4.2, we introduce our proposed additive mixed models (AMMs) and the corresponding fitting procedure. In Section 4.3, we present simulation studies designed to assess the performance of our new model in geospatial risk pattern recreation and parameter estimation. In Section 4.4, we use our proposed methods on PFOA data to estimate the geospatial pattern in serum PFOA concentration in the area of Lubeck, WV. Finally, Section 4.5 provides further discussion about the proposed work and considers avenues of future research.

4.2 Methods

4.2.1 Notations

Let $i = 1, 2, \ldots, N$ be the individual index where each i corresponds to one individual. For each individual i, measurements are taken at times $t_{i1}, t_{i2}, \ldots, t_{iJ_i}$. The measurements on individual i include the continuous response vector $Y_i = (y_{i1}, y_{i2}, \ldots, y_{iJ_i})$ and a length-p vector of adjustment covariates X_{ij} at time t_{ij}, $j = 1, \ldots, J_i$. In addition, each individual's geographical information (i.e. longitude and latitude) is tracked, labeled by (u_{ij}, v_{ij}) at time t_{ij}. Independence is assumed between individuals but not between the repeated measurements within each individual in statistical analysis.

4.2.2 LOESS with variance-covariance adjustment (LOESS-VCA)

We begin with a simple bivariate LOESS smoother for i.i.d. data given by

$$y_{ij} = lo(u_{ij}, v_{ij}) + \epsilon_{ij}, \quad i = 1, \ldots, N, \quad j = 1, ...J_i, \tag{4.1}$$

where $\epsilon \overset{i.i.d.}{\sim} N(0, \sigma^2)$ and $lo()$ denotes a LOESS smoother. In (4.1), the mean response at location (u^*, v^*), i.e. $lo(u^*, v^*)$, is the estimand of scientific interest. The local regression model involves k of the observed data points that are nearest to (u^*, v^*) where k is pre-specified according to the span size and distance is generally defined as Euclidean distance for determining nearest neighbors. Utilizing the nearest neighbor data, locally weighted regression is used to estimate $lo(u^*, v^*)$, where weights are assigned according to the distance between the neighbor and the target location (u^*, v^*). A traditional choice of weight function

is the tricube weight function given by

$$w(d) = \begin{cases} \left(1 - \frac{d^3}{[max(d)]^3}\right)^3, & \text{for } d < max(d) \\ 0, & \text{for } d > max(d) \end{cases}. \tag{4.2}$$

In (4.2), d denotes the distance and $max(d)$ is the maximum of the k distances corresponding to the nearest neighbors. Let L^* denote the local design matrix constructed by the local values of (u, v), W_1^* denote a diagonal matrix with tricube weights of the local observations, and Y^* denote the response values of the local observations. Then the fitted value $\hat{lo}(u^*, v^*)$ can be calculated using weighted least squares:

$$\hat{lo}(u^*, v^*) = (1 \ u^* \ v^*)(L^{*T} W_1^* L)^{-1} L^{*T} W_1^* Y^* \tag{4.3}$$

The above estimation procedure assumes $\epsilon \overset{i.i.d.}{\sim} N(0, \sigma^2)$. However, as we aim to estimate spatial patterns using longitudinal data, this assumption does not generally hold since measurements within each individual will tend to be correlated rather than independent. Thus, we extend the simple LOESS smoother to accommodate correlated data with a known, or assumed, correlation structure.

We again consider the mean model specified in (4.1) but release the i.i.d. assumption. Rather, we assume $\epsilon \sim N(0, \Sigma)$ where Σ is known and $\epsilon = (\epsilon_1, \epsilon_2, \ldots, \epsilon_N)'$. With known, or assumed, variance-covariance model Σ, the weighted least squares estimator given in (4.3) can be modified to account for within subject clustering via incorporation of inverse-variance weights. Specifically, let Σ^* denote the local components of Σ defined by the span size. Then a variance-covariance adjusted LOESS (LOESS-VCA) fit is given by

$$\hat{lo}(u^*, v^*) = (1 \ u^* \ v^*)(L^{*T} W_1^{*1/2} \Sigma^{*-1} W_1^{*1/2} L)^{-1} L^{*T} W_1^{*1/2} \Sigma^{*-1} W_1^{*1/2} Y^*. \tag{4.4}$$

Since Σ (and hence Σ^*) are generally not known in practice, a generalized estimating equations (GEE) approach (Liang and Zeger, 1986) that assumes a working covariance structure and iteratively replaces Σ with a method of moments estimator could be used. We, however, are interested quantifying potential random effects variance components. In the next section we consider a novel backfitting strategy for incorporating both random effects and adjustment covariates into the bivariate LOESS model.

4.2.3 An additive mixed model with kernel smoothers

As discussed in Section 4.1, additive models are popular tool among spatial epidemiologists seeking to estimate spatial disease risk patterns while simultaneously adjusting for potential confounding factors. Specifically, the model given in (4.1).

However, when the dataset of interest includes multiple measurements on some individuals, the independence assumption does not inherently hold. Since Model (4.1) does not take the potential correlation among the measurements into account, inefficient or incorrect inference could be yielded if Model (4.1) is adopted.

To account for within-individual correlation arising from longitudinal sampling of individuals over time, Laird and Ware (1982) proposed a class linear mixed effects models (LMEs) given by

$$y_{ij} = X_{ij}\beta + Z_{ij}b_i + \epsilon_{ij}, \tag{4.5}$$

where it is assumed that $b_i \sim N(0, D)$ and $\epsilon_{\mathbf{i}} = (\epsilon_{i1}, \epsilon_{i2}, \ldots, \epsilon_{iJ_i})' \sim N(0, R)$, with b_i and ϵ_i independent. Typically, the fixed effects $X_{ij}\beta$ component of the linear predictor is used to model the scientific association of interest and adjust for potential confounding covariates, the random effects $Z_{ij}b_i$ component is used to model individual-specific effects and the ϵ_{ij}'s

are assumed to be i.i.d. conditional upon the random effects.

A natural extension of (4.1) to the setting of longitudinal data is given by an additive mixed model of the form

$$y_{ij} = lo(u_i, v_i) + X_{ij}\beta + Z_{ij}b_i + \epsilon_{ij}. \tag{4.6}$$

While estimation procedures for generalized additive mixed models with spline smoothers have been proposed (cf. Lin and Zhang (1999)), to the best of our knowledge, no previous work covered model fitting procedures for an additive mixed model with kernel smoothers. Here we consider a novel generalizing of the backfitting algorithm proposed by Breiman and Friedman (1985) to estimate the model specified in (4.6).

We begin with the classic backfitting algorithm (presented in Algorithm 7). The basic idea of Algorithm 7 is to fit partial residuals using one component in the mean model given the fitted values of the rest components in an iterative way, and repeating until convergence. Based on this algorithm, we propose Algorithm 8 to fit (4.6). Specifically, we propose modifying Algorithm 7 by replacing the linear model with a LME utilizing a specified variance-covariance structure as determined by the specification of the random effects and the LOESS with the previously discussed LOESS-VCA estimator utilizing the induced working variance-covariance structre from the LME. As a result, estimation of both the non-parametric and semi-parametric components of the fixed effects linear predictor in (4.6) incorporate the assumed variance-covariance structure of the longitudinal data.

Algorithm 7 Backfitting algorithm for Model 4.1 (Gaussian response)

Initialize $\hat{lo}_{ij} = 0$, $\hat{f}_{ij} = 0$ for all i, j. ($\hat{lo}_{ij}$ will be the fitted values of the bivariate spatial LOESS smoother and $\hat{f}_{ij}$ will be the fitted values of the parametric component $X_{ij}\beta$.)

while at least one of the estimates $\hat{lo}_{ij}$ and $\hat{f}_{ij}$ change by 0.01%, **do**

Fit linear model $(y_{ij} - \hat{lo}_{ij}) \sim X_{ij}\beta$ and get fitted values $\hat{f}_{ij}$ for all i, j.

Centralize the fitted values using $\hat{f}_{ij} = \hat{f}_{ij} - \underset{i,j}{\text{mean}}(\hat{f}_{ij})$.

Fit LOESS smoother $(y_{ij} - \hat{f}_{ij}) \sim lo(u_{ij}, v_{ij})$ and get fitted values $\hat{lo}_{ij}$ for all i, j.

Centralize the fitted values using $\hat{lo}_{ij} = \hat{lo}_{ij} - \underset{i,j}{\text{mean}}(\hat{lo}_{ij})$.

end while

Algorithm 8 Backfitting algorithm for Model 4.6 (Gaussian response)

Initialize $\hat{lo}_{ij} = 0$, $\hat{f}_{ij} = 0$ for all i, j. ($\hat{lo}_{ij}$ will be the fitted values of the bivariate spatial LOESS smoother and $\hat{f}_{ij}$ will be the fitted values of the parametric component $X_{ij}\beta$.)

while at least one of the estimates $\hat{lo}_{ij}$ and $\hat{f}_{ij}$ change by 0.01%, **do**

Fit linear mixed model $(y_{ij} - \hat{lo}_{ij}) \sim X_{ij}\beta + Z_{ij}b_i$ and get fitted values $\hat{f}_{ij}$ for all i, j.

Centralize the fitted values using $\hat{f}_{ij} = \hat{f}_{ij} - \underset{i,j}{\text{mean}}(\hat{f}_{ij})$.

Calculate the estimated variance-covariance matrix V from the mixed model.

Fit LOESS-VCA smoother $(y_{ij} - \hat{f}_{ij}) \sim lo(u_{ij}, v_{ij})$ using V as the true variance-covariance matrix and get fitted values $\hat{lo}_{ij}$ for all i, j.

Centralize the fitted values using $\hat{lo}_{ij} = \hat{lo}_{ij} - \underset{i,j}{\text{mean}}(\hat{lo}_{ij})$.

end while

4.2.4 Quantification of uncertainty in spatial effects

Based on our proposed models, by using local weighted regression models, the estimated spatial effect at a specific location (u^*, v^*) can be written as

$$\hat{lo}(u^*, v^*) = H^*(\vec{\alpha})y^*, \tag{4.7}$$

where $\vec{\alpha} = (D, R)'$ stands for the variance component and y^* would be the corresponding local response vector, which could potentially be the local working partial residual vector within the backfitting procedure if an AMM is being fitted. Hence the variance of $\hat{lo}(u^*, v^*)$

can be expressed as

$$Var(\hat{lo}(u^*, v^*)) = H^*(\vec{\alpha})Var(Y^*)H^{*T}(\vec{\alpha}), \tag{4.8}$$

where $Var(Y^*) = V^*(\vec{\alpha})$ is a function of $\vec{\alpha}$ as well. We impute the estimated values of the variance component based on the linear mixed model and use

$$\widehat{Var}(\hat{lo}(u^*, v^*)) = H^*(\hat{\vec{\alpha}})V^*(\hat{\alpha})H^{*T}(\hat{\alpha}) \tag{4.9}$$

to quantify the uncertainty of the estimated spatial effects in a point-wise fashion.

4.3 Monte Carlo studies

To assess the performance of our proposed methods in spatial effects estimation, we conducted multiple simulation studies based on a 2×2 square map with a true spatial pattern given by

$$s_0(u, v) = -u + 0.1\log(1.3)v + 1.2\sin(3(u + 0.1)) + 2uv + 6\log(0.6)v^2 \tag{4.10}$$

and depicted in the top plot of Figure 4.1.

4.3.1 Spatial pattern recreation

In this section, we aimed to compare the performance of our proposed additive mixed models in terms of spatial pattern recreation. Responses were then simulated via the model

$$y_{ij} = s_0(u_{ij}, v_{ij}) + b_{0i} + \epsilon_{ij}, \tag{4.11}$$

where $b_{0i} \overset{i.i.d.}{\sim} N(0, \tau^2)$ and $\epsilon_{ij} \overset{i.i.d.}{\sim} N(0, 1)$. We repeated the simulation with $\tau^2 = 0, 1.5, 2.5$, respectively.

We sought to compare our proposed additive mixed model (given in (4.12)) with a naive model that assumes independent data (given in (4.13)):

$$y_{ij} = \beta_0 + \beta_t t_{ij} + lo(u_{ij}, v_{ij}) + b_{0i} + \epsilon_{ij} \tag{4.12}$$

$$y_{ij} = \beta_0 + \beta_t t_{ij} + lo(u_{ij}, v_{ij}) + \epsilon_{ij}. \tag{4.13}$$

We compared the estimated spatial risk patterns between the 2 models under varying random intercept conditions. Span sizes that minimize AIC (Akaike, 1998) were chosen for each model. Note that if $\tau^2 > 0$, there will be individual-specific intercepts hence the model given in (4.13) is misspecified. It follows that the likelihood of Model (4.13) used in the AIC calculation would be incorrect hence the chosen span size might not be the most appropriate, however a search across different span sizes did not return qualitatively different results from what is presented here.

The estimated spatial risk patterns are shown in the second and third rows of Figure 4.1. It is easily seen that that with correctly specified random effects, our proposed additive mixed models estimate the spatial patterns in each scenario (bottom row). In contrast, since the naive additive model fails to account for within-subject correlation, results for the model given in (4.13) tend to under-smooth the patterns using a relatively small span size (second row). This is due to the fact that the naive additive model treats correlated data as independent, resulting in an improper contribution to the total likelihood.

Similar simulations were conducted for scenarios where both random intercepts and random

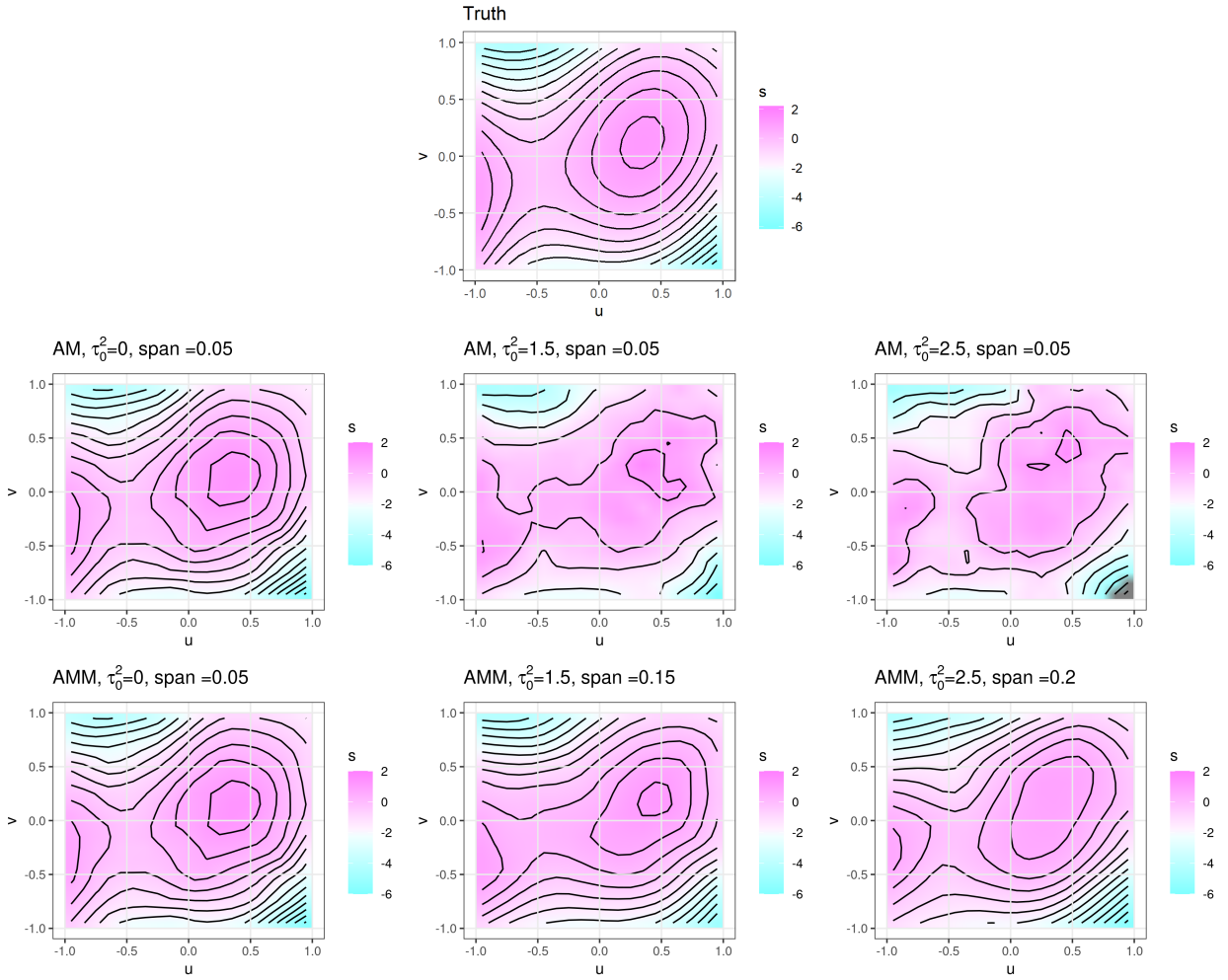

Figure 4.1: Top: Simulated spatial risk pattern; Middle: estimated patterns using additive models given differing correlation structures; Bottom : estimated patterns using additive mixed models given differing correlation structures.

slopes exist. With random slopes included, we simulated data using

$$y_{ij} = s_0(u_{ij}, v_{ij}) + b_{0i} + b_{1i}t_{ij} + \epsilon_{ij}, \tag{4.14}$$

where $b_{0i} \overset{i.i.d.}{\sim} N(0, 1.5)$, $b_{1i} \overset{i.i.d.}{\sim} N(0, \tau_1^2)$ and $\epsilon_{ij} \overset{i.i.d.}{\sim} N(0, 1)$. We repeated the simulation with $\tau_1^2 = 0, 0.04, 0.09$, respectively. To these simulated data, we applied our proposed AMM (given in (4.15)) and the same naive model that assumes independent data (given in (4.13)).

$$y_{ij} = \beta_0 + \beta_t t_{ij} + lo(u_{ij}, v_{ij}) + b_{0i} + b_{1i} t_{ij} + \epsilon_{ij}. \tag{4.15}$$

From the recreated spatial risk patterns shown in Figure 4.2, our proposed AMMs managed to deliver better performance with respect to selection of smoothing amount and pattern recreation when both random intercept and random slope exist and our model is accordingly specified.

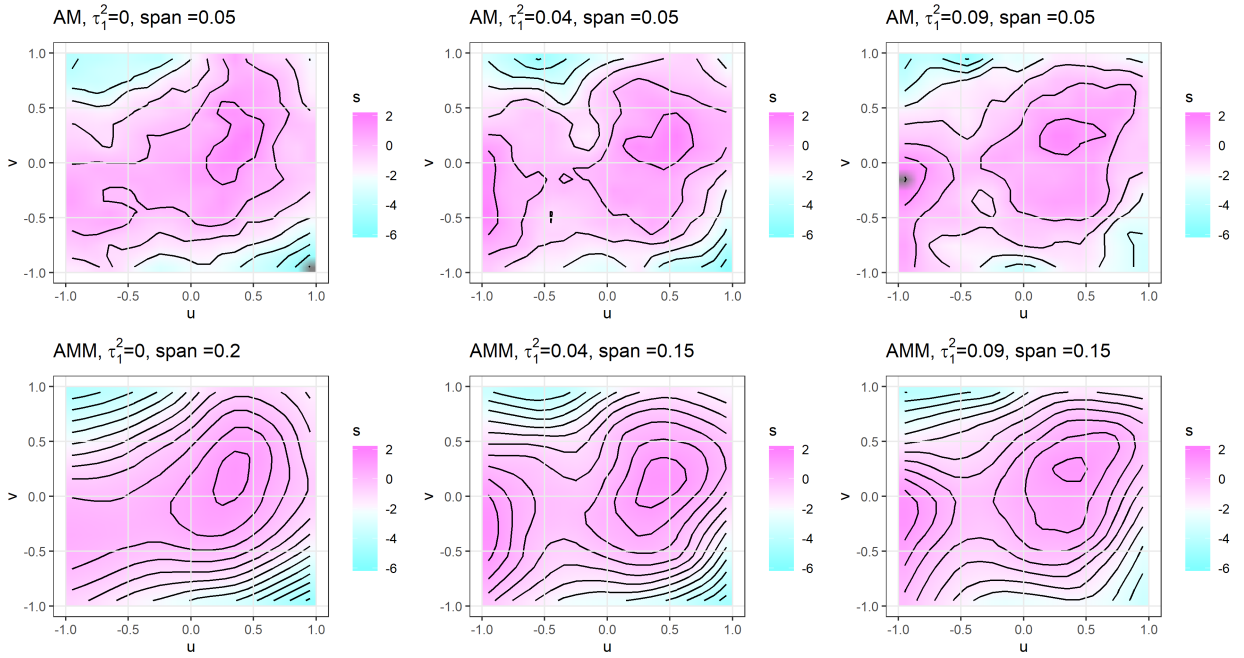

Figure 4.2: Top: estimated patterns using additive models given differing correlation structures; Bottom : estimated patterns using additive mixed models given differing correlation structures.

4.3.2 Quantification of uncertainty of estimated spatial effects

To assess the performance of our proposed methods for quantifying the uncertainty of estimated spatial effects discussed in 4.2.4, simulated data were created using (4.14), where

$b_{0i} \overset{i.i.d.}{\sim} N(0, 1.5)$, $b_{1i} \overset{i.i.d.}{\sim} N(0, \tau_1^2)$ and $\epsilon_{ij} \overset{i.i.d.}{\sim} N(0, 1)$. Two scenarios were designed with $\tau_1^2 = 0/$ and 0.09, respectively. Hence there was one scenario with random intercepts only and one with both random intercepts and random slopes. 500 repetitions were performed for each scenario. For each repetition, 95% CIs are derived using our proposed methods (Model 4.6 and the corresponding methods in Section 4.2.4) for every location on a uniformly designed 10×10 grid on the map. Coverages of corresponding CI of spatial effects are plotted in Figure 4.3. From the plots, empirical values of coverages could be less than 0.95 at the boundaries of the map as well as areas where the linearity assumption does not hold well. Locations on boundaries are generally estimated with less precision due to less available neighborhood information. Also, because not every single piece of the true pattern could be precisely estimated with universal smoothing as defined by one single span size, some areas were estimated with bias resulting in lower coverage probability.

Figure 4.3: Empirical coverages of 95% CI of spatial effects by locations based on correctly specified Model 4.6. Left: scenario with random intercepts (mean = 0.90); Right: scenario with random intercepts and slopes (mean = 0.91).

4.3.3 Parameter estimation

Here we investigate the performance of our proposed additive mixed model in terms of estimation of the parameters in both the fixed effects and the variance of random effects via another set of simulation studies where responses were simulated using model

$$y_{ij} = s_0(u_{ij}, v_{ij}) + \beta_t t_{ij} + \beta_x x_{ij} + b_{0i} + b_{1i} t_{ij} + \epsilon_{ij}, \tag{4.16}$$

with $b_{0i} \overset{i.i.d.}{\sim} N(0, \tau_0^2)$, $b_{1i} \overset{i.i.d.}{\sim} N(0, \tau_1^2)$ and $\epsilon_{ij} \overset{i.i.d.}{\sim} N(0, \sigma^2)$.

On the simulated dataset, we sought to access the performance of our proposed AMM given by

$$y_{ij} = \beta_0 + \beta_t t_{ij} + \beta_x x_{ij} + lo(u_{ij}, v_{ij}) + b_{0i} + b_{1i} t_{ij} + \epsilon_{ij}. \tag{4.17}$$

and a naively applied additive model given by

$$y_{ij} = \beta_0 + \beta_t t_{ij} + \beta_x x_{ij} + lo(u_{ij}, v_{ij})\epsilon_{ij}, \tag{4.18}$$

which assumes independence among the dataset.

The estimated values of β_x, β_t, τ_0 and τ_1 were recorded. Results presented in Table 4.1 were based upon a total of 500 simulations. The empirical mean and standard deviation of estimated model parameters, as well as the mean of estimated standard errors, were reported. From the results, it could be seen that our proposed AMMs managed to estimate parameters in both the mean and variance components in a consistent fashion.

Table 4.1: Average and standard deviation of parameter estimates across 500 simulated datasets.

model	para.	truth	$\bar{est.}$	relative bias	empirical sd	$\bar{est.sd}$
AM	β_x	0.10	0.10	4%	0.0568	0.0535
	β_t	-0.20	-0.20	0%	0.0157	0.0218
AMM	β_x	0.10	0.10	3%	0.0369	0.0359
	β_t	-0.20	-0.20	0%	0.0157	0.0152
	τ_0	1.23	1.21	-1.2%	-	-
	τ_1	0.20	0.20	-0.5%	-	-

4.4 Application to serum PFOA study

In this section, we apply our proposed additive mixed model to estimate the spatial pattern of residents' serum PFOA concentration with adjustment of relevant confounding covariates. Here we focus on an approximately square map defined within longitude 81°30′ W - 81°50′ W and latitude 39°07′ N - 39°27′ N. The are constitutes 23×23 square miles around the Lubeck, WV area. Within this area, 1070 records on 193 individuals are available where 140 of them have 6 measurements of serum PFOA concentration from May 2007 to August 2008. Among the 193 residents, 99 are female and 94 are male. Mean age at baseline is 54.6 years with a standard deviation 14.9 years. The baseline age of participants ranges from from 19 to 92 years.

According to the individual-specific trends of serum PFOA concentration values across time, visually significant individual-specific level of serum PFOA concentration were observed while the reducing trends of individuals did not show much variation hence a model that incorporate random intercepts would be reasonable. To estimate the spatial pattern of residents' serum PFOA concentration while controlling for gender, age and a linear trend in time, we fitted an AMM given by

$$log(PFOA)_{ij} = \beta_0 + \beta_1 female_i + \beta_2 age_i + \beta_3 t_{ij} + lo(u_{ij}, v_{ij}) + b_{0i} + \epsilon_{ij}, \tag{4.19}$$

where $b_{0i} \overset{i.i.d.}{\sim} N(0, \tau_0^2)$ and $\epsilon_{ij} \overset{i.i.d.}{\sim} \sigma^2$.

The fitted spatial pattern from the model given in (4.19), along with point-wise 95% confidence intervals, is shown in Figure 4.4. The estimated pattern is trimmed according to the locations of observations. From these results, it can be seen that potentially high PFOA risk areas exist in western and northern parts of the area. Further point-wise significance tests are performed at each location within a trimmed uniformly designed 20×20 grid on the map of interest. Specifically, 95% confidence intervals for the spatial effect at each location is compared with the mean estimated spatial effect over the 219 locations. A location is labeled as significant if the 95% CI at the location fails to cover the mean estimated effect. The results are plotted in Figure 4.5, from which significantly higher risks are observed at 42 locations (19.2%) while significantly lower risks are observed at 67 locations (30.6%), indicating significant geospatial disparity in residents' serum PFOA concentration over the map.

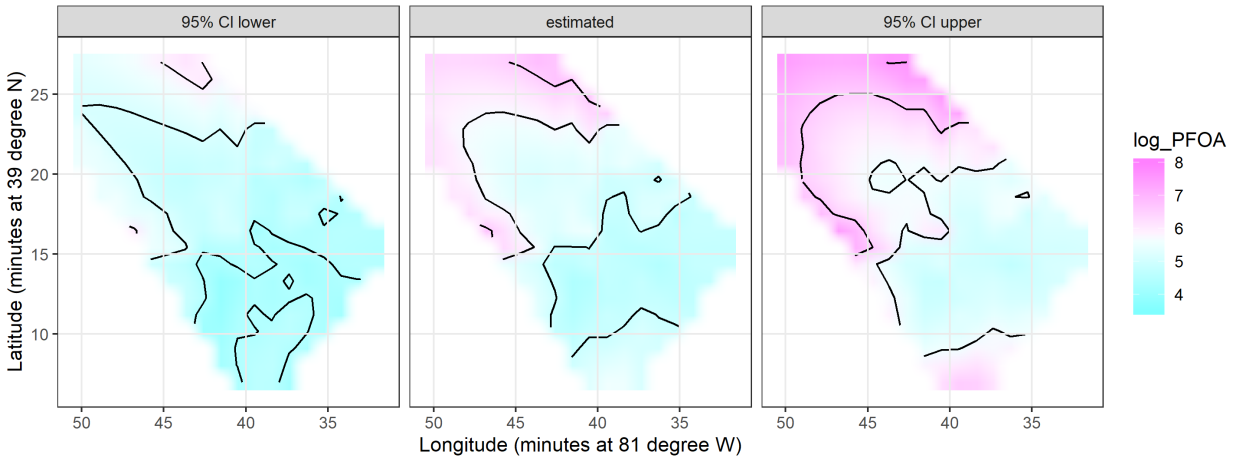

Figure 4.4: Left: pointwise lower bounds of 95% CIs ; Middle: estimated patterns using Model 4.19; Right: upper bounds of 95% CIs.

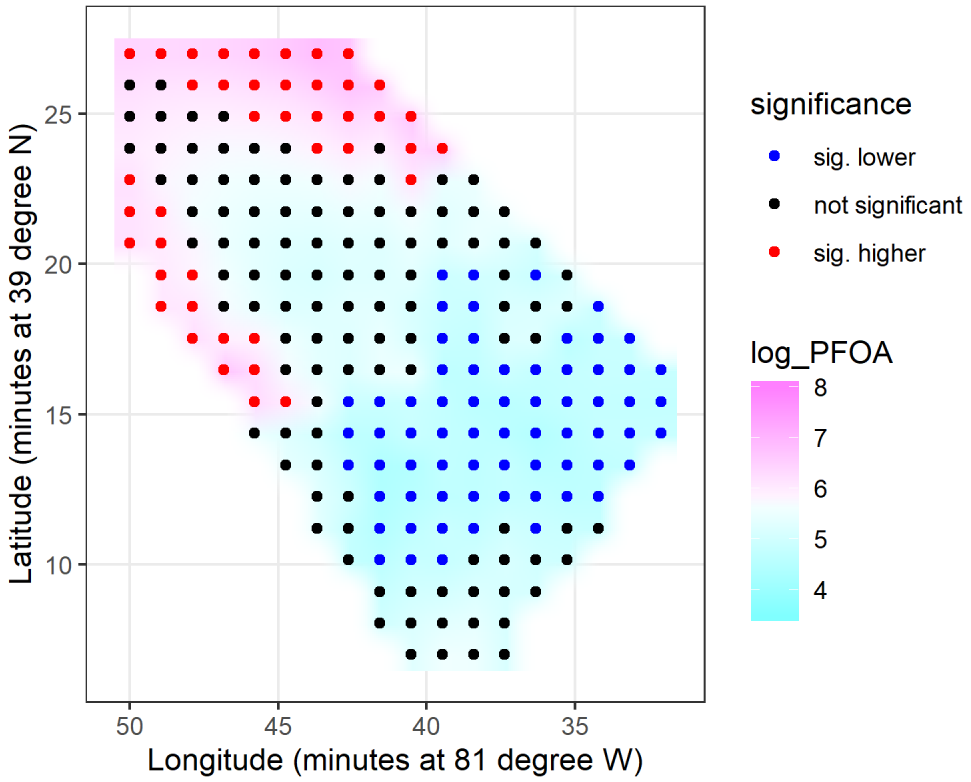

Figure 4.5: Significance of the grid locations on the estimated spatial effects map.

4.5 Discussion

In this work we have proposed a novel class of additive mixed models that incorporate kernel smoothers into the classic LME model. To achieve this, we first extended the LOESS smoother to adjust for a given variance-covariance structure and then proposed a new backfitting algorithm to fit the proposed additive mixed models using the extended LOESS (LOESS-VCA). Using Monte Carlo studies, we showed that our proposed additive mixed models managed to choose the proper amount of smoothing via AIC and resulted in accurate estimates of underlying spatial risk patterns. The performance of the model was shown to be superior to that of naive additive models that assume independence. Empirical results also showed that our model consistently estimates parameter values in both the systematic linear predictor as well as variance components, provided that the model is correctly

specified. Our methods were further used in a recent study of residents' serum PFOA concentration in Lubeck, WV and spatial disparities in serum PFOA concentration levels were identified.

In this work, we focused on additive models with kernel smoothers, utilizing LOESS in particular. The motivation for this is to meet the demand of various spatial epidemiology studies. We mentioned but did not elaborate on the use of spline smoothers in the context of mixed models, partially because those models are relatively well investigated in the existing literature. We did not consider Bayesian spatial estimation methods based on Gaussian processes but do recognize their popularity, as well.

We consider this work to be a first step to a complete set of generalized additive mixed models that incorporate kernel smoothing. In Chapter 5 of the dissertation we present work on these relevant extensions. Specifically, we extend the approach presented here to the setting of additive mixed models for exponential family outcomes, such as binary and count responses, by incorporating kernel smoothing into a generalized linear mixed model (GLMM) framework and propose a novel model fitting procedure that combined the backfitting algorithm presented here with a penalized quasi-likelihood (PQL) (Breslow and Clayton, 1993) approximation.

Chapter 5

Generalized additive mixed models with kernel smoothers

5.1 Introduction

In Chapter 4, a class of AMMs with kernel smoothers were proposed for continuous responses. Analogous to the extension of linear models to generalized linear models, when risk of a certain disease is to be mapped in a longitudinal study with a non-Gaussian outcome, such as a binary outcome or a counting outcome, the class of AMMs does not apply directly. As is mentioned in review of the existing literature in Chapter 4, using spline smoothers, Lin and Zhang (1999) accounted for exponential family outcomes in their class of GAMM. However, to the best of our knowledge, no existing literature covers GAMMs with kernel smoothers. To fill this gap, this chapter seeks to incorporate kernel smoothers and random effects into a generalized linear model and propose the corresponding model fitting and inference methods.

The remainder of the chapter is organized as follows: In Section 5.2, we introduce our proposed GAMMs and the corresponding model fitting procedure. In Section 5.3, we present

simulation studies designed to assess the performance of our new model in geospatial risk pattern recreation and parameter estimation. In Section 5.4, we applied our proposed methods on PFOA data to estimate the geospatial pattern in serum PFOA concentration in the area of Lubeck, WV. Lastly, Section 5.5 provides further discussion about the proposed work and considers avenues of future research.

5.2 Methods

5.2.1 Notations

Let $i = 1, 2, \ldots, N$ be the individual index where each i corresponds to one individual. For each individual i, measurements are taken at times $t_{i1}, t_{i2}, \ldots, t_{iJ_i}$. The exponential family outcome of individual i is denoted with a response vector $y_i = (y_{i1}, y_{i2}, \ldots, y_{iJ_i})$. x_{ij} stands for a length-p vector of adjustment variables for individual i at time t_{ij}. In addition, each individual's geographical information (i.e. longitude and latitude) is tracked, labeled by (u_{ij}, v_{ij}).

5.2.2 Generalized additive mixed models with kernel smoothers

Starting with generalized linear models (GLMs, McCullagh (2018)), in this section we provide a brief introduction to the formulation of generalized additive models (GAMs, Hastie and Tibshirani (1990)) and generalized linear mixed models (GLMMs, Breslow and Clayton (1993); Wolfinger and O'connell (1993)). We will then propose our novel class of generalized additive mixed models (GAMMs) by combining GAMs and GLMMs.

GLMs represent a ubiquitous class of regression models, used widely throughout many scientific fields. The broad utility of GLMs stems from a unified estimation and theoretical

framework that can be applied to response variables whose probability distribution function belongs to the exponential family. This includes the Bernoulli and binomial distributions for binary responses and the Poisson distribution for count responses. More specifically, GLMs assume that the outcome y_i follows a distribution of exponential family with $E(y_i) = \mu_i$ and $\text{var}(y_i) = v(\mu_i)$, where function $v()$ depends on the specific distribution of y_i. Commonly seen distributions of response y_i include Gaussian, Bernoulli and Poisson among others. The mean μ_i is linked to the linear predictor $x_i^T\beta$ by the link function $g()$. Hence the mean model for a GLM is commonly written as

$$g(\mu_i) = x_i^T\beta. \tag{5.1}$$

GLMs assume independence among data hence are appropriate in cross-sectional studies where linearity suffices to model the relationship between $g(\mu_i)$ and the explanatory variables x.

If the independence assumption is relaxed in GLMs by inducing random effects b_i to model cluster-specific effects, the linear predictor is then written as

$$g(\mu_{ij}) = x_{ij}^T\beta + z_{ij}b_i, \tag{5.2}$$

where $b_i \overset{i.i.d.}{\sim} \text{MVN}(0, D(\theta))$ and $D(\theta)$ is the variance-covariance matrix as a function of parameter vector θ. Due to the inclusion of the random effects $z_{ij}b_i$, this resulting model (Model (5.2)) is termed a generalized linear mixed model (GLMM).

Other than GLMMs, another class of generalization of the classic GLM framework is achieved by relaxing the linearity assumption io the mean model. Specifically, smooth functions $s_k()$'s are used to replace all or part of the linear terms. In this case the mean model can be written

as

$$g(\mu_i) = \sum_{k=1}^{p} s_k(x_i), \tag{5.3}$$

which is termed a generalized additive model (GAM) since arbitrary functions, $s_k()$, are combined in an additive fashion.

As noted in Section 5.1, since we aim to build a class of models that accommodate exponential family responses, random effects and smoothers simultaneously, combining the GLMM and GAM frameworks would be a natural approach. Specifically, we are interested in a class of models with mean function

$$g(\mu_{ij}) = \sum_{k=1}^{p} s_k(x_{ij}) + z_{ij}^T b_i. \tag{5.4}$$

Given our motivation of developing our methods for disease mapping in geospatial epidemiology studies and our collaboration group prefer kernel smoothers (LOESS smoother in particular) for spatial risk pattern estimation, the mean model of direct interest could be written as

$$g(\mu_{ij}) = x_{ij}^T \beta + lo(u_{ij}, v_{ij}) + z_{ij}^T b_i, \tag{5.5}$$

where $x_{ij}^T \beta$ models potential confounding effects, $lo(u_{ij}, v_{ij})$ models a flexible underlying spatial effect on the target disease risk and $z_{ij}^T b_i$ ($b_i \overset{i.i.d.}{\sim} \text{MVN}(0, D(\theta))$) stands for individual-specific random effects, such as random intercepts or random slopes.

5.2.3 Model fitting Procedure

As we stated in Section 5.1, although Lin and Zhang (1999) proposed a double penalized quasi-likelihood (DPQL) fitting procedure and approximate inference which considered both random effects and smoothing terms as a penalization when maximizing the likelihood of the full model for GAMMs with spline smoothers only, it was also recognized in their work that kernel smoothing was not accommodated due to the fact that kernel smoothers are generally not trivially parametrizable. Our goal is to fill this gap in current methodology by proposing a novel fitting and inferential procedure that incorporates LOESS kernel smoothing into the GLMM framework.

We propose a model fitting procedure based on the penalized quasi-likelihood (PQL) method for GLMMs. Briefly, to fit a GLMM, Breslow and Clayton (1993) proposed the PQL method to produce reasonable efficient inference in GLMM setting and Wolfinger and O'connell (1993) further elaborated the computation in a more detailed fashion. In practice, according to Wolfinger and O'connell (1993) PQL estimation is achieved using an iterative strategy. In particular, when a GLMM, such as that specified in (5.2), is to be fitted, working response y_{ij}^w is defined as

$$y_{ij}^w = g(\hat{\mu}_{ij}) + (y_{ij} - \hat{\mu}_{ij})g'(\hat{\mu}_{ij}), \tag{5.6}$$

with reasonably initialized $\hat{\mu}_{ij}$. Following Laird and Louis (1982) and Lindstrom and Bates (1990), the working response vector $\mathbf{y}_{\hat{\mu}}^w$ can then be approximated by a Gaussian distribution

$$N[X\beta + Zb,\ g'(\hat{\mu})R_{\hat{\mu}}g'(\hat{\mu})], \tag{5.7}$$

where $R_{\hat{\mu}}$ is the variance-covariance matrix defined by the assumed outcome distribution

given the estimated mean vector $\hat{\mu}$ and conditional upon the random effects. It follows that a weighted linear mixed effects (LME) model

$$\mathbf{y}^{w}_{\hat{\mu},ij} = x^{T}_{ij}\beta + z_{ij}b_i + \epsilon_{ij} \tag{5.8}$$

with working diagonal weight matrix

$$\hat{W}_{\hat{\mu}} = R^{-1}_{\hat{\mu}}[g'(\hat{\mu})]^{-2} \tag{5.9}$$

could be used to model the working response $\mathbf{y}^{w}_{\hat{\mu}}$. The PQL estimating procedure iteratively fits weighted linear mixed model with updated working response $\mathbf{y}^{w}_{\hat{\mu}}$ and working weight matrix $\hat{W}_{\hat{\mu}}$ based on the updated $\hat{\mu}$ at each iteration until the difference in parameter estimations are sufficiently small.

Based on the PQL procedure, Lin and Zhang (1999) developed double penalized quasi-likelihood (DPQL) by incorporating spline smoothers into the GLMM framework with an additional penalization term to control smoothness. While spline smoothing could be achieved using basis expansion functions of the design matrix, however, kernel smoothing could not be achieved in such ways. Consequently, in order to fit a GAMM with kernel smoothers, such as that specified in (5.5), the classic PQL framework must be modified to accommodate weighted additive mixed models (AMMs) that incorporate kernel-based smoothers at each iteration. One approach to fitting such a model is proposed in Chapter 4 of this dissertation and further details are provided in Tang et al. (2020).

Briefly, Tang et al. (2020) proposed a class of linear AMMs with kernel smoothers for Gaussian response and developed the details of a novel fitting procedure to allow for parameter estimation and response prediction. Specifically, this work proposed a modified backfitting algorithm which merged classical linear mixed model estimation with an iteratively estimated variance-covariance adjusted kernel smoother. The algorithm iteratively updates the

estimation of spatial effects and other parameters by fitting the partial residuals. Under the derivation that the working response $\mathbf{y}^w_{\hat{\mu}}$ is approximately Gaussian (see 5.7), their method serves as a suitable choice when one wishes to fit the working response at each iteration within a PQL fitting procedure.

To summarize, with the combination of the classic PQL procedure by Wolfinger and O'connell (1993) and the algorithm for additive mixed models by Tang et al. (2020), a model fitting procedure for our proposed GAMM (Model (5.5)) could be sketched as

1. Initialize $\hat{\mu}$ using a GAM $g(E(y_{ij})) = x_{ij}^T\beta + lo(u_{ij}, v_{ij})$.
2. Update working response $\mathbf{y}^w_{\hat{\mu}}$ using Eq. 5.6 and working weight matrix $\hat{W}_{\hat{\mu}}$ using (5.9) with the updated $\hat{\mu}$.
3. Fit the working response with AMM $y^w_{\hat{\mu},ij} = x_{ij}^T\beta + lo(u_{ij}, v_{ij}) + z_{ij}^T b_i + \epsilon_{ij}$ with weight matrix $\hat{W}_{\hat{\mu}}$, rendering inference on model parameters (including $\hat{\mu}$).
4. Repeat Step 2 and 3 until the difference in estimated parameters between iterations are satisfyingly small.

Estimation and inference of the model parameters is based on the final iteration of the working AMM. Since the backfitting algorithm is adopted, inference on β and θ can be achieved from classic linear mixed model theory using ML or REML while inference on spatial effects can be based on the kernel smoother. Specifically, using the local model, the estimated spatial effect at a specific location (u^*, v^*) can be written as

$$\hat{lo}(u^*, v^*) = H^*(\hat{\theta})y^*, \tag{5.10}$$

where $H^*(\hat{\theta})$ denotes the estimated variance-covariance adjusted local hat matrix and y^* would be the corresponding local partial residuals. Using a similar strategy as in Tang et al.

(2020), the variance of $\hat{lo}(u^*, v^*)$ can be estimated by

$$\widehat{Var}(\hat{lo}(u^*, v^*)) = H^*(\hat{\theta})\widehat{Var}(Y^*)H^{*T}(\hat{\theta}). \tag{5.11}$$

5.2.4 Out-of-sample likelihood for smoothing parameter selection

Under most smoothing frameworks, choosing the right amount of smoothing is vital to avoid potental over- or under-smoothing. Popular smoothing techniques in disease mapping, including splines (Wood, 2003) and a kernel-based methods (Cleveland, 1979), use one parameter to control the smoothness of the estimated surface, thereby making a search of the parameter space tractable. Since we are interested in the use of kernel-based LOESS smoothers, we focus on span size, the smoothness parameter. LOESS uses local weighted linear models to achieve a nonparametric estimation of the underlying risk surface. Span size (or span for short) indicates the proportion of data used in the local models.

Similar to model selection in general, span is commonly chosen according to criteria such as AIC and GCV. In Chapter 4 and Tang et al. (2020) we considered AIC for span selection and achieved satisfying results. However, based on the results from simulation studies in multiple synthesized scenarios, AIC does not have a stable performance in span selection for our proposed GAMMs. Conditional AIC, also known as cAIC, by Vaida and Blanchard (2005) were tested as well but no satisfying span selection was observed.

The reason for the failure of AIC and cAIC in span selection is, at least partially, due to the fact that the degree of freedom used by the LOESS smoother is merely approximated by the trace or similar metrics of the hat matrix. When the outcome follows an arbitrary exponential family distribution, we end up with decreased information from each sampling unit, compared to Gaussian distributed outcomes. The result is inaccurate estimation of the smoother degrees of freedom.

In this section, we propose an out-of-sample likelihood Monte Carlo (OLMC) method for span size selection. We first randomly divide the sample into 2 subsets: a training set $\mathbf{y}^{(t)}$ and a validation set $\mathbf{y}^{(v)}$. Since observations from one individual or cluster would commonly be correlated, the randomization should be performed based on individuals rather than observations. In other words, no individuals should have measurements in both the training and validation datasets. Given candidate span values, we train our GAMM, achieve the estimated fixed effects $\hat{\beta}^{(t)}$, estimated LOESS smoother $\hat{lo}^{(t)}$ and variance-covariance component $\hat{D}^{(t)}$. Based on the trained model, we aim to calculate the likelihood of the validation set using

$$\begin{aligned} L_{v|t} &= p_1(\mathbf{y}^{(v)}|\hat{\beta}^{(t)}, \hat{lo}^{(t)}, \hat{D}^{(t)}) \\ &= \int p_1(\mathbf{y}^{(v)}|\hat{\beta}^{(t)}, \hat{lo}^{(t)}, b^{(v)}) p_2(b^{(v)}|\hat{D}^{(t)}) db^{(v)}, \end{aligned} \tag{5.12}$$

where $p_1()$ stands for the likelihood of the response (from an exponential family), $p_2()$ is the likelihood of the random effects (from a Gaussian distribution $N(0, \hat{D}^{(t)})$), and $b^{(v)}$ stands for the corresponding individual-level random effects.

It worth noting that the integral in Equation (5.12) is not tractable, so we recommend a Monte Carlo method with a simulated sample of $b^{(v)}$. In particular, we simulate a relatively large sample (we used 500 in this work) from a $N(0, \hat{D}^{(t)})$ distribution, use each drawn sample $b^{(v),(k)},\ k = 1, \ldots, 500$, to calculate the likelihood using

$$p_1^{(k)} = p_1(\mathbf{y}^{(v)}|\hat{\beta}^{(t)}, \hat{lo}^{(t)}, b^{(v),(k)}) \tag{5.13}$$

and then calculate the Monte Carlo approximation of $L_{v|t}$ using

$$L_{v|t}^{*} = \underset{k}{\text{mean}}\ p_1^{(k)}. \tag{5.14}$$

This out-of-sample likelihood Monte Carlo approach provides approximate and convenient evaluation of the model's out-of-sample performance. As such, when comparing the proposed GAMMs with differing span sizes, it is advisable to choose the model with the greatest $L^*_{v|t}$ value since greater out-of-sample likelihood generally indicates less amount of over-fitting or under-fitting of the spatial effects when span is the only varying factor.

5.3 Monte Carlo studies

To assess the performance of our proposed method, we conducted multiple simulation studies based on a 2×2 square map with a true spatial pattern given by

$$s_0(u, v) = 1.7 + 0.25[1.2\sin(3u + 0.3) + 2uv + 6\log(0.6)v^2] \tag{5.15}$$

and depicted in the Figure 5.1.

5.3.1 Spatial pattern recreation

In this section, we aimed to compare the performance of our proposed generalized additive mixed models in terms of spatial pattern recreation. Binary responses were simulated via the model

$$\mathbf{P}(y_{ij} = m) = p_{ij}^m(1 - p_{ij})^{(1-m)},\ m = 0, 1 \tag{5.16}$$

where

$$\text{logit}(p_{ij}) = \beta_x x_{ij} + \beta_t t_{ij} + s_0(u_{ij}, v_{ij}) + b_{0i}, \tag{5.17}$$

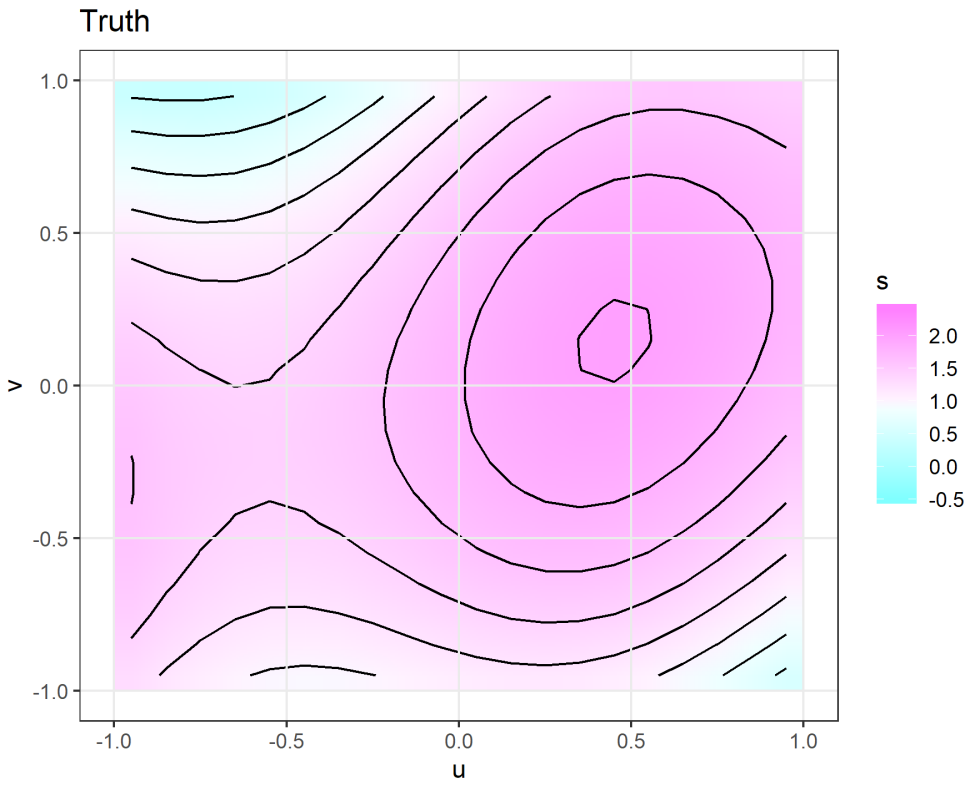

Figure 5.1: Simulated spatial risk pattern.

and $b_{0i} \overset{i.i.d.}{\sim} N(0, \tau_0^2)$. We repeated the simulation with $\tau_0 = 0, 0.3, 0.6, 0.9$, respectively.

We sought to compare our proposed GAMM (given in (5.18)) with a naive GAM that assumes independent data (given in (5.19)):

$$\text{logit}(p_{ij}) = \beta_0 + \beta_x x_{ij} + \beta_t t_{ij} + lo(u_{ij}, v_{ij}) + b_{0i} \tag{5.18}$$

$$\text{logit}(p_{ij}) = \beta_0 + \beta_x x_{ij} + \beta_t t_{ij} + lo(u_{ij}, v_{ij}) \tag{5.19}$$

We compared the estimated spatial risk patterns between the 2 models under varying random intercept conditions. Span sizes that minimize AIC (Akaike, 1998) were chosen for classic

GAM while our proposed OLMC method was used to choose the span size for the GAMM. Note that if $\tau_0^2 > 0$, there will be individual-specific intercepts hence the model given in (5.19) is a misspecified model due to its failure to account for the random effect. It follows that the likelihood of Model (5.19) used in the AIC calculation would be incorrect hence the chosen span size tends not to be the most appropriate. A search across different span sizes, however, did not return qualitatively different results from what is presented here.

The estimated spatial risk patterns are shown in Figure 5.2. It is easily seen that that with correctly specified random effects, our proposed additive mixed models estimate the spatial patterns in a more precise fashion. In contrast, since the naive additive model fails to account for within-subject correlation, results for the model given in (5.19) tend to under-smooth the patterns using a relatively small span sizes. This is due to the fact that the naive additive model treats correlated data as independent, resulting in an improper contribution of the correlated data to the total likelihood.

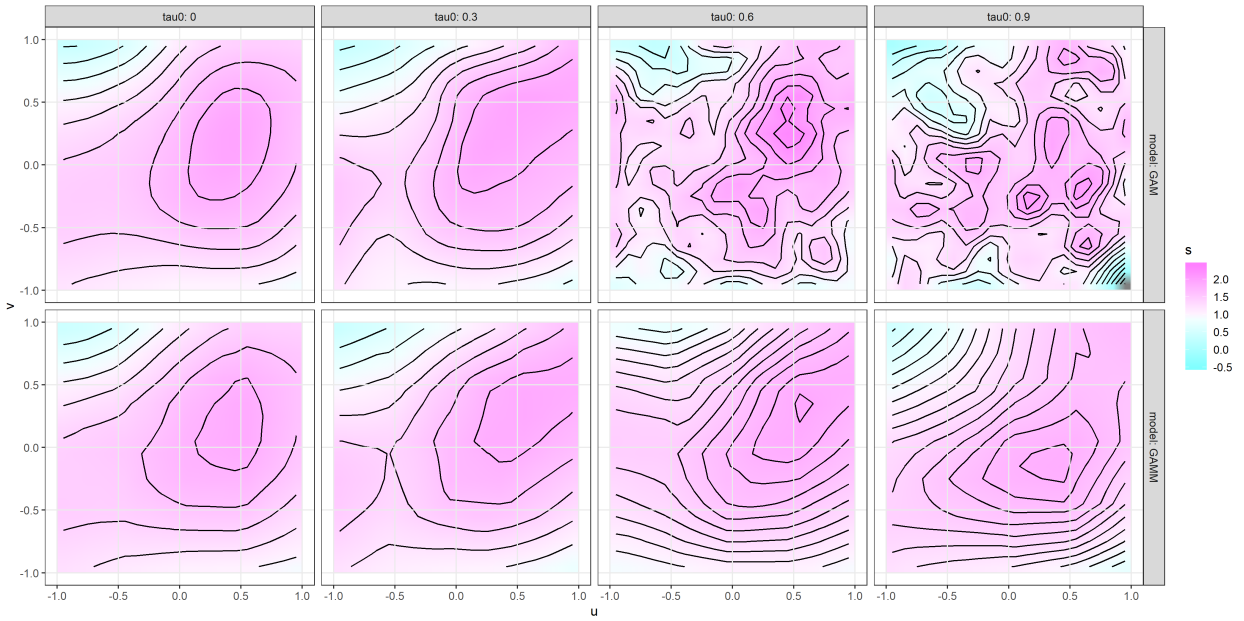

Figure 5.2: Top: estimated patterns using GAMs given differing true correlation structures; Bottom : estimated patterns using GAMMs given differing true correlation structures.

5.3.2 Quantification of uncertainty of estimated spatial effects

To assess the performance of our proposed method for quantifying the uncertainty of estimated spatial effects discussed in 5.2.3, simulated data were created using (5.17), where $b_{0i} \overset{i.i.d.}{\sim} N(0, \tau^2)$. 3 scenarios were designed with $\tau = 0.2,\ 0.4,\ 0.6$, respectively. 500 repetitions were performed for each scenario. For each repetition, 95% CIs were derived using our proposed method for every location on a uniformly designed 20×20 grid on the map. Both sample mean and sample standard deviation of the coverage were reported in Table 5.1. From the results, The uncertainty in spatial effects was well estimated and the 95% confidence intervals approximately managed to render the target coverage.

Table 5.1: Sample mean and sample standard deviation of the coverage proportion of 95% confidence intervals of spatial effects. Coverage proportions are computed based on 500 repetitions while mean and standard deviation are calculated over 400 locations on map.

τ_0	mean coverage	s.d.
0.2	0.945	0.057
0.4	0.908	0.062
0.6	0.976	0.032

5.3.3 Parameter estimation

In this section, we investigated the performance of our proposed GAMMs in terms of point estimation of the parameters in both the fixed effects and the variance of random effects via a new set of simulation studies where responses were simulated using model

$$\text{logit}(p_{ij}) = \beta_x x_{ij} + \beta_t t_{ij} + s_0(u_{ij}, v_{ij}) + b_{0i} + b_{1i} t_{ij}, \tag{5.20}$$

with $b_{0i} \overset{i.i.d.}{\sim} N(0, \tau_0^2)$ and $b_{1i} \overset{i.i.d.}{\sim} N(0, \tau_1^2)$.

On the simulated dataset, we sought to access the performance of our proposed GAMM

given by

$$y_{ij} = \beta_0 + \beta_x x_{ij} + \beta_t t_{ij} + lo(u_{ij}, v_{ij}) + b_{0i} + b_{1i} t_{ij}. \tag{5.21}$$

The estimates of β_x, β_t, τ_0 and τ_1 were recorded for each simulated example. The results presented in Table 5.2 were based upon a total of 500 simulations. The sample mean and standard deviation of estimated model parameters, along with the mean estimated standard deviations, were reported. From the results, it can be seen that by correctly accounting for the correlation structure, our proposed GAMMs managed to both estimate fixed effects with less bias relative to the naive GAM model as well to estimate the model variance components with reasonably low bias.

Table 5.2: Parameter estimation based on 500 repetitions

model	parameter	truth	mean of estimates	relative bias	empirical s.d.	mean of $\hat{s.d.}$
GAM	β_x	0.20	0.17	-16.5%	0.033	0.031
	β_t	-0.10	-0.09	12%	0.0034	0.0032
GAMM	β_x	0.20	0.19	-5.5%	0.035	0.032
	β_t	-0.10	-0.10	4%	0.004	0.0041
	τ_0	0.50	0.53	6.8%	-	-
	τ_1	0.07	0.07	-5.7%	-	-

5.4 Application to serum PFOA study

In this section, we apply our proposed GAMM to estimate the spatial risk pattern of high serum PFOA concentration with adjustment of relevant confounding covariates. Here we focus on a approximate square map defined within longitude $81°30' \ W$ - $81°50' \ W$ and latitude $39°07' \ N$ - $39°27' \ N$, which is approximately a 23×23 square in miles around Lubeck, WV area. Within this area, 1070 records on 193 individuals are available where 140 of them have 6 measurements of serum PFOA concentration from May 2007 to August 2008. Among the 193 residents, 99 are female and 94 are male. Mean age at baseline is 54.6

years with a standard deviation of 14.9 years. Age at baseline ranged from a minimum of 19 years to a maximum 92 years. Based upon prior scientific knowledge, we use 100 ng/mL as a threshold for "high concentration" where serum PFOA concentration values that are greater than 100 ng/mL are considered high.

According to the individual-specific trends of serum PFOA concentration values across time, we found visually significant individual-specific level of serum PFOA concentration while the reducing trends of individuals did not show much variation hence a model that incorporate random intercepts would be reasonable. To estimate the spatial pattern of the odds of residents' high serum PFOA concentration with control of gender, age and a linear trend in time, we fitted an GAMM given by

$$g(p_{ij}) = \beta_0 + \beta_1 female_i + \beta_2 age_i + \beta_3 t_{ij} + lo(u_{ij}, v_{ij}) + b_{0i}, \tag{5.22}$$

where $g(\cdot)$ is the logit link function, p_{ij} is the probability that individual i has a high serum PFOA concentration at time j, and $b_{0i} \overset{i.i.d.}{\sim} N(0, \tau_0^2)$.

The fitted spatial pattern of log-odds, along with point-wise 95% confidence intervals, is shown in Figure 5.3. The patterns are trimmed according to the locations of observations so that the shown estimated patterns are on the areas where residents were present. From the results, it could be seen that potentially high PFOA exposure areas exist in western, southern and north-eastern parts of the area. Further point-wise significance tests were performed at each location within a uniformly designed 20×20 grid on the map of interest. The grid was similarly trimmed. Specifically, 95% confidence interval of spatial effect at each location is compared with the mean estimated spatial effect over the 219 locations. A location was labeled as significant if the 95% CI at the location did not cover the mean estimated effect. The results were plotted in Figure 5.4, from which significantly higher risks

were observed at 57 locations (26.0%) while significantly lower risks were observed at 84 locations (38.4%), indicating significant geospatial disparity in residents' risk of high serum PFOA concentration over the map.

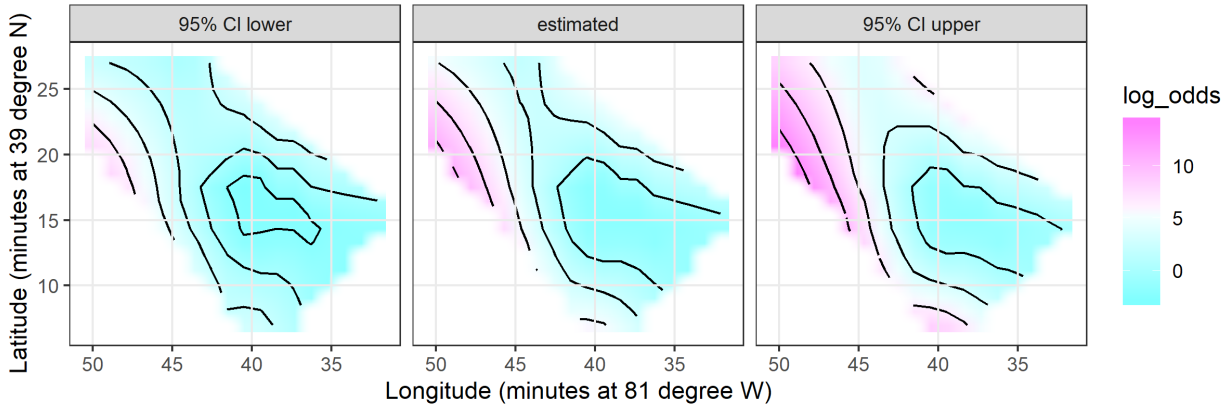

Figure 5.3: Left: point-wise lower bounds of 95% CIs ; Middle: estimated pattern of log odds ratio of high serum PFOA using Model (5.22); Right: point-wise upper bounds of 95% CIs.

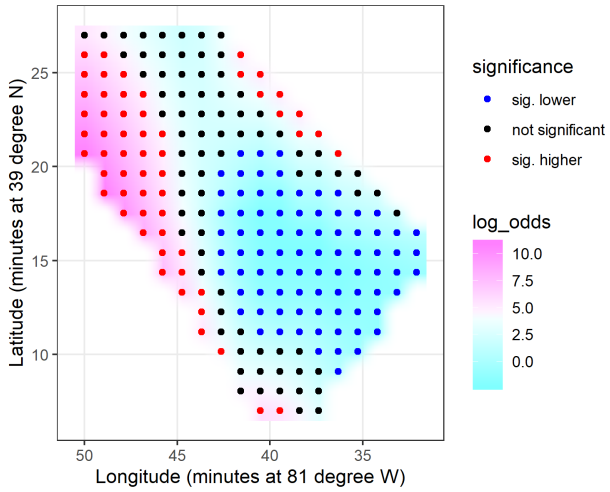

Figure 5.4: Significance of the grid locations on the estimated spatial effects map.

5.5 Discussion

In this chapter we proposed a novel class of generalized additive mixed models that incorporate kernel-based smoothers. To achieve this, we combined the PQL estimating procedure with a backfitting algorithm. To choose the proper amount of smoothing, we proposed a novel out-of-sample likelihood Monte Carlo method in order to avoid over-smoothing or under-smoothing. We showed that by adjusting for the correlation structure, our model better recreated the spatial risk pattern than a naively applied GAM. Empirical results also showed that our model managed to render estimation of parameters in both mean and variance components with less bias when the model was correctly specified. Our methods were further used in a recent study on residents' serum PFOA concentration in Lubeck, WV and spatial disparity in risk of high serum PFOA concentration was identified.

In this work, we focused on kernel smoothers, utilizing LOESS in particular in order to meet the demand of various spatial epidemiology studies. We mentioned but did not elaborate on the use of spline smoothers in the context, partially due to the fact that those models are relatively well investigated in existing literature such as Wood (2017) and Lin and Zhang (1999). We did not cover Bayesian spatial estimation methods based on stochastic processes but we do recognize their popularity, as well.

This work represents a novel extension of the work presented in Tang et al. (2020) and also found in Chapter 4 of this dissertation. This work could also be viewed as an alternative to Lin and Zhang (1999) when kernel smoothers are preferred over spline smoothers.

Throughout, we have considered a mixed effects modeling framework. One natural future direction is to develop a fitting procedure using marginal quasi-likelihood (Breslow and Clayton, 1993) for GAMM with kernel smoothers. While such an approach would be less useful for point-wise prediction of exposure, if inference for marginal model parameters is of scientific interest it may be preferred.

Chapter 6

Discussion

Using disease mapping problems in spatial epidemiology studies as motivation, this work developed in this dissertation provides multiple novel extensions of the GAM framework in order to accommodate studies that consider individual-level data measured over space and time. In Chapter 3, we proposed time-stratified bivariate kernel smoothers and incorporated them into a classic GAM framework. To test the significance of time-based stratification, we adopted a permutation strategy, resulting in a class of PMSD tests. Chapters 4 and 5 concentrated on disease mapping problems in longitudinal analysis where individual-level identifiers are available to track subjects over time. Chapter 4 filled a critical gap in the literature by proposing a class of AMMs that model fixed effects, random effects and spatial effects simultaneously for Gaussian distributed responses while incorporating kernel-based smoothers. Chapter 5 further relaxed the restriction on the distribution of response and constructed GAMMs that incorporate kernel-based smoothers. Chapters 4 and 5 combined could be viewed as a kernel-based alternative to the work of Lin and Zhang (1999), offering an option for researchers who aim to use LOESS or other kernel smoothers in mixed models.

We would also like to mention that, other than a frequentist GAM framework, Bayesian dis-

ease mapping techniques, which commonly utilize Gaussian process or other types of stochastic processes for spatial effects estimation, remain popular in spatial analyses. Bayesian methods can often be more flexible if a hierarchical structure is well adopted. Bayesian methods also enjoy direct inference on random effects via a unified framework for inference on the whole model without approximate derivation or backfitting procedures. Nevertheless, Bayesian methods require good experience in model setup, prior distribution selection and MCMC tuning, all of which are hardly trivial to non-statisticians or even statisticians who have little expertise in geospatial analysis. In addition, when handling large datasets, the methods presented in this dissertation are generally more scalable.

Future directions from the work proposed here are abundant and clear. The GAMMs developed in Chapter 5 utilized a PQL procedure. Hence to complete the whole framework, similar work could be done utilizing a MQL procedure (Breslow and Clayton, 1993; Goldstein, 1991). Also, we noticed that in mixed models with exponential family responses, and Bernoulli response in particular, the uncertainty in marginal trend parameter was often under-estimated. We believe a marginal modeling approach based upon generalized estimation equations are worth investigating is inference on marginal model parameters is of direct scientific interest. Finally, this work did not consider the modeling of censored data. Therefore incorporation of the proposed methods into a survival analysis framework, and a Cox proportional hazard modeling framework in particular, should be further investigated.

Bibliography

Hirotogu Akaike. Information theory and an extension of the maximum likelihood principle. In *Selected papers of hirotugu akaike*, pages 199–213. Springer, 1998.

Lu Bai, Scott Bartell, Robin Bliss, and Veronica Vieira. Mapgam-package: Mapping smoothed effect estimates from individual-level... 2019.

Sudipto Banerjee, Bradley P Carlin, and Alan E Gelfand. *Hierarchical modeling and analysis for spatial data*. CRC press, 2014.

Scott M Bartell, Antonia M Calafat, Christopher Lyu, Kayoko Kato, P Barry Ryan, and Kyle Steenland. Rate of decline in serum pfoa concentrations after granular activated carbon filtration at two public water systems in ohio and west virginia. *Environmental health perspectives*, 118(2):222–228, 2010.

Leo Breiman and Jerome H Friedman. Estimating optimal transformations for multiple regression and correlation. *Journal of the American statistical Association*, 80(391):580–598, 1985.

Norman E Breslow and David G Clayton. Approximate inference in generalized linear mixed models. *Journal of the American statistical Association*, 88(421):9–25, 1993.

Robert E Bristow, Jenny Chang, Argyrios Ziogas, Daniel L Gillen, Lu Bai, and Veronica M Vieira. Spatial analysis of advanced-stage ovarian cancer mortality in california. *American journal of obstetrics and gynecology*, 213(1):43–e1, 2015.

Chris Brunsdon, A Stewart Fotheringham, and Martin E Charlton. Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical analysis*, 28(4): 281–298, 1996.

William S Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association*, 74(368):829–836, 1979.

Noel Cressie. The origins of kriging. *Mathematical geology*, 22(3):239–252, 1990.

Noel Cressie. Statistics for spatial data. *Terra Nova*, 4(5):613–617, 1992.

Peter J Diggle and Paulo J Ribeiro Jr. Bayesian inference in gaussian model-based geostatistics. *Geographical and Environmental Modelling*, 6(2):129–146, 2002.

Peter J Diggle, Paulo J Ribeiro, and Ole F Christensen. An introduction to model-based geostatistics. In *Spatial statistics and computational methods*, pages 43–86. Springer, 2003.

Jean Duchon. Splines minimizing rotation-invariant semi-norms in sobolev spaces. *Constructive theory of functions of several variables*, pages 85–100, 1977.

Andrew O. Finley, Sudipto Banerjee, and Bradley P. Carlin. spBayes: An R package for univariate and multivariate hierarchical point-referenced spatial models. *Journal of Statistical Software*, 19(4):1–24, 2007. URL `http://www.jstatsoft.org/v19/i04/`.

Andrew O. Finley, Sudipto Banerjee, and Alan E.Gelfand. spBayes for large univariate and multivariate point-referenced spatio-temporal data models. *Journal of Statistical Software*, 63(13):1–28, 2015. URL `http://www.jstatsoft.org/v63/i13/`.

Alan E Gelfand and Sudipto Banerjee. Bayesian modeling and analysis of geostatistical data. *Annual Review of Statistics and Its Application*, 4:245–266, 2017.

Mariam S Girguis, Matthew J Strickland, Xuefei Hu, Yang Liu, Scott M Bartell, and Verónica M Vieira. Maternal exposure to traffic-related air pollution and birth defects in massachusetts. *Environmental research*, 146:1–9, 2016.

Harvey Goldstein. Nonlinear multilevel models, with an application to discrete response data. *Biometrika*, pages 45–51, 1991.

Gene H Golub, Michael Heath, and Grace Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, 1979.

Mark S Handcock and Michael L Stein. A bayesian analysis of kriging. *Technometrics*, 35 (4):403–410, 1993.

Trevor Hastie and Robert Tibshirani. *Generalized additive models*. Wiley Online Library, 1990.

Robert E Kass and Adrian E Raftery. Bayes factors. *Journal of the american statistical association*, 90(430):773–795, 1995.

Nan M Laird and Thomas A Louis. Approximate posterior distributions for incomplete data problems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2): 190–200, 1982.

Nan M Laird and James H Ware. Random-effects models for longitudinal data. *Biometrics*, pages 963–974, 1982.

Kung-Yee Liang and Scott L Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986.

Xihong Lin and Daowen Zhang. Inference in generalized additive mixed modelsby using smoothing splines. *Journal of the royal statistical society: Series b (statistical methodology)*, 61(2):381–400, 1999.

Mary J Lindstrom and Douglas M Bates. Nonlinear mixed effects models for repeated measures data. *Biometrics*, pages 673–687, 1990.

Peter McCullagh. *Generalized linear models*. Routledge, 2018.

Henning Omre. Bayesian kriging—merging observations and qualified guesses in kriging. *Mathematical Geology*, 19(1):25–39, 1987.

Henning Omre and Kjetil B Halvorsen. The bayesian bridge between simple and universal kriging. *Mathematical Geology*, 21(7):767–786, 1989.

Michael L Stein. *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media, 2012.

Yannan Tang, Verónica M Vieira, Scott M Bartell, and Daniel L Gillen. Additive mixed models with kernel smoothers for disease mapping using individual-level data. 2020.

Florin Vaida and Suzette Blanchard. Conditional akaike information for mixed-effects models. *Biometrika*, 92(2):351–370, 2005.

Verónica Vieira, Thomas Webster, Janice Weinberg, and Ann Aschengrau. Spatial analysis of bladder, kidney, and pancreatic cancer on upper cape cod: an application of generalized additive models to case-control data. *Environmental Health*, 8(1):3, 2009.

Grace Wahba. Spline bases, regularization, and generalized cross-validation for solving approximation problems with large quantities of noisy data. *Approximation theory III*, 2, 1980.

Thomas Webster, Verónica Vieira, Janice Weinberg, and Ann Aschengrau. Method for mapping population-based case-control studies: an application using generalized additive models. *International Journal of Health Geographics*, 5(1):26, 2006.

Russ Wolfinger and Michael O'connell. Generalized linear mixed models a pseudo-likelihood approach. *Journal of statistical Computation and Simulation*, 48(3-4):233–243, 1993.

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

Simon N Wood. Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, 99(467):673–686, 2004.

Simon N Wood. 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, 2011.

Simon N Wood. *Generalized additive models: an introduction with R*. CRC press, 2017.

Simon N Wood, Mark V Bravington, and Sharon L Hedley. Soap film smoothing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):931–955, 2008.

Simon N Wood, Natalya Pya, and Benjamin Säfken. Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association*, 111 (516):1548–1563, 2016.

Robin L Young, Janice Weinberg, Verónica Vieira, Al Ozonoff, and Thomas F Webster. Generalized additive models and inflated type i error rates of smoother significance tests. *Computational statistics & data analysis*, 55(1):366–374, 2011.

Appendix A

PMSD tests under correlation

In Chapter 3, we described a class of PMSD tests that identify temporal heterogeneity in spatial effects. Since the PMSD test arises from the permutation distribution, the test relies on the strong exchangeability assumption of time labels. Under the strong null hypothesis, exchangeability across time is assumed in both mean and variance components and hence, it is not guaranteed that the class of test would perform as well on data generated with a non-exchangeable correlation structure. For instance, when only random intercepts exist in a longitudinal dataset, the correlation structure is exchangeable across time, which is not the case when random slopes exist. In addition, even if independence holds but the magnitude of the variance does not stay invariant over time, the exchangeability assumption is questionable.

A.1 Calibrated PMSD tests

In Figure A.1, 2 simulated examples violating the exchangeability assumption of the variance components are shown. Independent errors with time-varying variance are shown in the left

plot while random intercepts and random slopes are displayed in the right plot, resulting in time-correlated errors with inconstant variance.

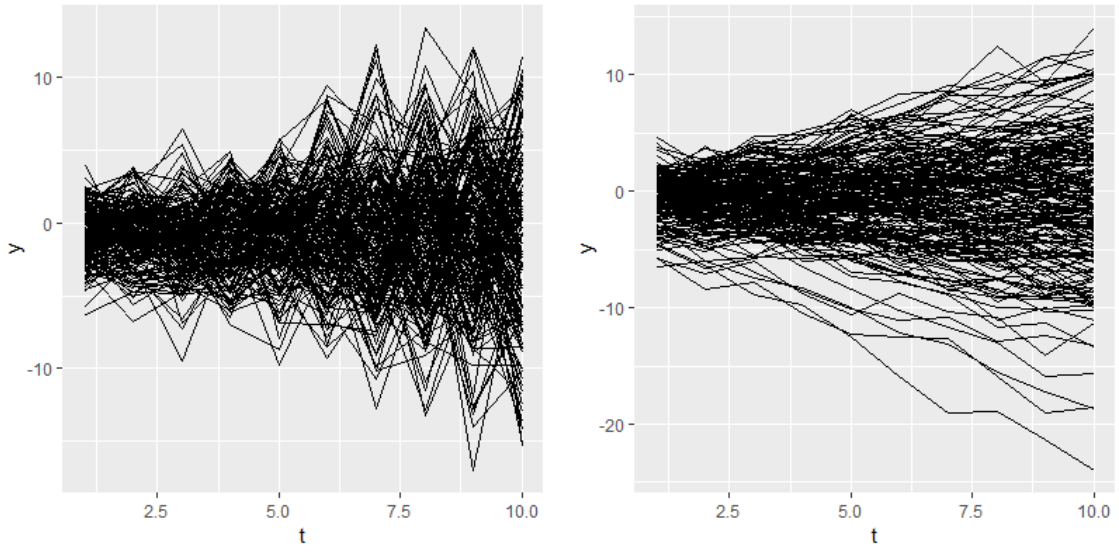

Figure A.1: Simulated random errors over time. Left: Inconstant variance; Right: Random intercept and slope.

Simulation studies using these simulated errors instead of i.i.d. errors, as considered in the Monte Carlo studies of Chapter 3, showed inflated type I errors if the PMSD test is naively applied (see columns labeled with "naive" in Table A.1). To calibrate the tests under these scenarios, we propose a standardization procedure to prepare working datasets for the PMSD test. The modified PMSD test is sketched as follows:

1. Fit $y_{ij} = X_{ij}\beta + lo(u_{ij}, v_{ij}) + \epsilon_{ij}$.
2. Residuals e_{ij}.
3. $\tilde{e}_{ij} = \underset{i}{std}(e_{ij})$.
4. Perform PMSD test using $\hat{lo}_{ij} + \tilde{e}_{ij}$ as response.

for non-constant variance over time. For a random effects variance structure, the following is proposed:

1. Fit $y_{ij} = X_{ij}\beta + lo(u_{ij}, v_{ij}) + \epsilon_{ij}$.

2. Fit $e_{ij} = Z_{ij}b_i + \epsilon_{ij}$

3. Residuals $e_{ij}^{(2)} = e_{ij} - Z_{ij}\hat{b_i}$

4. $\tilde{e}_{ij}^{(2)} = \underset{i}{std}(e_{ij}^{(2)})$

5. Perform PMSD test using $\hat{lo}_{ij} + \tilde{e}_{ij}^{(2)}$ as response

Notations are defined in a similar way to Chapters 3 and 4. The key idea of the procedures outlined above is to standardize the residuals at each time point so that approximate homoscedasticity is created across time. These procedures managed to calibrate the PMSD test in theory. This claim was further supported by the results of simulation studies (see columns with "calibrated" in Table A.1).

Table A.1: Empirical type I errors of naive and calibrated PMSD tests under differing covariance constructions. Correlated data were simulated using random effects (R.E.) while independent (ind.) data were simulated to match the magnitude of variance at each time point.

(quasi)τ_0^2	(quasi)τ_0^2	ind. naive	ind. calibrated	R.E. naive	R.E. calibrated
1	0.05	0.09	0.06	0.02	0.02
3	0.09	0.08	0.06	0.11	0.05
3	0.15	0.10	0.06	0.13	0.08
5	0.15	0.14	0.06	0.07	0.03
5	0.25	0.09	0.05	0.17	0.08
7	0.21	0.06	0.04	0.09	0.05
7	0.35	0.11	0.08	0.24	0.07
9	0.27	0.09	0.06	0.14	0.03
9	0.45	0.06	0.04	0.13	0.06

Further simulations were performed to investigate power of the calibrated PMSD tests. Similar to Chapter 3, shifted patterns were used to generate heterogeneity across 2 time points. The resulting powers were are in Table A.2.

Table A.2: Empirical powers of parametric LRT and calibrated PMSD given differing shift amount of spatial effects. $\tau_0^2 = 1$, $V_1^2 = 0.03$, $\sigma^2 = 1$. Both tests delivered valid type I errors (0.06 and 0.07). Powers becomes greater along with the shifting amount due to more heterogeneity. Calibrated PMSD test is more powerful, though not by much, in this scenario.

shift	parametric LRT	calibrated PMSD
0.00	0.07	0.06
0.04	0.08	0.13
0.08	0.22	0.30
0.12	0.43	0.53
0.16	0.71	0.74
0.20	0.90	0.94

A.1.1 Performance of calibrated PMSD given various settings of random effects

We considered multiple simulations to investigate the performance of the proposed calibrated PMSD test. Using various values of random intercepts and slopes, we present the study design and corresponding empirical p-values under the null hypothesis in Figure A.2.

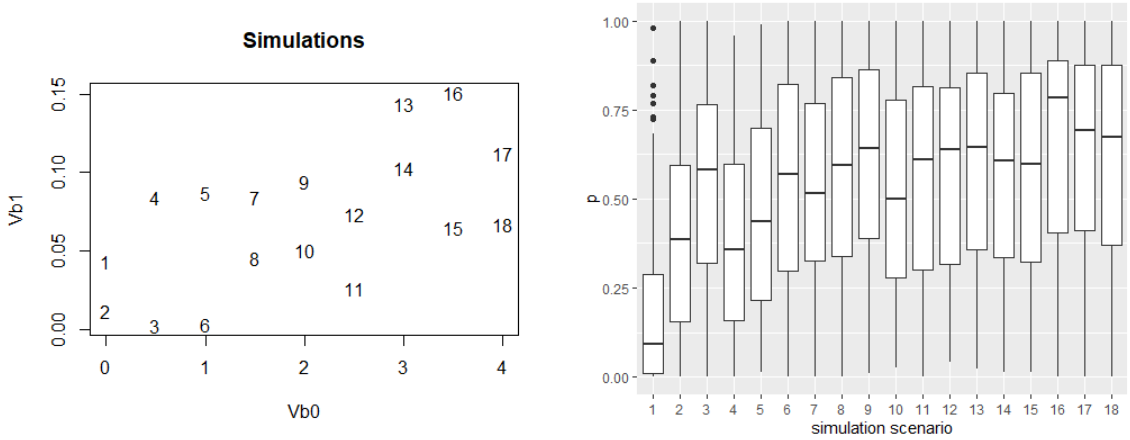

Figure A.2: Left: simulated random effects of each simulation scenario; Right: boxplots of p-values given each scenario. Scenario 1, 2, 4, and 5 could be potentially problematic.

As can be seen from Figure A.1, some of the scenarios could be problematic. We further show the empirical type I errors of the calibrated PMSD test and approximate F-tests given by Hastie and Tibshirani (1990), comparing models with a marginal smoother and a time-

stratified smoother.

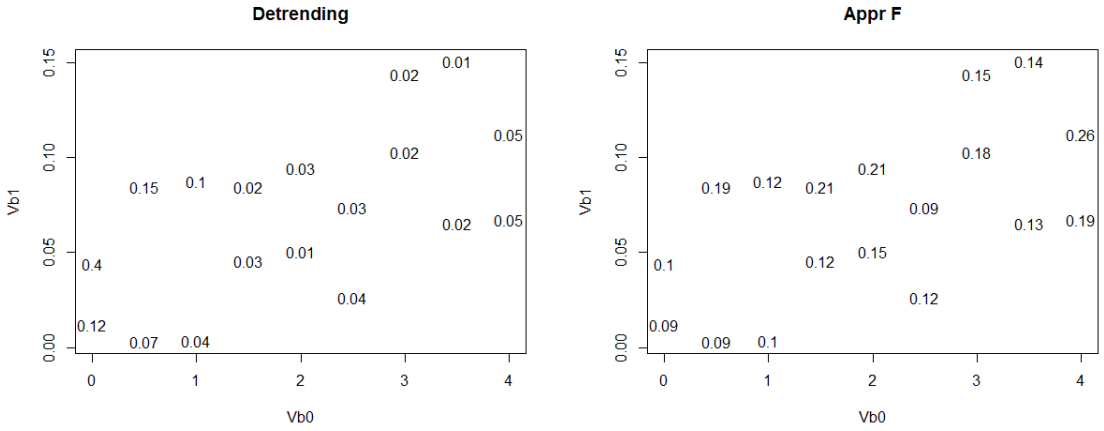

Figure A.3: Left: type I errors of calibrated PMSD tests under differing random effects scenarios. Scenario 1, 2, 4, and 5 have high $\frac{Vb1}{Vb0}$ values; Right: type I errors of approximate F tests under differing random effects scenarios. Inflated type I errors are observed.

From Figure A.3, a potential pattern emerges in that large values of $\frac{Vb1}{Vb0}$ ($\frac{\tau_1^2}{\tau_0^2}$) induced inflated type I errors. This pattern is further confirmed by another set of simulations shown in Figure A.4.

Further, according to a brief empirical assessment of studies where both random intercepts and random slopes exist, we found no examples where the $Vb1 > 0.05Vb0$ ($\tau_1^2 > 0.05\tau_0^2$) hence a new class of simulations were performed with control of τ_1^2. The results were presented in Figure A.5.

In comparison, results of the naive PMSD test and modified PMSD test with detrending but not standardization are shown in Figure A.6, showing that our calibrated PMSD tests appears to be the most promising when compared to these two methods.

Extended simulation studies with 4 time points rather than 2 were performed and results are displayed in Figure A.7 and Table A.3.

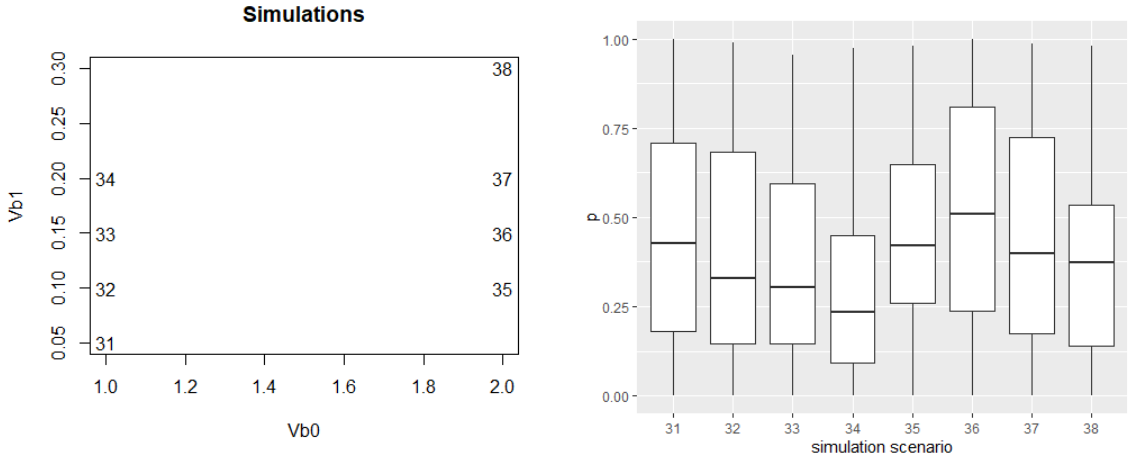

Figure A.4: More simulations to confirm the assumed pattern. Left: simulated random effects of each simulation scenario; Right: boxplots of p-values given each scenario. Hence it could be found that for a given $Vb0$, greater $Vb1$ values render greater empirical type I errors.

A.1.2 Power of calibrated PMSD test under multiple types of scenarios

In this section, we present our exploration on power of the calibrated PMSD test. Multiple simulation studies are presented. Datasets were created with spatial effects, random intercepts and random slopes. Our proposed calibrated PMSD test, a parametric LRT test and an approximate F-test based on a GAMM with thin-plate splines were used to test for temporal heterogeneity of spatial effects. One thing to note is that GAMM with stratified thin-plate splines from MGCV R package failed to fit the simulated data due to convergence issues in multiple cases.

First, we use the same spatial effects on a square map as Figure 3.1 in Chapter 3. Similar shifting amounts were used and the corresponding results were shown in Table A.4.

Further simulations are based on the map of California (cf. Figure A.8).

We induce a mitigation parameter, referred to henceforth as a "multiplier". The spatial

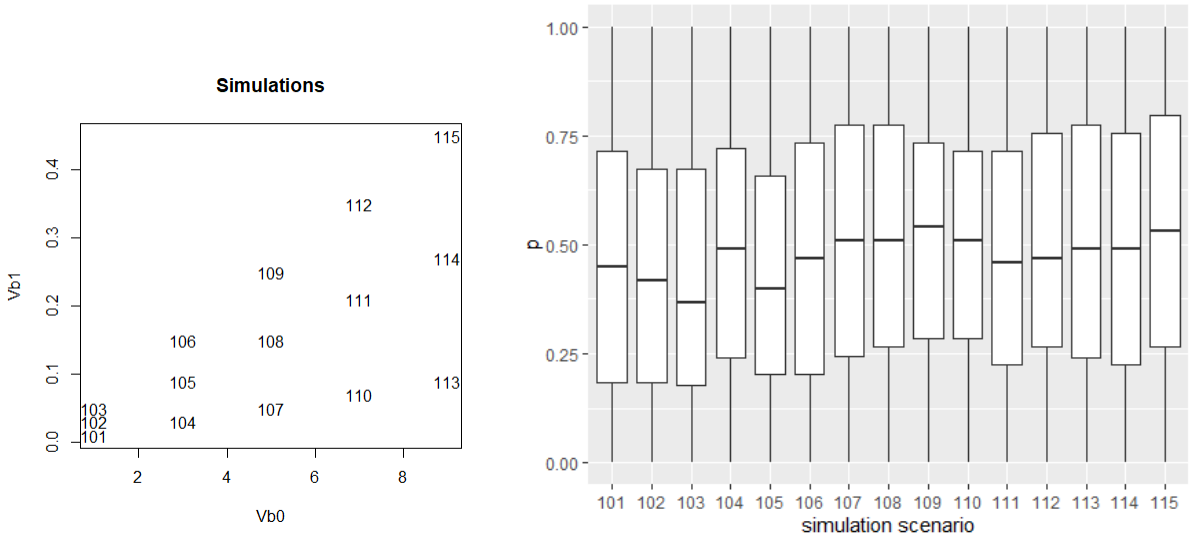

Figure A.5: Additional simulations with controlled τ_1^2 where $\tau_1^2 < 0.05\tau_0^2$. Left: simulated random effects of each simulation scenario; Right: boxplots of p-values given each scenario. Empirical p-values roughly follows a uniform distribution for each scenario and the type I errors are reasonably controlled.

effects at Time 2 would be the products of effects at Time 1 and the parameter hence a multiplier that is less than 1 would imitate the mitigation of spatial effects. The corresponding results could be found in Table A.5.

Another type of temporal heterogeneity was created by moving the spatial pattern rather than mitigating. Specifically, the pattern of Time 2 was created by moving the spatial pattern to the red point in the right plot of A.8 by a proportion of the distance between the Gaussian centers and the destination (red point). Hence proportion 0 indicates no moving at all while proportion 1 would result in a Time 2 pattern which is created by all 3 Gaussian pdfs effects centered at the red point. Results were shown in Table A.6.

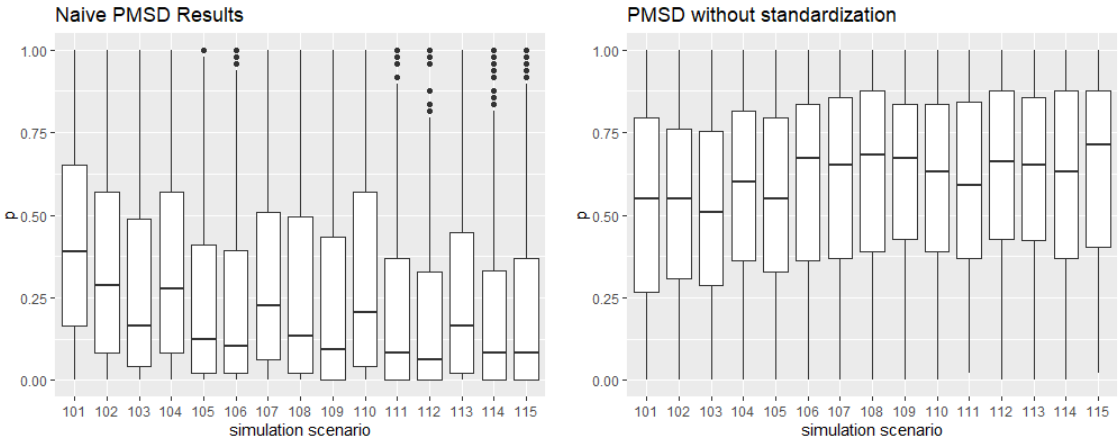

Figure A.6: Boxplots of empirical p-values. Left: 1st paper method, rendering inflated type I errors; Right: without standardization, rendering shrunk type I errors.

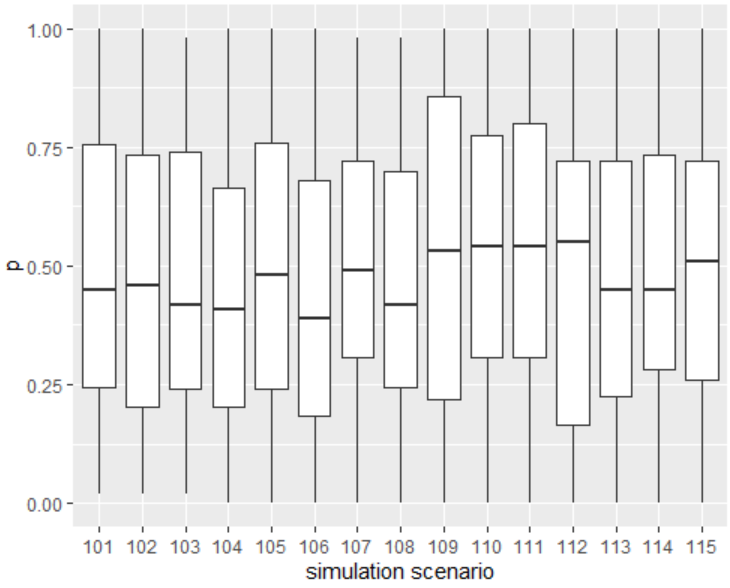

Figure A.7: Boxplots of empirical p-values from simulations with 4 time points.

Table A.3: Empirical type I errors of calibrated PMSD tests under differing simulation scenarios based on 100 repetitions. Type I errors were generally well controlled.

sim.	type I Er.
101	0.05
102	0.08
103	0.05
104	0.06
105	0.05
106	0.09
107	0.05
108	0.03
109	0.07
110	0.07
111	0.08
112	0.05
113	0.07
114	0.04
115	0.07

Table A.4: Power of the 3 tests using $\tau_0^2 = 1$, $\tau_1^2 = 0.05$ and $\sigma^2 = 1$ based on 600 repetitions. MGCV renders NA 25 times. In this scenario, approximate F-test based on thin-plate splines appears to be most powerful.

shift	calib. PMSD	para. LRT	TP-F
0.00	0.03	0.02	0.06
0.04	0.13	0.09	0.14
0.08	0.19	0.18	0.32
0.12	0.39	0.37	0.57
0.16	0.73	0.71	0.92
0.20	0.86	0.90	0.96

Table A.5: Power of the 3 tests using $\tau_0^2 = 2$, $\tau_1^2 = 0.08$ and $\sigma^2 = 1$. In this scenario, approximate F-test based on thin-plate splines appears to be most powerful.

calib. PMSD	LRT	TP-F	multiplier
0.08	0.10	0.05	1.00
0.20	0.05	0.13	0.97
0.13	0.08	0.58	0.94
0.15	0.08	0.82	0.91
0.30	0.17	0.90	0.88

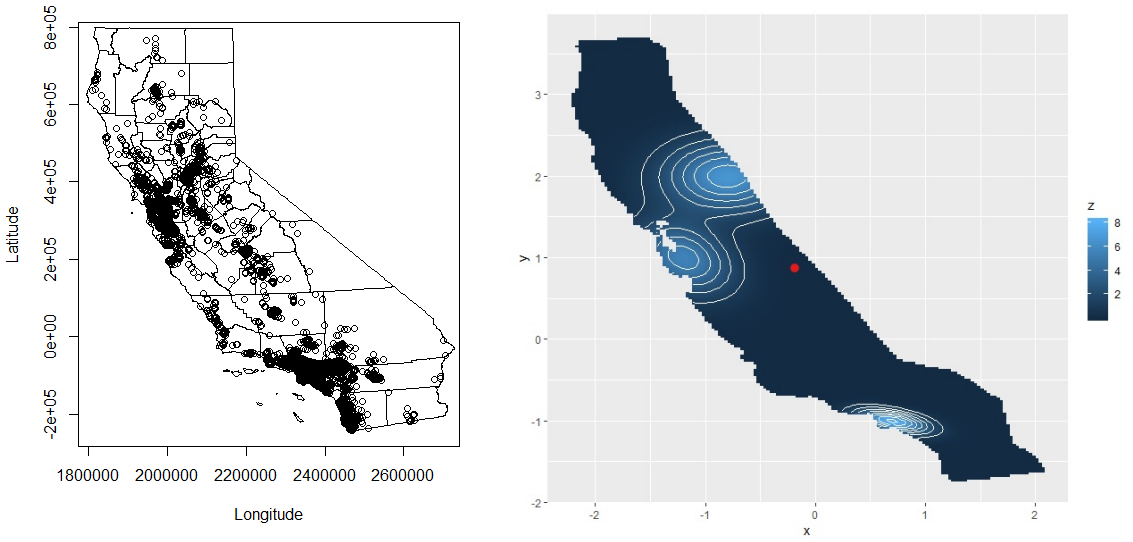

Figure A.8: Simulation settings using map of California. Left: spatial distribution of observations; Right: simulated spatial risk pattern (using pdf of bivariate Gaussian).

Table A.6: Powers of the 3 tests using $\tau_0^2 = 2$, $\tau_1^2 = 0.08$ and $\sigma^2 = 1$. In this scenario, calibrated PMSD test appears to be most powerful.

Moving proportion	calib. PMSD	LRT	TP-F
0.000	0.02	0.03	0.03
0.005	0.11	0.06	0.13
0.010	0.37	0.16	0.32
0.015	0.82	0.27	0.38
0.020	0.90	0.53	0.73

Appendix B

Smoothing parameter selection criteria in Chapter 5

In Chapter 5, we introduced a novel class of model selection criteria named out-of-sample likelihood Monte Carlo (OLMC) to choose the smoothing parameter (span size) for our proposed model. It was noted in Chapter 5 that neither AIC (Akaike, 1998) or conditional AIC (cAIC) (Vaida and Blanchard, 2005) rendered satisfying performance in span size selection. In this section we present relevant simulation results to support this claim. We consider a model of the form:

$$\text{logit}(p_{ij}) = 0.2x_{ij} - 0.1t_{ij} + cs_0(u_{ij}, v_{ij}) + b_{0i} + b_{0i}t_{ij}\epsilon_{ij} + d_{ij}, \tag{B.1}$$

where notations are similarly defined as in Chapter 5, c is a parameter to control the magnitude of spatial risk pattern, $d_{ij} \overset{i.i.d.}{\sim} Unif(-h, h)$ and $\epsilon_{ij} \overset{i.i.d.}{\sim} N(0, \sigma^2)$. The simulation scenarios are shown in Table B.1. For each simulation run under each scenario, 5 span sizes (ranging from 0.1 to 0.5) were tried and all three model selection criteria (OLMC, AIC, cAIC) were compared while the span size with the least squared spatial effects prediction

errors was considered "optimal". The optimal spans were generally 0.2 or 0.3 across all the simulations (28 out of 34). AIC was calculated using

$$AIC = 2k - 2log(\hat{L}|\hat{\beta}, \hat{\tau}), \tag{B.2}$$

where k is the model degree of freedom and L is the likelihood while cAIC was calculated using

$$cAIC = 2k - 2log(\hat{L}|\hat{\beta}, \hat{b}). \tag{B.3}$$

For each simulation, the three methods were labeled as "correct" if the optimal span was successfully chosen, "close" if the selected span differed from the optimal by 0.1 and "incorrect" otherwise. From the results given in B.2, our proposed OLMC rendered the best performance in selecting the "right" amount of smoothing.

Table B.1: Simulation scenarios for span selection criteria comparison.

Index	# of runs	$N_{subjects}$	# of time points	c	τ_0^2	τ_1^2	σ^2	h
1	6	600	12	2	0.3	0.0225	0.49	1.2
2	4	600	8	2	0.3	0.0225	0.49	1.2
3	6	700	20	2	0.4	0.01	0.64	0
4	6	700	20	2	0.4	0.01	0	0
5	6	700	20	1	0.25	0.0049	1	0
6	6	700	20	1	0.25	0.0049	0	0

Table B.2: Results of the 3 types of span selection criteria.

	correct	close	incorrect
OLMC	14	12	8
AIC	9	16	9
cAIC	2	12	20