UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Exatas
Programa de Pós-Graduação em Estatística

Lucas Azevedo Birro Michelin

Fast Mixture Spatial Regression: A Mixture in the Geographical and Feature Space Applied to Predict Oil in the Post-salt

Belo Horizonte
2023

Lucas Azevedo Birro Michelin

Fast Mixture Spatial Regression: A Mixture in the Geographical and Feature Space Applied to Predict Oil in the Post-salt

Final Version

Dissertation presented to the Graduate Program in Statistics of the Federal University of Minas Gerais in partial fulfillment of the requirements for the degree of Master in Statistics.

Advisor: Marcos Oliveira Prates

Belo Horizonte
2023

Michelin, Lucas Azevedo Birro.

M623f Fast mixture spatial regression: a mixture in the geographical and feature space applied to predict oil in the post-salt [recurso eletrônico] / Lucas Azevedo Birro Michelin. — 2023.
423f. il.; 29 cm.

Orientador: Marcos Oliveira Prates.
Dissertação (mestrado) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística.
Referências: f.40-43.

1. Estatística – Teses. 2. Estatística Espacial – Teses. 3. Teoria bayesiana de decisão estatística – Teses. 4. Petróleo – Pós-Sal – Brasil – Teses. I. Prates, Marcos Oliveira. II. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística. III. Título.

CDU 519.2(043)

Ficha catalográfica elaborada pela bibliotecária Irénquer Vismeg Lucas Cruz
CRB 6/819 - Universidade Federal de Minas Gerais - ICEx

FOLHA DE APROVAÇÃO

"Fast Mixture Spatial Regression: A Mixture in the Geographical and Feature Space Applied to Predict Oil in the Post-salt"

LUCAS AZEVEDO BIRRO MICHELIN

Dissertação submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em ESTATÍSTICA, como requisito para obtenção do grau de Mestre em ESTATÍSTICA, área de concentração ESTATÍSTICA E PROBABILIDADE.

Aprovada em 15 de agosto de 2023, pela banca constituída pelos membros:

Prof. Marcos Oliveira Prates - Orientador
DEST/UFMG

Profa. Thais Paiva Galletti
DEST/UFMG

Luis Mauricio Castro Cepero
PUC-Chile

Belo Horizonte, 15 de agosto de 2023.

Dedico este trabalho à memória eterna da minha querida avó, Maristela Coelho Azevedo. Ela sempre foi minha fonte de inspiração, meu farol de conhecimento e minha maior incentivadora na busca pela educação. Sua presença amorosa em minha vida é insubstituível, e é com profunda gratidão que celebro todas as minhas conquistas em sua honra. Seu legado de sabedoria e apoio permanecerá comigo para sempre.

Acknowledgments

Primeiramente, gostaria de expressar minha profunda gratidão às pessoas que tornaram possível a realização deste trabalho.

À minha mãe, Stela, e à minha irmã, Camila, agradeço por sempre estarem ao meu lado, me apoiando e incentivando na busca pelos meus objetivos. À Izabella, dedico um agradecimento especial por seu apoio inabalável, paciência incansável e carinho constante ao longo de todo o processo. Sem a sua presença e apoio, esta conquista não teria sido alcançada.

Ao meu orientador, Marcos Oliveira Prates, sou imensamente grato pela oportunidade de pesquisa, valiosos ensinamentos, compreensão e ajuda desprendida com este trabalho. Sua orientação foi essencial para o sucesso deste projeto. Gostaria também de estender meus agradecimentos a todos os meus amigos, colegas e professores que contribuíram de maneira significativa para que esta dissertação se tornasse realidade. Em especial, agradeço ao Lucas Godoy e ao professor Heitor Ramos por sua inestimável ajuda e orientação sempre que precisei.

Gostaria de agradecer ao suporte financeiro das agências de fomento CAPES, CNPq e FAPEMIG que auxiliam na infraestrutura do programa e facilitaram realização dessa dissertação.

Resumo

Extrair recursos geológicos, como fluidos de hidrocarbonetos, requer investimentos significativos e processos de tomada de decisão precisos. Para otimizar a eficiência do processo de extração, pesquisadores e especialistas da indústria têm explorado metodologias inovadoras, incluindo a previsão de locais de perfuração ótimos. A porosidade, um atributo fundamental das rochas de um reservatório, desempenha um papel crucial na determinação da sua capacidade de armazenamento de fluidos. Técnicas geoestatísticas, como a "krigagem", têm sido amplamente utilizadas para estimar a porosidade, capturando a dependência espacial em dados de amostras pontuais. No entanto, a dependência das coordenadas geográficas para determinar distâncias espaciais pode apresentar desafios em cenários de pequenas amostras e amplamente separadas. Neste artigo, desenvolvemos um modelo de mistura que combina a covariância gerada pelo espaço geográfico e a covariância gerada em um espaço de covariáveis (*features*) apropriado para aprimorar a precisão da estimativa. Desenvolvido no contexto Bayesiano, nossa abordagem utiliza métodos de Monte Carlo com Cadeia de Markov (MCMC) e aproveita a estratégia do Processo Gaussiano dos vizinhos mais próximos (NNGP) para atingir escalabilidade. Apresentamos uma comparação em um estudo de simulação, considerando várias configurações para geração dos dados, a fim de avaliar o desempenho do modelo de mistura em comparação aos modelos marginais. Além disso, a aplicação dos nossos modelos em uma simulação de reservatório tridimensional demonstra sua aplicabilidade prática e escalabilidade. Esta pesquisa apresenta uma abordagem inovadora para a melhoria da estimativa de porosidade, integrando informações espaciais e de covariáveis, oferecendo o potencial para otimizar atividades de exploração e extração de reservatórios.

Palavras-chave: Estatística espacial, cokriging, computação bayesiana, espaço das *features*, estimação de porosidade

Abstract

Extracting geological resources like hydrocarbon fluids requires significant investments and precise decision-making processes. To optimize the efficiency of the extraction process, researchers and industry experts have explored innovative methodologies, including the prediction of optimal drilling locations. Porosity, a key attribute of reservoir rocks, plays a crucial role in determining fluid storage capacity. Geostatistical techniques, such as kriging, have been widely used for estimating porosity by capturing spatial dependence in sampled point-referenced data. However, the reliance on geographical coordinates for determining spatial distances may present challenges in scenarios with small and widely separated samples. In this paper, we develop a mixture model that combines the covariance generated by geographical space and the covariance generated in an appropriate feature space to enhance estimation accuracy. Developed within the Bayesian framework, our approach utilizes flexible Markov Chain Monte Carlo (MCMC) methods and leverages the Nearest-Neighbor Gaussian Process (NNGP) strategy for scalability. We present a controlled empirical comparison, considering various data generation configurations, to assess the performance of the mixture model in comparison to the marginal models. Applying our models to a three-dimensional reservoir simulation demonstrates its practical applicability and scalability. This research presents a novel approach for improved porosity estimation by integrating spatial and covariate information, offering the potential for optimizing reservoir exploration and extraction activities.

Keywords: Spatial statistics, cokriging, computation bayesian methods, feature space, porosity estimation.

Resumo Estendido

A extração de recursos geológicos, especialmente fluidos de hidrocarboneto, é um empreendimento complexo e intensivo, exigindo investimentos substanciais com grandes períodos de retorno [4]. Essa busca por ganhos financeiros significativos tem despertado grande interesse e investimento, motivando pesquisadores e especialistas da indústria a explorar metodologias e estratégias inovadoras destinadas a otimizar a eficiência do processo de extração. A recuperação desses fluidos envolve o procedimento de perfuração do subsolo, exigindo execuções meticulosa devido à complexidade da construção envolvida e aos significativos custos de produção incorridos [15]. Consequentemente, pesquisadores têm se dedicado à tarefa de estimar locais ideais para perfuração de poços. Uma abordagem para alcançar esse objetivo é prever as regiões com alta probabilidade de conter reservatórios de petróleo e/ou gás natural. Um aspecto essencial que representa a capacidade e eficácia dos reservatórios reside em sua estrutura porosa. As características físicas dos reservatórios desempenham um papel fundamental na moldagem das atividades de exploração e extração nos campos de petróleo e gás. A porosidade, entre essas características, assume destaque como um indicador da capacidade de armazenamento de fluidos de uma rocha. Estimar a porosidade das rochas dentro de um reservatório envolve a utilização de várias metodologias, e entre elas, a técnica geostatística conhecida como "kriging" surge como um método proeminente [9, 25].

A técnica estatística conhecida como ordinary kriking, comumente referida apenas como kriging, é amplamente utilizada para interpolação espacial de dados de referência pontual, a fim de estimar os valores de uma variável em um campo espacial contínuo [20, 23]. Ao contrário dos métodos de interpolação tradicionais, o kriging considera não apenas a posição espacial entre os pontos observados adjacentes ao ponto a ser interpolado, mas também a relação posicional entre pontos vizinhos próximos a cada valor amostrado. Ao incorporar essa dependência espacial estruturada, o kriging facilita a estimativa e previsão de valores em locais onde não há observações diretas disponíveis [5]. Uma abordagem comum é assumir que o processo alvo segue um Processo Gaussiano (GP) caracterizado por uma média específica e uma função de covariância válida. Frequentemente, essas funções de covariância são modeladas como o produto de um parâmetro de variância e uma função de correlação que depende da distância Euclidiana entre as coordenadas geográficas dos pontos espaciais. No entanto, em muitos cenários, usar a métrica Euclidiana para determinar distâncias espaciais pode ser impraticável [7]. Isso se torna particularmente evidente em situações em que os pontos distribuídos espacialmente estão amplamente

separados por distâncias geográficas substanciais. Nesses casos, a força da dependência espacial diminui à medida que a distância entre os pontos aumenta, levando a uma perda de correlação dentro do processo. Isso pode representar desafios para prever valores em locais não amostrados. Além disso, a suposição de suavidade do processo, que é inerente ao kriging ao modelar com base na estrutura espacial gerada pelas coordenadas geográficas (ou seja, efeito de suavização), pode não condizer com a realidade [36].

Nosso trabalho introduz um modelo de mistura que combina contribuições tanto da covariância gerada pelo espaço geográfico (ou seja, o modelo geográfico) quanto da covariância gerada pelas covariáveis disponíveis (ou seja, o modelo de *features*). Essa abordagem permite uma integração mais eficaz das contribuições de cada modelo, adaptando-se à localização específica estimada. Implementamos nosso método no pacote `FUSE` no framework bayesiano utilizando o pacote `nimble` [8], disponível para o software `R` [26]. Isso possibilita uma amostragem eficiente dos parâmetros espaciais, proporcionando maior flexibilidade no processo de modelagem. Além disso, nossa metodologia proposta incorpora a escalabilidade, aproveitando a estratégia do processo Gaussiano dos vizinhos mais próximos (NNGP) [6]. Essa abordagem se beneficia da dependência espacial capturada por locais próximos à localização-alvo. Utilizando a esparsidade da matriz de covariância derivada dessas observações mais próximas, os modelos NNGP possibilitam inferências eficientes e escaláveis para grandes conjuntos de dados.

Resultados do estudo de simulação destacam a eficácia de nossa proposta em estimar com precisão os parâmetros espaciais e prever o processo-alvo. Especificamente, comparamos três modelos: geográfico, de *feature* e de mistura, em diferentes cenários. O modelo geográfico teve um bom desempenho quando a correlação espacial era estritamente dependente do espaço, enquanto o modelo de *features* se destacou ao utilizar as informações inerentes das covariáveis. No entanto, o modelo de mistura surgiu como uma escolha significativa, incorporando efetivamente contribuições tanto dos modelos geográficos quanto dos modelos de *features*, dependendo da localização estimada. O modelo de mistura obteve estimativas superiores da verdadeira configuração espacial, tornando-se uma ferramenta valiosa para capturar estruturas espaciais complexas onde os modelos tradicionais podem falhar.

Adicionalmente, aplicamos a abordagem proposta a um conjunto de dados simulados de um reservatório de petróleo, demonstrando sua aplicabilidade prática. O modelo de mistura integrou com sucesso características geológicas representadas por covariáveis, como a litologia do reservatório, na estimativa da porosidade, proporcionando previsões mais precisas em comparação com os outros modelos. Essa capacidade de incorporar características geológicas intrínsecas tornou o modelo de mistura particularmente adequado para modelar distribuições de porosidade neste contexto específico de reservatório.

De forma geral, modelo de mistura NNGP se destacou como uma estrutura robusta e versátil para modelagem geoestatística, capaz de lidar com grandes conjuntos de

dados ao mesmo tempo em que mantém previsões precisas. Sua capacidade de combinar informações tanto do espaço geográfico quanto do espaço de covariáveis abre novas oportunidades para diversas aplicações geoespaciais.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

The extraction of geological resources, particularly hydrocarbon fluids, is a complex and capital-intensive endeavor, necessitating substantial investments and extended pay-back periods [4]. This pursuit of significant financial gains has spurred considerable interest and investment in the field, motivating researchers and industry experts to explore innovative methodologies and strategies aimed at optimizing the efficiency of the extraction process. The retrieval of these fluids involves the intricate procedure of drilling into the subsurface, requiring meticulous execution due to the complexity of construction involved and the significant production costs incurred [15]. Consequently, researchers have undertaken the task of estimating optimal locations for well drilling. One approach to accomplish this objective is predicting regions with a high likelihood of containing oil and/or natural gas reservoirs. An essential aspect that signifies the capacity and effectiveness of reservoirs lies in their pore structure. The physical attributes of reservoirs play a pivotal role in shaping exploration and extraction activities within the oil and gas fields. Porosity, among these attributes, assumes prominence as an indicator of a rock's fluid storage capability. Estimating the porosity of rocks within a reservoir involves the utilization of various methodologies, and amidst them, the geostatistical technique known as "kriging" emerges as a prominent method [9, 25].

The statistical technique known as ordinary kriging, commonly referred to as kriging, is widely used for spatial interpolation of sampled point-referenced data to estimate the values of a variable across a continuous spatial field [20, 23]. In contrast to traditional interpolation methods, kriging considers not only the spatial position between adjacent observation data points and the point being interpolated but also the positional relationship among neighboring points near each sampled value. By incorporating the structured spatial dependence, kriging facilitates the estimation and prediction of values at locations where no direct observations are available [5]. This spatial dependence structure arises from the covariance matrix, which is obtained by considering the distances between pairs of observed points within the spatial domain. This spatial relationship is established by utilizing geographic coordinates and the subsequent computation of inter-coordinate distances. The covariance matrix assumes a crucial role in quantifying spatial dependence, serving as a fundamental component in diverse spatial statistical analyses [1]. When

multiple variables exhibit spatial interdependence, the kriging method can be naturally extended to a technique called cokriging. This extension shares similar properties with kriging and enables the inclusion of supplementary auxiliary information in the interpolation process. Through the exploitation of the interrelationship between the primary variable of interest and one or more auxiliary variables, cokriging enhances prediction accuracy and offers valuable insights into spatial patterns [19, 14]. Illustrative applications of these methodologies include the prediction of soil water storage [31], the estimation of population density in urban areas [35] and the interpolation of volume-weighted velocity statistics in cosmic velocity fields [37].

An usual approach is to assume that the target process conforms to a Gaussian Process (GP) characterized by a specific mean and a valid covariance function. Often, these covariance functions are modeled as the product of a variance parameter and a correlation function that relies on the Euclidean distance between the geographical coordinates of the spatial points. Nonetheless, in many scenarios, employing the Euclidean metric to determine spatial distances may be impractical [7]. This becomes particularly evident in situations where the spatially distributed points are widely separated by substantial geographical distances. In such cases, the strength of spatial dependence diminishes as the distance between points increases, leading to a loss of correlation within the process. This can pose challenges for predicting values in unsampled locations. Furthermore, the assumption of process smoothness, which is inherent in kriging when modeling based on the spatial structure generated by geographical coordinates (i.e. smoothing effect), may not align with reality [36]. Considering these factors, the exploration of alternative domains, not strictly limited to spatial characteristics, for determining similarity between points in space can be of significant interest. For instance, when covariates are highly indicative of the phenomenon under study, they can be employed to construct the spatial dependence structure instead of relying exclusively on geographical coordinates. The machine learning community has long explored the use of GP with features in the covariance function with mainly two purposes: 1) improve prediction; and 2) to allow for a non-linear relationship with the response [34, 22, 27]. The methodology was successfully applied to face recognition [17], drug discovery [13], and chemical synthesis [29], among others. In recent times, there has been relatively limited focus on constructing dependence structures within the domain generated by covariates in the spatial statistics literature [e.g., 21].

This paper introduces a mixture model that combines contributions from both the covariance generated by geographical space (i.e. *geographical model*) and the covariance generated by available covariates (i.e. *feature model*). This approach allows for more effective integration of the contributions from each model, adapting to the specific estimated location. We implemented our method in the `FUSE` package on the Bayesian framework using the `nimble` package [8] available for the `R` software [26]. This enables the efficient sampling of spatially correlated parameters, providing enhanced flexibility in the modeling

process. Moreover, our proposed methodology incorporates scalability by leveraging the Nearest-Neighbor Gaussian Process (NNGP) strategy [6]. This approach takes advantage of the spatial dependence captured by neighboring locations in close proximity to the target location. By utilizing the sparsity of the covariance matrix derived from these nearest observations, the NNGP models enable efficient and scalable inference for large datasets. The viability of employing covariates to determine the spatial dependence structure is showcased through a simulated study. This study highlights the ability of the mixture model to accurately recover the marginal models and its behavior in an intermediate configuration.

To illustrate the practical applicability and scalability of our approach, we utilized data from a three-dimensional reservoir simulation created by a geomechanical model. In our analysis, we will illustrate that for this dataset, the feature model exhibits greater sensitivity to the distribution of porosity. However, the mixture model effectively captures this relationship and yields competitive results for porosity prediction in new locations. Furthermore, we will demonstrate that despite the available number of points, the models demonstrate scalability and maintain robustness when employing the NNGP methodology.

The rest of the paper is organized as follows. Chapter 2 presents a comprehensive overview of the dataset used in this study, including details about the simulation of the oil reservoir and the available covariates. In Chapter 3, we revisit essential concepts from the literature that are essential for constructing our proposed model. Chapter 4 introduces and justifies the structure of the feature model in contrast to the geographical model, as well as the formulation of the mixture model that combines the two models. Chapter 5 summarizes the results of a simulation study where we analyze the performance of the three models across surfaces created by different scenarios. In Chapter **??**, we present the outcomes of fitting the models using the data from the simulated reservoir and compare their prediction performances in two specified regions of interest. Chapter **??** concludes the article.

Chapter 2

The Data

The dataset utilized in this study originates from a three-dimensional geomechanical model that simulates an oil reservoir. Specifically, the simulated data represents the Marimba oil field, which is a turbidite sandstone reservoir located in the post-salt area of Campos Basin, approximately 80 km from Macaé in the state of Rio de Janeiro, in the southern offshore region of Brazil [16]. In Figure 2.1(left), the geographical location of the Marimba field is indicated. The construction of this model involves the utilization of diverse data sources, which includes well drilling log data, measurements acquired from seismic wave sources, and crucially, laboratory studies providing reservoir-rock characterization of the reservoir. The resulting dataset obtained from the simulation consists of slightly over 70 million points distributed in a three dimension space (i.e. X,Y and Z), incorporating measurements from 10 wells. The variable of interest investigated in this study is the rock porosity at each of these points. Porosity refers to the proportion of pore volume in a rock compared to its total volume. It represents the rock's capacity to store fluids, and in the context of oil reservoirs, it indicates the potential for oil storage within the rock's pores. Consequently, higher porosity values signify a greater capacity for fluid storage [2].

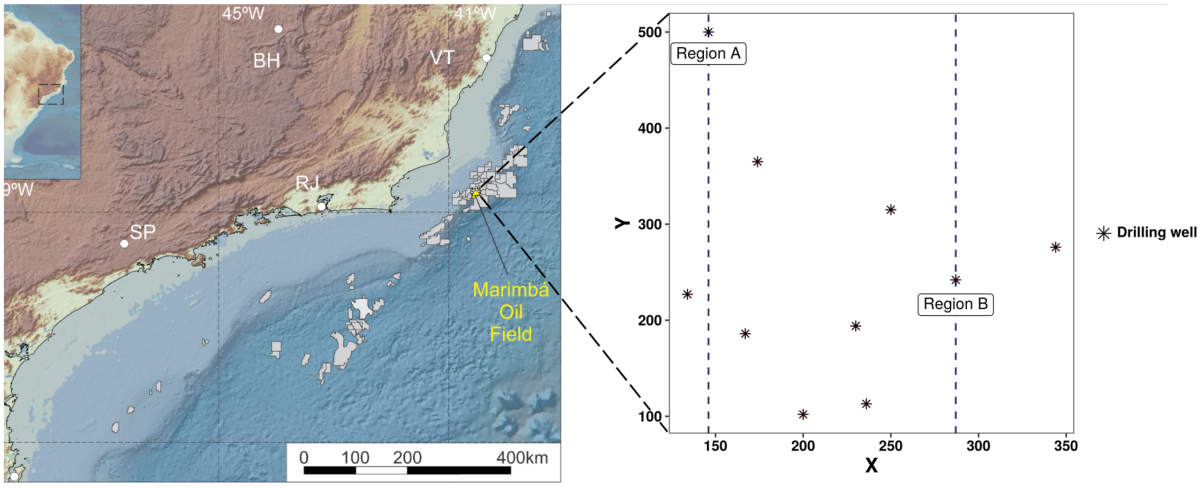

Figure 2.1: Left panel: Geographical location of the Marimba oil field. Right panel: Aerial two-dimensional representation of the ten well drilling locations used for model fitting. The line segments indicate two cross-sectional regions used to measure the porosity estimation capacity.

For each coordinate present in the dataset, alongside the rock porosity value, there are three available covariates that represents the reservoir lithology. These covariates consist of the rock density (ρ) as well as the velocities of the primary (V_p) and secondary (V_s) elastic waves. These wave velocities can be captured through various manners, such as the occurrence of seismic events (e.g., earthquakes) or by intentionally detonating an explosive source. For a more comprehensive understanding of the methodology used to obtain these wave measurements, as well as additional seismic quantities, see [10] and [30].

To evaluate the performance of the proposed model, the dataset was divided as follows: The portion of the dataset containing the well drilling data, which includes 10 available wells and a total of $2,510$ data points, was exclusively used for model fitting and estimation of the parameters of interest. For the assessment of porosity estimation capability, two distinct cross section regions within the remaining cube were selected, totalling $324,292$ locations. Each region includes a well, making them particularly relevant for examining the relationship between porosity prediction in the vicinity of the "ground truth" and the behavior of the models as the points deviate from the actual data. Figure 2.1(right) provides a two-dimensional (X and Y) visualization of the well drilling locations' arrangement, along with the two prediction regions.

Chapter 3

Preliminaries

In this chapter we revisit concepts of the literature necessary to construct and motivate our proposal.

3.1 Spatial Gaussian Process

A spatial Gaussian process (SGP) refers to a stochastic process frequently employed to model data exhibiting spatial, temporal, or spatio-temporal dependence [28]. It is fully specified by its mean function $\mu(\cdot)$ and a valid cross-covariance function $C_\theta(\cdot,\cdot)$. For a multivariate stochastic process $\{Y(s) : s \in D\}$, defined over a domain $D \subset \mathbb{R}^r$, the process is considered a spatial Gaussian process if, for any distinct choices of locations $s_1, \ldots, s_n \in D$, the random vector $\mathbf{Y} = [Y(s_1), \ldots, Y(s_n)]^\top$ follows a multivariate normal distribution with a mean $E[\mathbf{Y}] = \boldsymbol{\mu} = [\mu(s_1), \ldots, \mu(s_n)]^\top$ and covariance matrix $\Sigma_\theta = C_\theta\left(y(s_i), y(s_j)\right)$ [3]. The density function $f(\mathbf{Y})$ is defined by:

$$f(\mathbf{Y}) = (2\pi)^{-n/2}|\boldsymbol{\Sigma}_\theta|^{-1}\exp\left\{-\frac{1}{2}(\mathbf{Y}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}_\theta^{-1}(\mathbf{Y}-\boldsymbol{\mu})\right\}. \tag{3.1}$$

The selection of the covariance function is critical in modeling the process, as it encapsulates assumptions about the underlying surface [24]. Isotropic covariance functions are frequently selected for C_θ, wherein the covariance between two spatial points, s_i and s_j, depends solely on their Euclidean distance. In recent years, spatial statistics research has placed significant emphasis on exploring covariance functions belonging to the Matérn class. The Matérn class of covariance functions is defined as follows [18]:

$$C_\theta(d_{ij}) = \sigma^2\frac{1}{2^{\nu-1}\Gamma(\nu)}\left(\frac{d_{ij}}{\phi}\right)^\nu K_\nu\left(\frac{d_{ij}}{\phi}\right), \tag{3.2}$$

where $d_{ij} = \|s_i - s_j\|$ is the Euclidean distance between locations s_i and s_j, ϕ controls the decay in spatial correlation (also called the *range* parameter) , σ^2 denotes the variance of the spatial process and $K_\nu(\cdot)$ is the modified Bessel function of the second kind with

ν controlling the process smoothness. The Matérn class of covariance functions is known for its versatility. When $\nu \to \infty$, the covariance function converges to the Gaussian specification, while setting $\nu = 0.5$ yields the exponential model. This flexibility allows for the selection of an appropriate covariance function based on the desired behavior and characteristics of the spatial process being modeled.

3.2 Spatial Regression

Let $\{Y(s) : s \in D\}$ be a realization of a stochastic process defined over the domain $D \in \mathbb{R}^r$, where $r \geqslant 1$. A common approach to modeling the process Y is by utilizing a spatial linear mixed effects model:

$$Y(\boldsymbol{s}) = \mu(\boldsymbol{s}) + w(\boldsymbol{s}) + \epsilon(\boldsymbol{s}). \tag{3.3}$$

In this model, the residual component can be decomposed into two parts: a spatial component ($w(\boldsymbol{s})$) and a non-spatial component ($\epsilon(\boldsymbol{s})$). The non-spatial component, often referred to as the *nugget* effect τ^2, is typically modeled as a white noise process [3]. This effect can be attributed to various factors, such as measurement error or microscopic-scale variability. As discussed in 3.1, the spatial component $w(s)$ can be modeled as a *stationary* spatial Gaussian process with zero mean and a covariance function C_θ, i.e. $w(s) \sim SGP(0, C_\theta(s_i, s_j))$. The covariance between $w(s_i)$ and $w(s_j)$ is defined by C_θ, with the parameter vector $\boldsymbol{\theta}$ determining the characteristics of this covariance function. In the subsequent results presented in this study, we selected the covariance function specified in Equation (3.2) with $\nu = 0.5$, leading to the adoption of the exponential model

$$C_\theta(d_{ij}) = \sigma^2 \exp\left(\frac{-d_{ij}}{\phi}\right).$$

As a result, the entry i, j of the covariance matrix $\boldsymbol{\Sigma}$ is given by $\sigma^2 \exp(-d_{ij}/\phi) + \tau^2 \mathbf{1}_{(i=j)}$, where $\mathbf{1}$ is an indicator function. This covariance is derived from the chosen model with $\boldsymbol{\psi} = \{\boldsymbol{\theta}, \tau^2\}$ and $\boldsymbol{\theta} = \{\sigma^2, \phi\}$. This modeling framework provides the flexibility to incorporate both linear and nonlinear processes into the mean function $\mu(s)$. For example, by identifying p spatially referenced covariates, denoted as $\boldsymbol{X}(s) = [x(s_1), \ldots, x(s_p)]^\top$, we can express the relationship between the covariates and the mean of the process as a linear equation, given by $\mu(s) = \boldsymbol{X}^\top(s)\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ represents the vector of regression coefficients.

In the presence of the covariates $\boldsymbol{X}(s)$ and a new location $x(s_0)$ for which a prediction is desired, the classical approach of spatial prediction (kriging) can be employed

for spatial interpolation. In this context, the model described in Equation (3.3) adopts the general form

$$Y(\mathbf{s}) = \boldsymbol{X}^\top(s)\boldsymbol{\beta} + \epsilon(\mathbf{s}), \text{ where } \epsilon(\mathbf{s}) \sim N(\mathbf{0}, \boldsymbol{\Sigma}), \tag{3.4}$$

and the resulting likelihood

$$\mathbf{Y}|\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\psi} \sim N(\boldsymbol{X}^\top(s)\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\boldsymbol{\psi}}), \ \boldsymbol{\Sigma}_{\boldsymbol{\psi}} = \sigma^2\boldsymbol{R}(\phi) + \tau^2\mathbf{I}. \tag{3.5}$$

Here, $\boldsymbol{R}(\phi)$ is a valid correlation function over $\mathbb{R}^r$ subject to a dependency on pairwise Euclidean distance d_{ij} between locations. Consequently, the prediction problem involves finding a function $h(\boldsymbol{y})$, where $\boldsymbol{y}$ represents the collected data, that minimizes the mean squared prediction error,

$$E\left[(Y(s_0) - h(\boldsymbol{y}))^2 \mid \boldsymbol{y}\right]. \tag{3.6}$$

From the Bayesian perspective, the function $h(\boldsymbol{y})$ corresponds to the posterior mean of $Y(s_0)$ [3]. Within this framework, we can incorporate prior distributions for $\boldsymbol{\beta}$ and $\boldsymbol{\psi}$, allowing the construction of a full posterior predictive distribution $p\left(Y(s_0) \mid \boldsymbol{y}\right)$. Then, with respect to this distribution, any desired point or interval estimate and any desired probability statements can be computed.

3.3 Nearest Neighbor Gaussian Process

The growing availability of large georeferenced datasets presents a challenge in terms of computational requirements. When the number of locations n becomes very large, traditional Gaussian process modeling becomes impractical [11]. Evaluating the Gaussian density in Equation (3.1) involves storing and manipulating the covariance matrix $\boldsymbol{\Sigma}_\theta$, which becomes computationally burdensome when calculating its inverse and determinant. Models belonging to the Nearest Neighbor Gaussian Process (NNGP) class offer scalable inference capabilities as the number of spatial observations increases [6].

Consider a fixed finite set of locations in the spatial domain D, denoted as $\boldsymbol{S} = \{s_1, \ldots, s_n\}$, and let $w(s) \sim SGP(0, \boldsymbol{C}_\theta)$ represent a zero-centered Gaussian process. Thus, $\mathbf{w}_S$ follows a multivariate normal distribution $N(0, \boldsymbol{C}_\theta(S))$, where $\mathbf{w}_S = [w(s_1), \ldots, w(s_n)]^\top$. By factoring the joint probability of the spatial Gaussian process realization, we can express it as a chain of conditional probabilities, subject to a specific order. For example,

$$p(\mathbf{w}_S) = p\left(w(s_1)\right) \prod_{i=2}^{n} p\left(w(s_i) \mid w(s_1 : s_{i-1})\right). \tag{3.7}$$

Now, let's consider any location $\boldsymbol{s}$ in the spatial domain D, and define $K(s_i)$ as the collection of the m nearest neighbors of $s_i \in \boldsymbol{S}$. The NNGP is specified in a similar

manner to Equation (3.7), and it can be expressed as follows

$$p(\mathbf{w}_S) \sim \prod_{i=1}^{n} p\left(w(s_i) \mid w_{K(s_i)}\right),$$

with $w(\boldsymbol{s}) \mid \mathbf{w}_S \overset{\text{iid}}{\sim} p\left(w(\boldsymbol{s}) \mid \mathbf{w}_{K(\boldsymbol{s})}\right)$ for all $\boldsymbol{s} \in D$. In practice, the set $\boldsymbol{S}$ is commonly selected to include the locations presented in the available data. NNGP leverages the concept of nearest neighbors to create smaller conditioning sets. By utilizing the nearest neighbors, which correspond to points with higher spatial correlation, this approach provides a reliable approximation of the full Gaussian process. In [6] it is demonstrated that the construction described above results in a multivariate Gaussian distribution for $\mathbf{w} = \mathbf{w}_S$

$$\mathbf{w} \sim N(\mathbf{0}, \tilde{\boldsymbol{C}}_\theta), \tag{3.8}$$

where $\tilde{\boldsymbol{C}}_\theta$ is the NNGP covariance function. This function is constructed based on the original covariance function $\boldsymbol{C}_\theta$ and ensures that $\tilde{\boldsymbol{C}}_\theta^{-1}$, i.e. the inverse of $\tilde{\boldsymbol{C}}_\theta$, is sparse. As a result, the likelihood in Equation (3.8) can be evaluated at a linear cost, enabling the model to be scalable for handling massive datasets.

The size of the neighbor set m directly affects the storage and computation requirements of a NNGP model. Simulation experiments, as detailed in [6], demonstrate the possibility of running NNGP models for different choices of m, potentially in parallel. The optimal value of m can be determined by minimizing model evaluation metrics such as RMSPE (Root Mean Squared Prediction Error). However, the simulations indicate that models with very small values of m ($\approx$ 10 or 15), can yield inference results that are practically indistinguishable from those obtained using full geostatistical models. As specified in [32], one possible approach is to select $K(s_i)$ as the set of m nearest neighbors of s_i among $s_1, \ldots, s_{i-1}$ based on the Euclidean distance.

Chapter 4

Methodology

4.1 The geographical model

In a traditional approach, the spatial correlation structure is determined exclusively by the Euclidean distance between geographic coordinates. The spatial correlation structure in a geographical model is specified by its correlation function $\boldsymbol{R}(\phi)$, as defined in equation (3.5), which can be extended to

$$\boldsymbol{R}(\phi_G) = \boldsymbol{R}_{ij}(\phi_G; d_{Gij}), \tag{4.1}$$

where ϕ_G represents the correlation decay in *geographical* space and d_{Gij} is the Euclidean distance between two geographic coordinates of s_i and s_j, with both $s_i, s_j \in \boldsymbol{S}$. Here, as previously mentioned, the correlation function $\boldsymbol{R}(\phi)$ adopts the exponential correlation $\exp(-d_{G_{ij}}/\phi_G)$. One alternative is to consider the use of other distance metrics to specify spatial correlation. In many scenarios, relying solely on Euclidean distance may not capture the complex relationships present in the environment. Nonetheless, caution should be exercised when incorporating non-Euclidean distance metrics into models that were developed under the Euclidean paradigm, such as those used in kriging. Indiscriminate use of non-Euclidean distances does not guarantee that the resulting covariance matrix will be positive definite, which can lead to an invalid model [33].

4.2 The feature model

We suggest a distinct route, instead of changing the distance metric, we propose to incorporate useful covariates that assist a better representation of the dependence among observations. Let $\boldsymbol{Z}(\boldsymbol{s}) = [z(s_1), \ldots, z(s_k)]$ represent a set of features that represent the spatial characteristics of interest. Similarly to Equation (4.1), we can define the correlation

structure in the space generated by the covariates in $\boldsymbol{Z}(\boldsymbol{s})$ by

$$\boldsymbol{R}(\phi_F) = \boldsymbol{R}_{ij}(\phi_F; d_{Fij}), \tag{4.2}$$

where the parameter ϕ_F, the decay in spatial correlation, is regulated by the distances computed between covariates, i.e. the *feature* space. Similarly to the previous description, the distance d_{Fij} is defined as the Euclidean distance between two distinct covariates $z(s_i)$ and $z(s_j)$ to guarantee a valid correlation structure. As previously discussed, the use of other distance metrics can be explored but with care.

It is worth mentioning that in this specification, we have the flexibility to choose any set of relevant feature to define the space. For instance, we can select the same variables used in the mean function $\mu(\boldsymbol{s})$ in Equation (3.3) to compose the set $\boldsymbol{Z}(\boldsymbol{s})$, a subset of $\boldsymbol{X}$, or a new set of features. Therefore, we have the option to choose only the covariates that are deemed to possess intrinsic information about the dependence to the response variable analyzed. The objective of this approach is to enhance the interpolation of the spatial surface, particularly for distances where the traditional model may fail to capture any spatial correlation. By altering solely the spatial correlation structure, the construction of spatial models under the feature space configuration adheres to the same principles as the traditional model. The main distinction lies in the selection and utilization of the correlation function based on the covariates, while the other steps in the model construction, such as specifying the mean function, estimating parameters, and making predictions, remain unaffected. This facilitates a seamless integration of the modified spatial correlation structure into the geostatistical modeling framework. Additionally, the specification of the m nearest neighbors, based on Euclidean distance, employed in NNGP models, remains unaltered. However, it is important to note that the nearest neighbors of a location in $\boldsymbol{S}$ may not necessarily be in its immediate spatial proximity. This allows for determining similarity between two locations in space, even if they are separated by a long distance.

In the upcoming section, we will introduce the concept of the *"mixture model"*, which is derived from the two previously presented correlation structures. This model provides the flexibility to combine the contributions of information generated by the geographic space and the available covariates. By incorporating both sources of information, the mixture model aims to capture the spatial dependence structure in a comprehensive manner, leveraging the strengths of both the geographical and feature models. This integrated approach allows for a more accurate and robust representation of the underlying phenomenon of interest.

4.3 The mixture model

The mixture model is formulated as a weighted combination of the two "single" models. In this model, the spatial interpolation is performed by incorporating contributions from both the covariance of the geographical model and the covariance of the feature model. Formally, the covariance matrix defined in Equation (3.5) can be expressed as:

$$\boldsymbol{\Sigma}_M = \sigma^2 \left(\lambda(\boldsymbol{s})\boldsymbol{R}(\phi_G) + (1 - \lambda(\boldsymbol{s}))\boldsymbol{R}(\phi_F)\right) + \tau^2 \boldsymbol{I}, \tag{4.3}$$

and we define the transformation

$$\text{logit}(\lambda(\boldsymbol{s})) = \boldsymbol{B}(s)^\top \boldsymbol{\gamma}. \tag{4.4}$$

In this formulation, we allowed for an additional set of covariates $\boldsymbol{B}(s)$ to determine what are the important factors for the weight function. By doing so, we can allow the contribution of each model to vary spatially. It is worth noticing that since Equation (4.3) is a convex interpolation of valid correlation functions, the resulting one is also a valid correlation function. It is evident that as $\lambda(s) \to 0$, the feature model dominates ($\boldsymbol{\Sigma}_M = \boldsymbol{\Sigma}_F$), while as $\lambda(s) \to 1$, the geographical model becomes predominant ($\boldsymbol{\Sigma}_M = \boldsymbol{\Sigma}_G$). This flexibility allows the model to adapt and incorporate the appropriate contribution from each model depending on the specific location being estimated. By adjusting the value of $\lambda(s)$, the mixture model can effectively capture the spatial dependence and provide more accurate predictions in diverse regions of interest. The parameters $\boldsymbol{\gamma}$ are included as part of the vector of spatial parameters $\boldsymbol{\theta}$, which is estimated within the framework of the hierarchical model in (3.3). Similar to the other parameters in the Bayesian framework, we specify a prior distribution for $\boldsymbol{\gamma}$, and it is also updated at each step of the MCMC realization.

For our mixture proposal we define the dynamic NNGP model. In this context, we let the number of neighbors change according to the location s_i. Let $K_G(s_i)$ and $K_F(s_i)$ represent the sets of m nearest neighbors of s_i from the geographical model and the feature model, respectively. Henceforth, we write $K_G(s_i)$ as K_G and $K_F(s_i)$ as K_F, the dependence of a location s_i is implicit. We define $K_M = K_G \cup K_F$. It is reasonable to assume that $\{K_G \cap K_F\} \neq \emptyset$. Therefore, the resulting union may have a different number of components. As a result, the new set K_M will consist of $m^\star$ neighbors, where $m \leq m^\star \leq 2m$. As a consequence, we have a dynamic NNGP that enables the number of neighbors in the mixture model to vary spatially. This means that each location s_i among $s_1, \ldots, s_{i-1}$ may have a different number of neighbors, determined solely by the distinct neighbors present in K_G and K_F. By allowing this spatial variation, we can capture the specific local dependencies and incorporate the relevant information from both the

geographical and feature models in a flexible manner having both NNGP versions as special cases of the dynamic NNGP mixture.

Chapter 5

Simulation Study

To showcase the difference between the feature model and traditional kriging, and further validate the utility and performance of the mixture model, a comprehensive simulation study was conducted. The study aimed to assess the model's effectiveness in estimating the parameters of the geostatistical regression under various values of λ, as well as evaluating its prediction capability. To achieve these objectives, we generated synthetic datasets of three sizes, $n = 500,\ 1000$ and 2000. Each observation in these datasets included $p = 2$ covariates within a unit square domain. To demonstrate the flexibility of the feature model, we extended the set $\boldsymbol{Z}(\boldsymbol{s})$ to include an additional covariate that is exclusively used in the feature space, along with the second generated covariate. All the generated covariates were drawn from $N(0, 1)$. Further, we define $\boldsymbol{B}(s) = 1,\ \forall s$. We performed 100 replicates of model fitting for the three different models using the same datasets, without considering a *nugget* effect. For each iteration, we generated a response variable based on the true parameters, which were established under three different configurations under the transformation in (4.4). In the first scenario, we assume that the geographical model is the true underlying model ($\gamma_0 \to \infty$ or $\lambda = 1$). This means that the spatial dependence structure is solely determined by the geographical coordinates. In the second scenario, we consider the feature model as the true model ($\gamma_0 \to -\infty$ or $\lambda = 0$), where the spatial dependence structure is constructed based on the available covariates. Lastly, in the third scenario, we assume an intermediate configuration where both the geographical model and the feature model contribute to the spatial dependence structure ($\gamma_0 = 0$ or $\lambda = 0.5$). We also included an additional set of observations (out-of-sample), representing 10% of the total of each dataset, for evaluating the predictive power of the models. These observations were generated together with the response variable for each replication. Importantly, all three models used the same datasets for both model fitting and prediction within each replicate. In order to evaluate the performance and scalability of the models using the NNGP class, we also fitted their full Gaussian (*FullGP*) counterparts for the first two scenarios.

For all models, vague priors were assigned for the parameters. The intercept and regression coefficients were given a Normal$(0, 100)$ prior distributions. To the process standard deviation component, σ, was assigned a uniform prior distribution with sup-

port on the interval $U(0, 10)$, both the spatial decay parameters received uniform prior supports $U(0, 100)$, and the mixture weight parameter γ_0 followed a Normal$(0, 9)$ prior distribution. Model parameter estimates and performance metrics were obtained from a total of $10,000$ iterations, with the first $8,000$ iterations discarded as burn-in samples. All MCMC sampling process was conducted using the `FUSE` package available at https://github.com/lucasamich/FUSE. Convergence were verified by visual inspection of the chains traceplots.

Figure 5.1 displays examples of the surfaces generated by the likelihood described in Equation (3.5) for each scenario configuration. From the left to the right, we have geographic, feature, and mixture models, respectively. The contrast between them is clear. The geographic model has a smooth spatial pattern, while the feature set presents a noise pattern in the original spatial scale. Finally, as expected, the mixture surface is a combination of the single processes.

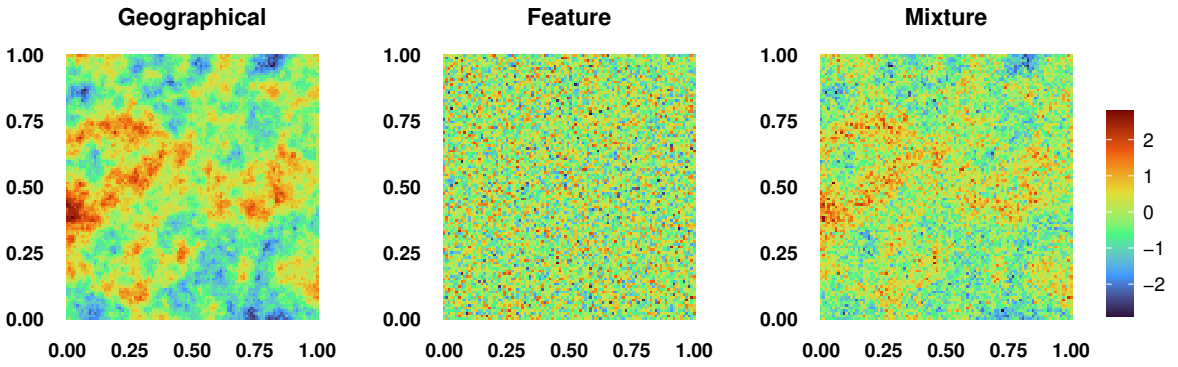

Figure 5.1: True spatial structures of the geographical, feature, and mixture models.

5.1 Parameter recovery capacity

We present the results of fitting the geostatistical model to demonstrate the accuracy and scalability of the proposed approach. All NNGP models were fitted using $m = 10$. The true parameter values utilized for data generation are presented in Table 5.1. The average computational times, in minutes, for estimating spatial parameters and predicting the out-of-sample dataset for both NNGP and FullGP methodologies are displayed in Table 5.2. The experiments were conducted on a computer equipped with an Intel Xeon processor featuring 16 cores operating at 3.40GHz, 126GB of available RAM, and running Ubuntu Linux version 20.04.1. It is evident from the results that the reduced

dimensionality and computational complexity of the NNGP models lead to significant reductions in computing time compared to the conventional FullGP models.

Table 5.1: True parameters used to generate the data in each scenario for $n = 1000$.

	Scenario 1	**Scenario 2**	**Scenario 3**
Parameter	True	True	True
ϕ_G	0.12	-	0.12
ϕ_F	-	0.48	0.46
σ^2	1.00	1.00	1.00
β_0	0.50	0.50	0.50
β_1	-1.00	-1.00	-1.00
β_2	1.50	1.50	1.50
λ	1.00	0.00	0.50

Table 5.2: Average duration (in minutes) for model fitting and prediction among different models and scenarios for NNGP and FullGP for all datasets sizes.

			Scenario 1		**Scenario 2**		**Scenario 3**	
n		Models	Estimation	Prediction	Estimation	Prediction	Estimation	Prediction
500		Geographical	0.91	0.22	1.01	0.20	0.97	0.23
	NNGP	Feature	1.18	0.31	1.07	0.29	1.15	0.33
		Mixture	8.06	0.23	8.03	0.23	9.01	0.24
		Geographical	43.17	10.03	23.52	7.49	31.5	9.42
	FullGP	Feature	28.13	7.43	29.73	9.5	28.9	8.78
		Mixture	39.81	6.76	45.99	6.85	51.94	6.75
1000		Geographical	2.04	0.4	1.85	0.38	2.00	0.44
	NNGP	Feature	2.53	0.67	2.69	0.77	2.69	0.70
		Mixture	18.19	0.51	18.03	0.51	20.16	0.51
		Geographical	232.63	101.07	300.59	307.94	232.55	103.41
	FullGP	Feature	223.36	89.45	454.96	291.48	225.08	90.49
		Mixture	336.73	62.46	632.51	253.68	380.87	62.49
2000		Geographical	4.17	0.74	3.78	0.72	4.13	0.75
	NNGP	Feature	5.23	1.53	5.52	1.79	5.49	1.66
		Mixture	39.72	1.13	39.69	1.12	43.89	1.11

The posterior median estimates and interval coverages for the three scenarios studied with $n = 1000$ are presented in Tables 5.3, 5.4, and 5.5. In the first scenario, where distances in the geographical space are used, the NNGP model provides a good estimate of the spatial decay parameter, ϕ_G. However, it slightly underestimates the process variance for both the geographic and feature model. The parameters of the geostatistical regression, including β_0, β_1, and β_2, are well estimated in this model. On the other hand, the feature model struggles to estimate the spatial range parameter accurately, but it still provides good estimates for the other parameters. The mixture model is able to estimate the geographical spatial range effectively, but it produces an unrealistic estimate for the range in the feature space. It slightly overestimates the process variance due to the introduction of more variability through dynamic neighbors. However, the other parameters, including λ, are well estimated. The FullGP models exhibit similar behavior to the NNGP models, with good estimates for the true parameters.

Although the NNGP provides good point estimate, it is worth to notice and mention that differently from the FullGP it shows a drastic reduction in the parameters coverage, even for the true generating models. To the best of our knowledge, this frequentist property was never explored for the NNGP approach and may need more attention when using it. It may suggest that the NNGP are over optimistic about its credible intervals.

Table 5.3: Posterior median regression parameter estimates for the geostatistical model using NNGP and FullGP methods for scenario 1 with $n = 1000$

			Geographical		**Feature**		**Mixture**	
	Parameter	True	Estimated	Coverage	Estimated	Coverage	Estimated	Coverage
NNGP	ϕ_G	0.12	0.10	0.37	-	-	0.14	0.71
	ϕ_F	-	-	-	0.03	-	61.66	-
	σ^2	1.00	0.92	0.28	0.92	0.25	1.25	0.67
	β_0	0.50	0.53	0.31	0.52	0.19	0.52	0.99
	β_1	-1.00	-1.01	0.89	-1.00	0.95	-1.00	0.96
	β_2	1.50	1.51	0.92	1.50	0.88	1.50	0.98
	λ	1.00	-	-	-	-	0.98	-
FullGP	ϕ_G	0.12	0.16	0.74	-	-	0.15	0.81
	ϕ_F	-	-	-	0.00	-	63.17	-
	σ^2	1.00	1.32	0.73	0.92	0.24	1.26	0.81
	β_0	0.50	0.53	0.97	0.52	0.18	0.56	0.98
	β_1	-1.00	-1.00	0.92	-1.00	0.95	-1.00	1.00
	β_2	1.50	1.50	0.97	1.50	0.91	1.50	0.98
	λ	1.00	-	-	-	-	0.99	-

In the second scenario, the behavior of the model estimates is similar to that of the first scenario. The geographical model faces challenges in estimating the range parameter in the feature space, while the feature model and the mixture model perform well in capturing the true parameter values. The estimation of λ also shows good convergence towards its true value. The coverage of credibility intervals continues to exhibit the same behavior in this scenario, similar to the previous scenario. In the third scenario, which represents an intermediate configuration, we observe that the geographical and feature models struggle to estimate the spatial parameters accurately. However, they still produce reasonable estimates for the other parameters. In contrast, the mixture model performs well in estimating both spatial decay parameters, providing good approximations to their true values. Additionally, the mixture model successfully recovers the correct value of λ, which corresponds to the configuration of this scenario. The consistent behavior observed in the other datasets across the scenarios further supports the finding that NNGP models serve as a reliable approximation to full Gaussian models, while offering the advantage of reduced computational time.

Table 5.4: Posterior median regression parameter estimates for the geostatistical model using NNGP and FullGP methods for scenario 2 with $n = 1000$

			Geographical		**Feature**		**Mixture**	
	Parameter	True	Estimated	Coverage	Estimated	Coverage	Estimated	Coverage
NNGP	ϕ_G	-	0.01	-	-	-	60.84	-
	ϕ_F	0.48	-	-	0.35	0.20	0.56	0.76
	σ^2	1.00	0.85	0.30	0.87	0.31	1.27	0.70
	β_0	0.50	0.43	0.12	0.46	0.30	0.46	0.99
	β_1	-1.00	-1.00	0.21	-1.02	0.29	-1.02	0.97
	β_2	1.50	1.50	0.95	1.50	0.91	1.50	0.96
	λ	0.00	-	-	-	-	0.03	-
FullGP	ϕ_G	-	0.00	-	-	-	62.26	-
	ϕ_F	0.48	-	-	0.60	0.83	0.57	0.85
	σ^2	1.00	0.85	0.29	1.22	0.81	1.28	0.79
	β_0	0.50	0.43	0.11	0.47	0.99	0.47	1.00
	β_1	-1.00	-1.00	0.22	-1.01	0.98	-1.02	0.96
	β_2	1.50	1.50	0.95	1.50	0.94	1.50	0.95
	λ	0.00	-	-	-	-	0.02	-

Table 5.5: Posterior median regression parameter estimates for the geostatistical model using NNGP and FullGP methods for scenario 3 with $n = 1000$

			Geographical		**Feature**		**Mixture**	
	Parameter	True	Estimated	Coverage	Estimated	Coverage	Estimated	Coverage
NNGP	ϕ_G	0.12	0.04	0.00	-	-	0.10	0.74
	ϕ_F	0.46	-	-	0.15	0.00	0.39	0.57
	σ^2	1.00	0.92	0.46	0.92	0.44	0.86	0.39
	β_0	0.50	0.46	0.14	0.46	0.15	0.43	0.17
	β_1	-1.00	-0.99	0.29	-0.98	0.33	-1.00	0.61
	β_2	1.50	1.50	0.92	1.50	0.93	1.50	0.96
	λ	0.50	-	-	-	-	0.51	0.66
FullGP	ϕ_G	0.12	0.02	0.00	-	-	0.15	0.80
	ϕ_F	0.48	-	-	0.05	0.00	0.54	0.91
	σ^2	1.00	0.97	0.55	0.95	0.51	1.30	0.72
	β_0	0.50	0.44	0.19	0.42	0.21	0.42	1.00
	β_1	-1.00	-0.98	0.22	-0.99	0.49	-1.00	0.99
	β_2	1.50	1.50	0.92	1.50	0.71	1.50	0.93
	λ	0.50	-	-	-	-	0.53	0.79

In Figure 5.2, we compare the ability of the three models to recover the true surface and examine their behavior across the three scenarios. In the first scenario (Figure 5.2, first row), the geographic model successfully reproduces the original process, albeit with slight smoothing in space. The feature model, constructed based on the feature space, is unable to accurately recover the true surface. Meanwhile, the mixture model performs well and estimates the true surface similar to the geographic model. From the Figure 5.2 second row, the geographic model oversmoothes the spatial structure in the feature space, while the feature model and the mixture model produce accurate estimates of the true surface. In the third scenario (Figure 5.2, last row), both models struggle to estimate the surface accurately. The geographic model again exhibits oversmoothing in space, and the

feature model fails to capture the spatial structure. However, the mixture model provides a more accurate estimation of the true region without excessive smoothing of the spatial structure.

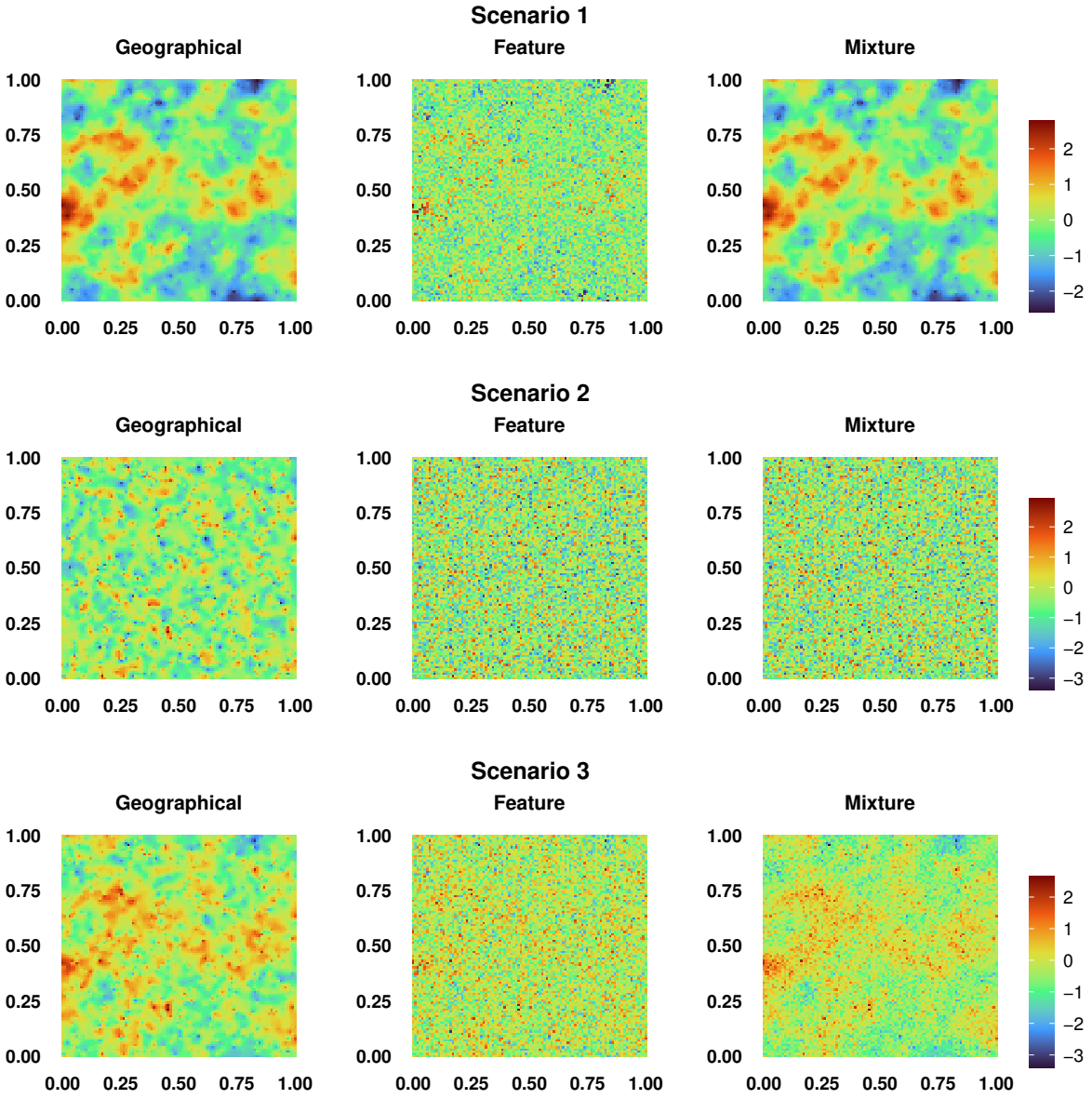

Figure 5.2: The recovered surface by each model for scenarios 1, 2 and 3.

5.2 Prediction capacity

To assess the models' performance in capturing the original surface, we employed two evaluation metrics

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|(y_i - \hat{y}_i)|, \quad \text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}.$$

In Scenario 1 (Figure 5.3a), both the NNGP and FullGP models demonstrate better accuracy in estimating the true multivariate random process for the geographical models. The mixture model performs competitively with the geographical model, providing highly accurate estimates. In Scenario 2 (Figure 5.3b), the feature model and the mixture model outperform the geographical model in capturing the true process, while the latter model struggles to produce accurate estimates. In the intermediate Scenario 3 (Figure 5.3c), the mixture model surpass the other models, producing lower prediction error estimates across both metrics. However, the error estimate of the mixture NNGP model is slightly worse than its FullGP counterpart. These findings highlight the NNGP models' ability to approximate the FullGP models effectively, and the mixture model's capacity to provide comparable or even superior estimates across all scenarios.

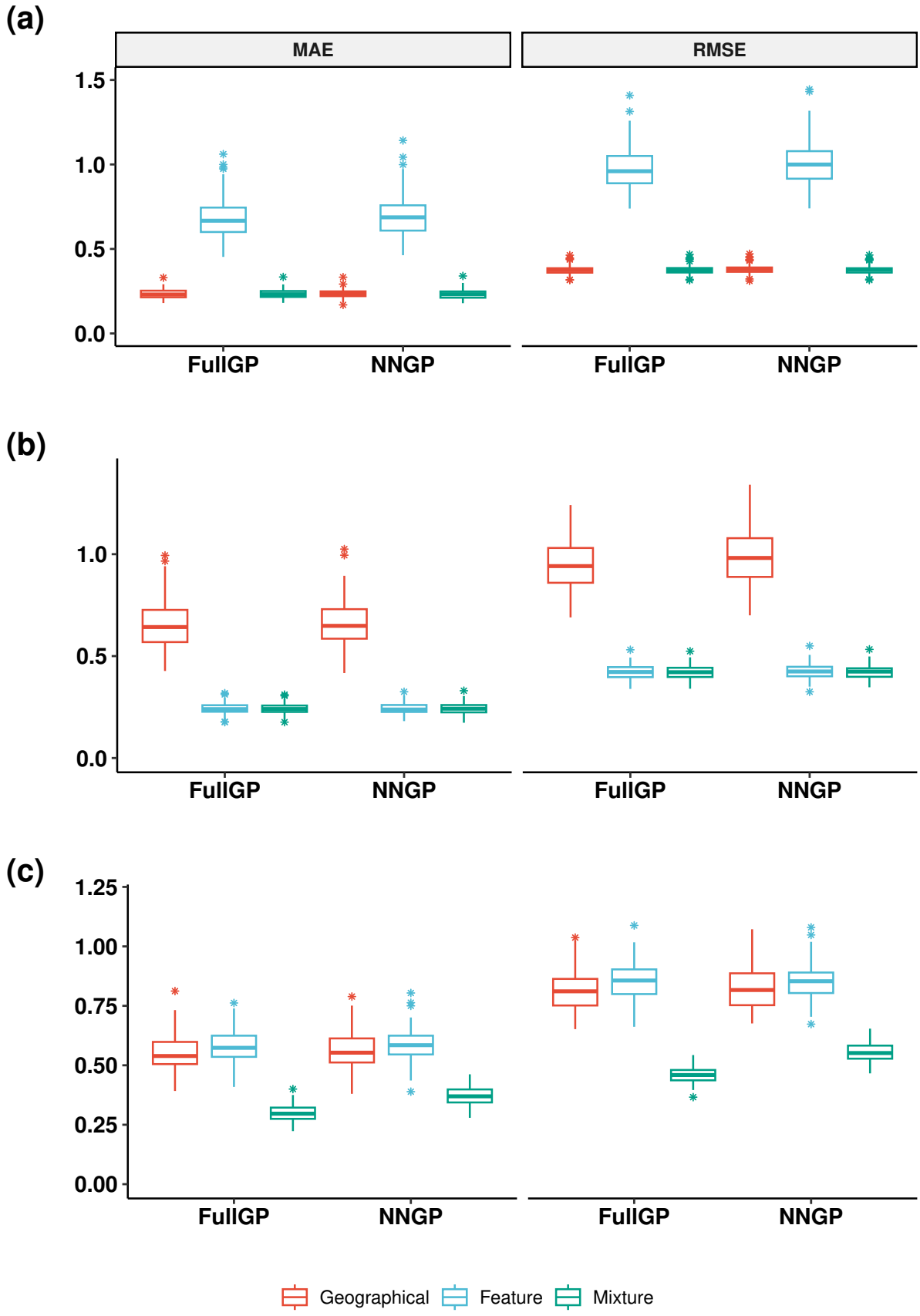

Figure 5.3: Comparison of prediction error metrics distribution between NNGP and FullGP models for $n = 1000$. (a) Scenario 1, (b) Scenario 2, and (c) Scenario 3.

Chapter 6

Porosity estimation results

We evaluated our proposed method on synthetic oil reservoir data, specifically focusing on estimating the porosity of rocks in the reservoirs. All the three models, geographical, feature, and mixture, were fitted using the available well data with the variables described in Section 2. For this application, we employed the same priors as those used in Section 5, with the exception of the spatial decay parameters ϕ_G and ϕ_F, which were assigned a uniform prior distribution over the interval $U(0, 500)$. Model parameter estimates and performance metrics were derived from a total of $750,000$ iterations, discarding the initial $500,000$ iterations as burn-in samples and applying a thinning interval of size 50, resulting in a final chain of size $5,000$ samples. For assessing convergence, we employed visual inspection of the chains traceplots and verified the [12] diagnostic. The evaluation focused on comparing the results for two specific regions within the available data volume, namely *Region A* and *Region B*, which represent cross-sections along the x-axis. To measure the prediction quality of the models, we provide a visualization of the surface of these two regions in Figure 6.1, which serves as the ground truth for comparison.

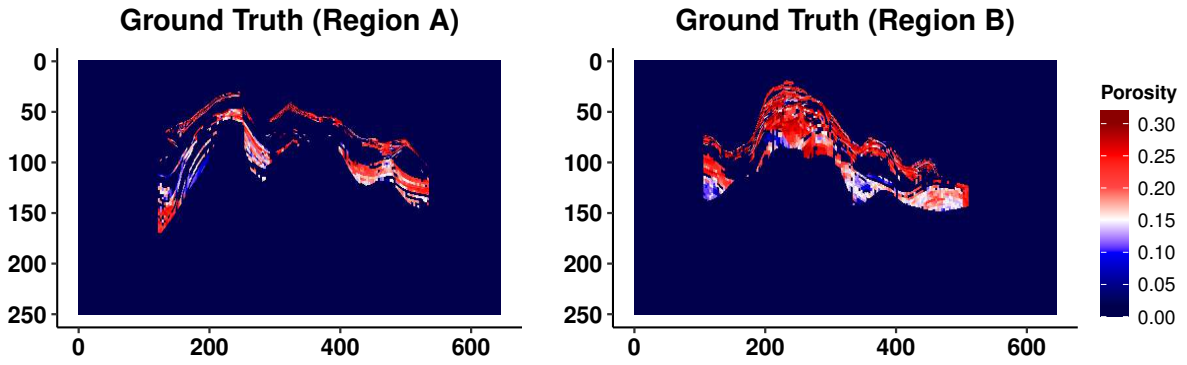

Figure 6.1: Porosity values in two regions extracted from the total volume of data. On the left is Region A, and on the right is Region B.

In all the models, we incorporated the covariates ρ, V_p, and V_s to specify the mean and, when suitable, the covariance of the process. This approach enables us to identify linear relationships between these variables and porosity through the mean function. Si-

multaneously, we can capture important spatial features related to porosity at distant locations by defining the covariance structure. In Table 6.1, we present the posterior estimates of the geostatistical models adjusted for this dataset. The covariates ρ and V_p show an inversely proportional relationship with porosity. Notably, the estimated spatial decay parameters by the geographical model and the feature model differ, which may raise doubt about model selection without knowledge of the true generative model. However, the mixture model estimates the value of λ very close to zero, indicating an almost exclusive contribution of the feature model in the composition of the target process. Therefore, when using the mixture model, we gain more information about the analyzed process and can make more appropriate decisions. The higher contribution of the feature model in the mixture model aligns with the nature of the data, as the studied reservoir is a result of a simulation process that mainly utilizes characteristics of reservoir lithology found in the field. This suggests that the geological features have a significant influence on the porosity distribution within the reservoir. The mixture model's ability to capture and incorporate these intrinsic geological features makes it a suitable choice for modeling the porosity distribution in this specific reservoir context.

Table 6.1: Estimates of the medians and Highest Posterior Density (HPD) 95% interval for the parameters of the geostatistical regression in the models using the data from the simulated oil reservoir.

	Geographical		**Feature**		**Mixture**	
Parameter	Estimate	CI 95%	Estimate	CI 95%	Estimate	CI 95%
ϕ_G	26.43	(17.11, 38.41)	-	-	329.80	(54.10, 499.96)
ϕ_F	-	-	63.47	(29.85, 119.39)	97.60	(38.62, 236.07)
β_0	0.03	(0.03, 0.03)	-0.45	(-0.74, -0.21)	-0.40	(-0.76, -0.14)
β_ρ	-0.04	(-0.04, -0.04)	-0.06	(-0.07, -0.06)	-0.07	(-0.07, -0.06)
β_{V_p}	-0.01	(-0.02, -0.01)	-0.01	(-0.01, -0.01)	-0.01	(-0.01, -0.00)
β_{V_s}	0.06	(0.06, 0.07)	0.05	(0.04, 0.05)	0.04	(0.04, 0.05)
σ^2	3.76×10^{-4}	$(2.69 \times 10^{-4}, 4.78 \times 10^{-4})$	0.02	(0.01, 0.03)	0.02	(0.01, 0.06)
τ^2	5.42×10^{-4}	$(4.35 \times 10^{-4}, 6.36 \times 10^{-4})$	2.94×10^{-9}	$(2.66 \times 10^{-12}, 1.29 \times 10^{-8})$	2.16×10^{-9}	$(1.31 \times 10^{-13}, 9.85 \times 10^{-9})$
λ	-	-	-	-	2.81×10^{-7}	$(8.59 \times 10^{-11}, 1.64 \times 10^{-6})$

Figure 6.2 displays the porosity estimates for the two regions of interest. It can be observed that the geographical model tends to oversmooth the porosity interpolation in space, leading to positive porosity values in regions where the true porosity is zero. In contrast, the feature and mixture models are able to recover the two regions well, providing more accurate and similar porosity estimates. Table 6.2 presents the prediction error measures for the three models across regions A and B. The prediction errors were computed for the entire region, encompassing locations with zero porosity, making it possible to measure the models' capability to estimate abrupt porosity changes present in the regions. The results clearly indicate that both the feature-based and mixture models outperform the geostatistical model in terms of prediction errors for both regions. The mixture model demonstrated equivalent or even lower error estimates compared to the feature model for both of the considered metrics.

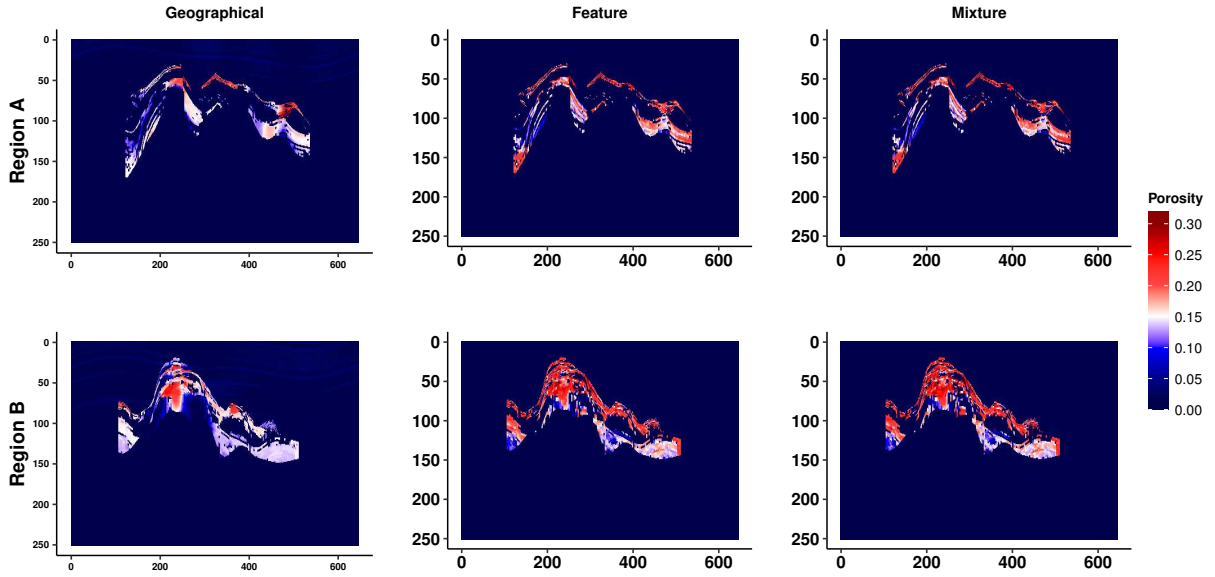

Figure 6.2: Top row: Estimates of porosity for the geographical, feature, and mixture models (in that order) for Region A. Bottom row: Estimates of porosity for the geographical, feature, and mixture models (in that order) for Region B.

Table 6.2: Prediction error estimates for the two regions among the models

		Geographical	**Feature**	**Mixture**
Region A	MAE	0.008	9.489×10^{-4}	6.415×10^{-4}
	RMSE	0.017	0.005	0.003
Region B	MAE	0.013	0.003	0.002
	RMSE	0.029	0.017	0.015

Chapter 7

Conclusion

Large spatial and spatiotemporal datasets pose challenges for fully model-based Bayesian inference due to computationally expensive matrix operations. Furthermore, for a broad spatial domain where the geographical distance among the observations can be substantial, the traditional spatial regression can be ineffective in capturing the spatial dependence relationship. This is the case in our application, where drilling an oil well is very expensive, and increasing the number of wells is not an option. Therefore, using lithological information to measure similarity among sites becomes a reasonable alternative. Beyond this, we introduce a mixture modeling framework that incorporates the spatial regression and feature model in one setting allowing data to determine the appropriate mixture among the two single models. We also leverage the computational capacity of our proposal based on a dynamic version of the NNGP approach where the number of neighbors can vary from site to site. Our methods are implemented in the `FUSE` package available at https://github.com/lucasamich/FUSE.

Our simulation study showcased the effectiveness of our proposal in accurately estimating spatial parameters and predicting the target process. Specifically, we compared three models: geographical, feature, and mixture models, under different scenarios. The geographical model performed well when the spatial correlation was solely space dependent, while the feature model excelled when using the covariates' inherent information. However, the mixture model emerged as a powerful choice, effectively incorporating contributions from both geographical and feature models depending on the location being estimated. The mixture model achieved superior estimates of the true spatial configuration, making it a valuable tool for capturing complex spatial structures where the traditional models may fail. We further applied the proposed approach to a simulated oil reservoir dataset, demonstrating its practical applicability. The mixture model successfully integrated geological features represented by covariates, such as reservoir lithology, into the porosity estimation, providing more accurate predictions compared to the other models. This ability to incorporate intrinsic geological features made the mixture model particularly suitable for modeling porosity distributions in this specific reservoir context. Overall, the dynamic NNGP mixture model emerged as a robust and versatile framework for geostatistical modeling, capable of handling large datasets while maintaining accu-

rate predictions. Its ability to combine information from both geographical space and covariates space opens up new opportunities for various geospatial applications.

The use of feature information can be extended to address other geostatistical challenges, e.g., ansiotropy. Another direct extension of this proposal is to spatiotemporal problems. Further, other distance metrics may be more appropriate and should be explored for the feature space. Finally, from the simulation studies empirical coverage, there is an indication that the NNGP produces narrow credible intervals for the parameters. Such a finding, although not in the scope of the paper, should be investigated since the NNGP methodology has gained enormous attention in the literature.

References

[1] Rachid Ababou, Amvrossios C Bagtzoglou, and Eric F Wood. On the condition number of covariance matrices in kriging, estimation, and simulation of random fields. *Mathematical Geology*, 26:99–133, 1994.

[2] A Adekanle and PA Enikanselu. Porosity prediction from seismic inversion properties over 'xld'field, niger delta. *American journal of scientific and industrial research*, 4(1):31–35, 2013.

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

[4] J. Benndorf and Jan Dirk Jansen. Recent developments in closed-loop approaches for real-time mining and petroleum extraction. *Mathematical Geosciences*, 2017.

[5] Noel A Cressie. The origins of kriging. *Mathematical Geosciences*, 1990.

[6] Abhirup Datta, Sudipto Banerjee, Andrew O Finley, and Alan E Gelfand. Hierarchical nearest-neighbor gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*, 111(514):800–812, 2016.

[7] Benjamin JK Davis and Frank C Curriero. Development and evaluation of geostatistical methods for non-euclidean-based spatial covariance matrices. *Mathematical geosciences*, 51(6):767–791, 2019.

[8] Perry de Valpine, Daniel Turek, Christopher J Paciorek, Clifford Anderson-Bergman, Duncan Temple Lang, and Rastislav Bodik. Programming with models: writing statistical algorithms for general model structures with nimble. *Journal of Computational and Graphical Statistics*, 26(2):403–413, 2017.

[9] Philippe M. Doyen. Porosity from seismic data: A geostatistical approach. *GEOPHYSICS*, 53(10):1263–1275, 1988.

[10] Ross Alan Ensley. Comparison of p-and s-wave seismic data; a new method for detecting gas reservoirs. *Geophysics*, 49(9):1420–1431, 1984.

[11] Andrew O Finley, Abhirup Datta, Bruce D Cook, Douglas C Morton, Hans E Andersen, and Sudipto Banerjee. Efficient algorithms for bayesian nearest neighbor gaussian processes. *Journal of Computational and Graphical Statistics*, 28(2):401–414, 2019.

[12] John Geweke. Evaluating the accuracy of sampling-based approaches to the calculations of posterior moments. *Bayesian statistics*, 4:641–649, 1992.

[13] José Jiménez-Luna, Francesca Grisoni, and Gisbert Schneider. Drug discovery with explainable artificial intelligence. *Nature Machine Intelligence*, 2(10):573–584, 2020.

[14] A G Journel and C J Huijbregts. Mining geostatistics, Jan 1976.

[15] Mark Kaiser. A survey of drilling cost and complexity estimation models. *International Journal of Petroleum Science and Technology ISSN Number*, 1:973–6328, 01 2007.

[16] Josenilda Nascimento Lonardelli, Rafael de Oliveira da Silva, Flávia de Oliveira Lima Falcão, Marco Antonio Cetale Santos, and Carlos Eduardo Borges de Salles Abreu. Evaluation of oil production related effects through geomechanical modeling: a case study from marimbá field, campos basin, brazil. *Journal of Petroleum Science and Engineering*, 158:186–201, 2017.

[17] Chaochao Lu and Xiaoou Tang. Surpassing human-level face verification performance on lfw with gaussianface. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

[18] Bertil Matérn. *Spatial variation: Stochastic models and their application to some problems in forest surveys and other sampling investigations.* 1960.

[19] G Matheron. Recherche de simplification dans un problème de cokrigeage. *Publication N-628, Centre de Géostatistique, Ecole des Mines de Paris, Fontainebleau*, 1979.

[20] Georges Matheron. *Les variables régionalisées et leur estimation: une application de la théorie de fonctions aléatoires aux sciences de la nature*, volume 4597. Masson et CIE, 1965.

[21] Fernando Gomes Moro. Estruturas de covariância definidas por covariáveis. Master's thesis, Universidade Federal do Paraná, 2019.

[22] RM Neal. Regression and classification using gaussian process priors (with discussion). *Bayesian statistics 6*, pages 475–501, 1999.

[23] M. Oliver and R. Webster. Kriging: a method of interpolation for geographical information systems. *Int. J. Geogr. Inf. Sci.*, 1990.

[24] Christopher Paciorek and Mark Schervish. Nonstationary covariance functions for gaussian process regression. *Advances in neural information processing systems*, 16, 2003.

[25] A. G. Pramanik, V. Singh, Rajiv Vig, A. K. Srivastava, and D. N. Tiwary. Estimation of effective porosity using geostatistics and multiattribute transforms: A case study. *Geophysics*, 69(2):352–372, 01 2004.

[26] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023.

[27] Carl Edward Rasmussen, Christopher KI Williams, et al. *Gaussian processes for machine learning*, volume 1. Springer, 2006.

[28] Brian D Ripley. *Statistical inference for spatial processes*. Cambridge university press, 1988.

[29] Benjamin J Shields, Jason Stevens, Jun Li, Marvin Parasram, Farhan Damani, Jesus I Martinez Alvarado, Jacob M Janey, Ryan P Adams, and Abigail G Doyle. Bayesian reaction optimization as a tool for chemical synthesis. *Nature*, 590(7844):89–96, 2021.

[30] William Murray Telford, WM Telford, LP Geldart, and Robert E Sheriff. *Applied geophysics*. Cambridge university press, 1990.

[31] M. Vauclin, S. R. Vieira, G. Vachaud, and D. R. Nielsen. The use of cokriging with limited field soil observations. *Soil Science Society of America Journal*, 47(2):175–184, 1983.

[32] Aldo V Vecchia. Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 50(2):297–312, 1988.

[33] Jay M Ver Hoef. Kriging models for linear networks and non-euclidean distances: Cautions and solutions. *Methods in Ecology and Evolution*, 9(6):1600–1613, 2018.

[34] Christopher Williams and Carl Rasmussen. Gaussian processes for regression. *Advances in neural information processing systems*, 8, 1995.

[35] Changshan Wu and Alan T. Murray. A cokriging method for estimating population density in urban areas. *Computers, Environment and Urban Systems*, 29(5):558–579, 2005. Remote Sensing for Urban Analysis.

[36] Jorge Kazuo Yamamoto. Correcting the smoothing effect of ordinary kriging estimates. *Mathematical geology*, 37:69–94, 2005.

[37] Yu Yu, Jun Zhang, Yipeng Jing, and Pengjie Zhang. Kriging interpolating cosmic velocity field. ii. taking anistropies and multistreaming into account. *Phys. Rev. D*, 95:043536, Feb 2017.