On Issues of Scale and Dependence in Spatial and Spatio-Temporal Data

by

Marco Henry Benedetti

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2019

Doctoral Committee:

Associate Professor Veronica J. Berrocal, Chair
Professor Roderick J. Little
Professor Bhramar Mukherjee
Professor Marie S. O'Neill

Marco Henry Benedetti
benedetm@umich.edu
ORCID iD: 0000-0002-6334-1975

This dissertation is dedicated to my parents,
Carolyn and Costantino Benedetti

ACKNOWLEDGEMENTS

First and foremost, I wish to express my sincere gratitude my committee, Professors Bhramar Mukherjee, Rod Little, Marie O'Neill, and especially my advisor Veronica Berrocal for their guidance throughout my dissertation. I am truly fortunate to have learned from this distinguished group of individuals and will carry their lessons forward in my career.

Next, there are a number of individuals at both the University of Michigan and Grand Valley State University who have played formative roles in my education and professional development, including Paul Stevenson, Kathleen Klinich, Carol Flannagan, Daniel Blower, Deb Mexicotte, and Timothy Johnson. I am extremely grateful for their support and instruction, and I have many great memories of the time we spent working together. In addition, I would like to thank Professors Kelley Kidwell, Alfred Franzblau, and Emily Youatt for their support and supervision of me as a Graduate Student Instructor. Teaching has been among my most rewarding experiences at the University of Michigan and I have learned so much from observing and interacting with them.

Next, I wish to thank the incredible friends I have had during my time in Ann Arbor, especially Jerry, Grace, Zack, Blake, Emily, Evan, Riley, Nick, Jeremy,

Lauren, Woody, Allison, Joe, Julie, and Chris. Their steadfast friendship and support is what I will remember most fondly when I reflect on the last six years.

I will conclude by expressing my love and gratitude to my family: my parents, Carolyn and Costantino Benedetti, to whom this dissertation is dedicated; my siblings Giulia, Cristina, Tino, and Sarah; my nephew Luca, and my nieces Sofia and Silvia. I could not have asked for better supporters and role models than my parents and siblings, and I am eager to return home to them. I especially want to thank my sisters Giulia and Cristina, both brilliant academics whom I look up to, and who have endured countless hours listening to me describe the many trials and tribulations of graduate school.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

ABSTRACT

Recent years have seen a massive increase in the number of publicly available spatial and spatio-temporal datasets. With these data comes a set of practical challenges, especially when researchers use spatial statistical models to generate predictions or synthesize datasets with differing spatial resolutions. At the basis of these models lies the notion of spatial scale which, for a stationary and isotropic covariance, is quantified through a range parameter which captures the distance at which observations are considered independent in space. In this dissertation, we propose a set of statistical methods to investigate issues related to the scale of spatial data, with the goal of providing a better characterization of the dependence structure of a spatial process. These methods are used to generate improved predictions and to generate estimates at the needed spatial resolution. Furthermore, several methods are developed to account for the sampling mechanism of the data, whether they are derived through surveys or from non-probabilistic samples such as electronic health records (EHRs).

In Chapter 2, building upon the Multi-resolution Approximation (M-RA) for large spatial data (Katzfuss, 2017), and leveraging the relationship between levels

of the M-RA and the scale of a spatial process, we develop a Bayesian hierarchical model that explores and accommodates non-stationarity in spatial processes. In contrast to several existing tests for global non-stationarity, our model can detect regions of local stationarity through the specification of a mixture of multivariate normal priors on the basis function weights of the M-RA. Furthermore, our model outperforms other standard spatial statistical models in terms of out-of-sample prediction.

In Chapter 3, we present a model for disaggregating to a fine spatio-temporal resolution estimates of proportions derived from the American Community Survey (ACS). We envision that disaggregated estimates will be better proxies of neighborhood exposure than the ACS estimates, which are resolved at either a fine spatial resolution and coarse temporal scale, or at a coarse spatial resolution and fine temporal scale. By characterizing the data as an aggregation of an underlying point-referenced process, we disaggregate the ACS estimates to the 1-year census tract resolution. Crucial to our methodological development is the incorporation of the surveys design effect. A secondary development is a spatio-temporal version of the M-RA.

In Chapter 4, we extend the disaggregation model of the previous chapter to accommodate estimates of count-valued characteristics. This chapter contains a comparison to the model of Bradley et al. (2016b) (the BWH model), which addresses a similar problem for purely spatial data. In addition to accommodating spatio-temporal data, our model differs from the BWH model by incorporating the survey design effect into the model specification. We find that our model outper-

forms the BWH model in terms of prediction accuracy and coverage probability.

In Chapter 5, we address the issue of sampling bias in EHR data, which can arise in studies of the association between disease and exposure when both the outcome variable and the exposure process are related to the process determining sample selection. Our method jointly models EHR and publicly available data to approximate sampling probabilities, which are then used to derive sampling weights. We show via simulation studies that we can recover data generating sampling probabilities and reduce bias compared to a naive analysis. To illustrate the utility of our model with clinical data, we present an analysis of smoking and lung cancer using subjects in the Michigan Genomics Initiative.

CHAPTER I

Introduction

With the widespread availability of geographical information systems (GIS), recent decades have witnessed a tremendous increase in the number of epidemiological studies investigating the effect on population health of individuals' social and physical environment. Environmental data for these types of studies are often gathered by leveraging multiple sources, each reporting data at different geographic, or *spatial*, resolutions. In addition, locations at which environmental data are collected rarely correspond directly to the residential locations of subjects in the study, thus requiring prediction of environmental exposures at unobserved locations. Because physical and social environmental data vary spatially, it is common practice to use spatial statistical methods to model these data, characterize their spatial dependence, generate predictions at unsampled locations, or provide estimates at the desired spatial resolution.

At the basis of spatial statistics is the concept that spatial dependence between random variables corresponding to different locations or areal units can be captured through a covariance function that depends on their geographical positioning; in

some cases (i.e. for stationary and isotropic covariance functions), it is simply a function of the separation or proximity between locations/areal units. In the case of point-referenced data and stationary covariance functions, separation between locations is quantified through their distance and the angle between locations. On the other hand, for areal data whose spatial dependence is characterized as stationary, the correlation depends on whether areal units are adjacent or not, that is, whether they share a common boundary or not. Closely related to the notion of covariance function/spatial dependence of a spatial process is the concept of scale of a spatial process or range of the covariance function, which indicates the distance at which sites or areal units can be considered independent. In most models used to analyze spatial data, the scale/range is assumed to be a global property of the spatial process and thus it is postulated to not vary spatially.

However, when dealing with point-referenced data, the assumption that the covariance function depends only on the distance between points and admits a global range, i.e. the assumption that the covariance function is stationary, might be an unrealistic one which does not represent the true, complex dependence structure of the process. This might be true particularly for geophysical processes, such as air pollution, temperature, etc., which are affected by external factors such as topography, weather, and urbanicity.

The overarching theme of this dissertation is to propose statistical methods to investigate issues related to the scale of spatial processes and spatial fields with the goal of providing a better characterization of their dependence structure. These methods will in turn be used to generate improved predictions of a point-referenced

spatial process at unsampled sites and generate estimates of areal-level data at the needed spatial resolution. In terms of application, the motivation for this work is to infer upon environmental and social factors that affect health at the scale of interest for epidemiological analyses.

1.1 Introduction to spatial statistical models

The following section provides a brief overview of topics in spatial statistics that are relevant to the projects presented in Chapters II through V. Specifically, we will quickly introduce statistical models for both point-referenced and areal data, and, in the case of point-referenced data, we will show how these models can be extended to accommodate for data over time. In addition, we present an introduction to spatial point process models. For the interested reader, more details on these topics can be found in Banerjee et al. (2004), Cressie (1993), Cressie and Wikle (2011), Gelfand et al. (2010), and Waller and Gotway (2004).

1.1.1 Point-referenced spatial data

Spatial data for which the geographical coordinates of the observation locations (e.g. latitude and longitude, or Easting and Northing) are available are typically referred to as *point-referenced* or *geostatistical data*. For this type of data, observation locations are assumed to be fixed and belonging to a continuous domain $\mathcal{S} \subset \mathbb{R}^d$ with d typically equal to 2 or 3. The common notation adopted for a point-referenced datum is $y(\mathbf{s})$ to explicitly indicate that the observation of a random variable Y was obtained at location $\mathbf{s}$ in $\mathcal{S}$. The classical modeling approach

used for point-referenced data decomposes $y(\mathbf{s})$, $\mathbf{s} \in \mathcal{S}$, into the sum of 3 terms:

$$y(\mathbf{s}) = \mu(\mathbf{s}) + w(\mathbf{s}) + \epsilon(\mathbf{s}) \qquad \epsilon(\mathbf{s}) \overset{iid}{\sim} N(0, \tau^2) \tag{1.1}$$

In (1.1), $\mu(\mathbf{s})$ represents the mean of the random variable $Y(\mathbf{s})$ and it accounts for the large-scale spatial trend in the spatial process $\{Y(\mathbf{s}) : \mathbf{s} \in \mathcal{S}\}$; typically, $\mu(\mathbf{s})$ is modeled to be a function of covariates $\mathbf{X}(\mathbf{s})$, e.g. $\mu(\mathbf{s}) = h(\mathbf{X}(\mathbf{s}), \boldsymbol{\beta})$. The second term in (1.1), $w(\mathbf{s})$, accounts for the small-scale spatial variation in the spatial process $\{Y(\mathbf{s}) : \mathbf{s} \in \mathcal{S}\}$ not accounted for by the mean function $\mu(\mathbf{s})$. Following nomenclature used in longitudinal data analysis, $w(\mathbf{s})$ is also deemed as a spatial random effect, and represents the deviation from the overall mean of the process, $\mu(\mathbf{s})$, associated with location $\mathbf{s} \in \mathcal{S}$. Finally, $\{\epsilon(\mathbf{s}) : \mathbf{s} \in \mathcal{S}\}$ is an independent error process, independent of $\{w(\mathbf{s}) : \mathbf{s} \in \mathcal{S}\}$, that accounts, among others, for measurement error. The variance τ^2 of $\epsilon(\mathbf{s})$ is interpreted as the non-spatial variability in the spatial process $Y(\mathbf{s})$ and is often referred to as *nugget effect* in the geostatistical literature.

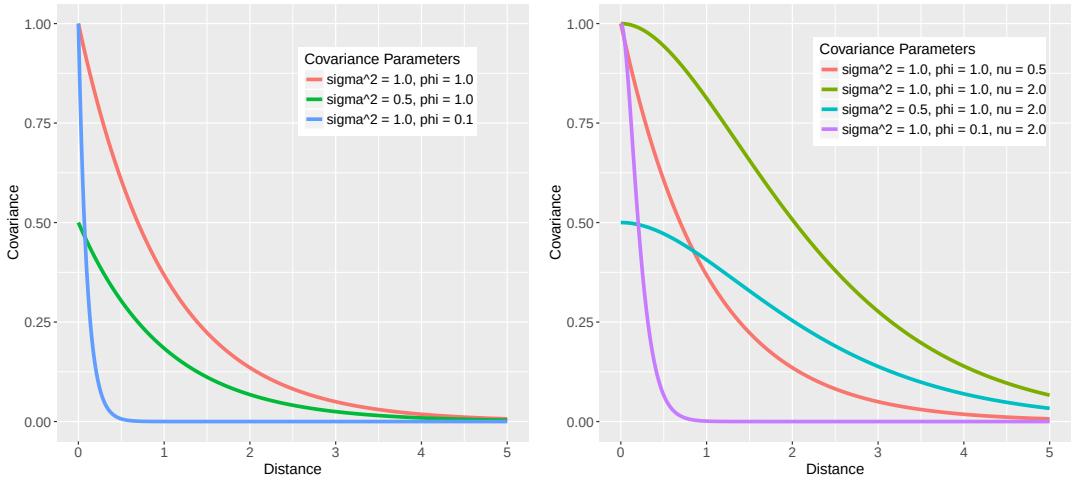
For spatial data that, marginally, at each location, can be thought of as following a Gaussian distribution, the spatial random field $\{w(\mathbf{s}), \mathbf{s} \in \mathcal{S}\}$ is modeled as a mean-zero Gaussian process. That is, $w(\mathbf{s})$, $\mathbf{s} \in \mathcal{S}$, is a stochastic process indexed by location, for which all finite-dimensional realizations are distributed according to a multivariate normal distribution. As a multivariate normal distribution is completely specified by its mean and covariance matrix, the mean-zero Gaussian process $w(\mathbf{s})$, $\mathbf{s} \in \mathcal{S}$, is completely determined once the covariance function

$$C(\mathbf{s}, \mathbf{s}') := \mathrm{Cov}\left(w(\mathbf{s}), w(\mathbf{s}')\right)$$

is defined for each pair of points $\mathbf{s}, \mathbf{s}' \in \mathcal{S}$.

A simplifying assumption that is often used for covariance functions is to assume that they depend only on the distance between the two locations, that is, only on $d(\mathbf{s}, \mathbf{s}')$. In this case, the covariance function is called *stationary* and *isotropic* and the spatial process is named a *stationary, isotropic spatial process*, more precisely a *second-order stationary, isotropic spatial process* if the mean of the spatial process is constant in space. In routine spatial statistical analysis of continuous point-referenced data, $w(\mathbf{s})$ in (1.1), is taken to be a mean-zero, second order stationary, isotropic Gaussian process.

Several parametric models exist for isotropic covariance functions: spherical, exponential, Gaussian, Matèrn, etc. (for more details, please refer to Banerjee et al. (2004), Cressie (1993), etc.); here we just present the exponential and Matèrn covariance function for illustration. The exponential covariance function states that the covariance between a spatial process at two locations $\mathbf{s}$ and $\mathbf{s}'$ decays exponentially with distance. On the other hand, the Matèrn covariance function is a general covariance function that admits many other parametric covariance functions as special cases, and expresses the covariance between the spatial process at locations $\mathbf{s}$ and $\mathbf{s}'$ as the product of a power function of the distance between the two sites times the modified Bessel function of the second kind, $\mathcal{K}_\nu$, applied to the distance between the two locations. The mathematical expressions for the



(a) Exponential covariance functions. (b) Matèrn covariance functions.

Figure 1.1: Examples of covariance functions for varying values of the parameters: σ^2, ϕ, and ν.

exponential and the Matèrn covariance functions are respectively:

$$C(\mathbf{s}, \mathbf{s}') = \sigma^2 \exp\left(-\frac{d(\mathbf{s}, \mathbf{s}')}{\phi}\right) \tag{1.2}$$

$$C(\mathbf{s}, \mathbf{s}') = \sigma^2 \cdot \frac{1}{2^{\nu-1}\Gamma(\nu)} \cdot \left(\frac{d(\mathbf{s}, \mathbf{s}')}{\phi}\right)^{\nu} \cdot \mathcal{K}_{\nu}\left(\frac{d(\mathbf{s}, \mathbf{s}')}{\phi}\right) \tag{1.3}$$

where $d(\mathbf{s}, \mathbf{s}')$ denotes the distance between $\mathbf{s}$ and $\mathbf{s}'$. Figure 1.1 shows examples of the two covariance functions in (1.2) and (1.3) for different choices of σ^2, ϕ and ν as the distance between sites varies.

As shown in (1.2) and (1.3), most isotropic covariance function models depend on two parameters: σ^2, also called the *marginal variance* of the spatial process $w(\mathbf{s})$, $\mathbf{s} \in \mathcal{S}$, or the *spatial variance* of $Y(\mathbf{s})$, $\mathbf{s} \in \mathcal{S}$, and represents the part of the variation in $Y(\mathbf{s})$, $\mathbf{s} \in \mathcal{S}$, not accounted for by $\mu(\mathbf{s})$ (in contrast with τ^2); and ϕ, also called the *range parameter*, which provides information on the scale of the spatial process/range of the spatial correlation, that is, the minimum distance at which two sites can be deemed independent or practically independent (e.g. correlation

equal to 0.05). As shown in (1.3), the Matèrn covariance function depends also on an additional parameter, ν, called the *smoothness parameter* which controls the smoothness of the realizations of the spatial process $w(\mathbf{s})$, $\mathbf{s} \in \mathcal{S}$. The larger ν, the more continuously differentiable the spatial process is. Specific values of ν in (1.3) lead to other particular parametric covariance functions: $\nu = 0.5$ in (1.3) yields the exponential covariance function, while for $\nu \to \infty$ the Matèrn covariance function becomes the Gaussian covariance function.

Inference for geostatistical data typically consists of fitting model (1.1), or a slight variant of (1.1), to observations $y(\mathbf{s}_1), y(\mathbf{s}_2), \ldots, y(\mathbf{s}_n)$ of a spatial process $\{Y(\mathbf{s}) : \mathbf{s} \in \mathcal{S}\}$ at a finite number of locations $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n$ to obtain estimates of the mean parameters $\boldsymbol{\beta}$, and the covariance parameters σ^2, ϕ, potentially ν, and τ^2. For observations at n locations, both maximum likelihood estimation and Bayesian inference via Markov Chain Monte Carlo (MCMC) algorithms require $\mathcal{O}(n^3)$ operations, making inference computationally intensive for large datasets.

Several methods have been proposed in the spatial statistical literature to ease the computational burden associated with fitting a spatial statistical model as (1.1). Among these, a modeling strategy that has been revisited and extended in different fashions is the approximation introduced by Vecchia (1988). Having observations at n locations and decomposing the joint distribution $f(y(\mathbf{s}_1), y(\mathbf{s}_2), \ldots, y(\mathbf{s}_n))$ of $y(\mathbf{s}_1), y(\mathbf{s}_2), \ldots, y(\mathbf{s}_n)$ into the product of a univariate marginal distribution - say, $f(y(\mathbf{s}_1))$ - times $n-1$ conditional distributions, Vecchia (1988) exploited the concept of range of a covariance function to simplify the conditioning set in each of the $n-1$ conditional distributions. Thus, for example, rather than using

the conditional distribution $f(y(\mathbf{s}_{n-k})|y(\mathbf{s}_1), y(\mathbf{s}_2), \ldots, y(\mathbf{s}_{n-k-1}))$ in the expression of the joint distribution $f(y(\mathbf{s}_1), y(\mathbf{s}_2), \ldots, y(\mathbf{s}_n))$, Vecchia (1988) replaced it with $f\left(y(\mathbf{s}_{n-k})|y(\mathbf{s}_{n-k-1}), y(\mathbf{s}_{n-k-2}), \ldots, y(\mathbf{s}_{n-k-l})\right)$, where $\{\mathbf{s}_{n-k-1}, \mathbf{s}_{n-k-2}, \ldots, \mathbf{s}_{n-k-l}\}$ denotes the set of observation locations that are within a certain distance of $\mathbf{s}_{n-k}$. Recently, Katzfuss and Guinness (2017) have demonstrated that many of the currently used, computationally efficient methods for large-dimension, spatial data are special cases of the Vecchia approximation. Examples include the nearest-neighbor Gaussian process (NNGP) (Datta et al., 2016), independent blocks (Stein, 2008), and the Multi-resolution Approximation (M-RA) (Katzfuss, 2017), for which we provide an extension in Chapter II.

1.1.2 Areal unit data

A different type of spatial data often encountered in several applications, including epidemiology, medicine (imaging), archaeology, atmospheric sciences, etc. is areal data. In this case, the data are measurements that refer to bounded subsets of the spatial domain $\mathcal{S} \subset \mathbb{R}^d$ (e.g. counties, zip-codes, census tracts, pixels/voxels, grid cells, etc.). Besides the name *areal data*, sometimes this type of data is referred to as *discrete spatial variation data* to highlight the difference with geostatistical data that represent data with continuous spatial variation.

Because with areal data it is not possible to associate an observation to a point with precise geographical coordinates, the spatial dependence of a discrete spatial random field is not described through a function that depends on the distance between areal units. Rather, the common approach employed to model these data is

to exploit the notion of adjacency, and assume that random variables that refer to areal units that are adjacent - for example, share a common border - are more similar and correlated than random variables that pertain to non-adjacent areal units. Specifically, let $A_1, A_2, \ldots, A_n$ denote non-overlapping subsets of $\mathcal{S} \subset \mathbb{R}^d$ and let $y(A_1), y(A_2), \ldots, y(A_n)$ indicate the corresponding observations; examples include average income over counties in a state, output by an air quality model over regular grid cells covering the US, etc.. If marginally it is appropriate to assume that each random variable follows a normal distribution, a classical approach to model the realization $\{y(A_1), y(A_2), \ldots, y(A_n)\}$ of the random field $\{Y(A_i) : i = 1, \ldots, n\}$ is to specify a Bayesian hierarchical model where the first stage is given by:

$$y(A_i) = \mu(A_i) + \varphi(A_i) + \epsilon(A_i) \qquad \epsilon(A_i) \overset{iid}{\sim} N(0, \tau_\epsilon^2), \qquad i = 1, ..., n \tag{1.4}$$

with $\mu(A_i)$ expected value of the random variable $Y(A_i)$ over areal unit A_i, $\varphi(A_i)$ random effect relative to areal unit A_i and $\epsilon(A_i)$, white noise or error at areal unit A_i. As for point-referenced data, $\varphi(A_i)$ and $\epsilon(A_i)$, $i = 1, \ldots, n$, are assumed to be independent for each A_i, $i = 1, \ldots, n$. In turn, the mean $\mu(A_i), i = 1, ..., n$ is usually modeled as a function of spatially-varying covariates and it is meant to capture the large scale trend in the observations, while $\varphi(A_i), i = 1, ..., n$ is assumed to capture any residual spatial variability in the data not accounted for by the covariates. In some applications, it is not uncommon to use a constant mean μ, that is, assume that $\mu(A_i) \equiv \mu$ for every A_i, $i = 1, \ldots, n$.

To account for spatial dependence in the data, the Bayesian hierarchical model should provide a joint prior distribution for $\varphi(A_1), \varphi(A_2), \ldots, \varphi(A_n)$. In his sem-

inal paper, Besag (1974) showed that for a particular class of (improper) multivariate distribution, specifying a joint prior distribution was equivalent to specifying the set of full conditionals, thus introducing the Conditionally AutoRegressive (CAR) prior, which provides an elegant and intuitive model to capture spatial dependence in areal data. Specifically, if $\boldsymbol{\varphi}(A_{-i})$ denotes the set of spatial random effects relative to *all other* areal units except areal unit A_i, i.e. $\boldsymbol{\varphi}(A_{-i}) = \{\varphi(A_1), \varphi(A_2), \ldots, \varphi(A_{i-1}), \varphi(A_{i+1}), \ldots, \varphi(A_n)\}$, the CAR prior of Besag (1974) states that the conditional distribution of $\varphi(A_i)$ given $\boldsymbol{\varphi}(A_{-i})$ is:

$$\varphi(A_i)|\boldsymbol{\varphi}(A_{-i}), \tau_\varphi^2 \sim N\left(\frac{\sum_j w_{ij}\varphi(A_j)}{w_{i+}}, \frac{\tau_\varphi^2}{w_{i+}}\right) \qquad i = 1, \ldots, n \tag{1.5}$$

where $w_{i+} := \sum_j w_{ij}$. The w_{ij} in (1.5) are weights that encode the spatial dependence between areal units in $\mathcal{S}$; they are subject to certain conditions to ensure that the joint distribution implied by (1.5) admits a symmetric and non-negative definite covariance matrix. In the CAR model, Besag (1974) used the following specification for the w_{ij}'s:

$$w_{ij} = \begin{cases} 1 & \text{if } A_i \text{ and } A_j \text{ share a boundary} \\ 0 & \text{otherwise} \end{cases} \tag{1.6}$$

Under such choice for the w_{ij}'s, the CAR prior of Besag (1974) implies that: conditionally on the other areal units, the spatial random effect $\varphi(A_i)$ at areal unit A_i follows a normal distribution with mean equal to the average of the spatial random effects at the neighboring areal units, while the variance is inversely proportional to the number of neighbors of areal unit A_i.

An application in spatial epidemiology where areal data are encountered very frequently and CAR models are routinely utilized is that of disease mapping. The goal of disease mapping is to estimate the spatially-varying risk of a disease given observations on the number of people being affected or deceased by the disease over areal units. Although generated with an epidemiological application in mind, disease mapping models can be used also for general applications, even in cases where the observations do not refer to a disease. In general, let $y(A_i)$ denote a count relative to areal unit A_i (i.e. number of deaths or disease occurrences in A_i), the classical disease mapping for these data specifies a Poisson likelihood, e.g.:

$$y(A_i)|\lambda(A_i) \overset{ind}{\sim} \text{Poisson}\left(E(A_i)\lambda(A_i)\right), \qquad i = 1, ..., n \tag{1.7}$$

Here, $E(A_i)$ denotes the expected number of cases in areal unit A_i, which can be obtained via internal or external standardization (Banerjee et al., 2004), while $\lambda(A_i)$ denotes the relative risk of disease in unit A_i. At the second stage of the disease mapping model, a function of the relative risk $\lambda(A_i)$ is expressed as a linear combination of covariates and a random effect, e.g.:

$$\log\left(\lambda(A_i)\right) = \mathbf{X}(A_i)\boldsymbol{\beta} + \varphi(A_i) \tag{1.8}$$

where $\mathbf{X}(A_i)\boldsymbol{\beta}$ quantifies the influence of the covariates $\mathbf{X}(A_i)$ on the relative risk, $\boldsymbol{\beta}$ denotes the vector of regression coefficients which includes an intercept term, while $\varphi(A_i)$ denotes the spatial random effect for unit A_i. To account for spatial dependence in the observed counts, and thus in the relative risks, the spatial random effects, $\{\varphi(A_i) : i = 1, \ldots, n\}$, are provided with the CAR prior in (1.5). While computationally convenient to fit, a drawback of this model is that any

overdispersion in the observed data is incorrectly modeled as spatial dependence (Riebler et al., 2016). To address this issue, Besag et al. (1991) slightly revised (1.8) and proposed a model, now commonly referred to as the BYM model, that adds to the spatial random effect $\varphi(A_i)$ an independent random effect term $\xi(A_i)$, e.g.

$$\log\left(\lambda(A_i)\right) = \mathbf{X}(A_i)\boldsymbol{\beta} + \varphi(A_i) + \xi(A_i), \qquad i = 1, ..., n \tag{1.9}$$

with $\xi(A_i) \overset{iid}{\sim} N(0, \tau_\xi^2)$, thus, decomposing the variance of the log relative risks into the sum of an independent overdispersion term and a structured spatial dependence term.

Although widely adopted in spatial epidemiology, it is important to note that in the BYM model the two sets of spatial random effects, $\{\varphi(A_i) : i = 1, \ldots, n\}$ and $\{\xi(A_i) : i = 1, \ldots, n\}$ are not identifiable.

1.1.3 Spatio-temporal statistical models

If spatial data are collected over time, the spatial modeling approaches presented previously have to be extended to allow for more complex modeling frameworks and to account for a potential temporal dependence. In this review, we will focus only on models for data that can be thought as continuous in space and discrete in time, that is, realizations of a time-series of spatial Gaussian processes. For this type of data, the spatial and temporal dependence is commonly accommodated via dynamic spatio-temporal models, which are an extension to the spatial setting of classic first order autoregressive ($AR(1)$) models, commonly used in time series analysis (Hamilton, 1994). In an AR(1) model, if $\{y_t : t = 1, \ldots, T\}$

denotes a realization of a stochastic process $\{Y_t : t = 1, \ldots, T\}$ at time t, the data are modeled as:

$$y_t = \alpha y_{t-1} + v_t \qquad v_t \overset{iid}{\sim} N(0, (1-\alpha^2)\tau_v^2) \qquad t = 1, \ldots, T \tag{1.10}$$

with α parameter that captures the autocorrelation in the time series, ν_t *innovation term*, and $y_0 \sim N(0, \tau_v^2)$. The formulation in (1.10) yields several distributional results, including for $t = 1, \ldots, T$:

$$y_t | y_{t-1}, ..., y_0 \sim N(\alpha y_{t-1}, (1-\alpha^2)\tau_v^2) \qquad \text{and}$$

$$\text{Cov}(y_{t+k}, y_t) = \alpha^k \tau_v^2 \tag{1.11}$$

This modeling framework has been recently used also for spatio-temporal data (see Cressie and Wikle (2011), Katzfuss and Hammerling (2017)). In particular, building on the representation of spatial data via basis functions (see Banerjee et al. (2004), Cressie (1993), Cressie and Wikle (2011), Gelfand et al. (2010)), recent papers have expressed spatio-temporal data using basis functions that do not vary in time, but whose basis function weights are provided with a dynamic vector autoregressive structure, similar in spirit to (1.10).

An alternative to the basis-function representation of spatio-temporal data is provided by Gelfand et al. (2005), who propose a Bayesian hierarchical model for this type of data. This modeling approach can be seen as the combination of (1.1) and (1.10) in the context of spatially-varying regression coefficient models (Gelfand et al., 2003). More specifically, let $y(\mathbf{s}, t)$ denote an observation collected at location $\mathbf{s}$ at time t, then, the hierarchical dynamic spatio-temporal model of

Gelfand et al. (2005) is given for $t = 1, \ldots, T$ by:

$$\begin{aligned} y(\mathbf{s},t) &= \mathbf{X}(\mathbf{s},t)\boldsymbol{\beta}(\mathbf{s},t) + \epsilon(\mathbf{s},t), \quad \epsilon(\mathbf{s},t) \overset{iid}{\sim} N(0,\tau_\epsilon^2) \\ \boldsymbol{\beta}(\mathbf{s},t) &= \boldsymbol{\beta}_t + \tilde{\boldsymbol{\beta}}(\mathbf{s},t) \\ \boldsymbol{\beta}_t &= \alpha\boldsymbol{\beta}_{t-1} + \boldsymbol{\eta}_t, \quad \boldsymbol{\eta}_t \overset{iid}{\sim} N(0,\Sigma_{\boldsymbol{\eta}}) \\ \tilde{\boldsymbol{\beta}}(\mathbf{s},t) &= \tilde{\alpha}\tilde{\boldsymbol{\beta}}(\mathbf{s},t-1) + w(\mathbf{s},t) \end{aligned} \tag{1.12}$$

where $\mathbf{X}(\mathbf{s},t)$ denote spatially and temporally-varying covariates and $\boldsymbol{\beta}(\mathbf{s},t)$ indicate p spatially and temporally-varying regression coefficients including an intercept term. Equation (1.12) decomposes the p regression coefficients $\boldsymbol{\beta}(\mathbf{s},t)$ into the sum of a purely temporal regression coefficient and a spatio-temporal component, which can be thought as a random deviation from the global, temporally-varying regression coefficient $\boldsymbol{\beta}_t$. In turn, each of the components evolve in time according to a dynamic model, with innovations $\boldsymbol{\eta}_t$, $t = 1, \ldots, T$, and $w(\mathbf{s},t)$, $\mathbf{s} \in \mathcal{S}; t = 1, \ldots, T$, respectively, independent replicates of a multivariate normal random vector and of a Gaussian process.

1.1.4 Spatial point processes

In the models presented above, space, either in the form of points where observations are collected, or in the form of areal units for which data are reported, is not random. Now, we present models for data for which the location where an event occurs is random. Such type of stochastic processes are called *spatial point processes* (van Lieshout, 2000). Examples of spatial point processes are encountered in ecology, brain imaging, geology, astronomy, etc. In some cases, besides the location of the event, we also have additional random quantities associated

with an event. In this case, the point process is called a *marked point process*. The marks, or the additional attributes of the event, may be either numeric or categorical. Interested readers may refer to van Lieshout (2000) and Gelfand et al. (2010) for more thorough introductions to spatial point processes.

A spatial point process is a stochastic mechanism that generates a countable (either finite of infinite) set of points on a spatial domain in $\mathbb{R}^d$ (Gelfand et al., 2010). More specifically, a spatial point process defined on the spatial domain $\mathcal{S} \subset \mathbb{R}^d$ is a mapping from the Borel σ-algebra $\mathcal{B}$ of subsets of $\mathcal{S}$ to $\mathbb{R}^d$, such that $\forall B \in \mathcal{B}$, $N(B)$ is an integer-valued random variable that represents the number of events falling in B (Gelfand et al., 2010). A spatial point process is characterized through the intensity function, $\lambda(\mathbf{s})$, defined for each $\mathbf{s} \in \mathcal{S}$ as:

$$\lambda(\mathbf{s}) \quad = \quad \lim_{|\Delta\mathbf{s}| \to 0} \frac{E(N(\Delta\mathbf{s}))}{|\Delta\mathbf{s}|}$$

where $\Delta\mathbf{s} \in \mathcal{B}$ denotes a region containing $\mathbf{s}$. Given $\lambda(\mathbf{s})$, for any $B \in \mathcal{B}, E(N(B)) \equiv \int_{\mathbf{s}\in B} \lambda(\mathbf{s}) d\mathbf{s}$. More colloquially, the larger the intensity function $\lambda(\mathbf{s})$ at $\mathbf{s}$, the higher the chance that location $\mathbf{s}$ hosts an event.

The building block of many models for spatial point processes is the *Poisson point-process*, defined by the following two conditions:

1. For any $B \in \mathcal{B}$, $N(B)$ is a Poisson random variable with mean $\mu(B) = \int_{\mathbf{s}\in B} \lambda(\mathbf{s}) d\mathbf{s}$.
2. For any integer n and for any $B \in \mathcal{B}$, with $0 < \mu(B) < \infty$, conditional on $N(B) = n$, the events are located independently and uniformly over B.

When $\lambda(\mathbf{s}) \equiv \lambda \ \forall \mathbf{s} \in \mathcal{S}$, that is, the intensity is constant over space, the Poisson

point process is referred to as a homogeneous Poisson process, while if $\lambda(\mathbf{s})$ varies in space, the process is called an inhomogeneous Poisson process. In an inhomogeneous Poisson process, it might be of interest to understand how the intensity function varies as a function of covariates. Hence it is customary to write the following model for $\lambda(\mathbf{s})$, $g(\lambda(\mathbf{s})) = h(\mathbf{X}, \boldsymbol{\beta})$, where $g(\cdot)$ denotes a link function, typically the log function, and $h(\mathbf{X}, \boldsymbol{\beta})$ may be a linear function of covariates with parameters $\boldsymbol{\beta}$.

Recently, spatial point processes have been incorporated in geostatistical models to account for what is called *preferential sampling*, that is the stochastic dependence between the spatial point process that governs where observations are collected and the spatial process of the response field. The idea of preferential sampling was introduced by Diggle et al. (2010), who modeled the sampling mechanism, specifically the sampling locations, using a log Gaussian Cox Process (a highly popular model for spatial point processes). Then, the responses, conditioned on the locations, were modeled through the classical geostatistical model in (1.1), with the two models sharing a common spatial random effect. Through simulation studies, Diggle et al. (2010) have shown that such modeling strategy leads to an unbiased estimation of the Gaussian process parameters, and a substantial reduction in bias for out-of-sample predictions. This modeling paradigm ties closely to the field of survey statistics, in which weighting by the inverse probability of selection into a survey is often used in order to correct for sampling bias.

1.2 Introduction to survey methods

Survey methodology is one of the most mature and widely studied branches of statistical science. Surveys are used to gather information about a sample of individuals, often with the goal of inferring about a larger population. Designing and conducting a sampling survey requires multiple phases, including the *a priori* steps of wording the survey instrument and selecting a sample, as well as the *a posteriori* steps of correcting for bias in the gathered sample through weighting, imputation, or other statistical methods. Due to the subject matter of this dissertation, we will forego discussion of survey wording and instead provide a brief introduction to survey coverage and sampling, as well as forms of bias and their correction through weighting. There is a wealth of literature available to interested readers, including Groves et al. (2009), Fowler (2014), and de Leeuw et al. (2008).

Surveys are almost always conducted with a larger target population in mind. The *sampling frame* of the survey consists of any members of the target population who have a non-zero probability of being selected for the survey (Fowler, 2014; de Leeuw et al., 2008). Sampling frames can range from a comprehensive or nearly comprehensive list of the target population, such as telephone records, to samples that are gathered mainly through convenience, such as hospital patients who consent to having their electronic health records used for research purposes. The quality of a survey can be tied directly to the quality of the sampling frame and the process by which it is obtained. More specifically, Fowler (2014) states that sampling frames should be evaluated based on three factors. The first of these

is comprehensiveness, or how well the sampling frame covers the true population. de Leeuw et al. (2008) note that a common form of error in sampling surveys is under coverage, meaning that the sampling frame does not include all units in the target population. The second factor for evaluating a sampling frame is whether or not the probability of selection for each sampled unit can be computed. The third factor is efficiency, which refers to the rate at which members of the target population can be found within the sampling frame.

After one defines a sampling frame, it is necessary to select a sample. The "prototypical" (Fowler, 2014) sampling technique upon which standard statistical methods are based is simple random sampling, in which elements of the sampling frame are sampled with equal probability and without replacement. While appealing for their simplicity, simple random samples can, by random chance, yield samples that differ from the target population and/or fail to adequately represent subgroups of the target population. This is especially true when limited resources necessitate a relatively small sample. This issue can be curtailed through stratification, in which the sampling frame is partitioned into disjoint groups or strata, for example based on a characteristic or residential location. A sample is then drawn from each stratum, possibly with differential sampling probabilities, thereby ensuring that each stratum is represented within the sample (Fowler, 2014).

When simple random or stratified random sampling are infeasible, perhaps due to size of the target population or its geographic dispersion, researchers will often instead use cluster sampling. In cluster sampling, the sample is constructed from large primary sampling units (PSUs) and, potentially, smaller secondary sampling

units from within those units. In contrast to stratified random sampling, PSUs do not form a partition of the domain, rather they are often naturally occurring subgroups in the target population (de Leeuw et al., 2008), for example cities, regions, or sub-populations that are readily available to researchers. With cluster sampling, within-unit variability tends to be low, whereas between-unit variability tends to be high (compared to, for example, a stratified random sample). Due to high within-cluster variability, cluster samples tend to have higher overall variability compared to a simple random or stratified random sample. This increased variability in a cluster sample can be quantified by the *design effect* (Kish, 1995), defined as the ratio of the design-based variance (i.e. the variance of the estimator that accounts for the survey design) divided by the variance of the estimator under simple random sampling.

The previously mentioned sampling techniques can alleviate certain practical and statistical pitfalls that arise in sampling surveys, particularly due to selection bias. Nevertheless, standard statistical techniques are rarely appropriate for survey data. This may be due to other forms of bias such as differential response rates, or deficiencies in the sampling technique. In the case of item non-response, one can impute missing values (see Little and Rubin (2002) for details). Alternatively, weighting is a commonly applied technique to reduce survey bias. The process of weighting introduces a new variable, often denoted w_i for survey datum i, $i = 1, ..., n$. Whenever the probability of selection can be quantified, the weight w_i is typically defined as the inverse of the selection probability for survey datum i, and can then be interpreted as the number of individuals in the target population

represented by datum i. *Inverse probability weighting* can produce estimators with very high variance, particularly when the sampling probabilities are themselves highly variable. The procedure of weight trimming (Little, 1991; Elliott, 2009) aims to curtail this problem by capping the survey weights at a certain fixed value, with the value of the weight cap determined so to yield a bias-variance trade-off in the estimator.

1.3 Dissertation summary

This dissertation is organized as follows. In Chapter II we extend the multi-resolution approximation of Katzfuss (2017) to explore characteristics of the spatial dependence of the spatial process that underlies the observed data. By decomposing the spatial random effect (see Section 1.1.1) into a linear combination of appropriately chosen basis functions, and by specifying a priori a mixture prior on the basis function weights, we can examine the posterior distribution of the basis function weights to detect inhomogeneities in the range of the spatial correlation function. Through numerous simulations and an analysis of Soil Organic Carbon, we demonstrate the utility of our model not only for its intended purpose as an exploratory tool to investigate the spatial scale of the process, but also as a non-stationary geostatistical modeling framework whose predictive performance matches or exceeds that of existing methodologies.

In Chapter III, we propose a method for spatio-temporal disaggregation of estimates of proportions over areal units derived from multi-year sampling surveys. A crucial contribution of our modeling framework is the incorporation of the survey's

design effect into the modeling framework, which allows to properly propagate the survey variance during inference. We examine the capability of our model through simulations and we illustrate its use on estimates of community characteristics derived from the American Community Survey (ACS). Specifically, we present disaggregations at the census tract level and at the yearly time scale of the ACS estimates for the proportion of families in poverty and the proportion of people who identify as Black/African-American in Michigan.

In Chapter IV, using a similar approach to Chapter III, we develop a modeling framework to disaggregate estimates of counts derived from multi-year sampling surveys. In addition to incorporating the design effect in our model to properly propagate the design-based variance, we introduce a flexible, spatio-temporal zero-inflation term to handle the excess number of zero counts. In this chapter, in addition to a simulation and a data analysis of the number of births in Michigan, we quantify the gain of our model compared to a spatio-temporal extension of the model of Bradley et al. (2016b), which does not explicitly account for the survey design. In addition, we derive metrics for Michigan hospitals, namely, the number of births and number of births per bed, that we posit will correlate with demand and available resources.

Chapter V presents a modeling framework to correct for sampling bias in Electronic Health Records (EHRs) data from the University of Michigan Hospital System. Specifically, inspired by the work of Diggle et al. (2010) on preferential sampling, we introduce a spatial point process to account for potential oversampling in the EHR data, and we derive an estimate of the association between incident

lung cancer and smoking for the state of Michigan. Furthermore, we present a simulation study assessing the bias that would result from failing to account for preferential sampling.

Finally, Chapter VI concludes the dissertation with a discussion of future research directions.

CHAPTER II

Identifying Regions of Non-stationarity in Spatial Processes via Multi-resolution Approximation and Mixture Priors

2.1 Introduction

In a typical spatial statistical analysis that uses a parametric modeling framework, a practitioner is faced with selecting a model for the covariance function, determining *a priori* whether the spatial process is stationary, i.e. the spatial dependence in the data is just a function of the separation between sites, potentially even simply a function of the distance between locations, or whether the process is non-stationary.

Lack of stationarity can be ascribed to various reasons, such as, for example, inhomogeneities in the strength of the spatial correlation, which can have a long range in some subregions of the spatial domain and a shorter one in others. Even though tests to assess whether a process is stationary (Bandyopadhyay and Subba Rao, 2017; Jun and Genton, 2012), isotropic (Guan et al., 2004) or symmetric (Li et al., 2008; Weller and Hoeting, 2016) have been proposed in the literature, to our knowledge the question of how to determine regions characterized by a similar

range in the spatial correlation has not yet been fully addressed. The goal of this chapter is to fill this gap. Specifically, we propose a statistical modeling approach that can handle processes with a constant, non-varying spatial correlation range as well as processes whose spatial dependence range varies over the spatial domain. In the latter case, our model allows one to identify regions with inhomogeneities in the effective range, that is, in the minimum distance at which the correlation between two sites is equal to 0.05 (Banerjee et al., 2004). Identifying such regions has important consequences from an application standpoint, a sampling design perspective, as well as from an inferential and computational efficiency point of view, as we elaborate in Section 2.4.

The decomposition of a spatial domain in areas with similar spatial dependence is one of the classical modeling approaches used to construct non-stationary covariance functions (Sampson, 2010); see, for example, the smoothing and kernel-based models for non-stationary spatial processes of Fuentes (2001); Fuentes and Smith (2001); Kim et al. (2005); Nott and Dunsmuir (2002), the piece-wise Gaussian process of Kim et al. (2005); Pope et al. (2018), or the tree-based Gaussian process model of Gramacy and Lee (2008); Konomi et al. (2014), where a spatial process is assumed to be globally non-stationary but locally stationary. Our model coheres with this literature for non-stationary spatial processes while breaking with previous approaches in that it provides an alternative, computationally less challenging way to determine the regions of local stationarity.

Within the literature for globally non-stationary, locally stationary spatial processes, three main approaches are commonly employed to partition the domain

into subregions: Voronoi tessellations as in the Bayesian hierarchical models of Kim et al. (2005) and Pope et al. (2018); treed partitioning processes as implemented in the Bayesian CART models of Chipman et al. (1998) and Denison et al. (1998), or in the treed Gaussian process (TGP) model of Gramacy and Lee (2008) (see also Konomi et al. (2014)); and model selection as in Fuentes (2001), where a global non-stationary model is fit to the data multiple times using different partition schemes. The final partition of the spatial domain is determined based on model-fitting criteria such as BIC or AIC. Recent work by Risser et al. (2018) leverages information in the covariates in order to define the domain segmentation. Although these approaches generate segmentations of the spatial domain, the Markov chain Monte Carlo algorithms associated with either the TGP or the Bayesian hierarchical model of Kim et al. (2005) are computationally demanding. In contrast, the statistical modeling framework that we propose here does not require specifying *a priori* the maximum number of possible segments and it keeps computation rather feasible.

Falling in the tradition of basis function expansions approaches (Nychka et al., 2002; Johannesson et al., 2007; Banerjee et al., 2008; Matsuo et al., 2011; Nychka et al., 2016), the M-RA model of Katzfuss (2017) alleviates computation by providing an approximation to the original covariance function of a Gaussian spatial process via a linear combination of basis functions obtained by recursively implementing a predictive process approximation.

Katzfuss (2017) noted that in the M-RA modeling framework, the magnitude of the basis function weights at each level is related to the strength of the spatial

dependence in the data. Exploiting this intuition, we propose a modification of the M-RA model that allows basis functions weights to be shrunk towards zero. We achieve this by specifying as a prior distribution for the basis function weights a mixture of normal distributions with one of the mixture components having mean zero and covariance matrix shrunk near zero. Examining the behavior of the basis function weights over space will provide us with information on whether the spatial dependence of the spatial process has the same strength across the domain. Because of the prior specification on the basis function weights, we call our approach the *mixture M-RA*. We show via simulation experiments that our modeling framework is flexible enough to accommodate both stationary and non-stationary data, and does not require a practitioner to choose *a priori* the form of the covariance function to use in a spatial data analysis.

The remainder of this chapter is organized as follows: in Section 2.2 we review the M-RA model and we present our modification of the M-RA approach to allow for the identification of regions of range parameter inhomogeneity. Section 2.3 presents applications of our model to simulated data as well as an application to Soil Organic Carbon. Finally Section 2.4 offers a discussion on limitations and future extensions of the proposed model.

2.2 Methods

2.2.1 The Multi-Resolution Approximation (M-RA)

This section offers a brief review of the Multi-Resolution Approximation (M-RA) approach of Katzfuss (2017). Interested readers are referred to Katzfuss

(2017) and Katzfuss and Hammerling (2017) for additional details. In the following, we adopt a notation that is slightly different from that used by Katzfuss (2017) and Katzfuss and Hammerling (2017), particularly with respect to the subscripts used to index domain partitions and levels.

Let $y(\mathbf{s}), \mathbf{s} \in \mathcal{S}$ denote a spatial process in $\mathcal{S}$ observed at locations $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n$. Using a geostatistical modeling approach (Banerjee et al., 2004; Cressie, 1993), we express $y(\mathbf{s})$ as

$$y(\mathbf{s}) = \mu(\mathbf{s}) + w(\mathbf{s}) + \epsilon(\mathbf{s}) \qquad \epsilon(\mathbf{s}), \overset{iid}{\sim} N(0, \tau^2), \tag{2.1}$$

where $\mu(\mathbf{s})$ denotes the mean, or large scale spatial trend in $y(\mathbf{s})$, $w(\mathbf{s})$ indicates spatial random effect, and $\epsilon(\mathbf{s})$ denotes an independent error process, independent of $w(\mathbf{s})$. Without loss of generality, we take $\mu(\mathbf{s})$ to be constant in space, e.g. $\mu(\mathbf{s}) \equiv \mu$. We will often refer to (2.1) as a Kriging model, using the expression Bayesian Kriging model if (2.1) is fit within a Bayesian framework.

In (2.1), the spatial process $w(\mathbf{s})$ is taken to be a mean-zero Gaussian process with covariance function $C_w(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ represents a vector of covariance parameters that, in the case of a stationary, isotropic covariance function, includes the marginal variance (σ^2) and the range parameter (ϕ). Our $\boldsymbol{\theta}$ does not include a nugget effect (τ^2) since that part of variability in the data is already accounted for by the term $\epsilon(\mathbf{s})$. In the case $C_w(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta})$ is the Matérn covariance function (Banerjee et al., 2004; Cressie, 1993), $\boldsymbol{\theta}$ also includes a smoothness parameter ν.

Using the M-RA framework, the spatial process $w(\mathbf{s})$ can be approximated by $\tilde{w}_M(\mathbf{s})$ defined as $\tilde{w}_M(\mathbf{s}) = \mathbf{B}(\mathbf{s})\boldsymbol{\eta}$, with $\mathbf{B}(\mathbf{s})$ matrix of basis functions up to level M evaluated at $\mathbf{s}$, and $\boldsymbol{\eta}$ set of basis function weights. Replacing $w(\mathbf{s})$ with $\tilde{w}_M(\mathbf{s})$

into (2.1) leads to the *M-RA model*

$$y(\mathbf{s}) \approx \mu + \tilde{w}_M(\mathbf{s}) + \epsilon(\mathbf{s}) = \mu + \mathbf{B}(\mathbf{s})\boldsymbol{\eta} + \epsilon(\mathbf{s}), \qquad \epsilon(\mathbf{s}) \overset{iid}{\sim} N(0, \tau^2), \tag{2.2}$$

which provides great computational efficiency for large dimensional spatial data as illustrated in Katzfuss (2017).

The basis functions $\mathbf{B}(\mathbf{s})$ are defined by recursively partitioning the spatial domain, introducing a new set of knots within each new partition and using a predictive process approximation each time. More precisely, at level 0, r knots are placed on the entire domain. No particular placement of the r knots is suggested, although placing them on an equidistant grid is probably the most convenient and easy-to-implement choice. We indicate with $\mathcal{Q}^{(0)}$ the set of r knots introduced at level 0. Using this first set of knots $\mathcal{Q}^{(0)}$, the original process $w(\mathbf{s})$ is approximated using the predictive process $\tau_0(\mathbf{s}) := E\left[w(\mathbf{s})|w(\mathbf{Q}^{(0)})\right]$, where $w(\mathcal{Q}^{(0)})$ denotes the r-dimensional realization of the spatial process $w(\mathbf{s})$ at the knot locations. After this first initial approximation at level 0, at level 1, the spatial domain is subdivided into J non-overlapping subregions, and r knots are placed in each new subregion (see Figure 2.1 for an illustration of the knots' placement). We indicate with $\mathcal{Q}^{(1)}$ the set of $J \cdot r$ knots introduced at level 1. The knots in $\mathcal{Q}^{(1)}$ are in turn used to construct the predictive process approximation $\tau_1(\mathbf{s})$ to the remainder process $\delta_1(\mathbf{s})$ obtained at level 0 and defined as $\delta_1(\mathbf{s}) := [w(\mathbf{s}) - \tau_0(\mathbf{s})]$, where $[\cdot]$ superimposes independence across subregions. We note that J and r do not need to be equal at each level, but it is assumed for convenience of notation. In addition, the knots do not need to lay on a grid at each level: Katzfuss (2017) uses the observation

locations as knots in the final level of the M-RA. This procedure of partitioning, introducing knots, and approximating the remainder term $\delta_m(\mathbf{s})$ with its predictive process approximation $\tau_m(\mathbf{s})$ is repeated M times leading to the following M-level M-RA approximation $w_M(\cdot)$ to $w(\mathbf{s})$:

$$w_M(\mathbf{s}) = \tau_0(\mathbf{s}) + \tau_1(\mathbf{s}) + \ldots + \tau_{M-1}(\mathbf{s}) + \delta_M(\mathbf{s}) \equiv \tilde{w}_M(\mathbf{s}) + \delta_M(\mathbf{s}), \qquad \mathbf{s} \in \mathcal{S}, \tag{2.3}$$

with $\delta_M(\mathbf{s})$ remainder at level M.

By construction, the individual terms in (2.3) are mutually independent processes. It is also important to note that in the M-RA framework, the remainder processes, $\delta_1(\mathbf{s})$, $\delta_2(\mathbf{s})$, ..., $\delta_M(\mathbf{s})$, are independent across subregions at the corresponding level, e.g. $\delta_1(\mathbf{s})$ is independent across the J subregions introduced at level 1. This leads to a convenient block-diagonal covariance matrix structure for the basis function weights which contributes to the computational savings associated with the M-RA.

As at each level m, the predictive process $\tau_m(\mathbf{s})$ can be rewritten as a basis function expansion (see Banerjee et al. (2008)), it follows that

$$\tilde{w}_M(\mathbf{s}) = \sum_{m=0}^{M} \sum_{j=1}^{J^m} \mathbf{b}_{m,j}(\mathbf{s}) \boldsymbol{\eta}_{m,j}, \tag{2.4}$$

where the sum is taken over partitions and levels (see Katzfuss (2017) for more details). In (2.4), $\mathbf{b}_{m,j}(\mathbf{s})$ denotes the set of basis functions corresponding to the j-th partition and the m-th level evaluated at $\mathbf{s}$, while $\boldsymbol{\eta}_{m,j}$ is the r-dimensional vector of basis function weights in the j-th partition of the m-th level. They are defined as follows: let $\mathcal{Q}^{(m,j)}$ denote the set of r knots in the j-th partition at the m-th level, $m = 0, ..., M, j = 1, ..., J^m$, with $\mathcal{Q}^{(m)} = \bigcup_{j=1}^{J^m} \mathcal{Q}^{(m,j)}$, then the basis

functions and the prior covariance of the basis function weights are defined by the following recursive formulas: for any $\mathbf{s}_1, \mathbf{s}_2 \in \mathcal{S}$,

$$\begin{aligned} v_0(\mathbf{s}_1, \mathbf{s}_2) &= C_w(\mathbf{s}_1, \mathbf{s}_2; \boldsymbol{\theta}) \\ v_{m+1}(\mathbf{s}_1, \mathbf{s}_2) &= \begin{cases} 0, \text{ if } \mathbf{s}_1 \text{ and } \mathbf{s}_2 \text{ are in different regions at resolution } m \\ v_m(\mathbf{s}_1, \mathbf{s}_2) - \mathbf{b}_{m,j}(\mathbf{s}_1)' \mathbf{K}_{m,j} \mathbf{b}_{m,j}(\mathbf{s}_2), \text{ otherwise} \end{cases} \\ \mathbf{K}_{m,j}^{-1} &= v_m(\mathcal{Q}^{(m,j)}, \mathcal{Q}^{(m,j)}) \\ \mathbf{b}_{m,j}(\mathbf{s}) &= v_m(\mathbf{s}, \mathcal{Q}^{(m,j)}), \end{aligned} \tag{2.5}$$

where for every m and j, $\mathbf{K}_{m,j}$ is a $r \times r$ covariance matrix, and

$$\boldsymbol{\eta}_{m,j} \sim N_r(\mathbf{0}, \mathbf{K}_{m,j}). \tag{2.6}$$

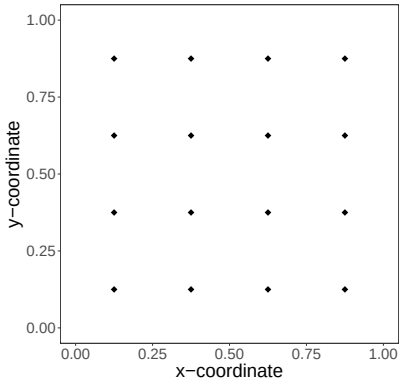
Replacing (2.4) into (2.2) yields

$$y(\mathbf{s}) \approx \mu + \sum_{m=0}^{M} \sum_{j=1}^{J^m} \mathbf{b}_{m,j}(\mathbf{s}) \boldsymbol{\eta}_{m,j} + \epsilon(\mathbf{s}), \qquad \epsilon(\mathbf{s}) \overset{iid}{\sim} N(0, \tau^2). \tag{2.7}$$

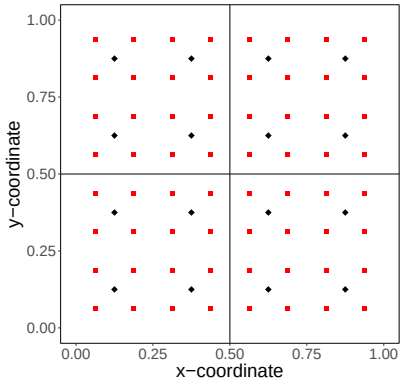
2.2.2 Illustration

Figure 2.1 illustrates the knots' placement for the M-RA models. While using a fine grid for the knots location could work well in many applications, a more general placement of the knots may be desired in some applications where we expect regions of local stationarity to display more irregular patterns. An example of such a procedure is provided in Appendix A.1.

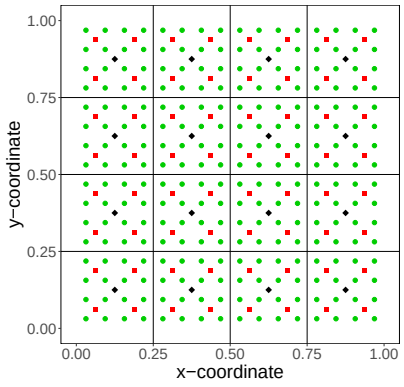
We now illustrate the relationship between the strength of the spatial dependence in a spatial process $w(\mathbf{s})$ and the number of levels needed in an M-RA to obtain a good approximation. Figure 2.2 and Figure 2.3 show this for $w_1(\mathbf{s})$ and



(a) Knot placement at level 0.



(b) Partition and knot placement at level 1.



(c) Partition and knot placement at level 2.

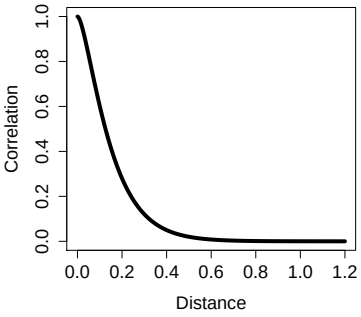
Figure 2.1: Illustration of the M-RA domain partitioning and knot placement at: (a) level 0; (b) level 1; and (c) level 2, respectively. Here: $r = 16$ and $J = 4$.

$w_2(\mathbf{s})$, realizations of mean-zero stationary Gaussian processes with Matérn correlation function with covariance parameters $\sigma^2 = 1$, $\nu = 1$ and $\phi = 0.1$, and $\sigma^2 = 1$, $\nu = 1$ and $\phi = 1.0$, respectively.

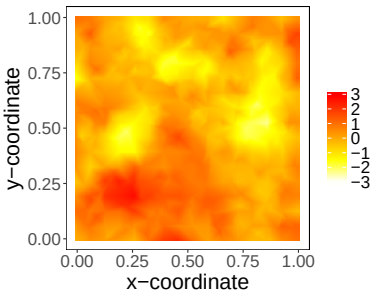
The key point of this illustration is that, when a spatial process is characterized by more rapidly decaying spatial dependence, a higher resolution approximation is required in order to capture fine spatial dependence. This concept is emphasized in Figure 2.2 (c)–(f), in which it is clear that the M-RA improves with each additional level. Alternatively, when a spatial process has slowly decaying spatial dependence, a low resolution approximation is sufficient, as illustrated in Figure 2.3 (c)–(f). We posit that this reasoning can be applied to a single non-stationary spatial process that is characterized by inhomogeneous rates of spatial decay. Specifically, regions of the domain with slowly decaying spatial dependence will require a lower-level approximation than regions with rapidly decaying spatial dependence.

2.2.3 The mixture M-RA

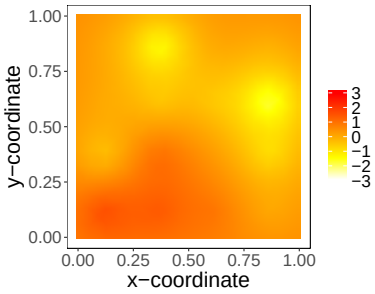
To allow for the possibility that a spatial process is characterized by a spatial correlation with a different range parameter in different subregions, we propose to slightly change the prior distribution on the basis function weights. Specifically, rather than placing a multivariate normal prior on them, as in the M-RA model, we provide them with a prior that allows them to be shrunk to zero from a level $\tilde{m}$ onward in certain subregions, if needed. Various alternatives are possible to shrink the basis function weights to 0: we could use a spike and slab prior on the basis function weights (Mitchell and Beauchamp, 1988; Ishwaran and Rao,



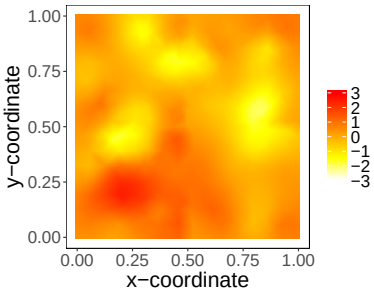
(a) Data generating covariance function.



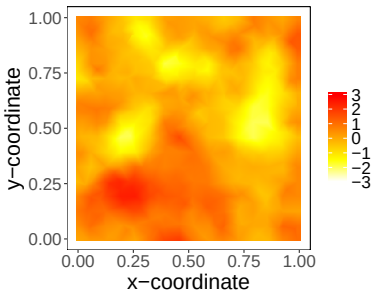
(b) Simulated $w_1(\mathbf{s})$.



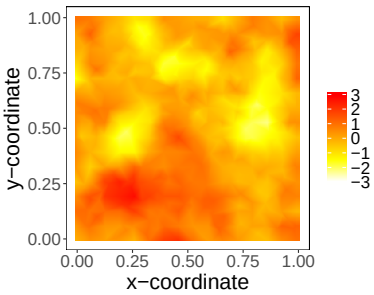
(c) Estimated $\hat{w}_M(\mathbf{s})$, $M = 0$.



(d) Estimated $\hat{w}_M(\mathbf{s})$, $M = 1$.

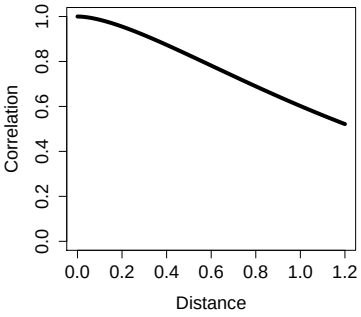


(e) Estimated $\hat{w}_M(\mathbf{s})$, $M = 2$.

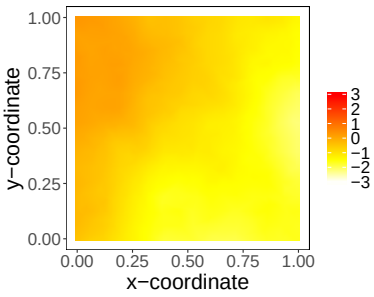


(f) Estimated $\hat{w}_M(\mathbf{s})$, $M = 3$.

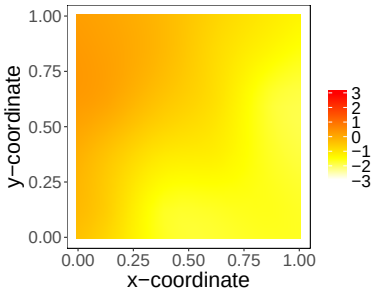
Figure 2.2: (a) Matérn correlation function with parameters $\sigma^2 = 1, \nu = 1$ and $\phi = 0.1$ used to simulate the mean-zero Gaussian process $w_1(\mathbf{s})$. (b) Simulated $w_1(\mathbf{s})$. (c)-(d)-(e)-(f) Estimated $\hat{w}_M(\mathbf{s})$ obtained by using an M-RA approximation with: (c) $M = 0$, (d) $M = 1$, (e) $M = 2$ and (f) $M = 3$. In each case, the number of subregions used was $J = 4$. Mean squared error defined as average of $(\hat{w}_M(\mathbf{s}_i) - w_1(\mathbf{s}_i))^2$ as $\mathbf{s}_i$ varies in a set of 756 points on the 1-unit square, are, respectively, (c) 0.42; (d) 0.16; (e) 0.05; and (f) 0.01 for the 4 M-RA approximations.



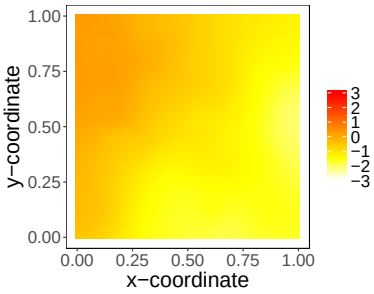
(a) Data generating covariance function.



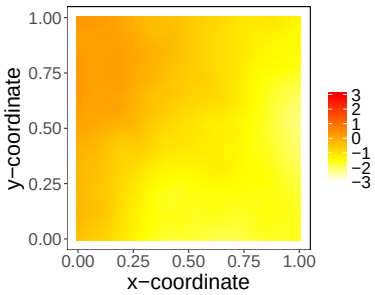
(b) Simulated $w_1(\mathbf{s})$.



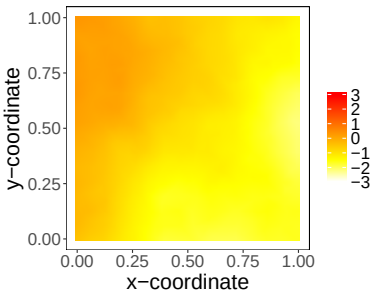
(c) Estimated $\hat{w}_M(\mathbf{s})$, $M = 0$.



(d) Estimated $\hat{w}_M(\mathbf{s})$, $M = 1$.



(e) Estimated $\hat{w}_M(\mathbf{s})$, $M = 2$.



(f) Estimated $\hat{w}_M(\mathbf{s})$, $M = 3$.

Figure 2.3: (a) Matérn correlation function with parameters $\sigma^2 = 1, \nu = 1$ and $\phi = 1.0$ used to simulate the mean-zero Gaussian process $w_2(\mathbf{s})$. (b) Simulated $w_2(\mathbf{s})$. (c)-(d)-(e)-(f) Estimated $\hat{w}_M(\mathbf{s})$ obtained by using an M-RA approximation with: (c) $M = 0$, (d) $M = 1$, (e) $M = 2$ and (f) $M = 3$. In each case, the number of subregions used was $J = 4$. Mean squared error, defined as average of $(\hat{w}_M(\mathbf{s}_i) - w_2(\mathbf{s}_i))^2$ as $\mathbf{s}_i$ varies in a set of 756 points on the 1-unit square, respectively, (c) 0.009; (d) 0.002; (e) 0.0006; and (f) 0.0001 for the 4 M-RA approximations.

2005), we could specify nonlocal priors (Johnson and Rossell, 2012), or employ some of the Bayesian methods for variable selection (in this case, leading to basis function weights selection), such as stochastic search variable selection (George and McCulloch, 1993, 1997), or empirical Bayes variable selection (George and Foster, 2000). For computational convenience, but mostly because we want to retain the hierarchical structure of the basis function weights and be able to shrink to 0 all the basis function weights nested within a given subregion and level, we elect to use the method proposed by Narisetty and He (2014) for Bayesian variable selection. Thus, we place mixture priors on the basis function weights $\boldsymbol{\eta}_{m,j}$. Specifically, using the prior distribution in (2.6) as a starting point, we specify the following mixture prior:

$$\boldsymbol{\eta}_{m,j} \sim p_m N_r(0, \mathbf{K}_{m,j}) + (1 - p_m) N_r(0, \mathbf{K}_{m,j}/L), \tag{2.8}$$

where L is a fixed large constant. The parameter $0 \leq p_m \leq 1$ in (2.8) indicates the probability that $\boldsymbol{\eta}_{m,j}$ is not shrunk to 0 and thus it is *active*. We call this model the *mixture M-RA*. In fitting the mixture M-RA model to data, we keep L fixed and we determine its value via cross-validation. In future implementations, L could be seen an additional parameter in the model, for which a prior distribution may be considered.

Continuous Gaussian spike and slab priors are commonly used in Bayesian hierarchical modeling to induce appropriate shrinkage while retaining efficient Gibbs sampling algorithms for computation (George and McCulloch, 1993; Ishwaran and Rao, 2005). However, one distinction that our prior specification on the $\boldsymbol{\eta}_{m,j}$'s has

in comparison with the commonly used spike and slab Gaussian priors is that the covariance matrix of both the spike and slab is proportional to $\mathbf{K}_{m,j}$ instead of the identity matrix. This is to reflect the dependence structure in the coefficients $\boldsymbol{\eta}_{m,j}$.

The grouping-preservation characteristic of our model specification can be seen in the following Bayesian hierarchical formulation. Let $Z_{m,j}$, for $m = 1, \ldots, M$, $j = 1, \ldots, J^m$ denote binary latent variables, then our model could be re-expressed as:

$$
\begin{aligned}
y(\mathbf{s}) = \mu(\mathbf{s}) + \textstyle\sum_{m=0}^{M} \sum_{j=1}^{J^m} \mathbf{b}_{m,j} \boldsymbol{\eta}_{m,j} + \epsilon(\mathbf{s}) \qquad \epsilon(\mathbf{s}) &\overset{iid}{\sim} N(0, \tau^2) \\
\boldsymbol{\eta}_{m,j} \mid Z_{m,j} = 1 &\sim N_r(0, \mathbf{K}_{m,j}) \\
\boldsymbol{\eta}_{m,j} \mid Z_{m,j} = 0 &\sim N_r(0, \mathbf{K}_{m,j}/L) \\
Z_{m,j} = 1 \mid (Z_{m-1,j^*} = 1,\ p_m) &\sim \text{Bernoulli}(p_m) \qquad (2.9) \\
p_m &= \rho^m \\
\rho &\sim \text{Beta}(\alpha_\rho, \beta_\rho) \\
P(Z_{m,j} = 1 \mid Z_{m-1,j^*} = 0) &= 0, \qquad (2.10)
\end{aligned}
$$

where j^* is the partition in level $m-1$ that contains the j-th partition at the m-th level.

In (2.9), p_m represents the probability that a set of basis function weights in the m-th level belongs to the first component of the mixture prior. Since we expect that for $m = 0$, all set of weights will belong to the first mixture component, whereas at higher level they are more likely to mix into the second component of the mixture prior, we set p_m to be equal to ρ^m. A more general framework will

define $p_m = \rho^{cm}$ with a positive parameter c to be estimated. To set the weights at a given level to be zero if the weights in the previous level are zero, we add (2.10) to our model specification. We note that by expressing $\tilde{w}_M(\mathbf{s})$ for each $\mathbf{s} \in \mathcal{S}$ as $\sum_{m=0}^{M} \sum_{j=1}^{J^m} \mathbf{b}_{m,j}(\mathbf{s})\boldsymbol{\eta}_{m,j}$ with the basis function $\mathbf{b}_{m,j}(\mathbf{s})$ defined as in (2.5) and the weights $\boldsymbol{\eta}_{m,j}$ distributed as in (2.8), the mixture M-RA model, as the M-RA model itself, defines a valid non-stationary Gaussian process (proof not included here).

We complete the specification of our model by providing priors to all the remaining model parameters. Choosing a stationary Matérn covariance function for $C_w(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta})$, the covariance parameter $\boldsymbol{\theta}$ is given by $(\sigma^2, \phi, \nu)'$. We place Inverse Gamma priors on the residual variance, or nugget effect, τ^2, and on the marginal variance σ^2 of $w(\mathbf{s})$. We choose hyperparameters α_{τ^2}, β_{τ^2} and α_{σ^2}, β_{σ^2}, corresponding to the shape and rate parameter, respectively, so that the priors on τ^2 and σ^2 are both weakly informative. Conversely, we specify a vague Gamma prior on the range parameter ϕ ($p(\phi) \propto 1/\phi$), while we place a Uniform prior on the interval $(0, 2)$ on the smoothness parameter ν. Finally, assuming $\mu(\mathbf{s}) \equiv \mu$, we place a vague mean-zero Normal prior on μ. In situations where $\mu(\mathbf{s})$ is modeled as a linear function of (spatial) covariates, the regression coefficients $\boldsymbol{\beta}$ are given similar, flat prior distributions.

2.2.4 Posterior inference

We fit our model within a Bayesian framework, approximating the posterior distribution using a Markov Chain Monte Carlo (MCMC) algorithm. The algorithm includes Gibbs sampling steps to generate posterior samples for the constant

mean μ, the basis function weights, $\boldsymbol{\eta}_{m,j}$, the nugget effect, τ^2, and the auxiliary binary variables $Z_{m,j}$. Metropolis-Hastings steps are used to generate posterior samples of the parameter ρ that defines the probabilities p_m, and of the covariance parameters σ^2, ϕ and ν. Specifically, to sample ρ we use a uniform proposal distribution bounded between 0 and 1 and centered at its current value in the MCMC algorithm. Similarly, we use uniform proposals to sample ϕ and ν, with the proposals being adaptively adjusted every 100 iterations until burn-in to achieve an acceptance rate of approximately 25%.

Although in the current implementation of the mixture M-RA we do not place a prior on L, we determine its value within the MCMC algorithm. Specifically, we start the MCMC algorithm with a large value for L for which we expect few no basis function to be drawn from the shrinkage prior. We typically use 1,000 as initial value for L based on the results obtained in Simulation Study 1, discussed in Section 2.3. We monitor the behavior of the basis function weights, decreasing the value of L every 1,000 iterations until we observe mixing of the basis function weights into the shrinkage prior. We continue monitoring the basis function weights, continuing to decrease the value of L in the burn-in until we do not see further significant changes in the proportion of basis function weights being shrunk to zero. We then keep L fixed for the rest of the MCMC iterations.

We implement our MCMC algorithm using `R` Version 3.4.1, incorporating the `Rcpp` package (Eddelbuettel and François, 2011) to increase speed. To compute the basis functions and covariance of the basis function weights we use code adapted from the supplementary material of Katzfuss (2017). Convergence of the chains

is determined by visually inspecting trace plots and marginal posterior density plots, and numerically by calculating Geweke's (Geweke, 1992) and Raftery Lewis' diagnostics (Raftery and Lewis, 1992) for every parameter.

2.3 Results

We now present results of the application of the mixture M-RA model to simulated data and observations of log Soil Organic Carbon in the conterminous US (CONUS).

2.3.1 Simulation results

To gain a better understanding of the mixture M-RA model, we designed multiple simulation studies. Here we report and discuss results for four simulation studies, with additional results and simulations available in the Supplementary Material:

1. *Simulation study 1*: data were generated according to the M-RA model in (2.7) with some basis function weights $\boldsymbol{\eta}_{m,j}$ set equal to 0;
2. *Simulation study 2*: data were generated on the unit square and non-stationarity was obtained by introducing two mean-zero stationary spatial processes with different spatial correlation ranges, each operating on one half of the square;
3. *Simulation study 3*: data were generated as in Simulation Study 2 except that the mean function $\mu(\mathbf{s})$ depended on a spatially varying covariate $x(\mathbf{s})$; and
4. *Simulation study 4*: data are generated with non-stationarity characterized by regions with four different spatial correlation ranges, with regions determined

via a latent spatial process. Such a data generation mechanism allows for more irregular regions of non-stationarity.

In addition, a fifth simulation is presented in Appendix A.3, which fits the mixture M-RA to randomly generated spatial data.

The goals of the simulation studies can be summarized as:

- to understand the role of L in the estimation of the basis function weights, and the magnitude of L needed to allow shrinkage to zero of the basis function weights, when a process is truly locally stationary (simulation study 1);
- to evaluate whether the mixture M-RA model can identify regions of local stationarity if they indeed exist, even in case of model mis-specification (simulation study 2, 3, 4, and 5 (Appendix A.3);
- to study how accounting for or ignoring non-stationarity in the residual dependence structure affects inference of the regression coefficients (simulation study 3 and 4);
- to compare the out-of-sample predictive performance of a non-stationary and a stationary model when data are realization of a non-stationary process (simulation study 2, 3, and 4); and
- to test our model in a setting where data deviates from our modeling framework. That is, data that are a realization of a spatial process characterized by more than two rates of spatial decay, and regions of non-stationarity have highly irregular shapes (simulation study 4).

For each simulation study, we generated multiple replicates $-$ 50 for simulation

study 1 and 30 for simulation studies 2, 3, and 4 − and results are averaged across the multiple realizations, unless otherwise noted. Posterior inference was based on samples yielded from an MCMC algorithm whose convergence was assessed using Geweke's (Geweke, 1992) and Raftery and Lewis' diagnostics (Raftery and Lewis, 1992). For the first, we tested whether the last 10% and 50% of each Markov chain post burn-in had significantly different means ($p = 0.05$, significance level); for the latter, we confirmed that the number of posterior samples employed for inference was greater than the number of iterations required to infer upon the 2.5^{th} percentile of each parameter within an accuracy of 0.01.

Simulation study 1

For this study, data were generated at $n = 756$ random locations in $\mathcal{S}$=[0,1] $\times$ [0,1] 50 times according to (2.7), where $\mu = 0$ and $\tau^2 = 0.05$. Basis functions weights were drawn either from the distribution in (2.6) with covariance matrix implied by a stationary Matérn covariance function with $\boldsymbol{\theta} = (\sigma^2, \phi, \nu)' = (1, 0.1, 1.0)$, $M = 3$, $J = 4$ and $r = 9$, or were set equal to 0. Figure 2.4(a) shows the regions with zero basis function weights.

To each of the 50 datasets, we fit our mixture M-RA model in (2.10) using $M = 3$, $J = 4$ and $r = 16$. As the goal of this simulation study is to determine whether our model is capable to identify regions with different strength of spatial dependence, in fitting the mixture M-RA model we did not estimate $\boldsymbol{\theta}$, rather we kept it fixed at its true value. However we varied the values of L and we used six different ones: L =10, 25, 50, 100, 200, and 10,000.

Table 2.1 presents summary statistics pertaining to the recovery of the basis

function weights averaged across levels, partitions, and the 50 simulations. As the table indicates, the true values of the basis function weights are contained in the 95% credible intervals with a frequency that is close to the nominal level, at times slightly over. The accuracy with which the basis function weights are estimated varies depending on the magnitude of the basis function weights. While on average across simulations, the average relative absolute error is around 0.40, when $L = 100$, the average relative absolute error of basis function weights whose true absolute value is less than 0.5 is 1.54. Meanwhile, for basis function weights whose true absolute value is greater than or equal than 0.5, is significantly smaller and equal to 0.28.

In terms of mixing into the shrinkage prior, when L=10,000 the zero-valued basis function weights do not mix into the shrinkage prior and we obtain a larger average MAE and MSE for those basis function weights. On the other hand, L=100 guarantees that the basis function weights mix into the shrinkage prior and they can be recovered with greater accuracy.

Simulation studies 2 and 3

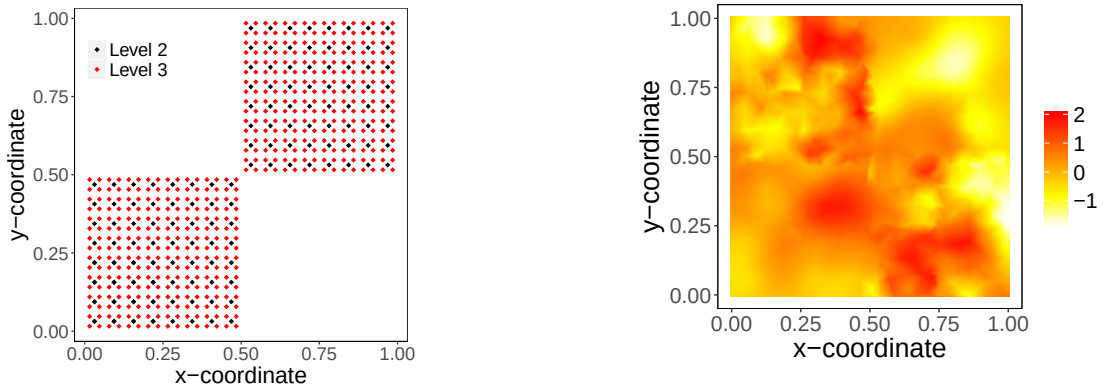
For both simulation studies, data were generated at $n = 1,012$ random locations in $\mathcal{S} = [0,1] \times [0,1]$ according to the following model:

$$
\begin{aligned}
(2.11) \qquad y(\mathbf{s}) &= \mu(\mathbf{s}) + I(\mathbf{s}_x < 0.5)w_1(\mathbf{s}) + I(\mathbf{s}_x \geq 0.5)w_2(\mathbf{s}) + \epsilon(\mathbf{s}) \\
\epsilon(\mathbf{s}) &\overset{iid}{\sim} N(0, \tau^2),
\end{aligned}
$$

where $\mathbf{s}_x$ indicates the first coordinate of the two-dimensional vector of geographical coordinates for point $\mathbf{s}$ (e.g. longitude or Easting). In (2.11), τ^2 was set equal

$\boldsymbol{\eta}_{m,j}$	L	Avg. $E[Z_{m,j}\|\mathbf{y}]$	Avg. MAE	Avg. MSE	Avg. bias	Avg. rel. MAE	Avg. rel. bias	Avg. rel. MSE	Avg. covg. of 95% CI
All	10	0.476	1.214	5.379	-0.021	NA	NA	NA	0.912
	25	0.518	1.227	5.394	-0.019	NA	NA	NA	0.925
	50	0.536	1.244	5.415	-0.017	NA	NA	NA	0.926
	100	0.600	1.341	5.453	-0.023	NA	NA	NA	0.935
	200	0.828	1.607	5.952	-0.023	NA	NA	NA	0.930
	1,000	0.884	1.661	6.029	-0.022	NA	NA	NA	0.930
	10,000	1.000	1.687	6.131	-0.024	NA	NA	NA	0.921
$= 0$	10	0.000	0.041	0.003	-0.000	NA	NA	NA	0.999
	25	0.001	0.041	0.003	-0.000	NA	NA	NA	0.999
	50	0.025	0.062	0.009	-0.000	NA	NA	NA	1.000
	100	0.152	0.302	0.377	-0.005	NA	NA	NA	0.999
	200	0.634	0.874	1.535	-0.008	NA	NA	NA	0.999
	1,000	0.754	0.912	1.719	-0.009	NA	NA	NA	0.999
	10,000	1.000	1.043	1.909	-0.010	NA	NA	NA	0.998
$\neq 0$	10	0.904	2.356	10.914	-0.033	0.491	−0.010	0.519	0.831
	25	0.978	2.301	10.493	-0.032	0.476	−0.008	0.491	0.866
	50	0.991	2.295	10.221	-0.032	0.401	−0.008	0.494	0.861
	100	0.999	2.264	9.966	-0.039	0.440	−0.011	0.481	0.927
	200	1.000	2.258	9.879	-0.037	0.488	−0.011	0.507	0.949
	1,000	1.000	2.259	9.887	-0.036	0.489	−0.011	0.507	0.950
	10,000	1.000	2.260	9.884	-0.037	0.450	−0.011	0.503	0.949

Table 2.1: Simulation study 1. Average posterior means of the $Z_{m,j}$, average Mean Absolute Error (Avg. MAE), average Mean Squared Error (Avg. MSE), average bias, average relative MSE, and average empirical coverage (covg.) of the 95% credible interval (CI) for the basis function weights, averaged across levels, subregions, and the 50 simulated datasets. Summary statistics are presented overall, and stratified based on whether the true basis function weights are equal to zero or not.



(a) Locations of zero-valued basis function weights.

(b) Realization of $y(\mathbf{s})$ generated under (2.7).

Figure 2.4: Simulation study 1: (a) locations of knots with zero-valued basis function weights at levels 2 and 3; (b) a realization of $y(\mathbf{s})$ generated according to (2.7), with $\mu = 0$, $\tau^2 = 0.05$, $\sigma^2 = 1.0$, $\nu = 1$, $\phi = 0.1$ and basis function weights at levels 2 and 3 set equal to zero as indicated in (a).

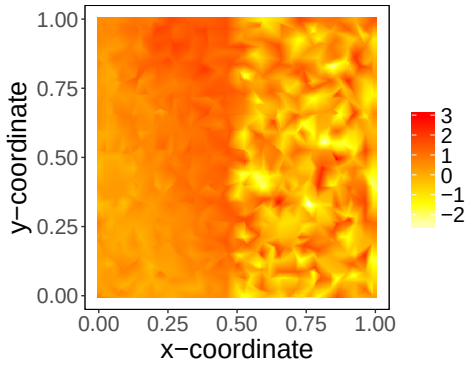
to 0.05, and $w_1(\mathbf{s})$ and $w_2(\mathbf{s})$ were taken to be two stationary Gaussian processes with Matérn covariance functions with parameters $\sigma^2 = 1$, $\nu = 1$, and ϕ equal to 1.0 and 0.01, respectively. For the 30 realizations of simulation study 2, $\mu(\mathbf{s}) \equiv 0$, $\forall \mathbf{s} \in \mathcal{S}$, while in simulation study 3, $\mu(\mathbf{s}) \equiv \beta_0 + \beta_1 x(\mathbf{s})$ with $\beta_0 = 2$, $\beta_1 = 3$, and $x(\mathbf{s})$ spatially-varying covariate, realization of a stationary Gaussian process with mean equal to 1 and Matérn covariance with parameters $\sigma^2 = 0.5, \nu = 0.5$ and $\phi = 0.2$. In simulation study 3, $x(\mathbf{s})$ was generated only once and kept constant across the 30 realizations of $y(\mathbf{s})$.

From each simulated dataset, we selected data at random from 756 locations. To these values, we fit (i) the mixture M-RA model with a stationary Matérn covariance function to define the basis functions, $M = 3$, $J = 4$, and $r = 16$, and L determined via tuning, and kept equal to 100 in the post burn-in iterations;

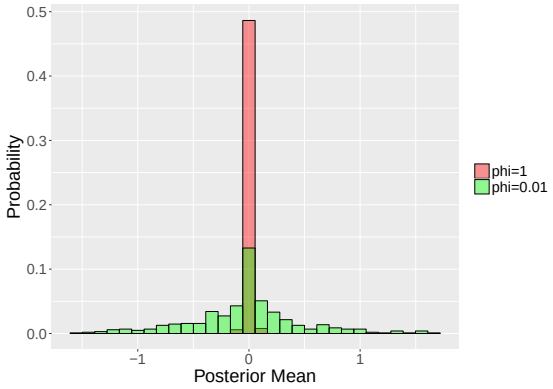
and (ii) a Bayesian Kriging model with a stationary Matérn covariance function. The out-of-sample predictive performance of the two models was evaluated from predictions at 256 hold-out sites.

Table 2.2 reports the out-of-sample predictive performance of the mixture M-RA model and the stationary Bayesian Kriging model, fitted to the 30 replicates of a mean-zero non-stationary spatial process generated under simulation study 2. As the table indicates, the mixture M-RA performs slightly better than the stationary model, even though the improvement is rather minimal.

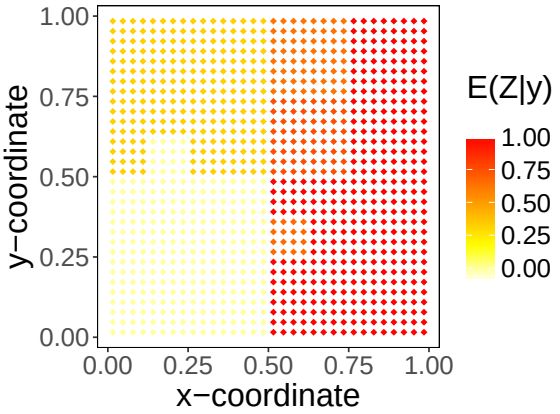
Table 2.3 evaluates whether, in simulation study 3, including the spatially-varying covariate $x(\mathbf{s})$, on which the mean trend function $\mu(\mathbf{s})$ of the spatial process depends on improves the out-of-sample predictive performance of the model fitted to the data regardless of whether the residual spatial covariance structure is allowed to be non-stationary (mixture M-RA model) or incorrectly specified as stationary (stationary Bayesian Kriging model). The table also investigates whether the out-of-sample predictive performance of the mixture M-RA and the stationary Bayesian Kriging model are affected by the exclusion of the spatially-varying covariate $x(\mathbf{s})$. As the results indicate, the inclusion of the spatially-varying covariate improves the predictive performance; nonetheless regardless of the inclusion or exclusion of the spatially-varying covariate, correctly specifying the non-stationary nature of the residual spatial dependence structure yields predictive intervals that have empirical coverage close to nominal. In this simulation setting, a stationary model yields posterior predictive distributions that are always underdispersive, and more so if the spatially-varying covariate is not included in the model.



(a) $y(\mathbf{s})$ generated according to (2.11).



(b) Histograms of posterior means of $\boldsymbol{\eta}_{m,j}$



(c) Posterior means of $Z_{3,j}$.

Figure 2.5: Simulation study 2. (a) One of the 30 realizations of $y(\mathbf{s})$ generated according to (2.11), with $\mu(\mathbf{s}) \equiv 0$, $\forall \mathbf{s} \in \mathcal{S} = [0,1] \times [0,1]$, $\tau^2 = 0.05$, and $w_1(\mathbf{s})$ and $w_2(\mathbf{s})$ mean-zero stationary Gaussian processes with Matérn covariance function with parameters, $\sigma_1^2 = 1.0$, $\nu_1 = 1$, $\phi_1 = 0.01$ and $\sigma_2^2 = 1.0$, $\nu_2 = 1$, and $\phi_2 = 1.0$, respectively. (b) Histograms of posterior means of basis function weights $\boldsymbol{\eta}_{m,j}$ in the third level ($m = 3$) of the mixture M-RA, grouped by values of ϕ, the range parameter. (c) Posterior mean of the latent binary variables $Z_{m,j}$ at the third level.

Model	Average MSPE	Standard deviation of MSPE	Average empirical coverage 95% PI
Mixture M-RA	0.51	0.06	93.4%
Stationary Bayesian Kriging	0.55	0.07	92.7%

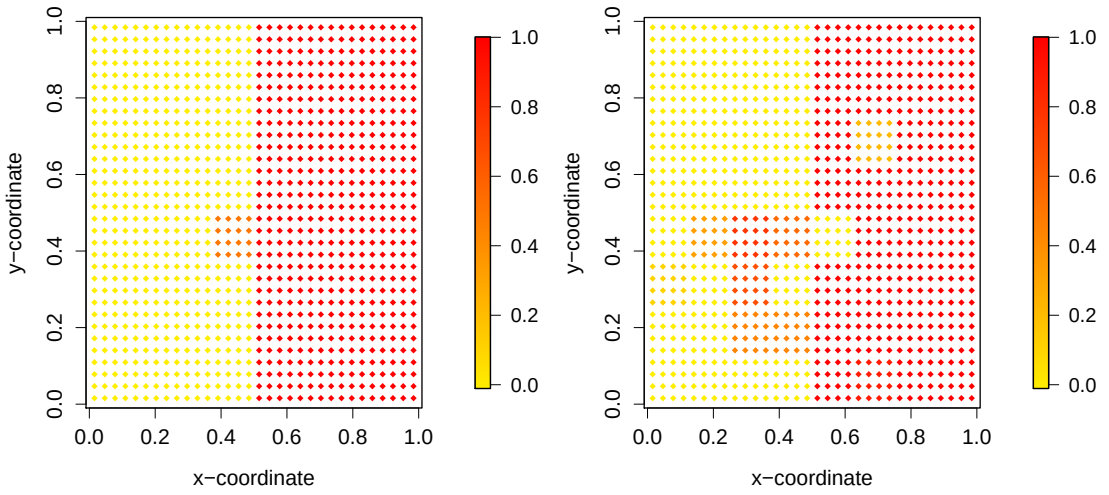
Table 2.2: Simulation study 2. Average Mean Squared Prediction Error (MSPE), standard deviation of the Mean Squared Prediction Errors, and average empirical coverage of the 95% prediction intervals for the mixture M-RA model and the stationary Bayesian Kriging model. The summary statistics are averaged over the 30 simulations.

Model	Bias β_1	Average coverage of 95% CI for β_1	Average MSPE	SD of MSPE	Average coverage 95% PI
Mixture M-RA with $x(\mathbf{s})$	-0.017	93.3%	0.53	0.08	92.6%
Stationary Bayesian Kriging with $x(\mathbf{s})$	-0.012	60.0%	0.51	0.08	70.1%
Mixture M-RA without $x(\mathbf{s})$	NA	NA	0.93	0.09	94.0%
Stationary Bayesian Kriging without $x(\mathbf{s})$	NA	NA	1.39	0.11	32.0 %

Table 2.3: Simulation study 3. Results averaged across 30 simulations: bias of the posterior mean of β_1; empirical probability that a 95% credible interval covers the true value for β_1; average Mean Squared Prediction Error (MSPE); standard deviation of MSPE; and empirical coverage of the 95% prediction intervals.

Figure 2.6 evaluates whether the mixture M-RA model correctly identifies the two regions of non-stationarity when data are generated according to (11), with $\mu(\mathbf{s}) \equiv \beta_0 + \beta_1 x(\mathbf{s})$, $\beta_0 = 2$, $\beta_1 = 3$, and $x(\mathbf{s})$ spatially-varying covariate, simulated as described in Section 3.1.2. Specifically, Figure 2.6 presents, for two of the 30 simulated datasets, the posterior mean of the latent binary variables $Z_{m,j}$ for $m = 3$ indicating which basis function weights are likely to be shrunk to zero at the third level.

Inspecting the results for simulation study 2, we observe that the basis function weights $\boldsymbol{\eta}_{m,j}$ mix into the two priors at different rates in the two halves of the spatial domain: in the part of the domain where $\phi = 1.0$, the average posterior mean of the binary latent variables $Z_{m,j}$ at the third level ($m = 3$), averaged



(a) Posterior mean of $Z_{3,j}$ for simulation 1 of 30

(b) Posterior mean of $Z_{3,j}$ for simulation 2 of 30

Figure 2.6: Simulation study 3. Posterior mean of the latent binary variables $Z_{m,j}$ at the third level for two of the 30 simulations.

across the 30 simulations, is 0.296, while in the region where $\phi = 0.01$, it is 0.968. In simulation study 3, we observe that ignoring non-stationarity in the residual spatial dependence structure does not have consequences in terms of point inference for β_1 as both the approaches result in similar levels of bias (bias of -0.017 for mixture M-RA and -0.012 for stationary Bayesian Kriging). However, it leads to an underestimation of the variability of the estimate: the 95% CI's for Bayesian Kriging are too short and do not provide the nominal coverage (actual coverage of 60.0%), which is instead achieved by the 95% CI's for β_1 of the mixture M-RA model (having an actual coverage of 93.3 %).

Quartile of $v(\mathbf{s})$	Data Generation	$w_i(\mathbf{s})$ covariance parameters
1^{st} quartile	$y(\mathbf{s}) = w_1(\mathbf{s}) + \epsilon(\mathbf{s})$	$\sigma^2 = 1, \phi = 1, \nu = 1$
2^{nd} quartile	$y(\mathbf{s}) = w_2(\mathbf{s}) + \epsilon(\mathbf{s})$	$\sigma^2 = 1, \phi = 0.5, \nu = 1$
3^{rd} quartile	$y(\mathbf{s}) = w_3(\mathbf{s}) + \epsilon(\mathbf{s})$	$\sigma^2 = 1, \phi = 0.1, \nu = 1$
4^{th} quartile	$y(\mathbf{s}) = w_4(\mathbf{s}) + \epsilon(\mathbf{s})$	$\sigma^2 = 1, \phi = 0.01, \nu = 1$

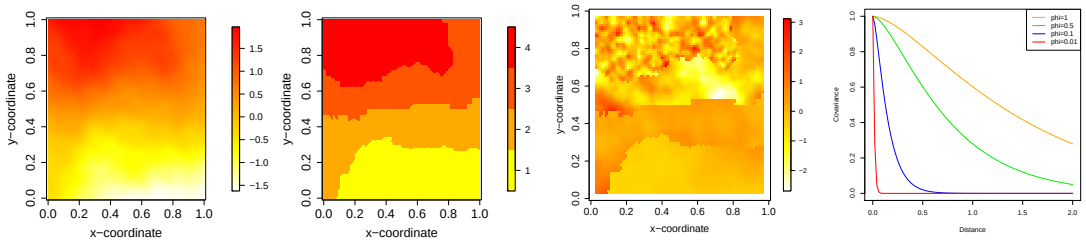
Table 2.4: Simulation study 4. Data generation mechanism used to generate 30 realization of a non-stationary spatial process $y(\mathbf{s})$ in the unit square $\mathcal{S}$.

Simulation study 4

In simulation study 4, the main objective is to determine whether the mixture M-RA model is able to detect regions of non-stationarity when the spatial process is strongly non-stationary and the regions of local stationarity are irregular. For this purpose we have generated two sets of simulations, each made of 30 replicates, with the only difference among the two being that one set is characterized by regions of local stationarity with more close-to-rectangular boundaries.

To generate realizations of a mean-zero locally stationary spatial process with irregular boundaries, we first simulated a mean-zero stationary Gaussian process $v(\mathbf{s})$ with an exponential covariance function and covariance parameters $\sigma^2 = 1$, $\phi = 1$ and then we truncate it using its quartiles to define four regions of non-stationarity. We then generated 30 realizations of a non-stationary spatial process at 1,374 locations in the unit square according to the data generation mechanism presented in Table 2.4, with $w_i(\mathbf{s})$, $i = 1, \ldots, 4$ mean-zero stationary Gaussian processes with Matérn covariance function and parameters as displayed in Table 2.4.

Figure 2.7 panels (a)-(d) show the four regions of non-stationarity, a realization of the four spatial processes $w_1(\mathbf{s})$, $w_2(\mathbf{s})$, $w_3(\mathbf{s})$, and $w_4(\mathbf{s})$, and their correlation



(a) Latent process $v(\mathbf{s})$. (b) Regions of non-stationarity. (c) $w_1(\mathbf{s})$-$w_4(\mathbf{s})$ (d) Data-generating spatial correlation functions.

Figure 2.7: Simulation study 4, first set. (a) Latent process $v(\mathbf{s})$ used to identify the four regions of non-stationarity. (b) Regions on non-stationarity, displayed in 4 different colors. (c) A realization of $w_1(\mathbf{s}), w_2(\mathbf{s}), w_3(\mathbf{s}), w_4(\mathbf{s})$ in the four regions of local stationarity. (d) Correlation functions of $w_1(\mathbf{s}), w_2(\mathbf{s}), w_3(\mathbf{s}), w_4(\mathbf{s})$.

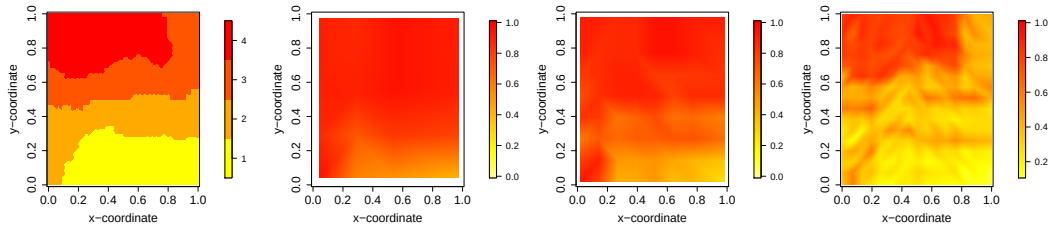
Quartile	Spatial Range	$E(Z_{2,j}\|\mathbf{y})$	$E(Z_{3,j}\|\mathbf{y})$	$E(Z_{4,j}\|\mathbf{y})$
1^{st} quartile	$\phi = 1$	0.45	0.40	0.23
2^{nd} quartile	$\phi = 0.5$	0.83	0.70	0.33
3^{rd} quartile	$\phi = 0.1$	0.93	0.85	0.45
4^{th} quartile	$\phi = 0.01$	0.94	0.93	0.85

Table 2.5: Simulation study 4, first setting. Average posterior expectation of the latent binary variables $Z_{m,j}$ for $m = 2, 3, 4$ in the four regions of non-stationarity averaged across the 30 simulations.

functions.

For each simulated dataset, we used data from 1,124 randomly selected locations and fitted a mixture M-RA model with $M = 4$, $J = 4$, and r=16, Matérn covariance function for the basis functions, and L determined via tuning during the burn-in period (kept equal to 100 post burn-in). Knowing *a priori* of the irregularity of the regions of non-stationarity, in lieu of an equidistant grid, at the highest multi-resolution level, we have used the observation locations as knots locations, following the procedure of Katzfuss (2017).

Table 2.5 presents the posterior probability that a basis function weight at levels 2, 3, and 4 is drawn from the shrinkage prior distribution averaged across parti-

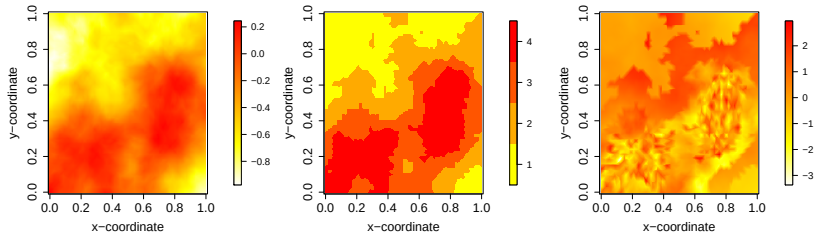


(a) Regions of non-stationarity. (b) Average $E(Z_{2,j}|\mathbf{y})$ (c) Average $E(Z_{3,j}|\mathbf{y})$ (d) Average $E(Z_{4,j}|\mathbf{y})$

Figure 2.8: Simulation study 4, first set. (a) Regions of non-stationarity. (b)-(d) Average $E(Z_{2,j}|\mathbf{y})$, $E(Z_{3,j}|\mathbf{y})$, and $E(Z_{4,j}|\mathbf{y})$, averaged across the 30 simulations.

tions and simulations, while Figure 2.8 presents a plot of the posterior mean of the latent binary variables $Z_{m,j}$ averaged across the 30 simulations. Even though not perfectly identified, we can see that in the region where the spatial correlation persists at large distances ($\phi = 1$), the basis function weights have a higher probability of being drawn from the shrinkage prior. Furthermore, the region with the second slowest rate of decay in the spatial correlation has the second highest probability of shrinkage, and so on, with least pronounced shrinkage at level 2, and most pronounced at level 4.

As we acknowledge that the regions in Figure 2.7(b) may still be considered somewhat rectangular, we have generated a second realization of $v(\mathbf{s})$, presented in Figure 2.9, which yields more volatile sub-regions. Using these newly identified regions of local stationarity, we have produced 30 simulated datasets for $w(\mathbf{s})$ using the data generation mechanism described in Table 2.4. Table 2.6 presents the posterior mean of the latent binary variables at levels $m = 2, 3$ and 4 for this new set of simulated datasets. As the table indicates, the mixture M-RA model struggled more to identify the regions of local stationarity than in the previous



(a) Latent process $v(\mathbf{s})$. (b) Regions of non-stationarity. (c) $w_1(\mathbf{s})$-$w_4(\mathbf{s})$

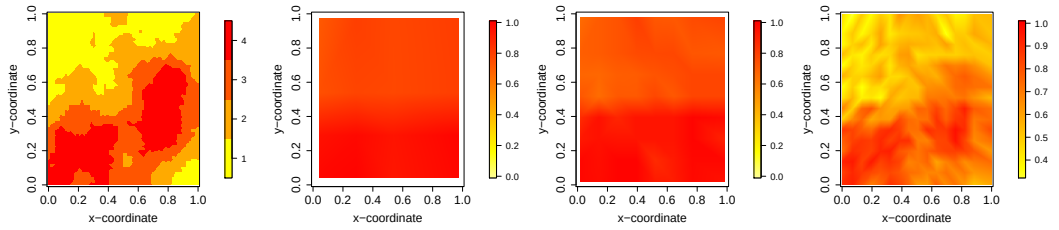
Figure 2.9: Simulation study 4, second set. Data generation example: (a) Latent variable $v(\mathbf{s})$. (b) Regions formed by truncating $v(\mathbf{s})$. (c) Spatial random effect $\mathbf{w}(\mathbf{s})$.

Quartile	Spatial Range	$E(Z_{2,j}\|\mathbf{y})$	$E(Z_{3,j}\|\mathbf{y})$	$E(Z_{4,j}\|\mathbf{y})$
1^{st} quartile	$\phi = 1$	0.50	0.42	0.37
2^{nd} quartile	$\phi = 0.5$	0.75	0.66	0.49
3^{rd} quartile	$\phi = 0.1$	0.90	0.84	0.81
4^{th} quartile	$\phi = 0.01$	0.99	0.93	0.89

Table 2.6: Simulation study 4, second set. Average posterior expectation of the latent binary variables $Z_{m,j}$ for $m = 2, 3, 4$ in the four regions of non-stationarity averaged across the 30 simulations.

simulated set of realizations generated under simulation study 4. However, we observe the same gradient effect in Figure 2.8 as we previously observed, and the rates of shrinkage are still consistent with our intuition.

Reporting results only for the first set of 30 simulated datasets generated in simulation study 4, Table 2.7 summarizes the out-of-sample predictive performance, averaged across the 30 simulations, of the mixture M-RA and the stationary Bayesian Kriging model in terms of average Mean Squared Predictive Error (MSPE), average empirical coverage of the 95% Predictive Intervals (PI), and average length of the 95% Predictive Intervals. As the table indicates, despite the strong non-stationarity in the spatial process, there is little difference in predictive performance between the mixture M-RA and the stationary Bayesian Kriging



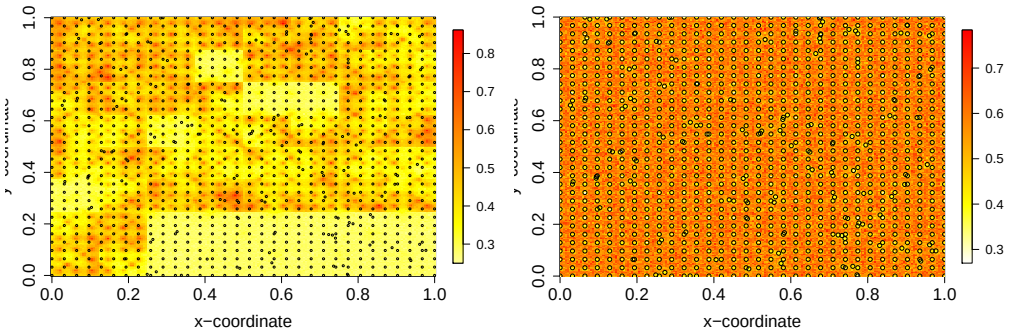
(a) Regions of non-stationarity. (b) Average $E(Z_{2,j}|\mathbf{y})$ (c) Average $E(Z_{3,j}|\mathbf{y})$ (d) Average $E(Z_{4,j}|\mathbf{y})$

Figure 2.10: Simulation study 4, second set. (a) Regions of non-stationarity. (b)-(d) Average $E(Z_{2,j}|\mathbf{y})$, $E(Z_{3,j}|\mathbf{y})$, and $E(Z_{4,j}|\mathbf{y})$, averaged across the 30 simulations.

model, with the stationary Bayesian Kriging model yielding a slightly lower MSPE. On the other hand, the mixture M-RA achieved a higher, closer to nominal level, empirical coverage. These results are not unexpected, as several papers in the spatial statistical literature (Risser et al., 2018; Fuglstad et al., 2015; Neto et al., 2014; Schmidt et al., 2011; Paciorek and Schervish, 2006) have indicated that non-stationary spatial models do not yield *point* predictions that are extremely different from those obtained using stationary models. The largest difference between the two class of models lies in the prediction variances and the spatial patterning of those variances. To illustrate this point, Figure 2.11 shows posterior predictive standard deviations at 40,000 locations on the unit square for the mixture M-RA and the stationary Bayesian Kriging model for a data set generated using the procedure of the first set of simulations in simulation study 4. As the figure illustrates, the posterior predictive standard deviations are quite constant over space under the stationary Bayesian Kriging model with lower prediction uncertainty only at the sites with data. On the other hand, the posterior predictive standard deviation surface for the non-stationary mixture M-RA model shows greater spatial

Model	Average MSPE	Average SD of predictions	Average empirical coverage of 95% PI	Average length of 95% PI
Mixture M-RA	0.395	0.629	0.964	2.466
Stationary Bayesian Kriging	0.386	0.596	0.929	2.336

Table 2.7: Simulation study 4, first set. Summary of the out-of-sample predictive performance averaged across 30 simulations for the mixture M-RA model and the stationary Bayesian Kriging model: average Mean Squared Predictive Error (MSPE), average standard deviation (SD) of the out-of-sample predictions, average empirical coverage of the 95% predictive intervals (PI), and average length of the 95% predictive intervals.



(a) Posterior predictive SD: Mixture M-RA.

(b) Posterior predictive SD: Kriging.

Figure 2.11: Simulation study 4, first set. (a) Posterior predictive standard deviations of predicted values at 40,000 locations on the unit-square for the mixture M-RA model. (b) Posterior predictive standard deviations for predictions obtained using the stationary Bayesian Kriging model. In each panel, circles or dots indicate the 1,124 locations with observation values used for model fitting.

variability. We have noticed such behavior for the stationary Bayesian Kriging model also in the analysis of log SOC in the main manuscript (see Figure 4(c)).

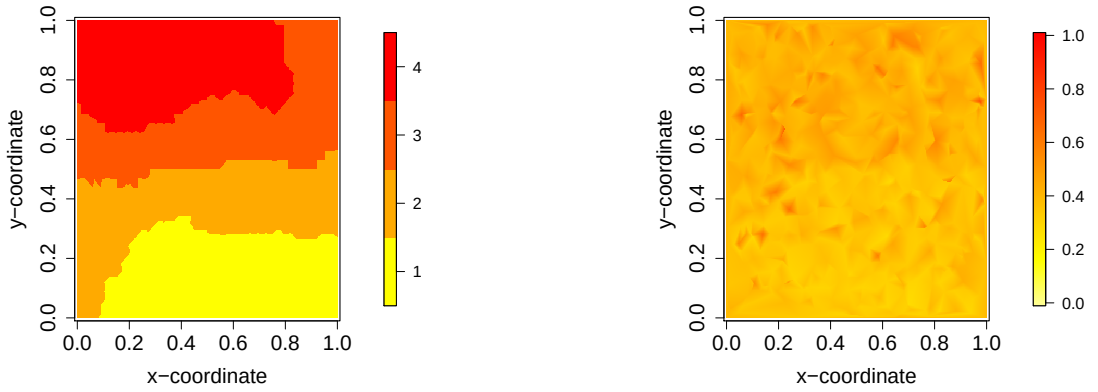
We view these results as highly encouraging that our modeling framework has utility for non-stationary spatial data that (1) exhibit non-stationarity that can be characterized by more than just two range parameters and (2) have non-rectangular regions of non-stationarity. These results prompted us to consider how the model would perform if we did not impose the somewhat informative and hierarchical prior distributions on the basis function weights in (2.8). Thus, we re-run the above

simulation study for the first set of 30 simulated datasets generated in simulation study 4 with the following prior distribution on the basis function weights, now denoted as $\boldsymbol{\eta} := \{\eta_l\}_{l=1,...,\sum_{m=0}^{M} J^m \times r}$:

$$
\begin{aligned}
\eta_l &\overset{iid}{\sim} pN(0, \tau_\eta^2) + (1-p)N(0, \tau_\eta^2/L), l = 1, ..., \sum_{m=0}^{M} J^m \times r \\
\eta_l \mid Z_l = 1 &\sim N_r(0, \tau_\eta^2) \\
\eta_l \mid Z_l = 0 &\sim N_r(0, \tau_\eta^2/L) \\
Z_l = 1 &\overset{iid}{\sim} \text{Bernoulli}(p) \\
p &\sim \text{Beta}(1, 1) \\
\tau_\eta^2 &\sim IG(1, 1),
\end{aligned} \tag{2.12}
$$

where L is once again a large constant that shrinks the variance of η_l toward zero. Under this prior construction, we impose no dependence *a priori* on the basis function weights, nor do we impose that the basis function weights at higher levels are more likely to mix into the shrinkage prior. Furthermore, there is no hierarchical structure to the shrinkage of the basis function weights. This can be viewed as a highly simplified version of the mixture M-RA, in which we place a more traditional Bayesian spike and slab prior distribution on the basis function weights.

Under this prior construction, we find that the mixing of the basis function weights was highly uninformative of the range of the spatial process. The average posterior mean of the latent binary variables Z_l in the region characterized by the most rapidly decaying spatial dependence is 0.49, whereas in the region character-



(a) Regions of non-stationarity.

(b) Average $E(Z_{4,j}|\mathbf{y})$

Figure 2.12: Simulation study 4, first set, under the modeling framework in (2.12) that specifies a mixture of univariate normal prior distributions on the basis function weights. (a) Regions of non-stationarity. (b) $E(Z_{4,j}|\mathbf{y})$, averaged across the 30 simulations.

ized by the most slowly decaying spatial dependence, the average posterior mean of Z_l is 0.40. Figure 2.12 presents a plot of the average over 30 simulations of the posterior means of the latent binary variables Z_l, which display a great degree of heterogeneity over space, in contrast to Figures 2.8 and 2.10. Despite this, the predictive performance of the model under this more simple prior construction was competitive with the other models in Table 2.7, with an average posterior predictive mean of 0.391. The results pertaining to the mixing of the basis function weights suggest that the mixture M-RA performs well as an exploratory tool when the active basis function weights maintain the same the prior construction of the M-RA of Katzfuss (2017).

2.3.2 Analysis of Soil Organic Carbon in Continental US

Soils contain a massive proportion of the Earth system's carbon (Lefèvre et al., 2017): soil organic carbon (SOC), the carbon stored within soil organic matter

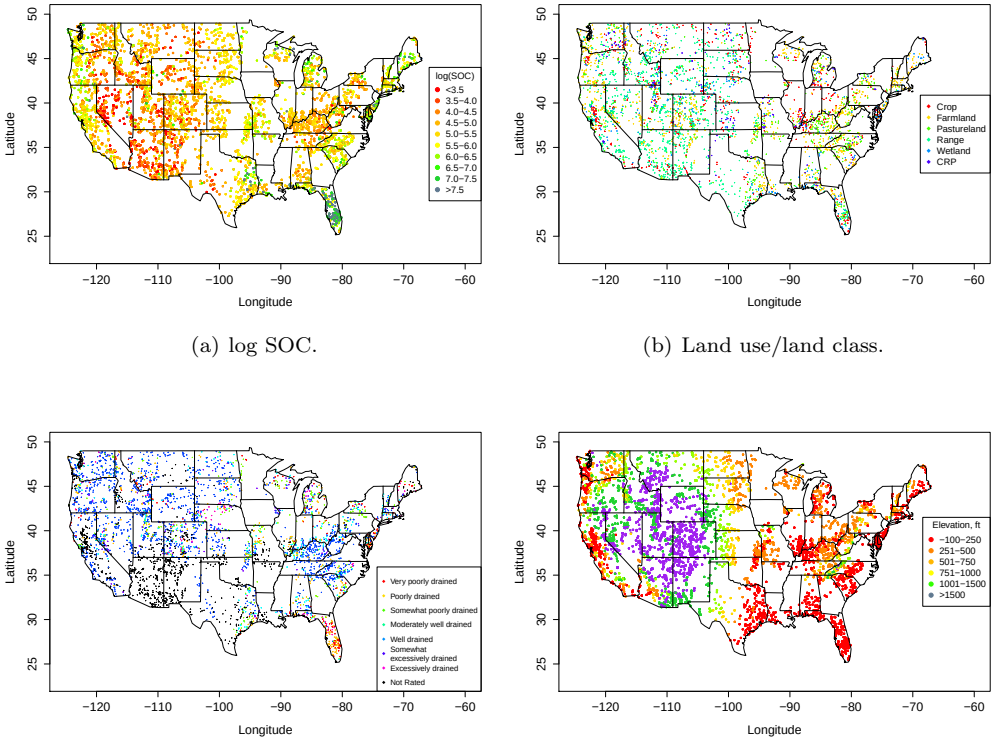
(e.g. within plant or animal residual matter), contains more carbon within the first meters below surface than the atmosphere and terrestrial vegetation combined. Because of this, SOC plays a key role in the global carbon cycle: human activity can transform soils into either a net sink for carbon in the atmosphere, thus contributing to climate change mitigation efforts, or alternatively into a net source of greenhouse gases, contributing to climate change. In addition, SOC is an indicator of soil quality, which greatly influences food productivity (Lefèvre et al., 2017), making SOC a key indicator of the Earth ecosystem and human inhabitants' well being. Because of this, as per recommendation of the Intergovernmental Panel on Climate Change, SOC should be carefully monitored (Lefèvre et al., 2017).

To better understand SOC dynamics, the National Resource Conservation Service (NRCS) began the Rapid Carbon Assessment (RaCA) project in 2010, in which SOC stocks, that is the amount of SOC in a volume of soil, were measured at tens of thousands of fixed locations throughout the CONUS. As collection of SOC data is costly and time consuming (Sleutel et al., 2003; Goidts and Wesemael, 2007), there is a great interest in understanding spatial variability of (log) SOC for data collection purposes as well as for generating maps of estimated (log) SOC. In the past, spatial models used for the analysis of (log) SOC assumed second-order stationarity (Mishra et al., 2009). Risser et al. (2018) effectively demonstrated that this assumption was inappropriate, and introduced a covariate-driven domain segmentation non-stationary spatial statistical model for prediction of log SOC. In this section, we use the mixture M-RA to examine the spatial dependence structure of log SOC in the CONUS with the goal of gaining useful insights that could

inform future sampling campaigns of (log) SOC. Unlike Risser et al. (2018), we incorporate information on land use/cover, drainage class (categorical variables), and elevation (a continuous variable) in the model for the mean function $\mu(\mathbf{s})$, instead of using them to partition the spatial domain. The former two covariates were selected based on the paper by Risser et al. (2018), while the latter was chosen due to its use in an analysis of SOC by Mishra et al. (2009).

For our analysis, we use 1-meter measurements of SOC collected at 20,087 locations in the CONUS for which information on land use/cover, drainage class, and elevation was also available. Figure 2.13 presents a plot of the 20,087 measurements of log SOC. Exploratory analysis for these data in Figure 2.14 indicated extreme right skewness in raw SOC, which can be remedied using a log transformation, as well as a strong non-stationarity highlighted by the empirical semi-variograms within each of the 48 conterminous US states. The variograms are fit to the residuals from linear models that regress log(SOC) on elevation, land use/land class, and drainage class. They provide empirical evidence that the underlying process is non-stationary.

We present a similar exploratory analysis in Figure 2.15, which displays six separate sub-regions for which empirical-semi variograms are fit, again to the residuals of linear models that regress log(SOC) on elevation, land use/land class, and drainage class. We construct confidence intervals around these variograms by fitting the model to bootstrap samples of the data. This results further confirm the assumption of stationarity is untenable, and that a non-stationary spatial statistical model would be more appropriate for these data.



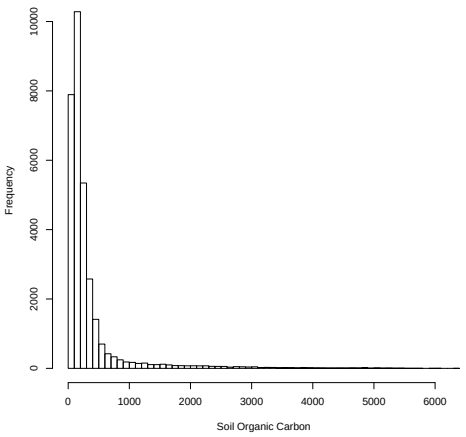
(a) log SOC.

(b) Land use/land class.

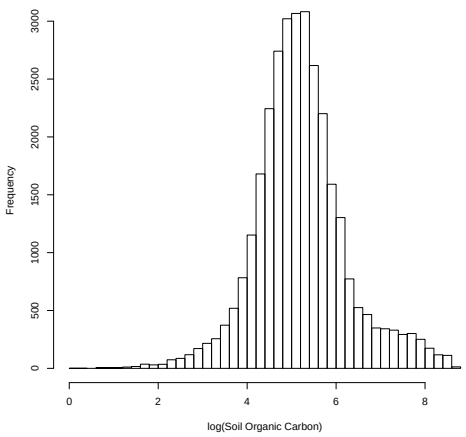
(c) Drainage class.

(d) Elevation.

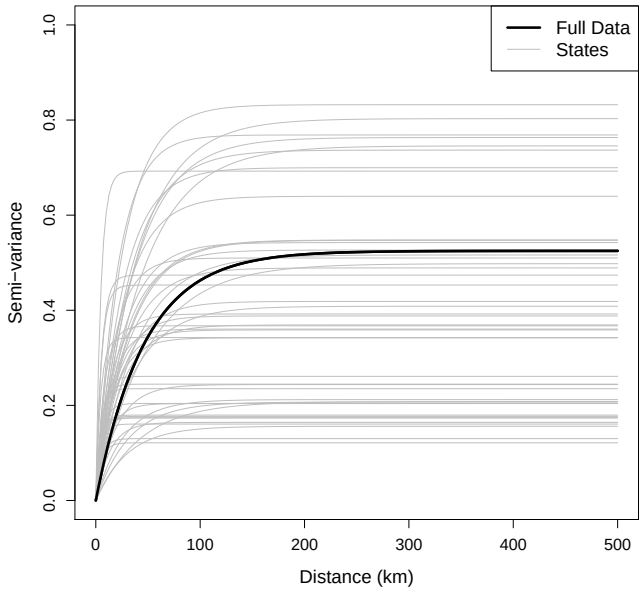
Figure 2.13: Soil Organic Carbon (SOC) exploratory analysis: (a) Measurements of log SOC; (b) land use/land class; (c) drainage class; (d) elevation.



(a) Histogram of SOC.

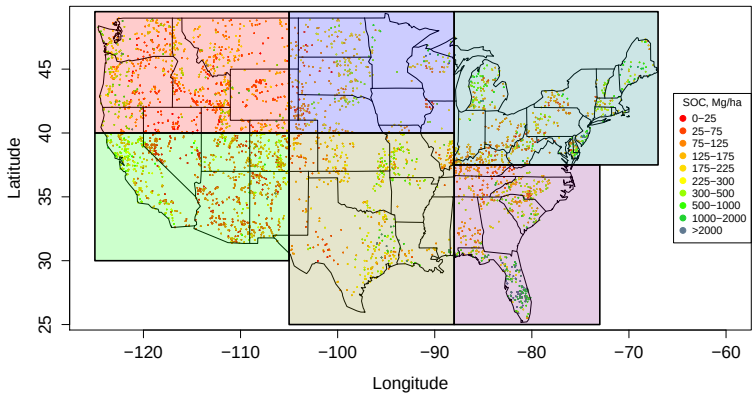


(b) Histogram of log SOC.

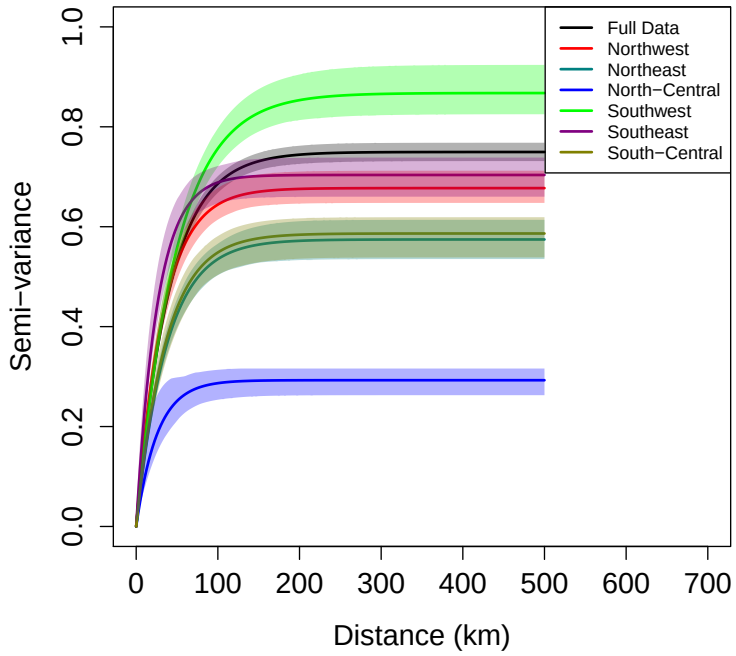


(c) Fitted semi-variograms of 48 states.

Figure 2.14: SOC exploratory analysis: (a) Histogram of SOC; (b) histogram of log(SOC); (c) fitted semi-variograms for each of the 48 conterminous states. Semi-variograms are fit to the residuals of a linear model regressing log SOC on elevation, land use/land cover, and drainage class.



(a) log SOC with sub-regions denoted.



(b) Fitted semi-variograms for 6 sub-regions.

Figure 2.15: SOC Exploratory analysis: (a) map of regions for variogram analysis (b) fitted semi-variograms for 6 sub-regions of the CONUS with confidence bands. Semi-variograms are fit to the residuals of a linear model regressing log SOC on elevation, land use/land cover, and drainage class.

We model log SOC as in (2.10), with $\mu(\mathbf{s})$ a linear function of land use/land cover, drainage class and elevation, a stationary Matérn covariance function to define the basis functions, and M, J, and r set equal to 4, 4 and 16, respectively. We run the MCMC algorithm for 10,000 iterations, keeping L=100 after burn-in, assessing convergence via Geweke's (Geweke, 1992) and Raftery and Lewis' (Raftery and Lewis, 1992) diagnostics.

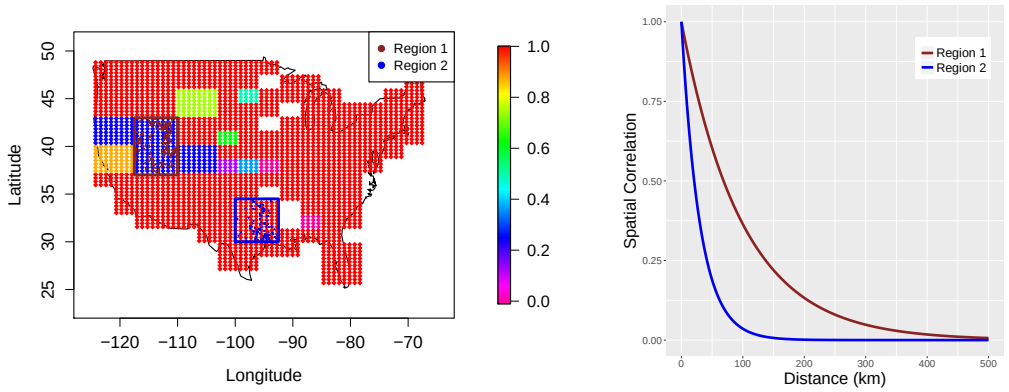
After discarding the first 5,000 iterations for burn-in, we derive posterior inference and generated predictions of log SOC at 2,000 hold-out sites according to the posterior predictive distribution. We evaluate the out-of-sample predictive performance of our model and compare it to that of a stationary Bayesian Kriging model that we fit using a predictive process approximation (Banerjee et al., 2008) due to the large size of the dataset. As another benchmark, we utilize the convolution-based non-stationary model of Risser and Calder (2017), which we select among all the non-stationary models proposed in the literature partly because statistical software to implement it is readily available. To assess the impact of the covariates on predictions of SOC, we also fit the three models without covariates in the mean function $\mu(\mathbf{s})$.

Figure 2.16 plots the posterior means of the latent binary variables $Z_{m,j}$ at their respective knots locations. The regions in which the basis function weights are shrunk to zero are expected to contain observations whose residual spatial correlation decays more slowly. To validate this, using land use/land cover, drainage class, and elevation as covariates, we fit a likelihood-based spatial statistical model to log SOC in the two regions denoted in Figure 2.16 (a): Region 1, further West, which

contains knots corresponding to basis function weights shrunk towards zero and Region 2, southeast of Region 1, in which basis function weights remain "active" in the model. Figure 2.16 (b) shows the estimated Matérn correlation function in the two regions. Consistent with our intuition, Region 1, in which basis function weights are shrunk towards zero, exhibits more slowly decaying spatial correlation than Region 2, in which the basis function weights are active in the model.

Figure 2.17 presents maps of the predicted log SOC at observation sites along with the corresponding posterior predictive standard deviations as yielded by the mixture M-RA and the stationary Bayesian Kriging model. As noted by Fuglstad et al. (2015) and references therein, using a non-stationary covariance function when modeling a spatial process provides pronounced differences particularly in terms of prediction uncertainty. As Figure 2.17 (c) and (d) illustrate, posterior predictive standard deviations are typically larger and more spatially homogeneous under the stationary Bayesian Kriging model than under the mixture M-RA. Figure 2.17(c) also identifies regions where the residual spatial correlation in log SOC persists at long distances. This is particularly useful for planning future SOC data collection campaigns: more sampling efforts should be concentrated in regions with large prediction uncertainty and where the spatial correlation has a short effective range. Based on Figure 2.17(c), more intensive SOC monitoring should occur in Western Texas, Eastern Montana-Wyoming/Western North and South Dakota, and Northern Mississippi/Alabama/Georgia, among others.

Table 2.8 presents results on the out-of-sample predictive performance of the various models. As the table shows, the mixture M-RA models with and without



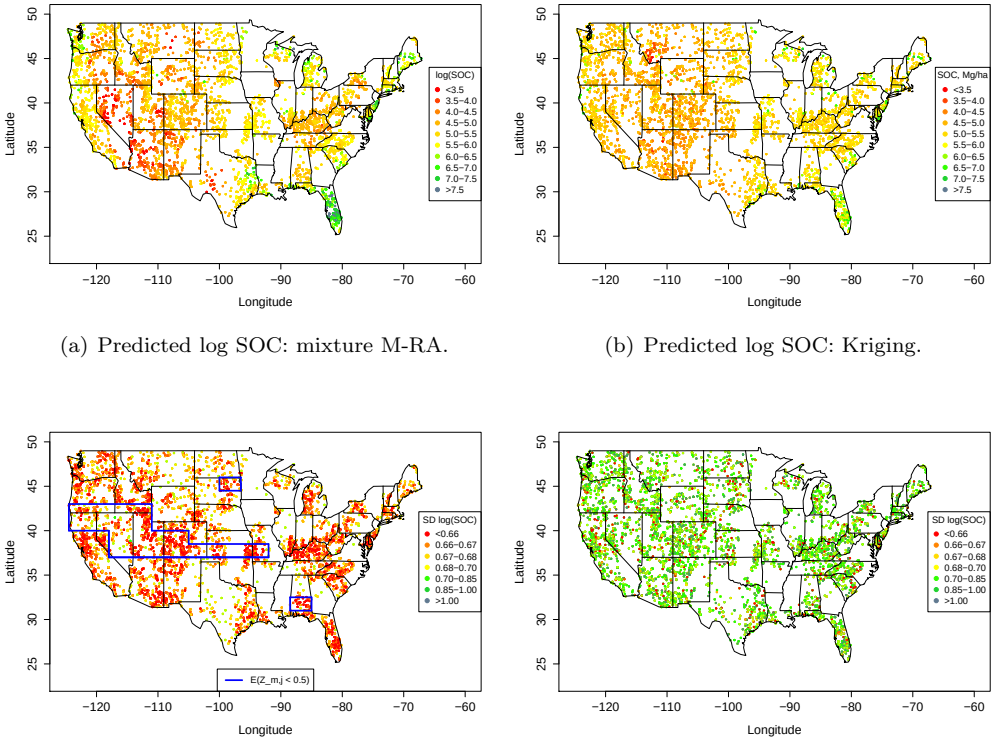
(a) $\mathrm{E}(Z_{4,j}|\mathbf{y})$.

(b) Fitted spatial correlation for two regions.

Figure 2.16: SOC analysis. (a) Posterior means of $Z_{m,j}$ at the highest level (m=4). Magenta-pink regions are ones in which the basis function weights at level 4 were shrunk towards zero. Regions with no posterior means of $Z_{m,j}$ are regions where no SOC observations are collected and thus no knots were placed in those locations, as per Katzfuss (2017) recommendation. (b) Estimated Maérn correlation functions in the selected subregions. Results indicate that the residuals in the region in which basis function weights are shrunk towards zero have spatial correlation with slower rates of decay.

covariates outperform the predictive process models across all predictive performance criteria. Comparing the two mixture M-RA models, we find only a moderate drop in predictive performance when covariates are excluded from the model. While it is preferable to have covariate information whenever possible, the results are reassuring to researchers who wish to make predictions of log SOC at locations where covariates, as well as SOC, are unobserved.

Our model has similar predictive accuracy to the model of Risser and Calder (2017), indicating that our model has utility beyond just identifying regions of non-stationarity: it is a viable non-stationary spatial statistical model in its own right. This is further supported by examining the performance of the stationary Bayesian Kriging model, which has worse prediction accuracy than both our model



(a) Predicted log SOC: mixture M-RA.

(b) Predicted log SOC: Kriging.

(c) PPSD: mixture M-RA.

(d) PPSD: Kriging.

Figure 2.17: Soil Organic Carbon (SOC) analysis: (a)-(b) Predicted log SOC as yielded (a) by the mixture M-RA model and (b) by the stationary Bayesian Kriging model . (c)-(d) Posterior predictive standard deviation as yielded (c) by the mixture M-RA model and (d) by the stationary Bayesian Kriging model. In (c) the blue lines delineate regions where the posterior mean of the $Z_{m,j}$'s at the highest level, m=4, is less than 0.5. We identify these as regions of local stationarity.

Model	MSPE	Rel. MSPE	Coverage 95% PI
Mixture M-RA with covariates	0.42	0.10	0.951
Mixture M-RA without covariates	0.46	0.11	0.957
Non-stationary convolution model with covariates	0.46	0.11	0.951
Non-stationary convolution model without covariates	0.50	0.13	0.940
Stationary Bayesian Kriging model with covariates	0.60	0.16	0.937
Stationary Bayesian Kriging model without covariates	0.64	0.18	0.920

Table 2.8: SOC analysis. Assessment of out-of-sample predictive performance of the various models reported in terms of Mean Squared Prediction Error (MSPE), Relative Mean Squared Prediction Error (Rel. MSPE), and empirical coverage of 95% prediction intervals.

and that of Risser and Calder (2017).

In terms of computation time, while a direct comparison between the mixture M-RA and the model of Risser and Calder (2017) is not possible due to the fact that latter is not implemented in a fully Bayesian inferential framework, compared to the predictive process our model took approximately 3.07 times less per MCMC iteration. Like the M-RA modeling framework itself, the mixture M-RA model is amenable to parallel computing (see Katzfuss and Hammerling (2017) for parallel inference with the M-RA), which will render the mixture M-RA model faster to implement even with massive spatial datasets.

Finally, we present MCMC convergence diagnostics (Geweke's diagnostics and Raftery and Lewis' diagnostics) for the covariance parameters of the Matérn covariance function used to define the basis functions in the mixture M-RA model fitted to the log SOC data (Table 2.9). On the other hand, Table 2.10 provides a comparison of the mixture M-RA model and the stationary Bayesian Kriging

model based on the Posterior Predictive Loss, one of the most common criteria for model selection used in the book by Daniels and Hogan (2008). The PPL for both models, is defined by Gelfand and Ghosh (1998) as:

$$\mathcal{L}_k(\mathbf{y}_{\text{rep}}, a; \mathbf{y}) = \mathcal{L}(\mathbf{y}_{\text{rep}}, a) + k\mathcal{L}(\mathbf{y}, a) \tag{2.13}$$

with $\mathbf{y}_{\text{rep}}$ a new replicate of the data drawn from the distribution of the data likelihood and $\mathcal{L}(\cdot)$ a loss function. In (2.13) k and a are constants, with the latter chosen to minimize the expected loss with respect to the posterior predictive distribution (Daniels and Hogan, 2008; Gelfand and Ghosh, 1998). Following Finley and Banerjee (2013), who use PPL as a model selection criteria in a geostatistical setting, we take $\mathcal{L}(\cdot)$ to be the squared error loss function (i.e. the squared difference between the predicted value and true value at held-out locations). In this setting, it can be shown that the PPL takes the form:

$$\sum_{i=1}^{n} \sigma_i^2 + \frac{k}{k+1} \sum_{i=1}^{n} (\mu_i - y_i)^2$$

where μ_i and σ_i^2 denote, respectively, the posterior predictive mean and standard deviation of $y_{\text{rep},i}$. Table 2.10 tabulates values of the posterior predictive loss for the mixture M-RA model and for the stationary Bayesian Kriging model applied to log SOC for different values of k, as suggested in Gelfand and Ghosh (1998).

Parameter	Geweke's diagnostic	Raftery and Lewis' diagnostic
σ^2	0.71	3,305
ϕ	-1.23	4,116
ν	1.09	4,329

Table 2.9: Geweke's and Raftery and Lewis' diagnostics for covariance function parameters in the log SOC analysis. The Raftery and Lewis' diagnostic reports the required sample size to infer upon the 2.5^{th} posterior percentile of the corresponding parameter with an accuracy of 0.01.

k	PPL Mixture M-RA	PPL Stationary Bayesian Kriging
1	11,629.31	23,911.64
3	13,415.44	28,065.90
9	14,187.12	30,558.45
∞	15,201.57	32,220.15

Table 2.10: Posterior predictive loss (PPL) comparing the mixture M-RA to the stationary Bayesian Kriging model in the analysis of Soil Organic Carbon.

2.4 Discussion

Analysis of a spatial process often involves as a first step the selection of a covariance function for the process. In applications, stationary covariance functions are often used, even though the assumption of stationarity might not be warranted for the data. In this chapter, we have proposed a modeling framework that is flexible enough to accommodate both stationary and non-stationary spatial data, when the non-stationarity in the dependence structure is due to inhomogeneities in the range of the spatial correlation. Application of our model to both stationary (result presented in the Appendix A.3) and non-stationary data have shown that our model has an out-of-sample predictive performance that is comparable to that of a stationary model when the data are indeed a realization of a stationary Gaussian process, and that of a non-stationary model when the data are indeed non-stationary. In addition, inference from our model allows one

to identify regions of local stationarity.

While tests for non-stationarity, isotropy and symmetry of the covariance function of a spatial process have been proposed in the literature (Guan et al., 2004; Li et al., 2008; Jun and Genton, 2012; Bandyopadhyay and Subba Rao, 2017; Weller and Hoeting, 2016), in the case of a locally stationary process, none of these methods allow for easy detection of regions of local stationarity. Determining whether there exist regions in the spatial domain where a spatial process displays varying strength of spatial dependence is extremely important for various reasons. From a sampling design perspective, knowing that in different regions the spatial process is characterized by a different range parameter, could lead to a differential strategy when collecting observations or when placing monitoring devices, as we have discussed in the analysis of SOC in Section 2.3.2. From a computational point of view, the decomposition of the spatial domain in regions of local stationarity can lead to computational savings as a spatial model can be fit to data within each region individually. Our model also allows one to determine the number of M-RA levels needed to approximate the covariance structure in the data, a question that is often only addressed empirically.

There are multiple ways in which our model could be extended and improved. First, for now we have only been considering Gaussian spatial processes: it would be interesting, and potentially not too difficult, to extend the mixture M-RA modeling framework to non-Gaussian spatial data. In its current form, our model only accommodates non-stationarity due to inhomogeneities in the range of the correlation function; extensions of this work could be geared towards accommodating

other types of non-stationarity. The use of an anisotropic spatial covariance function could also be explored. The prior specification on the M-RA basis function weights involves a mixture of two normal priors, with one of the two normal distributions introducing an additional parameter, the shrinkage parameter L. In our implementation, L is tuned through cross-validation. A further avenue of research could be to investigate how to provide a prior on L and infer upon it using the data.