Universidade Federal de Minas Gerais - UFMG
Instituto de Ciências Exatas - ICEX
Departamento de Estatística
Programa de Pós-Graduação em Estatística

Métodos e aplicações em Estatística Espacial para grandes bancos de dados

Zaida Cornejo Quiroz

Belo Horizonte, Brasil, 2018

Métodos e aplicações em Estatística Espacial para grandes bancos de dados
Zaida Cornejo Quiroz

Tese apresentada ao Programa de Pós-Graduação em Estatística da Universidade Federal de Minas Gerais como parte dos requisitos para a obtenção do grau de Doutora em Estatística.

Orientador: Marcos Oliveira Prates

Belo Horizonte, Brasil 2018

Thesis for the degree of Doctor in Statistics

On spatial statistical methods and applications for large datasets

Zaida Cornejo Quiroz

Department of Statistics
UNIVERSIDADE FEDERAL DE MINAS GERAIS
Belo Horizonte, Brazil, 2018

On spatial statistical methods and applications for large datasets
Zaida Cornejo Quiroz

Dissertation submitted in partial fulillment of the requirements for the degree of Doctor in Statistics in the Graduate School of Universidade Federal de Minas Gerais.

Advisor: Marcos Oliveira Prates

Belo Horizonte, Brazil 2018

Thesis for the degree of Doctor in Statistics

On spatial statistical methods and applications for large datasets

Zaida Cornejo Quiroz

Approved

Marcos Oliveira Prates (advisor) - UFMG - Brazil

Flávio Bambirra Gonçalves - UFMG - Brazil

Vinícius Diniz Mayrin - UFMG - Brazil

Håvard Rue - KAUST - Saudi Arabia

Sudipto Banerjee - UCLA - United States of America

Belo Horizonte, Brazil, 2018

To my family

Resumo

O foco deste trabalho está na aplicação de modelos inovadores para a análise espaço-temporal da biomassa de anchova em um grande banco de dados e no desenvolvimento de um novo campo aleatório Gaussiano adequado para a análise de grandes conjuntos de dados.

O primeiro artigo apresenta uma aplicação avançada da modelagem espaço-temporal através da Equação Diferencial Parcial Estocástica (SPDE) para estimar e prever a biomassa de anchova na costa do Peru. Foi introduzido um modelo espaço-temporal hierárquico Bayesiano completo, levando em consideração as possíveis dependências espaciais ou espaço-temporais dos dados. Estes modelos, computacionalmente eficientes e flexíveis, são também capazes de realizar previsões tanto da presença quanto da abundância de anchovas, em particular, quando o conjunto de locais é grande (> 500) e diferente ao longo do tempo. Eles são baseados em que os campos Gaussianos Matérn podem ser vistos como soluções de uma determinada SPDE que, em combinação com o INLA (Aproximação Integrada Aninhada de Laplace), tem uma melhora na eficiência computacional.

O segundo trabalho é dedicado a estender o Processo de vizinho mais próximo Gaussiano (NNGP), recentemente proposto. Uma nova classe de processos de campo aleatório Gaussiano foi construída e, também, mostrada sua aplicabilidade a dados com pequenas ou grandes dependências espaciais. A idéia-chave por trás do novo processo espacial é subdividir o domínio espacial em vários blocos, que são dependentes de alguns dos blocos "passados". A redução na complexidade computacional é obtida através da dispersão das matrizes de precisão e e na paralelização de extensos cálculos através de blocos de dados. Estes modelos são úteis para grandes conjuntos de dados espaciais, no qual os métodos tradicionaissão computacionalmente intensivos, tendo um alto custo para serem utilizados. Finalmente, para realizar a inferência, oi adotado o enfoque Bayesiano, no qual utilizou-se algoritmos de Monte Carlo via cadeias de Markov (MCMC). Além de demonstradas as capacidades inferenciais completas da modelagem, em termos de estimação, previsão e qualidade de ajuste, quando o novo processo espacial é incluído.

Palavras chave: Geostatística, INLA, GMRF, MCMC, NNGP, SPDE, estatística espacial, modelamento espaço-temporal.

Abstract

The focus of this work is on the application of novelty models for the spatio-temporal analysis of large anchovy biomass dataset, and the development of a new Gaussian random field suitable for the analysis of large datasets.

The first paper presents an advance application of spatio-temporal modeling through the Stochastic Partial Differential Equation (SPDE) for estimating and predicting anchovy biomass off the coast of Peru. We introduce a complete, and computationally efficient, flexible Bayesian hierarchical spatio-temporal modeling for zero-inflated positive continuous, accounting for spatial or spatio-temporal dependencies in the data. The models are capable of performing predictions of anchovy presence and abundance, in particular,in particular, when the set of observed sites is large (> 500) and different across the temporal domain. They are based on the fact that Gaussian Matérn field can be viewed as solutions to a certain SPDE, which combined with Integrated Nested Laplace Approximations (INLA) improves the computational efficiency.

The second paper is devoted to extend the newly proposed Nearest Neighbor Gaussian Process (NNGP). A new class of Gaussian random field process is constructed and, it is showed its applicability to simulated data with small or large spatial dependences. The key idea behind this new spatial process (or random field) is to subdivide the spatial domain into several blocks which are dependent on some of the "past" blocks. The new spatial process recovers the NNGP and independent blocks approach. Moreover, The reduction in computational complexity is achieved through the sparsity of the precision matrices and parallelization of many computations for blocks of data. It is useful for large spatial data sets where traditional methods are too computationally intensive to be used efficiently. Finally, to perform inference we adopt a Bayesian framework, we use Markov chain Monte Carlo (MCMC) algorithms and demonstrate the full inferential capabilities of the modeling including the new spatial process, in terms of estimation, prediction and goodness of fit.

Keywords: Geostatistics, INLA, GMRF, MCMC, NNGP, SPDE, spatial statistics, spatio-temporal modeling.

Acknowledgments

First, I give thanks to God for giving me the strength, knowledge, ability and opportunity to undertake this research study and to persevere and complete it satisfactorily.

I am so grateful to my dear family for their love and guidance, always providing me through moral and emotional support in my life. I would like to express my deepest gratitude to my sister Karen and also to Mrs. Zelia, who throughout all this years have provided encouragement, sound advice, good company and the best friendship I could ever have imagined.

I would like thank my advisor Marcos Prates for the continuous support of my graduate study and related research, for his guidance, enthusiasm, motivation and suggestions along the way of making this thesis. My appreciation must also be dedicated to Dipak Dey for his patience, generosity and supervision during my visiting scholar research.

My thanks also goes to the rest of my thesis committee: Håvard Rue, Sudipto Banerjee, Flávio Gonçalvez and Vinícius for the time they spent reading the thesis, and their critical examination to improve my research. With a special mention to Håvard and Sudipto, it was fantastic to have the opportunity to have both of you in my denfense, here in Brazil. My thanks also goes to Flávio Gonçalves for all your dedicated lectures and of couse your suggestions in connection with my work on the first paper, and to Vinícius for numerous comments on the writing of all parts of this thesis.

I am also grateful to the university Profesors and staff at UFMG, for the unfailing assistance. My sincere thanks goes to Lourdes Montenegro for all your support and generosity all these years.

I would like to thank all colleagues at UFMG and BH, for their feedback, cooperation and of course friendship. My special thanks are extended to my dear friends: Nívea and Francisco for all your support in the good and bad days!, Livinha, for all your help and friendship, thanks Wagner, Marcia, Luis, Larissa, Rodrigo, Pedrinho, Debora, Alejandro and Jenny, who have helped in one way or another during the

last years, specially when I fell down “literally”. Thanks to everyone in the LESTE!, it was great sharing laboratory with all of you during last years. Thanks to Rose, dona Dalva and Bernadette for all your kindness.

I would like to express my gratitude to Sophie and Arnaud Bertrand for introducing me to the field of spatial statistics. Also I specially thank all the dear friends I made in the IRD, specially to Rocio Joo.

I would like to thank all colleagues at PUCP and Perú, especially to Cristian Bayes and Luis Valdivieso. I especially thank Jaime for all the advices and support during the last months.

And last but not least, I would like to thank the financial support from CAPES (Brazil) all these years and ProUNI (Peru) who makes it possible my visiting scholar research to UCONN.

Belo Horizonte, 2018 Zaida Cornejo Quiroz

Contents

List of Figures

List of Tables

Part I

Introductory chapters

Chapter 1

Introduction

Spatio-temporal data were always been essential for humans, for instance, Cressie and Wikle (2011) stated that nomadic tribes of early civilization used them to return to seasonal hunting grounds, early explorers seeking to map new lands collected data of locations, weather, plants, animals, among others, and the indigenous people also did it. In some sense we all collect and analyze spatio-temporal data, in fact there would be no History without spatio-temporal data. With the recent computational advances, the availability of spatio-temporal data sets in many areas is growing, generating considerable interest in statistical models, in particular, for point-referenced (geostatistical) data. Spatial and spatio-temporal modeling often involve expensive matrix decompositions whose computational complexity increases with the number of spatial locations and temporal points, being a challenge for large spatial and spatio-temporal data sets.

One approach to proceed involves a construction of specific spatial and spatio-temporal models based on Gaussian Random Markov Fields (GRMF). Lindgren et al. (2011) suggested a link between Gaussian random fields with Matérn covariance function and a GRMF through stochastic partial differential equations (SPDE). In particular, they used the finite element method (FEM) to discretize complex geometries to get an approximation of the SPDE's solution using basis functions. As a consequence, the continuous interpretation of space is not lost, while the computational algorithms only see discrete structures with Markov properties. A great variety of applications using the SPDE approach for geostatistical data can be found in Bolin and Lindgren (2011), Blangiardo et al. (2013) and Cameletti et al. (2013). Further, it is relatively simple to extend the SPDE approach from spatial to spatio-temporal models with separable covariance function Cameletti et al. (2013). In this context, Lindström and Lindgren (2008) used a spatio-temporal model based on the SPDE approach to interpolate yearly precipitation data over African Sahel,

inference was performed through the Markov chain Monte Carlo (MCMC) method. On the other hand, Cameletti et al. (2013) consider a spatio-temporal model based also on the SPDE approach to estimate particular matter concentration and inference was performed through the Integrated Nested Laplace Approximation (INLA) method (Rue et al., 2009). Following this work, Paper 1 presents an application to study the spatial distribution of peruvian anchovy across years, as well as to predict anchovy abundance data. We assume that abundance of anchovy follows a mixture of a discrete probability mass at zero and some established continuous distribution for nonzero values, like it was proposed by Quiroz et al. (2015). The main contribution of this paper is to provide estimations and predictions of anchovy presence and abundance, in particular, when the set of locations is huge and differ across the temporal domain.

Another approach to proceed involves a kind of "low rank" models proposed by Datta et al. (2016), which are called Nearest Neighbor Gaussian Process (NNGP) models. Every spatial (or spatio-temporal) process induces many NNGPs models, which can project process realizations of the former to a lower dimensional subspace. They explore the spatial dependence/independence of data through conditional distributions, in particular using Vecchia (1988) approximation, to build a new valid stochastic process that works for any valid covariance function. Furthermore, when the number of neigbors is small, the NNGP process is also a GRMF, reducing the computational time requirements. Nevertheless, it is difficult to define the number of neighbors that guarantee goodness of fit. Paper 2 presents a generalization of NNGP to solve some specific limitations. Instead of using Vecchia's approximation, we propose to use the Stein's likelihood (Stein et al., 2004) approximation which works with block of data. We compare different scenarios to show the main benefits of block-NNGP.

This thesis consists of two parts. Part I is a general introduction to the field and puts the appended papers into context. Part II contains the appended papers.

Bibliography

Blangiardo, M., Cameletti, M., Baio, G., and Rue, H. (2013). Spatial and spatio-temporal models with R-INLA. *Spatial and Spatio-temporal Epidemiology*, 4:33–49.

Bolin, D. and Lindgren, F. (2011). Spatial models generated by nested stochastic partial differential equations, with an application to global ozone mapping. *Annals of Applied Statistics*, 5(1):523–550.

Cameletti, M., Lindgren, F., Simpson, D., and Rue, H. (2013). Spatio-temporal modeling of particulate matter concentration through the SPDE approach. *Advances in Statistical Analysis*, 97:109–131.

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

Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E. (2016). Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*.

Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: The SPDE approach. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 73(4):423–498.

Lindström, J. and Lindgren, F. (2008). A Gaussian Markov random field model for total yearly precipitation over the African Sahel. Master's thesis, Lund University.

Quiroz, Z. C., Prates, M. O., and Rue, H. (2015). A Bayesian approach to estimate the biomass of anchovies off the coast of Perú. *Biometrics*, 71(1):208–217.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society B*, 71(2):319–392.

Stein, M. L., Chi, Z., and J.Welty, L. (2004). Approximating likelihoods for large spatial data sets. *Journal of the Royal Statistical Society, Series B*, 66(2):275–296.

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

Chapter 2

Background and Challenges in Geostatistics

Fisher (1926) was studying crop growth at Rothamsted Experimental Station when noted that measurements close to each other were correlated. Then, he developed experimental design theory to reduce the effects caused by the spatial correlation (Fisher, 1935). Some years later, Krige (1951) and Matheron (1963) used the dependence in data to improve interpolations of spatial data, such statistical field nowdays is known as geostatistics. In this chapter we present the essential elements of geostatistical models.

Most of the modeling developed for point-referenced data is based on random fields. And in particular, Gaussian random fields have been widely used in geostatistics. For this reason, this chapter gives a brief summary of some theoretical background for Random fields and Gaussian random fields. Since the Gaussian Markov Random fields have also had an enormous impact on geostatistics, some theory of them is presented in this chapter.

2.1 Random fields

Definition 1. ***(Random field)*** *Let a probability space, (Ω, F, P), and $D \subset \Re^d$ a d-dimensional Euclidean space. A random field is defined by $\{Y(s,\omega) : s \in D \subset \Re^d, d > 1, \omega \in \Omega\}$, where $Y(s,\omega)$ is a real valued function, which, for every fixed s, is a measurable function of $\omega \in \Omega$. And it is specified by its finite-dimensional joint distributions*

$$F(y(s_1), \ldots, y(s_n)) = P(Y(s_1, w) \leq y(s_1), \ldots, Y(s_n, w) \leq y(s_n)) \qquad (2.1)$$

for every finite n and every collection $s_1, \ldots, s_n \in D \subset \Re^d$ for $d > 1$.

This definition means that for any fixed, finite set of spatial locations $\{s_1, \ldots, s_n\} \in D \subset \Re^d$, $(Y(s_1), \ldots, Y(s_n))^T$ is a random vector. For a fixed $w \in \Omega$, the function $Y(s)$ is a deterministic function of s, that is, a realization of the random field, which is denoted by $y(s)$. Then $(y(s_1), \ldots, y(s_n))^T$ are realizations of the random field and the observations of data are considered one realization. For geostatistical data D is a continuous, fixed set. And, the spatial dimension d is tipically either two or three, and s represents some location or position. When s is a position in space-time the random field is called spatio-temporal random field.
A random field must necessarily satisfiy two consistency requirements of the finite-dimensional distributions: Symmetry and Compatibility conditions.

Definition 2. ***(Symmetry condition)*** *Consider a permutation π of the index set $\{1, \ldots, n\}$ as $\{\pi_1, \ldots, \pi_n\}$, then*

$$F(y(s_1), \ldots, y(s_n)) = F(y(s_{\pi 1}), \ldots, y(s_\pi n)) \tag{2.2}$$

Definition 3. ***(Compatibility condition)***

$$F(y(s_1), \ldots, y(s_{n-1})) = F(y(s_{\pi 1}), \ldots, y(s_{n-1}), \infty)) \tag{2.3}$$

The first condition means that F is invariant under reordering of the sites, and the second contidion means that F is consistent under marginalization.

2.1.1 Mathematical construction of random fields

One approach involves defining a collection of random variables to have specific finite-dimensional distributions, and then using the Kolmogorov's existence theorem to prove that the corresponding stochastic process exists.

Theorem 1. ***(Komogorov's Existence theorem)*** *If a system of finite-dimensional distributions, $F(y(s_1), \ldots, y(s_n))$, for $s \in D \subset \Re^d, d > 1$, satisfies the symmetry condition (Equation (2.2)) and the compatibility condition (Equation (2.3)), then there exists a random field $\{Y(s, \omega) : s \in D \subset \Re^d, d > 1, \omega \in \Omega\}$ on some probability space (Ω, F, P).*

So, this theorem says that if any finite-dimensional distribution satisfy both conditions, then there exists a random field with those finite-dimensional distributions.

2.2 Gaussian random field

An important special class of random fields is the class of Gaussian random fields, usually called Gaussian fields.

Definition 4. ***(Gaussian Random field)*** *A random field* $\{Y(s), s \in D\}$ *is a Gaussian random Field (GF) if for any* $n \geq 1$ *and for any location* $s_1, s_2, \ldots, s_n \in D \subset \Re^d, d > 1$, $(Y(s_1), ...Y(s_n))^T$ *follows a multivariate Gaussian distribuion. The mean function and covariance function of Y are:*

$$\mu(s) = E(Y(s)); s = (s_1, s_2, \ldots, s_n)^T,$$

$$C(s_i, s_j) = cov(Y(s_i), Y(s_j)) = \sigma^2 \rho(s_i, s_j); i, j = 1, \ldots, n,$$

which are assumed to exist for all s_i *and* s_j*.*

Multivariate Gaussian distributions are specified by their mean vector and covariance matrix, thus Gaussian random fields are specified by their mean and covariance functions since they are especified by their finite-dimensional multivariate Gaussian distributions. The covariance function must be positive definite, that is for any set of locations the covariance matrix is positive definite. Moreover, if the covariance function is positive definite, then the finite dimensional distributions fulfills the consistency conditions of the Kolmogorov existence theorem.

Definition 5. *A random field is weakly stationary if* $\mu(s) = \mu$ *for all* $s \in D$ *and if the covariance function only depends on* $s_i - s_j$*.*

Definition 6. *A random field weakly stationary is called isotropic if the correlation function* ($\rho(s_i, s_j)$)*, and thus the covariance function, only depends on the Euclidean distance h between* s_i *and* s_j*, i.e.,* $\rho(s_i, s_j) = \rho(h)$ *with* $h = \parallel s_i - s_j \parallel$*.*

These definitions are also true for Gaussian random fields. Specifying a Gaussian random field through its covariance function is the most popular method in geostatistics, which means to use some specific correlation function. One of the most used correlation functions, for stationary and isotropic random fields, is the Matérn correlation function defined as follows

$$\rho(h) = \frac{(s_\nu h)^\nu K_\nu(s_v h)}{\Gamma(\nu) 2^{\nu-1}},$$

where $\nu > 0$ is a shape parameter and determines the smoothness of the process, K_ν is the modified Bessel function of order and s_v is a scale parameter. The effective

range ($r = \frac{\sqrt{(8\nu)}}{s_\nu}$), is the distance at which the correlation is approximately 0.1. The correlation function can be re-defined depending on the effective range by:

$$\rho(h) = \frac{1}{\Gamma(\nu)2^{\nu-1}}(\frac{\sqrt{(8\nu)}h}{r})^{\nu} K_\nu(\frac{\sqrt{(8\nu)}h}{r}).$$

2.2.1 Positive definite matrices

For existence of a Gaussian random field, with specific mean and covariance functions, it is enough to ensure that the covariance function is positive definite.

Definition 7. *The $n \times n$ matrix A is positive definite iff $x^T Ax > 0, \forall x \neq 0$, and denoted as $A > 0$.*

So, suppose that $\{Y(s) : s \in D\}$ is a weakly stationary random field with covariance function C. Given any finite set of positions $s_1, \ldots, s_n \in D$, the covariance $n \times n$ matrix C_n of the n-finite dimensional joint distribution (Equation (2.1))

$$C_n = \begin{pmatrix} C_0 & Cs_1 - s_2 & \cdots & Cs_1 - s_n \\ Cs_2 - s_1 & C_0 & \cdots & Cs_2 - s_n \\ \vdots & \vdots & \ddots & \vdots \\ C_{s_n - s_1} & a_{s_n - s_2} & \cdots & C_0 \end{pmatrix}$$

is positive definite, that is $x^T C_n x > 0, \forall x \neq 0$.

Conversely, given any positive definite function C that generates positive covariance matrices there exists a Gaussian random field with covariance function C (Gelfand et al., 2010).

This condition ensures, among other things, that every linear combination of random variables in the collection will have positive variance. For Gaussian random fields, the consistency conditions of the Kolmogorov existence theorem are reduced to the requirement that the covariance function is positive definite.

Another way to prove that a covariance function is postitive definite is through Bochner's Theorem.

Theorem 2. ***(Bochner's Theorem)*** *A real valued continuous function C is positive definte if and only if it is the Fourier transformation of a symmetric, nonnegative measure F on $\Re^d$, that is, if and only if*

$$C(h) = \int_{\Re^d} \exp(ih^T x) dF(x) = \int_{\Re^d} cos(h^T x) dF(x)$$

2.3 Gaussian Markov Random Fields

In this section a briefly introduction of some basic theory about graphs is given, to then define a Gaussian Markov Random Field (GMRF).

Definition 8. *A graph $G = (V, E)$ is defined by a group of V vertices, usually called nodes, joined between them by a group of lines called edges E. If two nodes $i, j \in V$ are joined by an edge, they are said to be neighbors ($i \sim j$).*

From this definition it is implicit that $i \sim j \Leftrightarrow j \sim i$. This definition of graph is very general, in fact many "things" can be seen like graphs, for instance in the spatial context, a regular or irregular lattice can represent a graph (Rue and Held, 2005).
If all edges have no direction this graph is called *undirected graph*. If all edges have direction this graph is called *directed graph* (DAG), where the direction of the edge is chosen based on the idea that something happening at the edge-head node has been "caused" by something happening at the edge-tail node. The DAGs are used to define the NNGP. And if the graphs have a combination of undirected and directed edges, they are called *chain graphs*, which are used to define the block-NNGP. All of these graphs can be used to show conditional independence assumptions.
A Markov Random Field is a random field satisfying conditional independence assumptions with respect to a graph, hence the term Markov. Markov random fields that are also Gaussian are called Gaussian Markov Random Fields and have had an enormous impact on spatial statistical modeling (Rue and Held, 2005).

Definition 9. *A random field $\{Y(s) : s \in D\}$ is a Gaussian Markov Random Field (GMRF) with respect to a graph G=(V,E) with mean μ and precision matrix $Q > 0$ (positive definite), if and only if, for any finite set of locations $s_1, s_2, \ldots, s_n \in D$, the joint distribution of $Y = (Y(s_1), ...Y(s_n))^T$ is given by*

$$f_Y(y) = (2\pi)^{(-n/2)}|Q^{1/2}|exp\left(-\frac{1}{2}(y-\mu)^T Q(y-\mu)\right)$$

where

$$Q_{ij} \neq 0 \Longleftrightarrow i, j \in E, \forall i \neq j.$$

Here the vertex set V corresponds to the nodes $\{s_1, \ldots, s_n\}$ and the edge set E specifies the dependencies between the random variables $Y(s_1), Y(s_2), \ldots, Y(s_n)$. Furthermore, if Q is a symmetric and positive definite matrix $n \times n$, then Q_{ij} is equal to zero if and only if, the nodes i and j are not connected by an edge. Then, for $i \neq j$,

$$Y(s_i) \perp Y(s_j)|Y(s_{-ij}) \Longleftrightarrow Q_{ij} = 0,$$

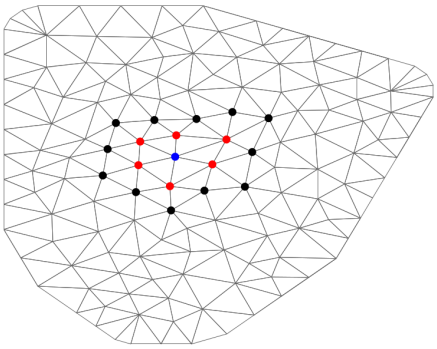

Figure 2.1: An example of a GMRF. The red and black points are neighbors to the blue point. Q will be more sparse if we only consider the red points as neighbors to the blue point

.

which implies that $Y(s_i)$ and $Y(s_j)$ are conditionally independent and it means that the conditional distribution of observed variable at some node only depends on its neighbors (Figure (2.1)). In other words, this definition says that we are able to know if two nodes are conditionally independent "reading off" the precision matrix Q, where Q determines the graph G by its non-zero values.

Another important feature about GMRF's is that due to their preserved Markov properties, the precision matrix Q is sparse i.e., it will have a few non-null elements. Therefore, working with a sparse precision matrix instead of a dense covariance matrix allow us to obtain much quicker inference. Thus, the benefit of using a GMRF it is purely computational and lies in the sparsity of the precision matrix, because there are many numerical methods which use this feature for fast computing. For a thorough approach to GMRFs, see Rue and Held (2005).

2.4 Methods

Point referenced data are usually analyzed through Gaussian random fields. However, it is well-known that computations can be prohibitive when the number of observations is large because calculations over a Gaussian field depend on the covariance and precision matrix, which are usually dense. For instance, inverting the covariance matrix is not computationally feasible in such case. This drawback is an important open problem that has led to a large number of new statistical methods, and two of them are introduced in this section.

2.4.1 SPDE

Recently, Lindgren et al. (2011) derived a method for producing approximations to the Matérn covariance family. They used the fact that a Gaussian field x with stationary Matérn family is the unique solution to the following stochastic partial differential equation (SPDE),

$$(k^2 - \triangle)^{\alpha/2} x(s) = W(s); s \in \Re^n; \alpha = \nu + d/2; \triangle = \sum_{i=1}^{d} \frac{d^2}{dx_i^2}, \tag{2.4}$$

where $(k^2 - \triangle)^{\alpha/2}$ is the fractional Laplacian operator and W is a spatial Gaussian white noise (Whittle, 1954). The weak formulation of Equation (2.4) is

$$[< \phi_i, (k^2 - \triangle)^{\alpha/2} \widetilde{x} >_\Omega] \stackrel{D}{=} [< \phi_i, W >_\Omega], \tag{2.5}$$

where $a \stackrel{D}{=} b$ denotes equality in distribution, and ϕ_i is any arbitrary well-behaved test function. The solution of Equation (2.4) should guarantee that the left side and the right side of Equation (2.4) are equal, in the sense that the left side should have the "same properties" as the stochastic process W. Since the solution of Equation (2.5) guarantees the left side of Equation (2.4) has the same mean and covariance functions of a Gaussian White noise, it is enough to find the solution of this last expression. Of course, x, the solution of Equation (2.4), now is approximated by $\widetilde{x}$, the solution of Equation (2.5).

A common approach to solve SPDEs like Equation (2.5) involves numerical methods. Lindgren et al. (2011) used the finite element method (FEM). Their general procedure involves the following main steps: (i) Generate a Delaunay triangulation over the domain, (ii) Construct basis functions (ψ_j) over the triangulation, such that the approximation of the solution $x(s)$ has the form,

$$\widetilde{x}(s_k) = \sum_{j=1}^{n} \psi_j^T(s_k) w_j, \tag{2.6}$$

where w_j are Gaussian weights and n is the number of vertices in the triangulation. The explicit choice for ψ_j are piecewise linear basis functions, where $\psi_j(s_v)$ is 1 at vertex $v = j$ and zero at all other vertices. To understand how are obtained the piecewise linear basis functions on $\Re^2$, assume that some triangle Tr of the mesh has vertices s_j, s_k and s_l, in such case, by definition, the basis functions of this specific triangle Tr are $\psi_j^T r(s_j) = 1$, $\psi_j^T r(s_k) = 0$ and $\psi_j^T r(s_l) = 0$. Thus, the piecewise basis function of $\psi_j^T r(s_v)$ for any location $s_v = (x_v, y_v)$ inside the triangle Tr, which is not

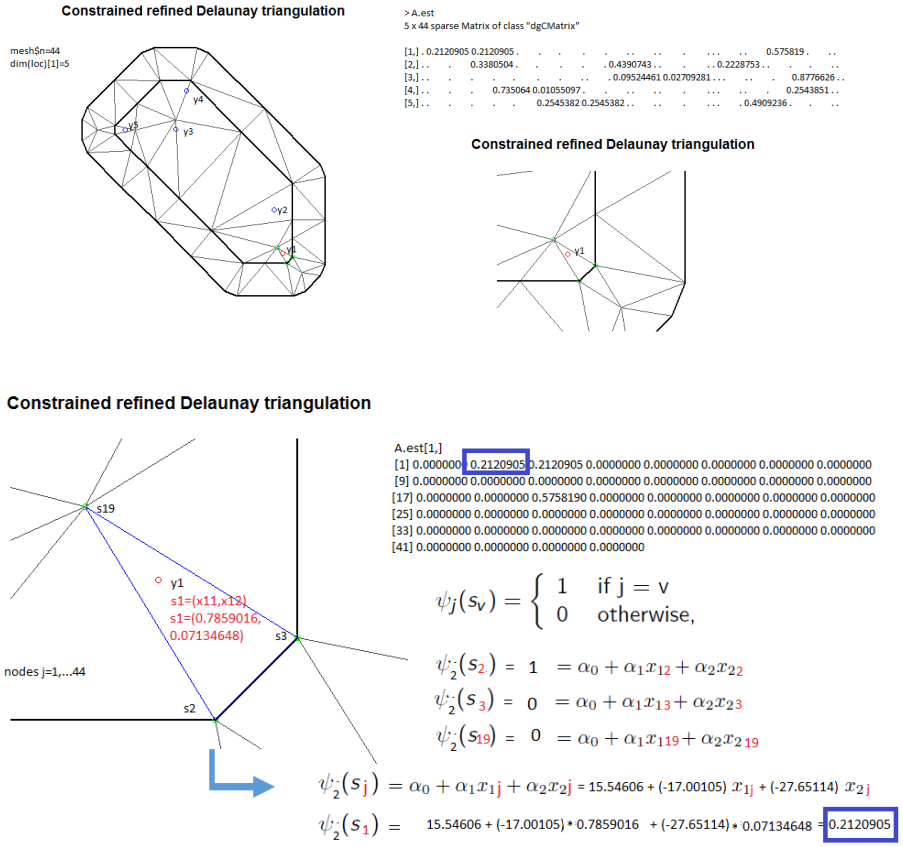

Figure 2.2: Example of computation of piecewise linear basis functions on $\Re^2$.

a vertex of such triangle, is constructed as linear combination, where,

$$\psi_j^T r(s_j) = 1 = \alpha_{j0}^T r + \alpha_{j1}^T r x_j + \alpha_{j2}^T r y_j$$

$$\psi_j^T r(s_k) = 0 = \alpha_{j0}^T r + \alpha_{j1}^T r x_k + \alpha_{j2}^T r y_k$$

$$\psi_j^T r(s_l) = 0 = \alpha_{j0}^T r + \alpha_{j1}^T r x_l + \alpha_{j2}^T r y_l.$$

The values of $\alpha^T r_0$, $\alpha^T r_1$ and $\alpha^T r_2$ are computed using the Cramer's Rule. Therefore, $\psi_j^T r(s_v) = \alpha_{j0}^T r + \alpha_{j1}^T r x_v + \alpha_{j2}^T r y_v$. The basis function for the other vertices of the triangle can be computed using the same procedure, such that, $\psi_k^T r(s_v) = \alpha_{k0}^T r + \alpha_{k1}^T r x_v + \alpha_{k2}^T r y_v$ $\psi_l^T r(s_v) = \alpha_{l0}^T r + \alpha_{l1}^T r x_v + \alpha_{l2}^T r y_v$, (Figure (2.2)).

From Equations (2.5) and (2.6),

$$[< \phi_i, (k^2 - \triangle)^{\alpha/2} \sum_{j=1}^{n} \psi_j(s) w_j >_\Omega] \stackrel{D}{=} [< \phi_i, W >_\Omega]; i = 1, ..., n,$$

$$[\sum_{j=1}^{n} < \phi_i, (k^2 - \triangle)^{\alpha/2}\psi_j(s) >_\Omega w_j] \stackrel{D}{=} [< \phi_i, W >_\Omega]; i = 1, ..., n.$$

(iv) The choice of the test function is related to the basis functions. Choosing test functions equal to the basis functions themselves ($\phi_i = \psi_i$) leads to Galerkin solutions. Lindgren et al. (2011) chose this specific test function, for $\alpha = 2$. (v) Solve the system of equations. For $\alpha = 2$ the weak formulation can be written as

$$[\sum_{j=1}^{n} < \psi_i, (k^2 - \triangle)\psi_j(s) >_\Omega w_j] \stackrel{D}{=} [< \psi_i, W >_\Omega]; i = 1, ..., n.$$

The finite dimensional weak SPDE solution is reduced to find the distribution for the weights w_j that fulfills the last formulation. The stochastic weights follow a Gaussian distribution with mean zero and a precision matrix (Q_{α,κ^2})=$K_{\kappa^2}C^{-1}K_{\kappa^2}$ that can be written directly as a function of the parameters. Such precision matrix is computed with the help of Green's first identity, which is a multi-dimensional version of integration-by-parts, and Garlekin solutions (for more details see appendix B). Matrices C, G e K are defined by:

$$C_{i,j} = \langle \psi_i, \psi_j \rangle, \quad G_{i,j} = \langle \nabla\psi_i, \nabla\psi_j \rangle, \quad (K_{\kappa^2})_{i,j} = \kappa^2 C_{i,j} + G_{i,j} \tag{2.7}$$

These matrices are computed using the geometry of each triangle defined by its vertices, angles and edges.

For the basis function chosen, C, G and K are sparse matrices. Nevertheless, the precision matrix C^{-1} is dense, as consequence, Q_{α,κ^2} is also dense. To solve this problem, Lindgren et al. (2011) included a Markov approximation replacing the C-matrix with a diagonal matrix $\widetilde{C}$ with diagonal elements $C_{ij} =< \psi_i, 1 >$. As a result, the stochastic weights have a sparse precision matrix, they are approximations of GMRFs and can be written directly as a function of the parameters.

2.4.2 NNGP

Cressie (1993) stated that it would be of great interest to construct a spatial model partly from local specifications of the conditional probabilities (Markov random-field approach) and partly from global specifications of joint probabilities (classic geostatistical approach). Recently, Datta et al. (2016) have found a way to make such specifications, so that the new random process satisfying them exists. They called it the Nearest neighbor Gaussian Process (NNGP).

To built the new process, it is assumed that w is a Gaussian field, $w(s) \sim \text{GP}(0, C(\theta))$, defined for all $s \in D \subset \Re^2$, where $C(\theta)$ is any valid covariance function. Let

$S = (s_1, \ldots, s_n)'$ be a fixed set of locations in D. Then the joint density of $w_s = (w(s_1), \ldots, w(s_n))$ for $i = 1, \ldots, n$, can be written as

$$p(w_s) = p(w(s_1)) \prod_{i=2}^{M} p(w(s_i)|w(s_1), w(s_2), \ldots, w(s_{i-1})). \tag{2.8}$$

Vecchia (1988) propose to replace the conditioning sets on the right-hand side of Equation (2.8) with conditioning sets of size at most m, where m<<n. Datta et al. (2016) propose to use the m nearest neighbors observations from the "past", then the approximated joint density of Equation (2.8) is $\widetilde{p}(w_s) = p(w(s_1)) \prod_{i=2}^{n} p(w(s_i)|w(s_{i_m}))$, where $w(s_{i_m})$ are the neighbor observations of $w(s_i)$. Throughout the last statement and theory of Section (2.1.1), Datta et al. (2016) built a valid spatial process called NNGP. This process is a particular case of the general process we are proposing in Paper 2.

Bibliography

Cressie, N. (1993). *Statistics for Spatial Data*. John Wiley & Sons Ltd., NJ.

Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E. (2016). Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*, 111(514):800–812.

Fisher, R. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain*, 33:503–513.

Fisher, R. (1935). *The Design of Experiments*. Hafner publishing company, New York.

Gelfand, A., Fuentes, M., Guttorp, P., and Diggle, P. (2010). *Handbook of Spatial Statistics*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. Taylor & Francis, Boca Raton.

Krige, D. (1951). A statistical approaches to some basic mine valuation problems on the Witwatersrand. *Journal of the Chemical, Metallurgical and Mining Society of South Africa*, 52:119–139.

Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: The SPDE approach. *Journal of the Royal Statistical Society. Series B.*, 73(4):423–498.

Matheron, G. (1963). Principles of geostatistics. *Economic Geology*, 58:1246–1266.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and applications*. Chapman & Hall/CRC.

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

Whittle, P. (1954). On stationary processes in the plane. *Biometrika*, 44:434–449.

Chapter 3

Bayesian Inference for geostatistical models

Let $Y = (Y(s_1), \ldots, Y(s_n))$ be a realization of a random field defined for all $s_i \in D \subset \Re^2$, $i = 1, \ldots, n$. The basic geostatistical Gaussian regression model is of the form

$$Y(s_i) = X'(s_i)\beta + w(s_i) + \epsilon(s_i), \tag{3.1}$$

where β is a coefficient vector (or regression parameter), the covariates $X(s_i) = (1, x_i)$, the gaussian random field $w = (w(s_1, \ldots, w(s_n)))' \sim N(0, C(\theta))$ is a spatial structured effect, it captures the spatial association, and a common assumption is that $\epsilon(s_i)$ for $i = 1, \ldots, n$ are independent identically distributed, $\epsilon(s_i) \sim N(0, \tau^2)$, which models the measurement error.

In geostatistics, one often must develop models in the presence of complicated spatial processes, multiple sources of data, uncertainty in parameterizations, among other challenges. One can approach such complex problems from either a joint or conditional viewpoint. Spatial associations between observed data might be captured using models that build dependencies in different stages or hierarchies. In particular, hierarchical models are especially characterized on the fact that the joint probability distribution of a collection of random variables can be descomposed into conditional distributions and a marginal distribution. Although these models can be considered from either a classical or Bayesian perspective, as the level of complexity increases, the Bayesian paradigm of statistical inference, which uses posterior distributions of model parameters, becomes a necessity.

For instance, the simple geostatistical model, we have just described, can be written

hierarchically as follows:

$$\begin{aligned} \text{Data model} &: Y|. \sim N(X\beta + w, D(\theta)), \\ \text{Gaussian model} &: w \sim N(0, C(\theta)); \beta \sim N(\mu_\beta, \Sigma_\beta) \\ \text{Parameter model} &: \theta = [\phi, \sigma^2, \tau^2]. \end{aligned}$$

So, the joint posterior pdf is given by

$$p(\theta, \beta, w|y) \propto p(\phi) \times p(\sigma^2) \times p(\tau^2) \times f_N(\beta|\mu_\beta, \Sigma_\beta) \times f_N(w|0, C(\theta)) \times f_N(y|X\beta + w, D(\theta)), \tag{3.2}$$

where $p(.)$ is some pdf, and f_N is a pdf of a multivariate Normal distribution. In particular, this model fits in the class of Latent Gaussian models. Statistical inference for this model can be achieved using Markov Chain Monte Carlo (Casella and Robert, 1999) techniques or Integrated Nested Laplaple approximation (Rue et al., 2009), which are the introduced in the next sections.
Often a marginalized likelihood is used, which is obtained by integrating out the spatial effects w. This yields

$$p(\theta, \beta|y) \propto p(\phi) \times p(\sigma^2) \times p(\tau^2) \times f_N(\beta|\mu_\beta, \Sigma_\beta) \times f_N(y|X\beta, C(\theta) + D(\theta)). \tag{3.3}$$

This marginal formulation is used to facilitate estimation of parameters (Gelfand et al., 2010).

3.1 INLA

Eidsvik et al. (2009) proposed fast approximate methods for computing posterior marginals in spatial generalized linear mixed models, in particular for geostatistical data with a high dimensional latent spatial variable. Their approximations were very fast, in contrast to MCMC runs. Rue et al. (2009) extend this fast inference to complex spatial and spatio-temporal models, in particular for LGMs. Inference for these models was usually performed through MCMC methods, but such methods are computationally expensive, specially when dealing with big datasets.

3.1.1 Latent Gaussian models

Latent Models are a subclass of structured additive models, which can also be seen as a representation of a hierarchical model. First, let us assume that for $I = \{i : i = 1, ..., n\}$, we have n observed variables y_i. The linear predictor is defined

by

$$\eta_i = \beta_0 + \sum_{j=1}^{\eta_f} f_S^{(j)}(u_i) + \sum_{k=1}^{\eta_{\beta_k}} \beta_k z_{ki} + \epsilon_i, \tag{3.4}$$

where $\beta_k' s$ are coefficients for linear effects on a vector of covariates z, $f_S^{(j)}$ incorporates dependence between observations, which can be of various kind like spatial, temporal or spatiotemporal. In geostatistical models, the gaussian random field $w(s)$ is a component of $f_S^{(j)}$. And ϵ represents unstructured random effects. The latent field x is composed by a vector: $x = \{\{\beta_0\}, \{\beta_k\}, \{f_S^{(j)}\}\}$. If the distribution of the latent field is set as Gaussian such model becomes a Latent Gaussian Model (LGM).

A typical Hierarchical model is defined by: a first stage, where a distributional assumption is formulated for the observations, which depend on the latent field. Here, we assume observations conditionally independent given the latent field. A second stage, is a latent field, which might follow a Multivariate Gaussian distribution with mean μ and covariance matrix $\Sigma(\theta)$. And a third stage is composed by all the unknown parameters (called hyperparameters). A prior model is assigned for these unknown parameters. Thus, a LGM can be defined like a Hierarchical model with the following structure:

i) a likelihood model for the response variables, which are assumed to be independent given the latent parameters x: $y|x, \theta \sim \prod_{i \in I} \pi(y_i|x_i, \theta)$,

ii) a latent Gaussian field: $x|\theta \sim N(\mu, \Sigma(\theta))$, and

iii) hyperparameters θ: $\theta \sim \pi(\theta)$.

In many LGM's and hierarchical models, like geostatistical models, the latent Gaussian field is also a Gaussian Markov Random Field (GMRF), or it can be approximated by GMRF's.

The joint posterior of the LGM can be calculated using the likelihood function, latent Gaussian distribution and the distribution of hyperparameters as follows:

$$\pi(x, \theta|y) \propto \pi(\theta)\pi(x|\theta) \prod_{i \in I} \pi(y_i|x_i, \theta).$$

Let $x|\theta \sim N(0, \Sigma(\theta))$ and $Q^{-1} = \Sigma(\theta)$ be the precision matrix, then

$$\pi(x, \theta|y) \propto \pi(\theta)|Q^{1/2}| \exp\left(-\frac{1}{2}x^T Q x + \sum_{i \in I} \log\{\pi(y_i|x_i, \theta)\}\right).$$

3.1.2 Bayesian Inference with INLA

INLA (Rue et al., 2009) works out with LGM's that satisfy two properties: (i) The latent field x is a GMRF, and (ii) The number of hyparameters m is small. These properties make it possible to obtain fast and accurate Bayesian inference.

More specifically, in geostatistical models, set j=1 and let $f_S = w$ in Equation (3.4) be a Gaussian field with dense covariance structure $C(\theta)$. To improve computational time, INLA approximates f_S by $\widetilde{f}_S$, where $\widetilde{f}_S$ is a GMRF, which allows a sparse precision (inverse covariance) matrix representation. Rue and Tjelmeland (2002) showed that for a regular lattice, a Gaussian field with Matérn correlation function can be well approximated by a GMRF. Therefore, although f_S is not exactly the same as $\widetilde{f}_S$, it is well approximated by $\widetilde{f}_S$. Such result, combined with the analytical results presented in Lindgren et al. (2011), can improve computational performance dramatically (for more details, see Rue and Tjelmeland (2002); Rue and Held (2005); Lindgren et al. (2011)). Thus, the latent field x is composed by a vector: $x = \{\{\beta_0\}, \{\beta_k\}, \{\widetilde{f}_S\}\}$. Therefore, the classical geostatistical model is a LGM that satisfies properties (i) and (ii).

The posterior marginals of the latent variables $\pi(x_i|y)$ and the posterior marginal of hyperparameters $\widetilde{\pi}(\theta_j|y)$ are not easy to calculate, and that is the main aim of INLA. The general idea of INLA is divided into the next tasks:

- First, it provides an approximation of $\widetilde{\pi}(\theta|\text{y})$ to the join posterior of hyperparameters given the data $\pi(\theta|\text{y})$.

- Second, it provides an approximation of $\widetilde{\pi}(x_i|\theta,\text{y})$ to the marginals of the conditional distribution of the latent field given the data and the hyperparameters $\pi(x_i|\theta,\text{y})$.

- And third, it explores $\widetilde{\pi}(\theta|\text{y})$ on a grid and use it to integrate out θ in $\widetilde{\pi}(x_i|y)$ and θ_{-j} in $\widetilde{\pi}(\theta_j|y)$.

Approximating $\pi(\theta|y)$

In the first case, the denominator $\pi(x|\theta, y)$ is not available in closed form but it can be approximated using a Gaussian approximation, that is:

$$\pi(\theta|y) = \frac{\pi(x, \theta|y)}{\pi(x|\theta, y)} \propto \frac{\pi(x, \theta, y)}{\pi(x|\theta, y)}$$

which is approximated by:

$$\widetilde{\pi}(\theta|y) \propto \frac{\pi(x,\theta,y)}{\widetilde{\pi}_G(x|\theta,y)}|_{x=x^*(\theta)} \tag{3.5}$$

where $\widetilde{\pi}_G$ denotes a Gaussian approximation to the full conditional density of x. In particular, the Gaussian approximation was contructed by matching the mode and the curvature at the mode to ensure a good approximation of the true marginal density. Here $x^*(\theta)$ is the mode of the full conditional for x for a given θ, and it is obtained by using some optimization method like Newton-Raphson. In additon, Equation (3.5) is a Laplace approximation.

Approximating $\pi(x_i|\theta,y)$

In order to approximate $\pi(x_i|\theta,y)$, three options are available. The first option, is to use the marginals of the Gaussian approximation $\pi_G(x|\theta,y)$. The extra cost to obtain $\pi_G(x_i|\theta,y)$ is to compute the marginal variances from the sparse precision matrix (matrix with many null elements) of $\pi_G(x|\theta,y)$. The second and third options solve the fact that even if the Gaussian approximation often gives aceptable results, there still can be errors in the location and/or errors due to the lack of skewness. Then, the second option is to do again a Laplace approximation, this approximation is more accurate and it is denoted by $\widetilde{\pi}_{LA}(x_i|\theta,y)$:

$$\widetilde{\pi}_{LA}(x_i|\theta,y) \propto \frac{\pi(x,\theta,y)}{\widetilde{\pi}_{GG}(x_{-i}|x_i,\theta,y)}|_{x_{-i}=x^*_{-i}(x_i,\theta)}, \tag{3.6}$$

where $\widetilde{\pi}_{GG}$ is the Gaussian approximation to $\pi(x_{-i}|x_i,\theta,y)$ and $x^*_{-i}(x_i,\theta)$ is the mode. The third option is the simplified Laplace approximation $\pi_{SLA}(_{xij}|\theta,y)$, which is obtained by doing a Taylor expansion on the numerator and denominator of Equation (3.6). It corrects the Gaussian approximation for location and skewness with a moderate extra cost when compared to the Laplace approximation.

Approximating $\pi(\theta_j|y)$

It can be calculated from $\widetilde{\pi}(\theta|y)$, however, this solution has a high computational cost. Then, an easier approach is to select good evaluation points for the numerical solution of $\widetilde{\pi}(\theta_j|y)$. To find these points, two approaches are proposed: the GRID and the central composit design (CCD) strategies (Rue et al., 2009).
(i) the GRID strategy is more accurate but also time consuming, it defines a grid of points covering the area where most of the mass of $\widetilde{\pi}(\theta|y)$ is located, (ii) on the other hand, the CCD strategy consists in laying out a small amount of points in a

m-dimensional space in order to estimate the curvature of $\widetilde{\pi}(\theta|y)$. For this reason this last one requires much less computational power compared to the GRID strategy. Then using approximations $\widetilde{\pi}(x_i|\theta, y)$ and $\widetilde{\pi}(\theta_j|y)$ the posterior marginal for latent variables $\widetilde{\pi}(x_i|y)$ can be computed via numerical integration:

$$\widetilde{\pi}(x_i|y) = \int \widetilde{\pi}(x_i|\theta, y)\widetilde{\pi}(\theta|y)d\theta$$

$$\widetilde{\pi}(x_i|y) = \sum_j \widetilde{\pi}(x_i|\theta_j, y)\widetilde{\pi}(\theta_j|y)\triangle\theta_j.$$

For more details on Bayesian spatial modelling through the SPDE approach using INLA see Cameletti et al. (2011); Lindgren and Rue (2015).

3.2 MCMC

Markov Chain Monte Carlo (MCMC) is a class of Monte Carlo methods. It was popularized by a paper of Gelfand and Smith (1990). The method relies on simulate dependent samples that are approximately from a posterior probability distribution. This method has revitalized Bayesian statistics, has also transformed the sciences, especially fitting spatial and spatio-temporal models. In Bayesian statistics, there are two popular algorithms that are useful for Markov Chain simulation: the Gibbs Sampling and the Metropolis-Hastings algorithm.
The Gibbs sampling is another algorithm useful for MCMC. Suppose that we are interested in sampling from the posterior $\pi(\theta|y)$, where θ is divided into k subvectors, $\theta_1, \theta_2, \ldots, \theta_k$. The steps to a Gibbs Sampling are summarized as follows:

Algorithm (The Gibbs sampling)

1: Pick a vector of starting values $\theta^{(0)}$.
2: Draw a value of $\theta_1^{(1)}$ from the full conditional $\pi(\theta_1^{(1)}|\theta_2^{(0)}, \ldots, \theta_k^{(0)}, y)$.
3: Draw a value of $\theta_2^{(1)}$ from the full conditional $\pi(\theta_2^{(1)}|\theta_1^{(1)}, \theta_3^{(0)}, \ldots, \theta_k^{(0)}, y)$.
4: Repeat 3 for $j = 3, \ldots, k$ to draw a value of $\theta_j^{(1)}$ from the full conditional $\pi(\theta_j^{(1)}|\theta_1^{(1)}, \ldots, \theta_{j-1}^{(1)}, \theta_{j+1}^{(0)}, \ldots, \theta_k^{(0)}, y)$.
5: Repeat step 2 to 4 to draw $\theta^{(2)}$, using $\theta^{(1)}$ and recursively using the most updated.
6: Repeat untill you have M draws with each draw being a vector $\theta^{(t)}$, $t = 1, \ldots, M$.

The Metropolis–Hastings (MH) algorithm was an original contribution of Metropolis et al. (1953). Some years later it was generalized by Hastings (1970). But it was rediscovered by Tanner and Wong (1987) and Gelfand and Smith (1990). Suppose that we are interested in sampling from the posterior $\pi(\theta|y)$. The algorithm requires

a starting point $\theta^{(0)}$ and a proposal distribution $q(.|.)$ which must contain the support of the posterior distribution. The algorithm is summarized as follows:

Algorithm (The Metropolis-Hastings)

1: Choose starting value $\theta^{(0)}$), such that $\pi(\theta^{(0)}|y) > 0$.
2: At iteration t, draw a candidate $\theta^{\star}$ from a proposal distribution $q(\theta|\theta^{(t-1)})$.
3: Compute the Metropolis-Hastings acceptance probability

$$r = \frac{\pi(\theta^{\star}|y)q(\theta^{(t-1)}|\theta^{\star})}{\pi(\theta^{(t-1)}|y)q(\theta^{\star}|\theta^{(t-1)})}.$$

4: Generate $U \sim$ Uniform$(0, 1)$
5: **if** $U < \min(r, 1)$ **then**
6: accept $\theta^{(t)} = \theta^{\star}$
7: **else**
8: $\theta^{(t)} = \theta^{(t-1)}$.
9: **end if**
10: Repeat steps 2, 3 and 4, until the chain converges and you have M samples.

In the geostatistical context, Diggle et al. (1998) show the power of geostatistical Hierarchical models in diverse settings, while Banerjee et al. (2004) and Gelfand et al. (2010) give a modern perspective on these models.
Finley et al. (2015) uses the marginalized Equation (3.3) to estimate the parameters. Only the full conditional distribution of β is available analytically. So they use Gibbs sampling to obtain the posterior samples of β and Metropolis-Hastings to sample from θ. Then we can draw w from its full-conditional distribution, given both θ and β. The package spBayes (Finley et al., 2015) implements this algorithm.
It is worthwhile to mention that irrespective of whether we use Equation (3.2) or Equation (3.3), estimation and prediction will require matrix factorizations involving the dense $n \times n$ matrix $C(\theta)$ which may become prohibitively expensive for large number of observations. It is often more efficient in such cases to consider various dimension reduction or decorrelation approaches for modeling the Gaussian random fields Gelfand et al. (2010).

Bibliography

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall/CRC, Boca Raton.

Cameletti, M., Ignaccolo, R., and Bande, S. (2011). Comparing spatio-temporal models for particulate matter in Piemonte. *Environmetrics*, 22:985–996.

Casella, G. and Robert, C. P. (1999). *Monte Carlo Statistical Methods*. Springer, New York.

Diggle, P. J., Tawn, J. A., and Moyeed, R. A. (1998). Model based geostatistics. *Journal of the Royal Statistical Society. Series C.*, 47(3):299–350.

Eidsvik, J., Martino, S., and Rue, H. (2009). Approximate Bayesian inference in spatial generalized linear mixed models. *Scandinavian Journal of Statistics*, 36:1–22.

Finley, A. O., Banerjee, S., and P.Carlin, B. (2015). spBayes: An R package for univariate and multivariate point-referenced spatio-temporal data models. *Journal of Statistical Software*, 63(13):1–28.

Gelfand, A., Fuentes, M., Guttorp, P., and Diggle, P. (2010). *Handbook of Spatial Statistics*. Taylor & Francis.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85(410):398–409.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov Chains and their applications. *Biometrika*, 57(1):97–109.

Lindgren, F. and Rue, H. (2015). Bayesian spatial modelling with R-INLA. *Journal of Statistical Software*, 63:19.

Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: The SPDE approach. *Journal of the Royal Statistical Society. Series B.*, 73(4):423–498.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., and Teller, A. H. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and applications*. Chapman & Hall/CRC.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society . Series B.*, 71(2):319–392.

Rue, H. and Tjelmeland, H. (2002). Fitting Gaussian Markov random fields to Gaussian fields. *Scandinavian Journal of Statistics*, 29:31–49.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398):528–540.

Part II

Appended papers

Chapter 4

Bayesian spatio-temporal modeling of anchovy abundance through the SPDE Approach

Zaida C. Quiroz and Marcos O. Prates

Submitted to Spatial Statistics (2017).

.

Bayesian spatio-temporal modeling of anchovy abundance through the SPDE Approach

Z. C. Quiroz and M. O. Prates

Abstract

The Peruvian anchovy is an important species from an ecological and economical perspective. Some important features to evaluate fisheries management are the relationship between the anchovy presence/abundance and covariates with spatial and temporal dependencies accounted for, the nature of the behavior of anchovy throughout space and time, and available spatio-temporal predictions. With these challenges in mind, we propose to use flexible Bayesian hierarchical spatio-temporal models for zero-inflated positive continuous data. These models are able to capture the spatial and temporal distribution of the anchovies, to make spatial predictions within the temporal range of the data and predictions about the near future. To make our modeling computationally feasible we use the stochastic partial differential equations (SPDE) approach combined with the Integrated Nested Laplace Approximation (INLA) method. After balancing goodness of fit, interpretations of spatial effects across years, prediction ability, and computational costs, we suggest to use a model with a spatio-temporal structure. Our model provides a novel method to investigate the Peruvian anchovy dynamics across years, giving solid statistical support to many descriptive ecological studies.

Keywords: Bayesian method, GMRFs, marine ecology, INLA, spatio-temporal model, SPDEs.

1 Introduction

The Northern Humboldt Current System (NHCS) is a marine ecosystem highly dominated by the anchovy (*Eugralis ringens*), species heavily exploited by industrial and artisanal fisheries (Fréon et al., 2005), and one of the ecosystems most affected by intense climatic variability (Chavez et al., 2008). Patchiness is a rule for living marine organisms (Bertrand et al., 2014), thus, nested aggregation structures of

anchovy explain the inherent spatial dependence among individuals. Moreover, the anchovy reveals a fast response to environmental variability, in fact, anchovy populations can vary rapidly in both space and time (Bertrand et al., 2008a). As a result, the spatial anchovy distribution might be different across years, seasons (summer, spring and winter) and El Niño/La Niña events. Hence, efficient models taking into account for spatial and temporal dynamics of anchovy are essential to understand and interpret its behavior, while it contributes to decisions making, that guarantee its perpetuation.

Anchovy abundance data are non-negative and continuous, thus following Quiroz et al. (2015), we propose to model the probability of anchovy presence as well as the positive anchovy abundance using a mixture of a discrete probability mass at zero and some established continuous distribution for nonzero values. This kind of two-stage modeling was also used to analyze precipitation data (Stidd, 1973; Stern & Coe, 1984; Wilks, 1990; Sloughter et al., 2007; Berrocal et al., 2008). Regarding the spatio-temporal modeling of zero-inflated data, the main focus was on discrete data either with areal unit or point-referenced structure (Wikle & Anderson, 2003; Hoef & Jansen, 2007; Fernandes et al., 2009; Ross et al., 2012; Cosandey-Godin et al., 2014; Wang et al., 2015). On the other hand, the spatio-temporal modeling for zero-inflated continuous data with point-referenced structure was less explored, some references are Sansó & Guenni (2000); Fuentes et al. (2008); Fernandes et al. (2009) and Sigrist et al. (2012).

There are two major purposes of spatio-temporal modeling. First, to describe the past behavior of some process through the estimation of the spatial pattern in the data domain for each time of period (prediction within the range of the data - PWD). Second, to make assessments of possible different scenarios to prevent future adverse events through prediction of the spatial pattern into the future (prediction outside the range of the data - POD). However, from a statistical perspective, there is still a lack of modeling to provide POD, either due to the complex modeling or unavailable information to make predictions. In particular, Sigrist et al. (2012) applied a two-stage spatio-temporal model based on a temporal autoregressive convolution with spatially colored and temporally white innovations, to get POD of precipitation at 26 stations across 720 time periods, via transformation of the non-zero values. Fernandes et al. (2009) discussed a hurdle Gamma model to provide PWD and POD for non-negative continuous rainfall data at 32 monitoring stations across 75 weeks. They assumed the same spatial process for all weeks and the temporal pattern was only captured through temporal covariates, instead of a spatio-temporal process. Paradinas et al. (2015) used a hurdle Gamma spatio-temporal to analyze non-negative continuous European hake recruits data at 40 specific stations across

13 years. They included spatial or spatio-temporal processes, nevertheless the spatio-temporal process was not significant. They projected estimations of the mean posterior instead of computing PWD. And POD were not computed. In all of these applications, the data were collected in the same site at each time and have a relatively small number of observations per time (<80).

In this context, we introduce a complete, yet computationally efficient, spatio-temporal model that is capable of performing PWD and POD of anchovy presence and abundance, in particular, when the set of sites is large (> 500) and different across the temporal domain. Our approach includes several novel features and interpretations for these big data. We propose a Bayesian hierarchical hurdle model for positive continuous data with point-referenced structure, incorporating spatial or spatio-temporal processes, and temporal random processes, to seize the necessity of any of these structures. It is well-known that the computational efficiency of these models is highly limited by the covariance function (CF) used to introduce the spatial or spatio-temporal random processes. Thus, the Gaussian random field (GRF) is approximated to a Gaussian Random Markov field (GMRF). In particular, through the SPDE approach proposed by Lindgren et al. (2011), we approximated the GRF with Matérn CF to GMRFs. We performed statistical inferences through INLA (Rue et al., 2009), making the proposed modeling available for practitioners. A variety of applications using INLA and the SPDE approach for spatio-temporal modeling are found in Ross et al. (2012); Blangiardo et al. (2013); Cameletti et al. (2013); Musenge et al. (2013); Pennino et al. (2014); Cosandey-Godin et al. (2014); Paradinas et al. (2015). Finally, PWD and POD were obatined through samples from the posterior predictive distributions on the triangulation required for the SPDE approximation, reducing computational time requirements.

The rest of this paper is organised as follows. Section 2 presents some description of the data. Section 3 describes the models proposed to provide PWD and POD over the spatial domain. It outlines the Gaussian field approximation to a GMRF through SPDEs and Bayesian inference using INLA. Section ends with a variety of model assessment criteria. In Section 4 we applied the proposed modeling on anchovy abundance data. Final remarks are discussed in Section 6.

2 Description of data

The data used in this paper were collected from acoustic surveys by the Peruvian Marine Institute (IMARPE) usually on board the research vessels “Humboldt”, “José Olaya Balandra” and “SNP-1”, during the summer season, from 1999 to 2007. These

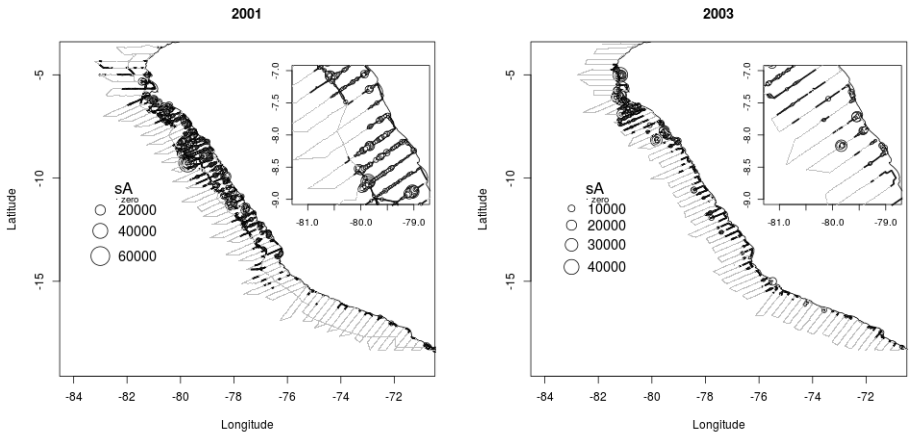

Figure 1: The observed data of anchovy abundance for the years 2001 (left) and 2003 (right). The trajectory of survey tracks is represented by parallel cross-shore transects (black circles and gray dots). The size of the circles corresponds to the abundance of anchovy higher than zero. The gray dots correspond to abundance of anchovy equal to zero. The upper right panels show a zoom of each plot.

surveys consisted on parallel cross-shore transects off the Peruvian coast with length of approximately 170 km and inter-transect distance varying between 26 and 30 km. Simmonds & MacLennan (2005) describes general methods to analyze acoustic data. In particular, the acoustic backscattered energy by surface unit (s_A) was recorded through an echosound in each geo-referenced sample unit (ESDU) of approximately 1.852 km. Data selection and classification of the species were also carried out by IMARPE. The s_A, corresponding to anchovy at each ESDU, is an indicator of "anchovy abundance". Figure 1 presents the samples of anchovy abundance equal to zero (the gray dots) and anchovy abundance higher than zero (the black circles) for specific years. There is evidence of high proportion of anchovy absence, spatial dependence on anchovy distribution, and a highly different spatial pattern of anchovy abundance across years.

The potential relevant covariates in this study are the orthodromic distance to the Peruvian coast (DC in km), the sea surface temperature (SST in °C), the latitude (Lat in °C) and the year of each survey. In addition, the depth (in km < 0) was provided by the General Bathymetric Chart of the Oceans (http://www.gebco.net/, last accessed on August 16th of 2016). These variables are particularly related to the anchovy behavior. Bertrand et al. (2011) stated that the relative anchovy biomass decreased when the distance to the coast is increased. Swartzman et al. (2008) performed an analysis of anchovy presence probability as a function of the latitude, water column depth and years. Escudero & Rivera (2011) described the relationship between the SST and site of fishing operations. The structure of SST fields is one of the main factors to identify El Niño and La Niña events, for this

reason, we believe that the time-varying environmental covariate SST may represent the irregular frequency of these events.

3 Models, inference and assessment

Define D_s as a continuous spatial domain in $\Re^2$ and D_t as a finite discrete temporal domain in $\aleph$. Let $Y(s_i,t)$ be the observational variable and $y(s_i,t)$ be the observed response with site $s_i \in D_s$ and time $t \in D_t$, where $i = 1,...,N$ and $t = 1,...,T$. For each site s_i and time t, the abundance of anchovy is zero ($y(s_i,t) = 0$) or a positive value ($y(s_i,t) > 0$). Let $p(s_i,t)$ be the probability of anchovy absence and, using the results presented in Quiroz et al. (2015), the anchovy abundance, given that anchovies are present, follows a gamma distribution, $Y(s_i,t)|Y(s_i,t) > 0 \sim \text{gamma}(\phi, \phi/\mu(s_i,t))$, with mean $\mu(s_i,t)$ and precision parameter ϕ. So, the distribution for $Y(s_i,t)$ has the finite mixture density, $\pi(y(s_i,t)|p(s_i,t),\mu(s_i,t),\phi) = p(s_i,t)\delta_0 + (1-p(s_i,t)) \times h(y(s_i,t)|\mu(s_i,t),\phi)I_{[y(s_i,t)>0]}$, where δ_0 is the Dirac delta function and h is the probability density function (pdf) of a gamma distribution.

Let $\boldsymbol{y}(.,t) = (y(s_1,t),\ldots,y(s_N,t))'$ and $\boldsymbol{y} = (\boldsymbol{y}(.,1),\ldots,\boldsymbol{y}(.,T))'$. Assuming conditional independence of $Y(s_i,t)$s given $p(s_i,t)$, $\mu(s_i,t)$ and ϕ, $\forall i \in D_s$ and $\forall t \in D_t$, then, the likelihood function can be represented as follows

$$L(\boldsymbol{y}|\boldsymbol{p},\boldsymbol{\mu},\phi) = \prod_{t=1}^{T}\prod_{i=1}^{N} \pi(y(s_i,t)|p(s_i,t),\mu(s_i,t),\phi). \tag{4.1}$$

A logit function links the linear predictor $\eta(s_i,t)^{(1)}$ with the probability of anchovy absence $p(s_i,t)$, while a logarithmic function links the linear predictor $\eta(s_i,t)^{(2)}$ to the mean $\mu(s_i,t)$,

$$\begin{aligned} \text{logit}(p(s_i,t)) &= \eta(s_i,t)^{(1)} = \mathbf{Z}^{(1)}\boldsymbol{\beta}^{(1)} + f^{(1)}(t) + f_s^{(1)}(s_i,t), \\ \log(\mu(s_i,t)) &= \eta(s_i,t)^{(2)} = \mathbf{Z}^{(2)}\boldsymbol{\beta}^{(2)} + f^{(2)}(t) + f_s^{(2)}(s_i,t). \end{aligned} \tag{4.2}$$

For each linear predictor ($k = 1,2$): $\mathrm{Z}^{(k)}$ is a covariate matrix, $\boldsymbol{\beta}^{(k)}$ is a coefficient vector (or regression parameters), $f^{(k)}(t)$ is a temporal effect and $\boldsymbol{f}_s^{(k)}(.)$ is a spatial or a spatio-temporal structured effect. The definition of a spatial or spatio-temporal effect in Equation (4.3) gives rise to S or ST models, respectively.

In S model, the structured spatial effect $f_s^{(k)}(s_i,t)$ follows a Gaussian distribution

with mean zero and it is the same across time, thus $\forall i \neq j$,

$$\text{Cov}(f_s^{(k)}(s_i,t), f_s^{(k)}(s_j,t')) = \Sigma^{(k)}(i,j) = \begin{cases} 0 & \text{if } t \neq t'; \\ \sigma_f^{(k)2} \times \rho^{(k)}(.) & \text{if } t = t', \end{cases} \tag{4.3}$$

where, $\sigma_f^{(k)2}$ is the marginal variance of $f_s^{(k)}(s_i,t)$, $\forall i,t$, and $\rho^{(k)}(.)$ is a spatial correlation function. Then $\boldsymbol{f}_s^{(k)}(.,t) = (f_s^{(k)}(s_1,t),\ldots,f_s^{(k)}(s_N,t))'$ is a GRF with mean zero and spatial covariance matrix $\boldsymbol{\Sigma}^{(k)}$ composed by $\boldsymbol{\Sigma}^{(k)}(i,j)$. Therefore, $\boldsymbol{f}_s^{(k)} = (\boldsymbol{f}_s^{(k)}(.,1),\ldots,\boldsymbol{f}_s^{(k)}(.,T))'$ is a Gaussian field with mean zero and covariance matrix $\boldsymbol{\Sigma}_f^{(k)} = \boldsymbol{I_T} \otimes \boldsymbol{\Sigma}^{(k)}$, where $\boldsymbol{I_T}$ represents the identity T-diagonal matrix and $\otimes$ is the Kronecker product of matrices.

In ST models, the structured spatio-temporal effect $f_s^{(k)}(s_i,t)$ evolves in time with a first order autoregressive dynamics AR(1),

$$f_s^{(k)}(s_i,t) = a^{(k)} f_s^{(k)}(s_i,t-1) + w^{(k)}(s_i,t), \tag{4.4}$$

where the coefficient $|a^{(k)}| < 1$, $w^{(k)}(s_i,t) \perp f_s^{(k)}(s_i,1)$, and $w^{(k)}(s_i,t)$ follows a Gaussian distribution with mean zero and are supposed to be temporally independent,

$$\text{Cov}(w^{(k)}(s_i,t), w^{(k)}(s_j,t')) = \Sigma_w^{(k)}(i,j) = \begin{cases} 0 & \text{if } t \neq t'; \\ \sigma_w^{(k)2} \times \rho^{(k)}(.) & \text{if } t = t', \end{cases} \tag{4.5}$$

where $\sigma_w^{(k)2}$ is the marginal variance of $w^{(k)}(s_i,t)$, $\forall i,t$, $\rho^{(k)}(.)$ is a correlation function, $\boldsymbol{w}_s^{(k)}(.,t) = (w_s^{(k)}(s_1,t),\ldots,w_s^{(k)}(s_N,t))'$ is a GRF with mean zero and spatial covariance matrix $\boldsymbol{\Sigma}_w^{(k)}$ composed of $\Sigma_w^{(k)}(i,j)$, and $\boldsymbol{f}_s^{(k)}(.,1) = (f_s^{(k)}(s_1,1),\ldots,f_s^{(k)}(s_N,1))'$ is a GRF with mean zero and covariance matrix $\boldsymbol{\Sigma}_w^{(k)}/(1-a^{(k)2})$. Then,

$$\text{Cov}(f(s_i,t)^{(k)}, f(s_j,t')^{(k)}) = \Sigma_f^{(k)}(i,j) = \frac{a^{(k)|t-t'|}}{1-a^{(k)2}} \times \Sigma_w^{(k)}(i,j).$$

Therefore, $\boldsymbol{f}_s^{(k)} = (\boldsymbol{f}_s^{(k)}(.,1),\ldots,\boldsymbol{f}_s^{(k)}(.,T))'$ is a Gaussian field with mean zero and covariance matrix $\boldsymbol{\Sigma}_f^{(k)} = \boldsymbol{\Sigma}_T^{(k)} \otimes \boldsymbol{\Sigma}_w^{(k)}$, where $\boldsymbol{\Sigma}_T^{(k)}$ is the temporal correlation function of an AR(1) process.

Further, the S and ST models are subclassified depending on the inclusion (or not) of temporal effects in the linear predictors, giving rise to the following models: S1 and ST1 do not include any temporal effect, S2 and ST2 incorporate an autoregressive dynamic AR(1) temporal effect, S3 and ST3 incorporate a seasonal component, and S4 and ST4 incorporate a second-order random walk (rw2) temporal effect.

The structured temporal effect $f^{(k)}(t)$ following an AR(1) is defined as,

$$f^{(k)}(t) = b^{(k)} f^{(k)}(t-1) + \epsilon_{ar}^{(k)}(t),$$

where $|b^{(k)}| < 1$, $\epsilon_{ar}^{(k)}(t) \perp f^{(k)}(1)$ and $\epsilon_{ar}^{(k)}(t) \sim \mathrm{N}(0, 1/\tau_{ar}^{(k)})$. Moreover, $f^{(k)}(1) \sim \mathrm{N}(0, (\tau_{ar}^{(k)}(1 - b^{(k)2}))^{-1})$. Therefore, $\boldsymbol{f}_t^{(k)} = (f^{(k)}(1), \ldots, f^{(k)}(T))'$ is a GMRF with mean zero and precision matrix $\boldsymbol{Q}_{ar}^{(k)}$.

The seasonal component is included throughout the first Fourier harmonics, $\cos(2\pi t/P)$ and $\sin(2\pi t/P)$, where P is a fixed number representing the time periods required to complete a single cycle. This seasonal trend varies with time and it is identical for all sites. Both components are included as covariates in each linear predictor.

The temporally structured effect, modeled dynamically by using a random walk of order 2 (rw2), defined by

$$f^{(k)}(t) = 2f^{(k)}(t+1) - f^{(k)}(t+2) + \epsilon_{rw2}^{(k)}(t)$$

where $\epsilon_{rw2}^{(k)}(t) \sim \mathrm{N}(0, 1/\tau_{rw2}^{(k)})$, such that, for $t = 1, \ldots, n-2$, $f^{(k)}(t) - 2f^{(k)}(t+1) + f^{(k)}(t+2) \sim \mathrm{N}(0, 1/\tau_{rw2}^{(k)})$ are second order independent increments. Therefore, $\boldsymbol{f}_t^{(k)} = (f^{(k)}(1), \ldots, f^{(k)}(T))'$ is a GMRF with mean zero and precision matrix $\boldsymbol{Q}_{rw2}^{(k)}$. Hellton (2011) showed that the rw2 is an approximation to a smoothing spline, so it can be seen as a non-parametric approximation to the temporal trend giving great flexibility to this model.

In summary, our explicit formulation of spatio-temporal models (ST) requires samples collected on the same sites across years. Hence, the data need to be aggregated inside a grid, regular or irregular, which depending on the case. It can be beneficial due to the large number of sites and to relax the assumption of stationarity (Sherman, 2011), while allowing for finer resolution. Stroud et al. (2001) introduced a spatio-temporal model imposing minimum constraints on the format of the data, for instance non-stationarity and sites that move over time, but they only assume a normal observational variable. If such approach is used assuming another distribution like the mixture, we are assuming here, the computational time requirements would be far expensive. On the other hand, models S2 and S4 present an additive form in purely temporal and purely spatial random structures, while models ST2 and ST4 present temporal evolution at each site plus some purely temporal structure. Gelfand et al. (2003) and Gelfand et al. (2004) performed a comparison of some of these forms, called spatially varying coefficients with normal point-referenced spatio-temporal data, adopting a Bayesian inference framework. In particular, the AR(1) form was introduced in Gelfand et al. (2004) to capture short-term temporal changes. Gelfand,

Banerjee & Gamerman (2005) extended the approach of spatially varying coefficients to accommodate temporal dependence, achieving a class of dynamic models for normal point-referenced spatio-temporal data. They considered the AR(1) structure for the purely temporal component plus a spatio-temporal component. We follow this approach to propose models ST2 and ST4. Following time series analysis studies, possible seasonal variability in the data is considered through harmonic functions in models S3 and ST3. To add flexibility, the purely temporal component of models S4 and ST4 are rw2 Markovian structures, commonly used in age-period-cohort and disease mapping models (Knhorr-Held & Rainer, 2001; Rue & Held, 2005; Riebler & Held, 2010; Bauer et al., 2016).

To complete the definition of Equations (4.3) and (4.5), we assume the Matérn correlation function, $\rho^{(k)}(d) = \frac{1}{\Gamma(\nu)2^{\nu-1}}(\kappa^{(k)}d)^{\nu}K_{\nu}(\kappa^{(k)}d)$, where d is the Euclidean distance between two sites, ν is a shape parameter controlling the smoothness of the process, K_{ν} is the modified Bessel function of order ν, and $\kappa^{(k)}$ is a scale parameter associated with the range parameter. The effective range $r^{(k)} = \sqrt{8\nu}/\kappa^{(k)}$ is the distance d at which the correlation is approximately 0.1. Hence, $\Sigma^{(k)}$ and $\Sigma_w^{(k)}$, in Equations (4.3) and (4.5), are Matérn covariance functions (CF) with marginal variances $\sigma_f^{(k)2}$ and $\sigma_w^{(k)2}$, respectively.

3.1 SPDE for spatial models

Whittle (1954) asserted that a Gaussian field X with stationary Matérn Covariance is the unique solution to the next SPDE equation,

$$(\kappa^2 - \triangle)^{\alpha/2}(\tau X(s)) = G(s), s \in R^n, \tag{4.6}$$

where $\triangle$ is the Laplacian, $(\kappa^2 - \triangle)^{\alpha/2}$ is the fractional Laplacian operator, G is a Gaussian white noise, $\alpha = \nu + 1$ for two-dimensional domains, it controls the smoothness, while τ and κ jointly controls the marginal variance of X given by $\sigma^2 = \frac{\Gamma(\nu)}{\Gamma(\alpha)(4\pi)^{d/2}\kappa^{2\nu}\tau^2}$. The weak formulation of Equation (4.6) is

$$[< \phi_i, (k^2 - \triangle)^{\alpha/2}\widetilde{X} >_{\Omega}] \stackrel{D}{=} [< \phi_i, G >_{\Omega}], \tag{4.7}$$

where $a \stackrel{D}{=} b$ denotes equality in distribution, and ϕ_i is any arbitrary well-behaved test function. Thus X can be approximated by $\widetilde{X}$, the solution of Equation (4.7), which guarantees that the left side of Equation (4.6) has the same mean and covariance function of G.

Lindgren et al. (2011) proposed to solve Equation (4.7) throughout the finite element method (FEM). The general procedure for two-dimensional domains involves:

(i) generate a Delaunay triangulation over the spatial domain, composed of n_v nodes.

(ii) Construct piecewise linear basis functions (ψ_j) over the triangulation. Each $\psi_j(.)$ defined on a triangle Tr is uniquely determined by its values at the three vertices. Thus, if j and s are vertices, then $\psi_j(s) = 1$ if $j = s$ and $\psi_j(s) = 0$ if $j \neq s$. For each site s inside the triangle Tr, $\psi_j(s)$ is determined by a linear interpolation, then $0 < \psi_j(s) < 1$. While for each site s outside Tr, $\psi_j(s) = 0$. Then, the approximated solution of $x(s)$ is a linear combination of these basis functions, $\widetilde{X}(s) = \sum_{j=1}^{n_v} \psi_j(s) g_j$, where g_js are Gaussian weights with mean zero.

(iii) Choose the test function. In particular, for $\alpha = 2$ they assumed test functions equal to the basis functions themselves ($\phi_i = \psi_i$) which leads to Galerkin solutions.

(iv) Solve the system of equations. The solution of Equation (4.7) is reduced to find the precision of the Gaussian weights g_j. Let $\boldsymbol{g} = (g_1, \ldots, g_{n_v})'$, then $\boldsymbol{g}$ is a GRF with mean zero and precision matrix $\boldsymbol{Q}_g$ which is computed with the help of Green's first identity and Galerkin solutions (for $\alpha = 2$), hence, $\boldsymbol{Q}_g = \tau^2(\boldsymbol{K}_{\kappa^2}\boldsymbol{C}^{-1}\boldsymbol{K}_{\kappa^2})$. The matrices $\boldsymbol{C}$ and $\boldsymbol{K}_{\kappa^2}$ depend on the geometry of each triangle. For the basis function chosen, $\boldsymbol{C}, \boldsymbol{G}$, and $\boldsymbol{K}$ are sparse matrices, nevertheless $\boldsymbol{C}^{-1}$ is dense, then $\boldsymbol{Q}_g$ is also dense. To keep computational convenience, $\boldsymbol{C}$ is replaced with a diagonal matrix $\widetilde{\boldsymbol{C}}$ where $\widetilde{C}_{ii} =< \psi_i, 1 >$. As result, $\boldsymbol{g}$ is a GMRF with zero mean and the sparse precision matrix $\boldsymbol{Q}_g = \tau^2(\boldsymbol{K}_{\kappa^2}\widetilde{\boldsymbol{C}}^{-1}\boldsymbol{K}_{\kappa^2})$. For further details see Lindgren et al. (2011) and Lindgren & Rue (2015).

3.2 SPDE for our models

This section describes the approximation of GRFs $\boldsymbol{f}_s^{(k)}$ to GMRFs for S and ST models, that is, when $f(s_i, t)^{(k)}$ is a spatial or a spatio-temporal effect. The approximation is performed for all the subclassified models, depending on $f(s_i, t)^{(k)}$. Assuming a two-dimensional spatial domain, let fix $\nu = 1$, which implies $\alpha = 2$.

For S models, $\boldsymbol{f}_s^{(k)}(., t)$ is a GRF with Matérn CF, $\boldsymbol{\Sigma}^{(k)}$, with marginal variance $\sigma_f^{(k)2} = \sigma^{(k)2}$. Then using the SPDE approach, $\sigma^{(k)2} = \frac{1}{(4\pi)\kappa^{(k)2}\tau^{(k)2}}$ and let $f_s^{(k)}(s_i, t)$ be approximated by

$$\widetilde{f}_s^{(k)}(s_i, t) = \sum_{j=1}^{n_v} \psi_j(s_i) g_j^{(k)}, \tag{4.8}$$

where $\boldsymbol{g}^{(k)} = (g_1^{(k)}, \ldots, g_{n_v}^{(k)})'$ is a GMRF with mean zero and sparse precision matrix $\widetilde{\boldsymbol{Q}}_g^{(k)} = \tau^{2(k)}(\boldsymbol{K}_{\kappa^{(k)2}}[\widetilde{\boldsymbol{C}}^{(k)}]^{-1}\boldsymbol{K}_{\kappa^{(k)2}})$. So, $\widetilde{\boldsymbol{f}}_s^{(k)}(.,t) = (\widetilde{f}_s^{(k)}(s_1,t), \ldots, \widetilde{f}_s^{(k)}(s_N,t))'$ is a GMRF with mean zero and precision matrix $\widetilde{\boldsymbol{Q}}_s^{(k)}$, $\forall t$, which depends on $\widetilde{\boldsymbol{Q}}_g^{(k)}(\theta)$. Thus, $\widetilde{\boldsymbol{f}}_s^{(k)} = (\widetilde{\boldsymbol{f}}_s^{(k)}(.,1), \ldots, \widetilde{\boldsymbol{f}}_s^{(k)}(.,T))'$ is a GMRF with mean zero and precision $\widetilde{\boldsymbol{Q}}^{(k)} = \boldsymbol{I}_T \otimes \widetilde{\boldsymbol{Q}}_s^{(k)}$.

For ST models, $\boldsymbol{w}_s^{(k)}(.,t)$ is a GRF with Matérn CF, $\boldsymbol{\Sigma}_w^{(k)}$, with marginal variance $\sigma_w^{(k)2} = \sigma^{(k)2}$. Using the SPDE approach, $\sigma^{(k)2} = \frac{1}{(4\pi)\kappa^{(k)2}\tau^{(k)2}}$ and $w^{(k)}(s_i,t)$ is approximated by

$$\widetilde{w}^{(k)}(s_i,t) = \sum_{j=1}^{n_v} \psi_j(s_i) g_j^{(k)}, \tag{4.9}$$

where $\boldsymbol{g}^{(k)} = (g_1^{(k)}, \ldots, g_{n_v}^{(k)})'$ is a GMRF with mean zero and sparse precision matrix, $\widetilde{\boldsymbol{Q}}_g^{(k)} = \tau^{2(k)}(\boldsymbol{K}_{\kappa^{(k)2}}[\widetilde{\boldsymbol{C}}^{(k)}]^{-1}\boldsymbol{K}_{\kappa^{(k)2}})$. So, $\widetilde{\boldsymbol{w}}_s(t)^{(k)} = (\widetilde{w}^{(k)}(s_1,t), \ldots, \widetilde{w}^{(k)}(s_N,t))'$ is a GMRF with mean zero and precision $\widetilde{\boldsymbol{Q}}_s^{(k)}$, $\forall t$, which depends on $\widetilde{\boldsymbol{Q}}_g^{(k)}$. From Equation (4.4),

$$\widetilde{f}_s^{(k)}(s_i,t) = \widetilde{a}^{(k)}\widetilde{f}_s^{(k)}(s_i,t-1) + \sum_{j=1}^{n_v} \psi_j(s_i) g_j^{(k)}, \tag{4.10}$$

where $|\widetilde{a}^{(k)}| < 1$, $\widetilde{w}^{(k)}(s_i,t) \perp \widetilde{f}_s^{(k)}(s_i,1)$. Thus, $\widetilde{\boldsymbol{f}}_s^{(k)}(t) = (\widetilde{f}_s^{(k)}(s_1,t), \ldots, \widetilde{f}_s^{(k)}(s_N,t))'$, such that $\widetilde{\boldsymbol{f}}_s^{(k)}(1)$ is a GMRF with mean zero and precision $[\widetilde{\boldsymbol{Q}}_s{}^{(k)}]^{-1}/(1-a^{(k)2}))$. Then, $\widetilde{\boldsymbol{f}}_s^{(k)} = (\widetilde{\boldsymbol{f}}_s^{(k)}(1), \ldots, \widetilde{\boldsymbol{f}}_s^{(k)}(T))'$, is a GMRF with mean zero and precision matrix $\widetilde{\boldsymbol{Q}}^{(k)} = \boldsymbol{Q}_T^{(k)} \otimes \widetilde{\boldsymbol{Q}}_s^{(k)}$, where $\boldsymbol{Q}_T^{(k)}$ is the inverse of $\boldsymbol{\Sigma}_T^{(k)}$.

3.3 Bayesian Inference and prediction

From Equations (4.8) and (4.10), the linear predictors of Equation (4.3) can be rewritten as

$$\begin{aligned} \text{logit}(p(s_i,t)) &= \eta(s_i,t)^{(1)} = \text{Z}^{(1)}\boldsymbol{\beta}^{(1)} + f^{(1)}(t) + \widetilde{f}_s^{(1)}(s_i,t), \\ \log(\mu(s_i,t)) &= \eta(s_i,t)^{(2)} = \text{Z}^{(2)}\boldsymbol{\beta}^{(2)} + f^{(2)}(t) + \widetilde{f}_s^{(2)}(s_i,t). \end{aligned} \tag{4.11}$$

From the previous section, for $\boldsymbol{f}_t^{(k)}$ is assigned a Gaussian prior with zero mean and sparse precision matrix $\boldsymbol{Q}_{ar}^{(k)}$ or $\boldsymbol{Q}_{rw2}^{(k)}$, and for $\widetilde{\boldsymbol{f}}_s^{(k)}$ is assigned a Gaussian prior with zero mean and sparse precision matrix $\widetilde{\boldsymbol{Q}}^{(k)}$ depending on the spatial or spatio-temporal effect. For each fixed effect $\boldsymbol{\beta}^{(k)}$ is assigned independent vague Gaussian prior. Then, the latent field $\boldsymbol{x}$ can be composed of $\boldsymbol{x} = ((\boldsymbol{\beta}^{(1)})', (\boldsymbol{f}_t^{(1)})', (\widetilde{\boldsymbol{f}}_s^{(1)})', (\boldsymbol{\beta}^{(2)})', (\boldsymbol{f}_t^{(2)})', (\widetilde{\boldsymbol{f}}_s^{(2)})')'$. To complete the specification of the Bayesian hierarchical models we

assigned prior distributions to the hyperparameters

$$\boldsymbol{\theta} = (\phi, b^{(1)}, \tau_{ar}^{(1)}, \tau^{(1)}, \kappa^{(1)}, \widetilde{a}^{(1)}, \tau_{rw2}^{(1)}, b^{(2)}, \tau_{ar}^{(2)}, \tau^{(2)}, \kappa^{(2)}, \widetilde{a}^{(2)}, \tau_{rw2}^{(2)})'.$$

Hence, assuming that $\tau_0^{(k)}$ and $\kappa_0^{(k)}$ depend on $\sigma_0^{(k)}$ and $r_0^{(k)}$ which are the base-line deviation and range values, then, $\log(\tau^{(k)}) = \log(\tau_0^{(k)}) - \theta_1^{(k)} + \nu\theta_2^{(k)}$ and $\log(\kappa^{(k)}) = \log(\kappa_0^{(k)}) - \theta_2^{(k)}$. Let $\theta_{b1} = \log\left(\frac{1+b^{(1)}}{1-b^{(1)}}\right)$, $\theta_{b2} = \log\left(\frac{1+b^{(2)}}{1-b^{(2)}}\right)$, $\theta_{a1} = \log\left(\frac{1+a^{(1)}}{1-a^{(1)}}\right)$ and $\theta_{a2} = \log\left(\frac{1+a^{(2)}}{1-a^{(2)}}\right)$. For $\log(\phi)$, $\log(\tau_{ar}^{(k)}(1-b^{(k)}))$ and $\log(\tau_{rw2}^{(k)})$ are assigned independent log-gamma prior distributions. For θ_{ak} and θ_{bk} are assigned independent normal prior distributions. While for $(\theta_1^{(k)}, \theta_2^{(k)})$ is assigned a joint normal prior distribution suggested by Lindgren et al. (2011).

Therefore, $\pi(y(s_i,t)|p(s_i,t),\mu(s_i,t),\phi) = \pi(y(s_i,t)|\boldsymbol{x},\boldsymbol{\theta})$, $Y(s_i,t)$'s are conditionally independent given $\boldsymbol{x}$ and $\boldsymbol{\theta}$, $\forall i \in D_s, \forall t \in D_t$, and from Equation (4.1),

$$L(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta}) = \exp\left[\sum_{t=1}^{T}\sum_{i=1}^{N}\log\left\{\pi(y(s_i,t)|\boldsymbol{x},\boldsymbol{\theta})\right\}\right].$$

The latent field $\boldsymbol{x}$ given the hyperparameters $\boldsymbol{\theta}$ is a GMRF with block diagonal sparse precision matrix $\boldsymbol{Q}(\boldsymbol{\theta})$. Therefore, all models fit into the latent Gaussian model framework and the joint posterior distribution can be computed as follows,

$$\pi(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y}) \propto \pi(\boldsymbol{\theta})\exp\left[-\frac{1}{2}x'\boldsymbol{Q}(\boldsymbol{\theta})x + \sum_{t=1}^{T}\sum_{i=1}^{N}\log\{\pi(y(s_i,t)|\boldsymbol{x},\boldsymbol{\theta})\}\right].$$

The marginals of $\boldsymbol{x}$ and $\boldsymbol{\theta}$ can be obtained throughout the joint posterior distribution, however, due to their high dimension, either integration or sampling approaches as MCMC would be computationally expensive. INLA (Rue et al., 2009) overcomes this issue by using a variety of deterministic approximations, primarily, the Laplace approximation. It is restricted to Latent Gaussian models where the latent field $\boldsymbol{x}$ is a GMRF. Thus, given that our models fulfill such requirements, INLA computes numerical approximations of the marginals densities as follows,

$$\widetilde{\pi}(x_j|\boldsymbol{y}) = \int \widetilde{\pi}(x_j|\boldsymbol{\theta},\boldsymbol{y})\widetilde{\pi}(\boldsymbol{\theta}|y)d\boldsymbol{\theta} \text{ and } \widetilde{\pi}(\theta_r|\boldsymbol{y}) = \int \widetilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})d\theta_{-r}, \tag{4.12}$$

where $j = 1,\ldots,dim(\boldsymbol{x})$, $r = 1,\ldots,dim(\boldsymbol{\theta})$, $\widetilde{\pi}(x_j|\boldsymbol{\theta},\boldsymbol{y})$ is an approximation for the marginal of the latent field and $\widetilde{\pi}(\boldsymbol{\theta},\boldsymbol{y})$ is an approximation for the marginal joint posterior of hyperparameters. In summary, first INLA computes a Gaussian approximation of the full conditional of the latent field, $\pi_G(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y})$, and the mode of this full conditional throughout an iterative method. Using the Laplace approximation it computes $\widetilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ and its mode to find a regular set of hyperparameters $\boldsymbol{\theta}^\star$ with the

higher mass of probability. Second, it computes $\widetilde{\pi}(x_j|\boldsymbol{\theta}, \boldsymbol{y})$ using another approximation, depending on the desired accuracy. Finally, it constructs an interpolation of $\boldsymbol{\theta}^\star$s to integrate out $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_{-r}$ in Equation (4.12).

The predictions of anchovy biomass at new site s_p inside or outside the range of data are defined as the mean of the posterior predictive, that is, $E[y_p|\boldsymbol{y}]$. In particular, (i) $E[y_p|\boldsymbol{y}] = E[y(s_p, t)|\boldsymbol{y}]$ is a PWD ($s_p \in D_s; t \in \{1, \ldots, T\}$) and (ii) $E[y_p|\boldsymbol{y}] = E[y(s_p, T+1)|\boldsymbol{y}]$ is a POD ($s_p \in D_s$). To compute $E[y_p|\boldsymbol{y}]$, we need to compute the predictive density $\pi(y_p|\boldsymbol{y}) = \int\int \pi(y_p|\boldsymbol{x}_p, \theta)\pi(\boldsymbol{x}_p, \boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{x}_p d\boldsymbol{\theta}$, where $\boldsymbol{x}_p$ is composed of $\boldsymbol{x}$ at sites $s_1, \ldots s_N$ and s_p, in times $t \in \{1, \ldots, T\}$ and $T+1$ depending on the case (i) or (ii), respectively. INLA is capable of providing i.i.d. samples from the joint posterior distribution $\pi(\boldsymbol{x}_p, \boldsymbol{\theta}_p|\boldsymbol{y})$. Therefore, instead of getting the exact analytic solution of $\pi(y_p|\boldsymbol{y})$, we simulate its values according to the next steps: (i) draw M i.i.d. samples $(\boldsymbol{x}_p^{(m)}, \boldsymbol{\theta}^{(m)})$ from $\pi(\boldsymbol{x}_p, \boldsymbol{\theta}|\boldsymbol{y})$, for $m = 1, \ldots, M$. (ii) simulate $y_p^{(m)}$ from $\pi(y_p|\boldsymbol{x}_p^{(m)}, \boldsymbol{\theta}^{(m)})$, where $y_p^{(m)}$ are actually draws from $\pi(y_p|\boldsymbol{y})$. First, generate $y_p^{\star(m)} \sim \text{Bernoulli}(p^{(m)}(s_p, t))$, if $y_p^{\star(m)} = 1$ then $y_p^{(m)} = 0$, otherwise generate $y_p^{(m)} \sim \text{gamma}(\phi_p^{(m)}, \phi^{(m)}/\mu^{(m)}(s_p, t))$. (iii) Finally, use the samples $y_p^{(1)}, \ldots, y_p^{(M)}$ for computing $E[y_p|\boldsymbol{y}] = \frac{1}{M}\sum_m^M y_p^{(m)}$. We also use these samples to calculate the standard deviation of the posterior predictive, $\text{sd}[y_p|\boldsymbol{y}] = \sqrt{\frac{1}{M-1}\sum_m^M (y_p^{(m)} - E[y_p|\boldsymbol{y}])^2}$. It is worth mentioning that, in step (i), we can calculate any $f_s(s_p, t)$ of $\boldsymbol{x}_p$, using the m-th sample $f_s(s_v, t)$ from $\pi(\boldsymbol{x_v}, \boldsymbol{\theta}|\boldsymbol{y})$, where s_v are the n_v nodes of the mesh. Then $f_s(s_p, t)$ can be calculated from Equation (4.6) or Equations (4.9 and 4.10) depending on the model S or ST, respectively. Therefore, the triangulation required for the SPDE approximation efficiently reduces computational time requirements to make predictions at high resolution, when $n_v \ll n_p$, where n_p is the number of observations to be predicted.

3.4 Model Assessment

The estimate accuracy rate (EAR), the logarithm of the pseudo marginal likelihood (LPML), the Watanabe-Akaike (or "widely applicable") information criterion (WAIC), and the root of mean squared estimation error (RMSEE) are considered to measure the performance of each model and to compare their fits.

The EAR is the percentage of observations estimated as presence of anchovy when the acnchovies are actually present and observations estimated as absence of anchovy when the anchovies are actually absent.

The WAIC was introduced by Watanabe (2010), it is based on the posterior predictive density, which from a Bayesian perspective is its main advantage over other similar measures. Gelman, Hwang & Vehtari (2014) stated that the WAIC is particularly

helpful for models with hierarchical and mixture structures and, proposed a slightly change of the Watanabe's WAIC original version,

$$\text{WAIC}=-2\times\sum_{t=1}^{T}\sum_{i=1}^{N}\Big[\log\Big\{\tfrac{1}{M}\sum_{m=1}^{M}\pi(y(s_i,t)|\boldsymbol{x}^{(m)},\boldsymbol{\theta}^{(m)})\Big\}-V_{m=1}^{M}\log\{\pi(y(s_i,t)|\boldsymbol{x}^{(m)},\boldsymbol{\theta}^{(m)})\}\Big],$$

where $(\boldsymbol{x}^{(m)},\boldsymbol{\theta}^{(m)})$ are samples from $\pi(x(s_i,t),\boldsymbol{\theta}|\boldsymbol{y})$ and $V_{m=1}^{M}(.)$ is the sample variance. The lower the value of WAIC, the better the model.

The LPML summarizes the goodness of fit of each observation in site s_i and time t throug the conditional predictive ordinate (CPO_{it}) introduced by Geisser & Eddy (1979). Dey, Chen & Chang (1997) use a Monte Carlo simulation to approximate the CPO_{it} and the LPML as follows:

$$\widehat{\text{CPO}}_{it}=\frac{1}{M}\sum_{m}^{M}\left[\frac{1}{\pi(y(s_i,t)|\boldsymbol{x}^{(m)},\boldsymbol{\theta}^{(m)})}\right]^{-1}\text{ and LPML}=\sum_{t=1}^{T}\sum_{i=1}^{N}\log(\widehat{\text{CPO}}_{it}),$$

such that higher values of LPML indicate better model fit.

The RMSEE evaluates the closeness between the estimation of anchovy abundance by the model and the observed anchovy abundance. This quantity is defined by:

$$\text{RMSEE}=\sqrt{\frac{1}{N\times T}\sum_{t=1}^{T}\sum_{i=1}^{N}(y(s_i,t)-E[y(s_i,t)|\boldsymbol{x},\boldsymbol{\theta}])^2}.$$

In order to evaluate the predictive performance of the models we have used a training data. For POD we use all samples from year $T+1$. This set is defined as $\boldsymbol{y}_{val}$, then we evaluated the predictive accuracy and the root of mean squared prediction error (RMSPE). The root of the mean of the squared difference between the observed value $y_p=y(s_p,T+1)\in y_{val}$ and the POD, is computed as follows,

$$\text{RMSPE}=\sqrt{\frac{1}{n_p}\sum_{y_p\in y_{val}}(y_p-E[y_p|\boldsymbol{y}])^2}.$$

4 Application

The study was carried out with the available data of anchovy abundance from years 1999 to 2007. Even though the survey design is similar across years, the trajectory is not exactly the same, the set of sites differs for each year (Figure 1). Our explicit formulation of the spatio-temporal models requires samples collected on the same sites across years. Hence, the coast of Peru was subdivided using a Delaunay triangulation (mesh). To provide adequate estimation in our modeling, the maximum side length of each triangle must be smaller than the effective range of

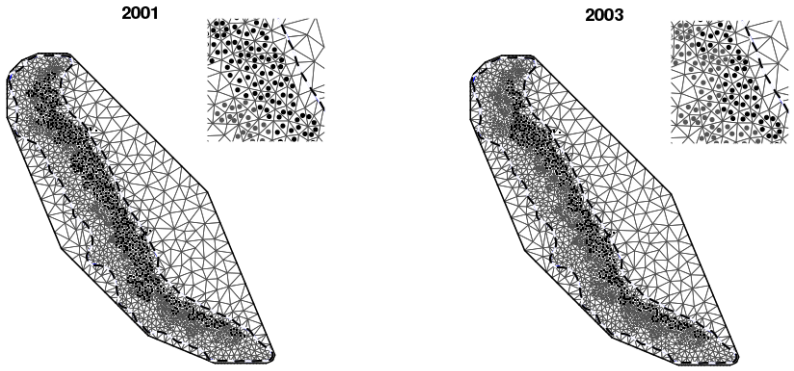

Figure 2: Triangulation off the coast of Peru composed by 1,147 nodes. The dots indicate the centroid of the $N = 785$ triangles with at least one sample of absence (gray) or presence (black) of anchovy for the years 2001 (left panel) and 2003 (right panel). The region of main interest is inside the inner boundary (dashed line). The upper panels show a zoom of each plot.

the data. So using the results presented by Quiroz et al. (2015) on the same data set in 2005, we set the maximum side length of each triangle as 0.2 degrees. This result also agrees with other studies about the range for these data from different years and seasons. Of course we could use a smaller side length, but a higher resolution increases the computational cost.

The mesh has $n_v = 1{,}147$ nodes (vertices of triangles) and $n_t = 2{,}168$ triangles (Figure 2). The location of each triangle is determined by its centroid. Only $N = 785$ triangles were selected, those with at least one sample of presence or absence of anchovy for all years. The dots in Figure 2 are the centroids of the N triangles for specific years with information about anchovy absence (gray), when all the samples inside the triangle are zero, or presence (black), when at least one sample inside the triangle is higher than zero. For each year, if a triangle exhibits absence of anchovy, the abundance of anchovy for the specific triangle and year will be considered zero. Otherwise, if a triangle exhibits presence of anchovy, it is computed the mean of anchovy abundance of all samples inside the triangle, which is considered the abundance of anchovy in the centroid of the specific triangle and year. We could compute the median or sum of anchovy abundance instead of the mean, but we do not have too many samples inside the triangle to lead with extreme values, and the sum does not represent the anchovy of abundance in the centroid of the triangle.

The $y(s_i, t)$ is the observed abundance of anchovy (≥ 0) in the triangle with centroid site $s_i \in D_s$ and time $t \in D_t$, where $i = 1, ..., 785$ and $t = 1, ..., 8$. The general model

Table 1: Summary of models according to the definition of the temporal structure $f^{(k)}(t)$ and spatial or spatio-temporal structures $\widetilde{f}_s^{(k)}(.)$

Model	S1	S2	S3	S4	ST1	ST2	ST3	ST4
$f^{(k)}(t)$	none	AR(1)	none (seasonal)	rw2	none	AR(1)	none (seasonal)	rw2
$\widetilde{f}_s^{(k)}(.)$	spatial	spatial	spatial	spatial	spatio-temporal	spatio-temporal	spatio-temporal	spatio-temporal

is,

$$\begin{aligned}
\pi(y(s_i,t)|\boldsymbol{x},\boldsymbol{\theta}) &= p(s_i,t)\,\delta_0 + (1-p(s_i,t))\,h(y(s_i,t)|\mu(s_i,t),\phi)\,I_{[y(s_i,t)>0]},\\
\text{logit}(p(s_i,t)) &= \eta(s_i,t)^{(1)} = \mathbf{Z}^{(1)}\boldsymbol{\beta}^{(1)} + f^{(1)}(t) + \widetilde{f}_s^{(1)}(s_i,t),\\
\log(\mu(s_i,t)) &= \eta(s_i,t)^{(2)} = \mathbf{Z}^{(2)}\boldsymbol{\beta}^{(2)} + f^{(2)}(t) + \widetilde{f}_s^{(2)}(s_i,t).
\end{aligned}$$

To evaluate the predictive performance of the models, we fit the models from $T = 1 to T = 7$, and use all the data from the last year ($T = 8$) to perform predictive assessment. Table 1 summarizes the temporal structure and spatial or spatio-temporal structure of the specific models fitted. To complete the definition of these models, the covariate matrix $\mathbf{Z}^{(2)}$ comprises the DC, depth and SST, while the covariate matrix $\mathbf{Z}^{(1)}$ includes these covariates, together with the latitude and the squared of latitude. In order to identify possible important periods or seasons we used a periodogram for the mean of anchovy abundance (on the observed original data) from 1999 to 2006 (Figure S1 in Supplementary Material). The peak corresponds to a period of 3 years, hence S3 and ST3 models set $P = 3$ and their covariate matrices $\mathbf{Z}^{(1)}$ and $\mathbf{Z}^{(2)}$ also comprise $\sin(2\pi t/3)$ and $\cos(2\pi t/3)$ as covariates.

4.1 Results and Analysis

Table 2 reports posterior parameter estimates of the covariates for each model. The positive or negative contribution of the covariates is not strongly affected by the model specification. In fact, each covariate contributes to explain the global mean trend of the probability of anchovy absence and presence. And with regard to the contribution of each covariate on the positive abundance of anchovy, the depth is significant in all models, while the SST and the DC are significant in some specific models. Moreover, there is evidence of a seasonality pattern, the absence/presence and abundance of anchovy is similar every 3 years.

The estimation performance and posterior parameter estimates of the hyperparameters for each model are reported in Tables 3 and 4, respectively. Models S2 and S4 have better estimation performance than model S1 (higher EAR, lower LPML

Table 2: Summary statistics: mean posterior, (95% credible interval [CI]) for the hyperparameters for each model. ($\star$) : 95% CI includes the zero value.

Model	S1	S2	S3	S4	ST1	ST2	ST3	ST4
Probability of anchovy absence								
Intercept	($\star$)1.54	3.82	($\star$)1.95	4.52	($\star$)8.95	($\star$)8.51	25.32	37.91
	$(-2.36, 5.61)$	$(0.49, 7.16)$	$(-2.21, 6.35)$	$(-0.07, 9.26)$	$(-31.16, 49.09)$	$(-14.24, 32.56)$	$(3.15, 47.62)$	$(4.33, 71.75)$
DC	0.04	0.05	0.04	0.05	0.17	0.20	0.27	0.51
	$(0.03, 0.05)$	$(0.04, 0.06)$	$(0.03, 0.05)$	$(0.04, 0.06)$	$(0.13, 0.21)$	$(0.15, 0.25)$	$(0.22, 0.32)$	$(0.45, 0.58)$
Lat	2.18	2.18	1.88	2.26	10.94	9.93	9.35	27.78
	$(1.62, 2.7367)$	$1.62, 2.74)$	$(1.10, 2.71)$	$(1.41, 3.14)$	$(3.03, 18.85)$	$(5.83, 14.66)$	$(4.79, 13.94)$	$(22.55, 33.12)$
Lat^2	0.10	0.10	0.09	0.10	0.50	0.4464	0.4295	1.23
	$(0.07, 0.12)$	$(0.07, 0.12)$	$(0.05, 0.12)$	$(0.06, 0.14)$	$(0.14, 0.87)$	$(0.26, 0.66)$	$(0.22, 0.64)$	$(0.99, 1.47)$
Depth	-0.47	-0.55	-0.35	-0.50	-2.26	-2.5673	-1.46	-7.27
	$(-0.66, -0.29)$	$(-0.80, -0.29)$	$(-0.53, -0.17)$	$(-0.71, -0.29)$	$(-3.1458, -1.39)$	$(-3.61, -1.67)$	$(-2.57, -0.35)$	$(-8.77, -5.81)$
SST	0.07	0.07	0.12	($\star$)0.07	1.21	0.89	-0.24	2.04
	$(0.00, 0.14)$	$(0.00, 0.14)$	$(0.07, 0.18)$	$(-0.01, 0.14)$	$(0.68, 1.75)$	$(0.02, 0.12)$	$(-0.33, -0.14)$	$(1.09, 2.99)$
$\sin(2\pi t/3)$			0.16				-3.58	
			$(0.05, 0.28)$				$(-3.67, -2.71)$	
$\cos(2\pi t/3)$			0.16				-2.89	
			$(0.05, 0.26)$				$(-3.67, -2.12)$	
Positive anchovy abundance								
Intercept	5.78	6.32	6.29	6.26	4.18	4.86	5.71	4.33
	$(5.06, 6.49)$	$(5.49, 7.15)$	$(5.55, 7.03)$	$(5.48, 7.04)$	$(3.38, 4.98)$	$(3.80, 5.93)$	$(4.54, 6.88)$	$(3.4941, 5.17)$
DC	($\star$)0.00	($\star$)0.00	($\star$)0.00	($\star$)0.00	($\star$)0.00	($\star$)0.00	($\star$)0.00	0.00
	$(0.00, 0.01)$	$(-0.00, 0.01)$	$(-0.00, 0.01)$	$(0.00, 0.01)$	$(0.00, 0.00)$	$(0.00, 0.00)$	$(0.00, 0.01)$	$(0.00, 0.00)$
Depth	0.26	0.22	0.26	0.24	0.42	0.3724	0.39	0.43
	$(0.15, 0.37)$	$(0.11, 0.32)$	$(0.14, 0.37)$	$(0.13, 0.35)$	$(0.36, 0.49)$	$(0.30, 0.45)$	$(0.31, 0.48)$	$(0.36, 0.49)$
SST	0.05	($\star$)0.02	($\star$)0.02	($\star$)0.02	0.11	0.07	($\star$)0.03	0.10
	$(0.02, 0.08)$	$(-0.02, 0.05)$	$(-0.01, 0.05)$	$(-0.01, 0.05)$	$(0.07, 0.15)$	$(0.02, 0.12)$	$(-0.03, 0.08)$	$(0.06, 0.15)$
$\sin(2\pi t/3)$			-0.20				($\star$) -0.23	
			$(-0.25, -0.14)$				$(-0.38, 0.89)$	
$\cos(2\pi t/3)$			0.12				($\star$)0.25	
			$(0.07, 0.18)$				$(-0.08, 0.03)$	

and RMSEE), being evidence of short-term-temporal variability. This result is confirmed by the 95% credible intervals of $f^{(k)}(t)$ not including zero for specific years (Figure S2). Further, as expected, $(1/\tau_{ar}^{(k)})$ is lower in model ST2 than model S2, because a great temporal variability was captured by the spatio-temporal process in model ST2. This last result is also true for $(1/\tau_{rw2}^{(k)})$ in models S4 and ST4. We conclude that ST models have better goodness of fit than S models. In fact, the marginal variance $\sigma^{(k)2}$ is higher for ST models, being able to capture more spatial variability of the distribution (absence/presence) and abundance of anchovy. The mean posterior estimation of ϕ provides a method for monitoring the model fit. Hence, the remaining variability of the positive anchovy abundance (μ^2/ϕ), is lower for ST models. For instance, these variabilities are approximately proportional to $1.24\mu^2$ and $0.18\mu^2$ for S1 and ST1 models, respectively. In addition, the EAR, WAIC, LPML and RMSEE values confirm that ST1 and ST4 models are the best ones to estimate the abundance of anchovy.

To evaluate the performance of POD of each model, we set $M = 1000$ samples. The predictive measure results are also reported in Table 3, the RMSPE favors model S2 from models S and ST1 from ST models. We compute the posterior predictive densities of S2 and ST1 for some randomly selected sites in 2007 (Figure S2 in Supplementary Material). Both models overestimate the anchovy abundance, being

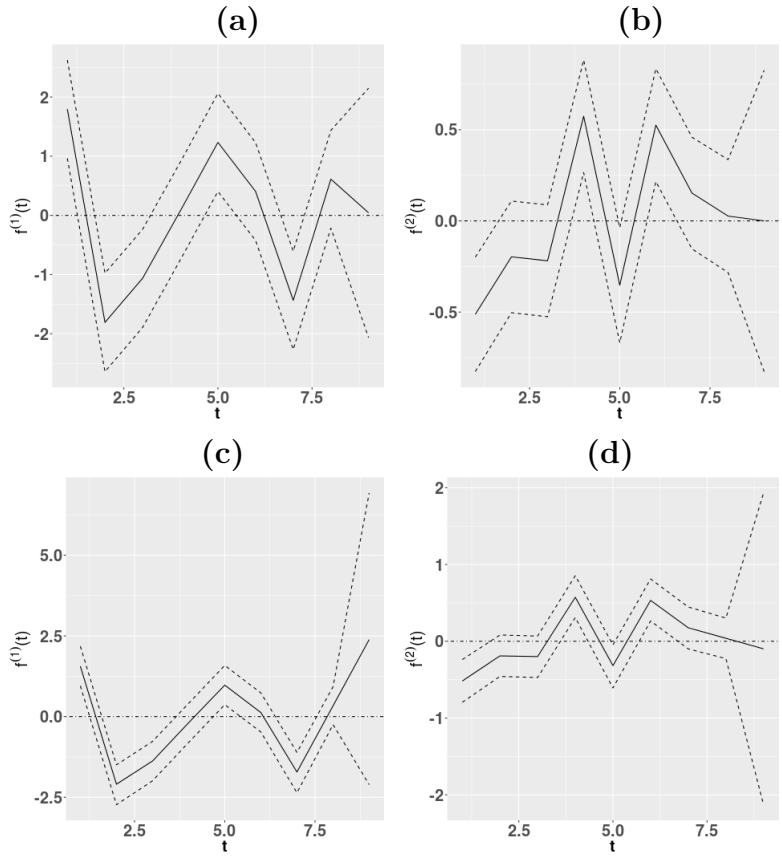

Figure 3: Posterior mean (solid line), upper and lower credible intervals (dashed lines) of the purely temporal structures $f^{(k)}(t)$ corresponding to Model S2 (a, b) and model S4 (c, d). (a) and (c): $f^{(1)}(t)$ related to the Probability of anchovy absence/presence, (b) and (d): $f^{(2)}(t)$ related to the Positive anchovy abundance.

Table 3: The selection criteria for the models proposed. The WAIC, LPML and RMSPE were computed using M = 1000 samples. The time is measured in minutes (min), hours (h) and days (d).

Model	**S1**	**S2**	**S3**	**S4**	**ST1**	**ST2**	**ST3**	**ST4**
EAR	85.3	88.1	85.0	88.1	99.6	99.6	99.8	99.7
WAIC	134541.8	144472.8	132382	157928.1	57851.78	1300572	1223885	59135.91
LPML	−223454.6	−201825.6	−224057.3	−183475.9	−33240	−573887.6	−560865.3	−33999.51
RMSEE	1387.8	1379.3	1329.3	1325.9	429.0	459.3	600.2	450.0
RMSPE	1071.8	1060.8	1098.7	1068.8	1090.5	2356.3	3962.3	1188.7
Time	49 min	3 h	50 min	2 h	3 d	4 d 21 h	3 d 3 h	4 d 13 h

Table 4: Summary statistics: mean posterior, (95% credible interval [CI]) for the hyperparameters for each model.

Model	**S1**	**S2**	**S3**	**S4**	**ST1**	**ST2**	**ST3**	**ST4**
Probability of anchovy absence								
$\sigma^{2(1)}$	1.16	1.31	1.22	1.88	343.85	112.05	502.23	4.38
	(0.70, 1.98)	(1.00, 1.85)	(0.68, 2.10)	(1.20, 2.98)	(341.59, 348.06)	(102.80, 119.73)	(490.97, 511.04)	(2.14, 8.20)
$r^{(1)}$	1.19	0.77	1.23	1.05	2.24	1.17372	0.92	1.83
	(0.76, 1.89)	(0.62, 1.04)	(0.73, 1.98)	(0.70, 1.64)	(2.23, 2.25)	(1.16, 1.19)	(0.91, 0.93)	(1.32, 2.46)
$\widetilde{a}^{(1)}$					0.78	0.51	0.24	0.30
					(0.78, 0.78)	(0.50, 0.52)	(0.24, 0.24)	(0.29, 0.31)
$b^{(1)}$		0.13				0.14		
		(−0.36, 0.63)				(0.12, 0.17)		
$\tau_{ar}^{(1)}$		0.75				13788.87		
		(0.20, 1.63)				(12422.72, 16165.80)		
$\tau_{rw2}^{(1)}$				0.20				13.46
				(0.06, 0.45)				(13.34, 13.56)
Positive anchovy abundance								
ϕ	0.81	0.79	0.80	0.86	5.47	5.85	2.87	5.16
	(0.77, 0.82)	(0.77, 0.81)	(0.77, 0.82)	(0.83, 0.90)	(5.45, 5.52)	(5.67, 6.01)	(2.84, 2.90)	(5.03, 5.28)
$\sigma^{2(2)}$	0.56	0.64	0.82	0.52	10.68	9.20	1.62’	0.92
	(0.34, 0.91)	(0.39, 1.08)	(0.41, 1.48)	(0.34, 0.82)	(9.37, 11.13)	(8.48, 10.00)	(1.62, 1.63)	(0.46, 1.71)
$r^{(2)}$	1.32	1.59	1.57	1.24	0.20	1.09	5.19	2.04
	(0.86, 2.02)	(1.10, 2.26)	(0.91, 2.53)	(0.83, 1.90)	(0.20, 0.21)	(1.05, 1.13)	(1.01, 5.29)	(1.47, 2.74)
$\widetilde{a}^{(2)}$					0.10	−0.28	−0.31	−0.09
					(0.10, 0.12)	(−0.33, −0.21)	(−0.31, −0.31)	(−0.10, −0.09)
$b^{(2)}$		0.08				0.07		
		(−0.51, 0.69)				(0.03, 0.09)		
$\tau_{ar}^{(2)}$		6.21				18692.77		
		(1.90, 14.91)				(15926.47, 23610.96)		
$\tau_{rw2}^{(2)}$				1.00				461.15
				(0.29, 2.31)				(458.44, 463.99)

slightly less uncertain for ST1. Further, both models systematically underestimate the anchovy abundance at sites where the anchovy abundance is higher, but it is slightly more underestimated in S2 than in ST1 (Figure 5 in Supplementary Material).

The analysis of the posterior mean of the spatio-temporal fields, $\widetilde{\boldsymbol{f}}_s^{(k)}(.,t)$, representing the true spatial distribution and abundance of anchovy (>0) per year. The first and second rows of Figure 6, shows the latent field $\widetilde{\boldsymbol{f}}_s^{(k)}(.,t)$ that represents clusters of anchovy aggregations, as well as their size. The third and fourth rows of the same figure, $\widetilde{\boldsymbol{f}}_s^{(k)}(.,t)$ represents their density (high or low abundance). In fact, a large patch does not necessarily mean a high anchovy abundance, and conversely, a small patch does not mean low anchovy abundance (Bertrand et al., 2008b). For instance, the pattern of anchovy presence is broadly similar in 1999, 2003 and 2006, nevertheless, the anchovy abundance is quite different. In 1999 there are dense small patches and medium patches with very low anchovy abundance, in 2003 the medium patches of anchovy abundance are not too dense, while in 2006 there are dense medium patches and small patches with low anchovy abundance. Indeed, if we assume the same spatial process for all years, the distribution (absence/presence) and positive abundance of anchovy are smooth patterns for overall years (Figure 7).

In summary, for interpretability, fitting and predicting performance, we should choose model S2 or ST1. If interpretations are really important, the usual case in ecology

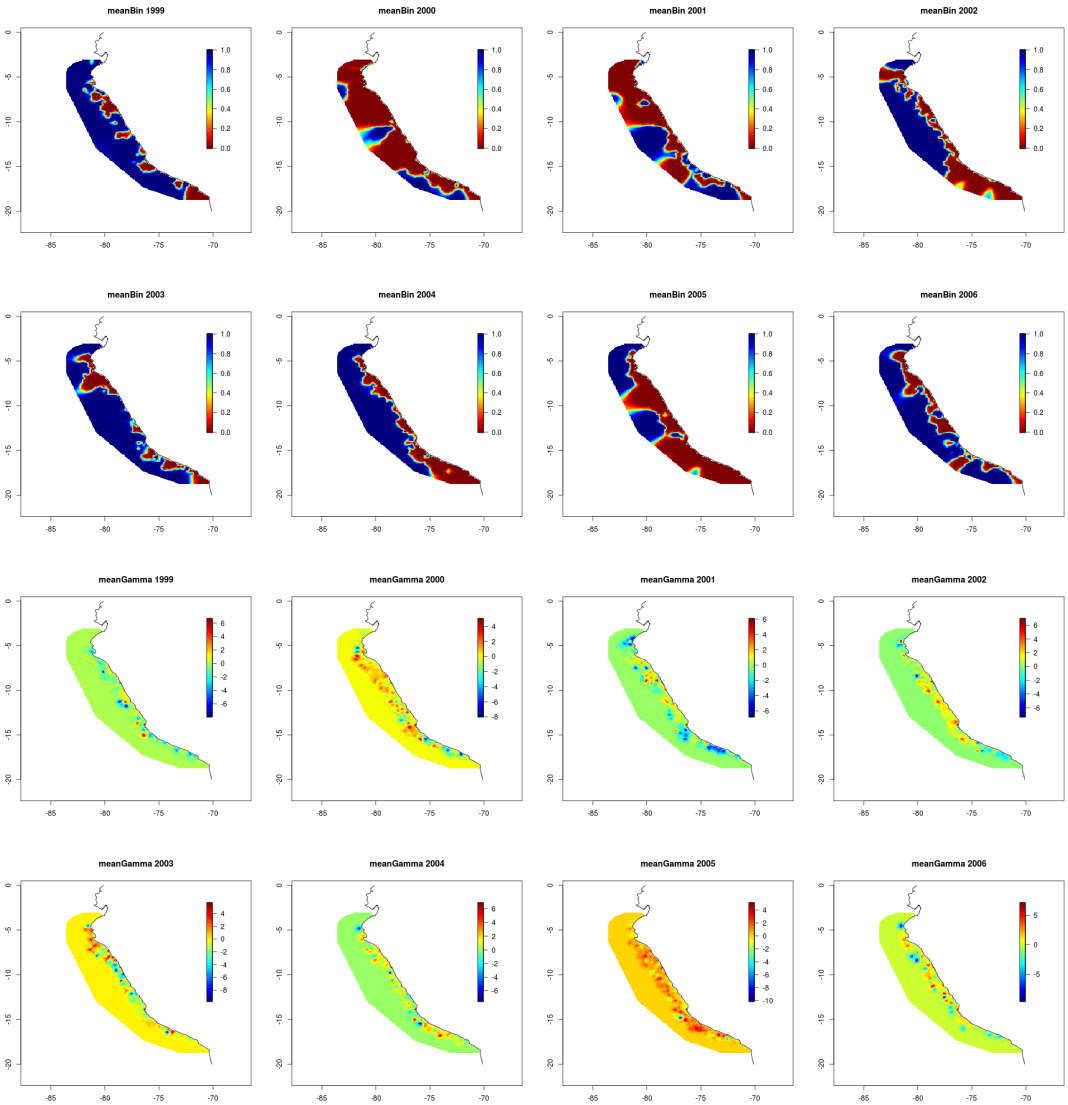

Figure 4: Projection of the posterior mean of the spatio-temporal fields $\boldsymbol{f}_s^{(k)}(.,t)^{(k)}$ for each year, corresponding to Model ST1, $\boldsymbol{f}_s^{(k)}(.,t)^{(1)}$ (in logarithmic scale) of probability of anchovy absence (first and second rows) and $\boldsymbol{f}_s^{(k)}(.,t)^{(2)}$ (in exponential scale) of positive anchovy abundance (third and fourth rows).

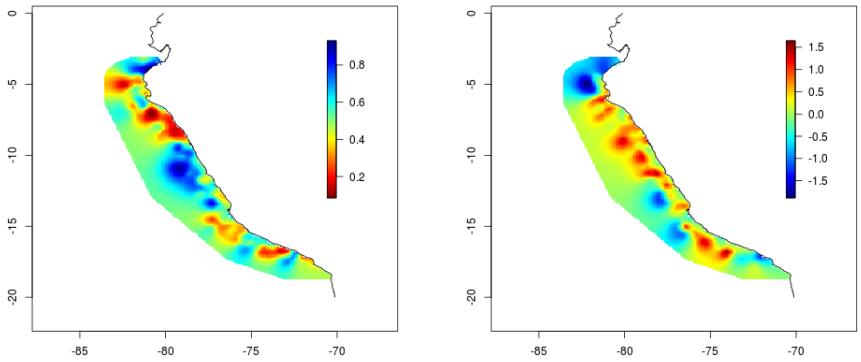

Figure 5: Projection of the posterior mean of the spatial fields $\boldsymbol{f}_s^{(k)}(.,t)^{(1)}$ (in logarithmic scale, left panel) and $\boldsymbol{f}_s^{(k)}(.,t)^{(2)}$ (in exponential scale, right panel) corresponding to Model S1.

studies, despite total computing time, we should choose model ST1. However, if the objective is only to bring predictions for the next year at low computational cost, model S2 can be used. Model ST1 provides better fitting and a richer interpretation.

The mean posterior fixed effects and hyperparameters, for model ST1, are reported in Table 2 and Table 4. These results point out that for the DC (0.17), the further is the distance, the higher the probability of anchovy absence. The linear and quadratic latitude terms (10.94 and 0.50, respectively) indicate that there is a higher probability of anchovy absence at the extremes. The statistical significance of the depth, on the probability of anchovy absence and positive abundance of anchovy (-2.26, 0.42), suggests that for deeper ocean regions, there is higher probability of anchovy absence, and lower positive anchovy abundance. The result for the SST on the probability of anchovy absence (1.21) states that the higher SST, the higher probability of anchovy absence. During El Niño, the surface waters of the central and eastern equatorial Pacific are warmer, and the anchovy apparently disappears in this region. In fact, anchovies live mainly in cool, coastal waters (Escudero & Rivera, 2011). In the positive abundance of anchovy, the SST coefficient (0.11) indicates that the higher SST, the higher anchovy abundance, considering that here the SST is in the tolerable range for anchovy. The mean posterior effective range $\mathrm{r}^{(1)}$ of the probability of anchovy absence is 2.24 degrees ($\approx$ 246 km), whilst the mean posterior effective range $\mathrm{r}^{(2)}$ of the positive anchovy abundance is 0.20 degrees ($\approx$ 22 km). The AR(1) coefficients $\widetilde{a}^{(1)}$ and $\widetilde{a}^{(2)}$ suggest that the probability of absence/presence and abundance of anchovy for each site depends positively on the previous year, being less dependent in the second case. Figure S4 (in Supplementary Material) provides evidence that the few years condidered in our study are enough to capture

temporal dependence, we show that the posterior and prior distributions for $\widetilde{a}^{(k)}$ are very different, proving that the data actually contribute to update the temporal dependence in this complex model.

One of our main interest lies in reconstructing reliable maps of anchovy abundance, as well as their corresponding uncertainty measures. The logarithm of PWD and POD of anchovy abundance, and standard deviation (sd), for each year are computed as described in Section 3.3. Figure 6 shows the POD and their sd, corresponding to model ST1. Finally, Figure 7 shows the POD for year 2007, using the model ST1, fitted with the data from 1999 to 2006 (left), and the PWD of year 2007 also using the model ST1, but fitted with observations from 1999 to 2007 (right). Both patterns agree closely, thus we could know where the anchovy would be present the next year, and where the higher abundance of anchovy is expected.

5 Discussion

In this study we presented an application of spatio-temporal modeling using the SPDE approach for a large data set of point-referenced abundance of anchovy, characterized by a large number of zero values. We compared a variety of spatio-temporal models taking into account the goodness of fit, the spatial predictive capability and the computational cost. Some models were recently proposed to deal with these kind of data (e.g., Fernandes et al., 2009; Paradinas et al., 2015). In particular, Fernandes et al. (2009) proposed the model S1 but using an exponential correlation function and MCMC. They fit a model defining $f(s,t)^{(1)}$ like in model S1 and $f(s,t)^{(2)}$ like in model ST1. However, the convergence was not achieved. They argued that the problem was that whenever there were observations for a particular time equal to zero, they were zero for all gauged sites. We did not have this problem because we can have observations equal to zero and different from zero, at the same site, across years, (see Figure 2). Moreover, we assumed that the zero values are missing data in $\eta(s_i,t)^{(2)}$ and they should not influence in our modeling. On the other hand, Paradinas et al. (2015) proposed the same models S1 and ST1 for a data set spatially small (40 sites), where model S1 was the best. They only evaluated the estimation performance of models to get projections of the posterior mean. We also explored other combinations of spatio-temporal models for the linear predictors, for instance, defining $f(s,t)^{(1)}$ like in model ST1 and $f(s,t)^{(2)}$ like in model S1, but any of these ones gave significant improvement to predictive measures. Thus, we omitted these results from the paper.

One of our main purpose was to provide an efficient model which brings PWD, POD

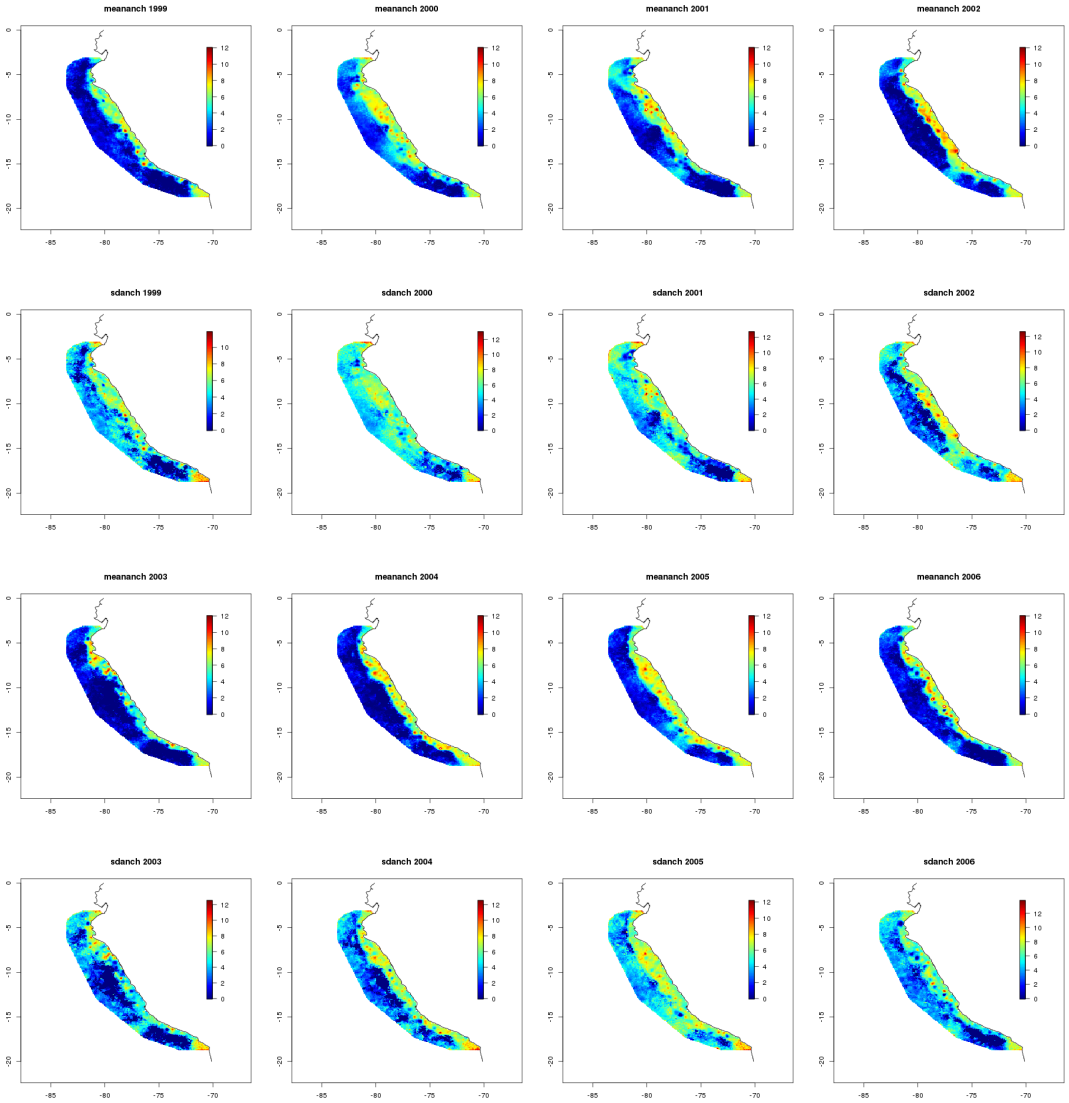

Figure 6: Model ST1; Logarithm of PWD of anchovy abundance (first and third rows) and standard deviation (second and fourth rows) for each year

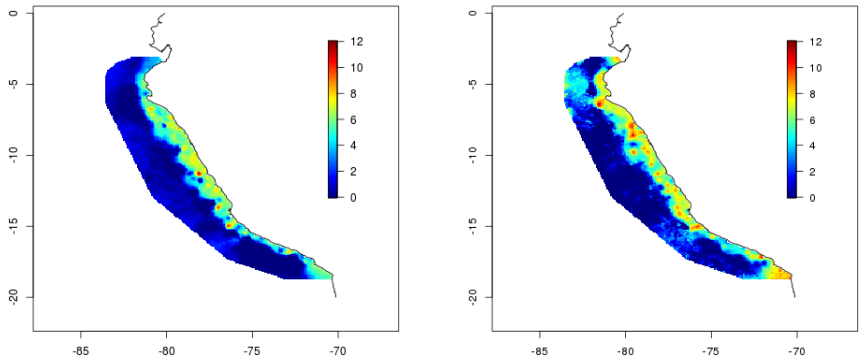

Figure 7: The logarithm of POD from Model ST1, fitted with data from 1999 to 2006 (left panel) and fitted with data from 1999 to 2007 (right panel).

and quantification of their uncertainty, to indicate where is the highest probability of anchovy presence and where we would expect it will be more anchovy abundance. To bring POD of anchovy abundance was challenging, since predicting future values is typically less accurate and riskier than predictions at sites for which we have data at the time of study (Gelman et al., 2014). Although the variability of anchovy abundance is very high, making POD very difficult, the chosen model seems to have potential to identify anchovy presence, as well as regions of high anchovy abundance. The complexity of the spatio-temporal models proposed, in particular of the winner model, is clearly justified also to bring interpretable summaries of the spatio-temporal anchovy distribution, in particular to understand how it is the true distribution of anchovy when lefting out the effect of the environmental covariates. The detection of spatial/temporal patterns and significance of covariates is also directly related to the spatial and temporal scale at which ecological data are measured (Fortin & Dale, 2005), then all results presented here should be interpreted carefully for this sub-mesoscale. For other scales the results would change dramatically. Finally, we had a large number of spatial sites with relatively few temporal observations, for this reason, less general conclusions should be made concerning to global temporal relationships.

Acknowledgments

The authors would like to thank IMARPE for the valuable aid in providing the data. The first author would like to thank ProUNI, and the second author would like to thank FAPEMIG and CNPq, for partial financial support.

Bibliography

Bauer, C., Wakefield, J., Rue, H., Self, S., Feng, Z., & Wang, Y. (2016). Bayesian penalized spline models for the analysis of spatio-temporal count data. *Statistics in Medicine*, *35*, 1848–1865.

Berrocal, V., Raftery, A. E., & Gneiting, T. (2008). Probabilistic quantitative precipitation field forecasting using a two-stage spatial model. *The Annals of Applied Statistics*, *2*, 1170–1193.

Bertrand, A., Chaigneau, A., Peraltilla, S., Ledesma, J., Graco, M., Monetti, F., & Chavez, F. P. (2011). Oxygen: A fundamental property regulating pelagic ecosystem structure in the coastal southeastern tropical Pacific. *PLoS ONE*, *6 (12)*.

Bertrand, A., Gerlotto, F., Bertrand, S., Gutiérrez, M., Alza, L., Chipollini, A., Díaz, E., Espinoza, P., Ledesma, J., Quesquén, R., Peraltilla, S., & Chavez, F. (2008a). Schooling behaviour and environmental forcing in relation to anchoveta distribution: An analysis across multiple spatial scales. *Progress in Oceanography*, *79*, 264–277.

Bertrand, A., Grados, D., Colas, F., Bertrand, S., Capet, X., Chaigneau, A., Vargas, G., Mousseigne, A., & Fablet, R. (2014). Broad impacts of fine-scale dynamics on seascape structure from zooplankton to seabirds. *Nature*, *55*, 2–39.

Bertrand, S., Díaz, E., & Lengaigne, M. (2008b). Patterns in the spatial distribution of Peruvian anchovy (Engraulis ringens) revealed by spatially explicit fishing data. *Progress in Oceanography*, *79*, 379–389.

Blangiardo, M., Cameletti, M., Baio, G., & Rue, H. (2013). Spatial and spatio-temporal models with R-INLA. *Spatial and Spatio-temporal Epidemiology*, *4*, 33–49.

Cameletti, M., Lindgren, F., Simpson, D., & Rue, H. (2013). Spatio-temporal modeling of particulate matter concentration through the SPDE approach. *Advances in Statistical Analysis*, *97*, 109–131.

Chavez, F., Bertrand, A., Guevara-Carrasco, R., Soler, P., & Csirke, J. (2008). The northern Humboldt current system: brief history, present status and a view towards the future. *Progress in Oceanography*, *79*, 95–105.

Cosandey-Godin, A., Krainski, E. T., Worm, B., & Flemming, J. M. (2014). Applying Bayesian spatiotemporal models to fisheries bycatch in the Canadian Arctic. *Canadian Journal Fisheries Aquatics Science*, *72*, 1–12.

Dey, D. K., Chen, M.-H., & Chang, H. (1997). Bayesian Approach for Nonlinear Random Effects Models. *Biometrics*, *53(4)*, 1239–1252.

Escudero, L., & Rivera, V. (2011). Handbook of satellite remote sensing image interpretation: Applications for marine living resources conservation and management. chapter Case14: Distribution of Peruvian Anchovy Fleets in Relation to Oceanic Parameters. (pp. 207–213). Canada: EU PRESPO Project and IOCCG. URL: `http://ioccg.org/what-we-do/ioccg-publications/eu-prespoioccg-handbook/`.

Fernandes, M., Schmidt, A., & Migon, H. (2009). Modelling zero-inflated spatio-temporal processes. *Statistical Modelling*, *9*, 3–25.

Fortin, M. J., & Dale, M. R. T. (2005). *Spatial Analysis: A Guide for Ecologists*. Cambridge: Cambridge University Press.

Fréon, P., Cury, P., Shannon, L., & Roy, C. (2005). Sustainable exploitation of small pelagic fish stocks challenged by environmental and ecosystem changes: a review. *Bulletin of Marine Science*, *76*, 385–462.

Fuentes, M., Reich, B., & Lee, G. (2008). Spatio-temporal mesoscale modeling of rainfall intensity using gage and radar data. *Annals of Applied Statistics*, *2*, 1148:1169.

Geisser, S., & Eddy, W. F. (1979). A Predictive Approach to Model Selection. *Journal of the American Statistical Association*, *74*, 153–160.

Gelfand, A. E., Banerjee, S., & Gamerman, D. (2005). Spatial process modelling for univariate and multivariate dynamic spatial data. *Environmetrics*, *16*, 465–479.

Gelfand, A. E., Ecker, M. D., Knight, J. R., & Sirmans, C. F. (2004). The dynamics of location in home price. *Journal of Real Estate Finance and Economics*, *29 (2)*, 149–166.

Gelfand, A. E., Kim, H.-J., Sirmans, C. F., & Banerjee, S. (2003). Spatial modeling with spatially varying coefficient processes. *Journal of the American Statistical Association*, *98*, 387–396.

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, *24*, 997–1016.

Hellton, K. H. (2011). *Stochastic Models for Smoothing Splines: A Bayesian Approach*. Master's thesis Norwegian University of Science and Technology. URL: `https://brage.bibsys.no/xmlui/handle/11250/258927`.

Hoef, J. M. V., & Jansen, J. K. (2007). Space–time zero-inflated count models of Harbor seals. *Environmetrics*, *18*, 697–712.

Knhorr-Held, L., & Rainer, E. (2001). Projections of lung cancer mortality in west Germany: a case study in Bayesian prediction. *Biostatistics*, *2*, 109–129.

Lindgren, F., & Rue, H. (2015). Bayesian spatial modelling with R-INLA. *Journal of Statistical Software*, *63 (19)*. doi:`10.18637/jss.v063.i19`.

Lindgren, F., Rue, H., & Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: The SPDE approach. *Journal of the Royal Statistical Society, Series B*, *73*, 423–498.

Musenge, E., Chirwa, T. F., Kahn, K., & Vounatsou, P. (2013). Bayesian analysis of zero inflated spatiotemporal HIV/TB child mortality data through the INLA and SPDE approaches: Applied to data observed between 1992 and 2010 in rural North East South Africa. *International Journal of Applied Earth Observation and Geoinformation*, *22*, 86–98.

Paradinas, I., Conesa, D., Pennino, M. G., Muñoz, F., Fernández, A. M., López-Quílez, A., & Bellido, J. M. (2015). A Bayesian spatio-temporal approach for identify fish nurseries by validating persistence areas. *Marine Ecology Progress Series*, *528*, 245–255.

Pennino, M. G., Muñoz, F., Conesa, D., López-Quílez, A., & Bellido, J. M. (2014). Bayesian spatio-temporal discard model in a demersal trawl fishery. *Journal of Sea Research*, *90*, 44–53.

Quiroz, Z. C., Prates, M. O., & Rue, H. (2015). A Bayesian approach to estimate the biomass of anchovies off the coast of Perú. *Biometrics*, *71*, 208–217.

Riebler, A., & Held, L. (2010). The analysis of heterogeneous time trends in multivariate age–period–cohort models. *Biostatistics*, *11*, 57–69.

Ross, B. E., Hooten, M. B., & Koons, D. N. (2012). An accessible method for implementing hierarchical models with spatio-temporal abundance data. *Plos One*, *7 (11)*. doi:`https://doi.org/10.1371/journal.pone.0049395`.

Rue, H., & Held, L. (2005). *Gaussian Markov Random Fields: Theory and applications*. Boca Raton: Chapman & Hall/CRC.

Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society, Series B*, *71 (2)*, 319–392.

Sansó, B., & Guenni, L. (2000). A non-stationary multisite model for rainfall. *Journal of the American Statistical Association*, *95*, 1089–1100.

Sherman, M. (2011). *Spatial Statistics and Spatio-Temporal Data*. West Sussex: Wiley Series.

Sigrist, F., Künsch, H. R., & Stahel, W. A. (2012). A dynamic nonstationary spatio-temporal model for short term prediction of precipitation. *The Annals of Applied Statistics*, *6 (4)*, 1452–1477.

Simmonds, E. J., & MacLennan, D. N. (2005). *Fisheries Acoustics: Theory and Practice*. Oxford: Blackwell Science.

Sloughter, J. M., Gneiting, A. E. A. E. R., & Fraley, C. (2007). Probabilistic quantitative precipitation forecasting using Bayesian modeling averaging. *Monthly weather Review*, *135*, 3209–3220.

Stern, R. D., & Coe, R. (1984). A model fitting analysis of daily rainfall data. *Journal of the Royal Statistical Society, Series A*, *147 (1)*, 1–34.

Stidd, C. K. (1973). Estimating the precipitation climate. *Water Resources*, *9*, 1235–1241.

Stroud, J. R., Müller, P., & Sansa'o, B. (2001). Dynamic models for spatiotemporal data. *Journal of the Royal Statistical Society, Series B*, *63*, 673.

Swartzman, G., Bertrand, A., Gutiérrez, M., Bertrand, S., & Vasquez, L. (2008). The relationship of anchovy and sardine to water masses in the Peruvian Humboldt current system from 1983 to 2005. *Progress in Oceanography*, *79*, 228–237.

Wang, X., Chen, M.-H., Kou, R. C., & Dey, D. K. (2015). Bayesian spatial-temporal modeling on ecological zero-inflated count data. *Statistica Sinica*, *25*, 189–204.

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning*, *11*, 3571–3594.

Whittle, P. (1954). On stationary processes in the plane. *Biometrika*, *44*, 434–449.

Wikle, C., & Anderson, C. (2003). Climatological analysis of tornado report counts using a hierarchical Bayesian spatiotemporal model. *Journal of Geophysical Research*, *108*, 1984–2012.

Wilks (1990). Maximum likelihood estimation for the gamma distribution using data contaning zeros. *Journal of Climate*, *3*, 1495–1501.

Appendix

A: Precision matrices in section 3

$$Q_{ar}^{(k)} = \tau_{ar}^{(k)} \begin{pmatrix} 1 & -b^{(k)} & 0 & . & . & . & 0 & 0 \\ -b^{(k)} & (1+b^{(k)2}) & -b^{(k)} & . & . & . & 0 & 0 \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \\ 0 & 0 & . & . & . & -b^{(k)} & (1+b^{(k)2}) & -b^{(k)} \\ 0 & 0 & . & . & . & 0 & -b^{(k)} & 1 \end{pmatrix}$$

$$Q_{rw2}^{(k)} = \tau_{rw2}^{(k)} \begin{pmatrix} 1 & -2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -2 & 5 & -4 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & -4 & 6 & -4 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & -4 & 6 & -4 & 1 & 0 & 0 & 0 \\ . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . \\ 0 & 0 & 0 & 1 & -4 & 6 & -4 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & -4 & 6 & -4 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & -4 & 5 & -2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -2 & 1 \end{pmatrix}$$

$$\Sigma_T = \frac{1}{1-a^{(k)2}} \begin{pmatrix} 1 & a^{(k)} & a^{(k)2} & . & . & . & a^{(k)|T-2|} & a^{(k)|T-1|} \\ a^{(k)} & 1 & a^{(k)} & . & . & . & a^{(k)|T-3|} & a^{(k)|T-2|} \\ a^{(k)2} & a^{(k)} & 1 & . & . & . & a^{(k)|T-4|} & a^{(k)|T-3|} \\ & . & . & . & . & . & . & . \\ & . & . & . & . & . & . & . \\ a^{(k)|T-2|} & a^{(k)|T-3|} & . & . & . & a^{(k)} & 1 & a^{(k)} \\ a^{(k)|T-1|} & a^{(k)|T-2|} & a^{(k)|T-3|} & . & . & a^{(k)2} & a^{(k)} & 1 \end{pmatrix}$$

Given $\widetilde{f}_s^{(k)}(t) = a^{(k)}\widetilde{f}_s^{(k)}(t-1) + \widetilde{w}_s^{(k)}(t); \widetilde{w}_s^{(k)}(t) \sim N(0, [Q^{(k)}]_s^{-1}),$ and, $Cov(\widetilde{f}_s(i), \widetilde{f}_s(j)) = \frac{a^{|i-j|}}{1-a^2}\widetilde{Q}_s^{-1}$.

Let $\widetilde{f}_s^{(k)} \sim (0, \widetilde{\Sigma}^{(k)} = [\widetilde{Q}^{(k)}]^{-1})$, where

$$\Sigma^{(k)} = \begin{pmatrix} \frac{1}{1-(a^{(k)})^2}\widetilde{Q}_s^{-1} & \frac{(a^{(k)})^2}{1-(a^{(k)})^2}\widetilde{Q}_s^{-1} & \frac{(a^{(k)})^3}{1-(a^{(k)})^2}\widetilde{Q}_s^{-1} & . & . & . & \frac{(a^{(k)})^{|T-2|}}{1-(a^{(k)})^2}\widetilde{Q}_s^{-1} & \frac{(a^{(k)})^{|T-1|}}{1-(a^{(k)})^2}\widetilde{Q}_s^{-1} \\ \frac{(a^{(k)})^2}{1-(a^{(k)})^2}\widetilde{Q}_s^{-1} & \frac{1}{1-(a^{(k)})^2}\widetilde{Q}_s^{-1} & \frac{(a^{(k)})}{1-(a^{(k)})^2}\widetilde{Q}_s^{-1} & . & . & . & \frac{(a^{(k)})^{|T-3|}}{1-(a^{(k)})^2}\widetilde{Q}_s^{-1} & \frac{(a^{(k)})^{|T-2|}}{1-(a^{(k)})^2}\widetilde{Q}_s^{-1} \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \\ \frac{(a^{(k)})^{|T-2|}}{1-(a^{(k)})^2}\widetilde{Q}_s^{-1} & \frac{(a^{(k)})^{|T-3|}}{1-(a^{(k)})^2}\widetilde{Q}_s^{-1} & . & . & . & \frac{(a^{(k)})^3}{1-(a^{(k)})^2}\widetilde{Q}_s^{-1} & \frac{(a^{(k)})^2}{1-(a^{(k)})^2}\widetilde{Q}_s^{-1} & \frac{(a^{(k)})}{1-(a^{(k)})^2}\widetilde{Q}_s^{-1} \\ \frac{(a^{(k)})^{|T-1|}}{1-(a^{(k)})^2}\widetilde{Q}_s^{-1} & \frac{(a^{(k)})^{|T-2|}}{1-(a^{(k)})^2}\widetilde{Q}_s^{-1} & \frac{(a^{(k)})^{|T-3|}}{1-(a^{(k)})^2}\widetilde{Q}_s^{-1} & . & . & \frac{(a^{(k)})^2}{1-(a^{(k)})^2}\widetilde{Q}_s^{-1} & \frac{(a^{(k)})}{1-(a^{(k)})^2}\widetilde{Q}_s^{-1} & \frac{1}{1-(a^{(k)})^2}\widetilde{Q}_s^{-1} \end{pmatrix},$$

$$\Sigma^{(k)} = \frac{1}{1-(a^{(k)})^2} \begin{pmatrix} 1 & a^2 & (a^{(k)})^3 & . & . & . & (a^{(k)})^{|T-2|} & (a^{(k)})^{|T-1|} \\ (a^{(k)})^2 & 1 & (a^{(k)}) & . & . & . & (a^{(k)})^{|T-3|} & (a^{(k)})^{|T-2|} \\ & . & . & . & . & . & . & . \\ & . & . & . & . & . & . & . \\ (a^{(k)})^{|T-2|} & (a^{(k)})^{|T-3|} & . & . & . & (a^{(k)})^3 & (a^{(k)})^2 & (a^{(k)}) \\ (a^{(k)})^{|T-1|} & (a^{(k)})^{|T-2|} & (a^{(k)})^{|T-3|} & . & . & (a^{(k)})^2 & (a^{(k)}) & 1 \end{pmatrix} \otimes [\widetilde{Q}_s^{(k)}]^{-1},$$

$$\Sigma^{(k)} = \begin{pmatrix} 1 & -(a^{(k)}) & 0 & . & . & . & 0 & 0 \\ -(a^{(k)}) & (1+(a^{(k)})^2) & -(a^{(k)}) & . & . & . & 0 & 0 \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \\ 0 & 0 & . & . & . & -(a^{(k)}) & (1+(a^{(k)})^2) & -(a^{(k)}) \\ 0 & 0 & 0 & . & . & 0 & -(a^{(k)}) & 1 \end{pmatrix}^{-1} \otimes [\widetilde{Q}_s^{(k)}]^{-1}.$$

Let

$$Q_T^{(k)} = \begin{pmatrix} 1 & -(a^{(k)}) & 0 & . & . & . & 0 & 0 \\ -(a^{(k)}) & (1+(a^{(k)})^2) & -(a^{(k)}) & . & . & . & 0 & 0 \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \\ 0 & 0 & . & . & . & -(a^{(k)}) & (1+(a^{(k)})^2) & -(a^{(k)}) \\ 0 & 0 & 0 & . & . & 0 & -(a^{(k)}) & 1 \end{pmatrix}.$$

Then, $\widetilde{\Sigma}^{(k)} = [Q_T^{(k)}]^{-1} \otimes [\widetilde{Q}_s^{(k)}]^{-1}$, and for invertible matrices property,

$$\widetilde{Q}^{(k)} = [\widetilde{\Sigma}^{(k)}]^{-1} = ([Q_T^{(k)}]^{-1} \otimes [\widetilde{Q}_s^{(k)}]^{-1})^{-1} = ([Q_T^{(k)}]^{-1})^{-1} \otimes ([\widetilde{Q}_s^{(k)}]^{-1})^{-1} = Q_T^{(k)} \otimes \widetilde{Q}_s^{(k)}.$$

B: Supplementary Figures

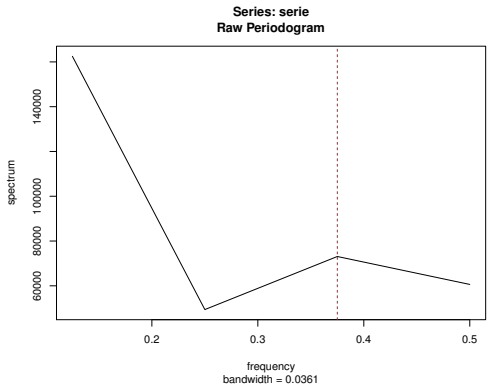

Figure S1: Periodogram of the yearly mean of anchovy abundance. The red line represents the period P = 3.

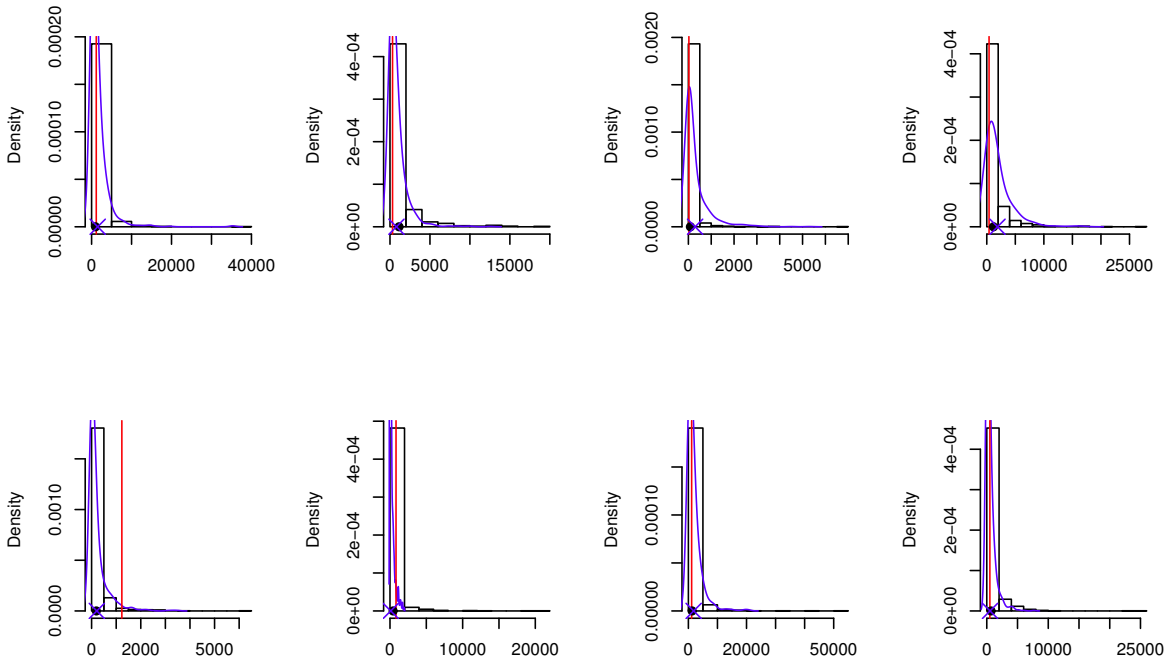

Figure S2: Posterior predictive histograms at random selected locations corresponding to Model ST1 and densities (blue line) corresponding to Model S2. POD corresponding to Model ST1 (black dot) and Model S2 (blue cross). The red vertical line is the observed anchovy abundance.

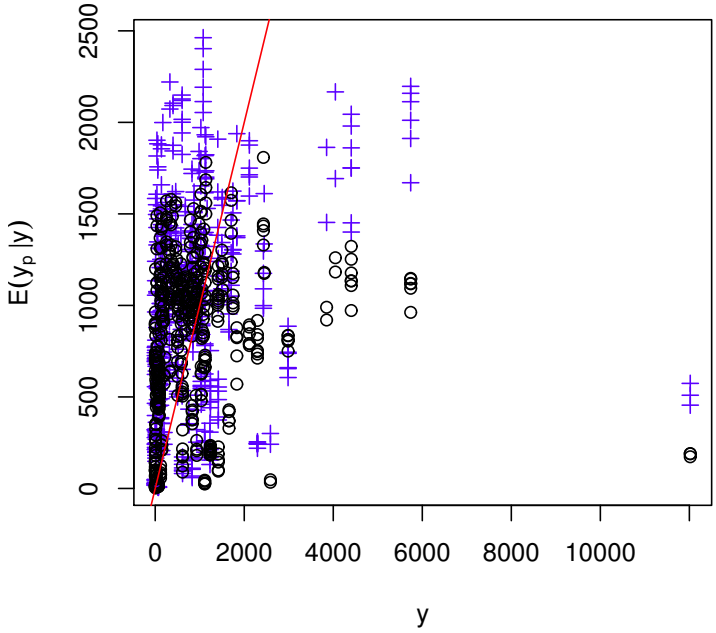

Figure S3: POD plotted against the observed anchovy abundance from Model S2 (black circles) and Model ST1 (blue cross).

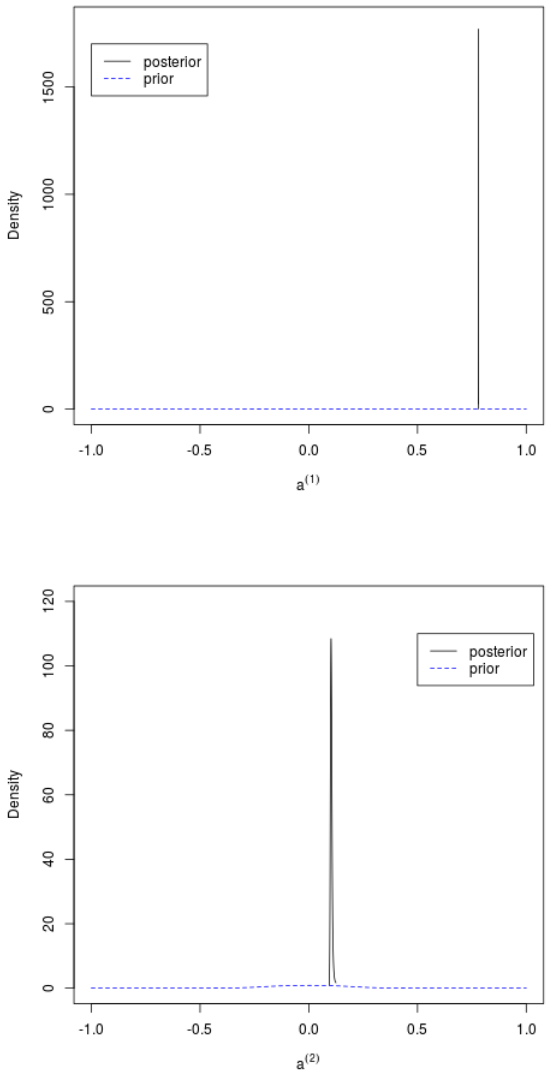

Figure S4: Posterior and prior distributions of the parameters: $a^{(1)}$ (left) and $a^{(2)}$ (right) corresponding to Model ST1 .

Chapter 5

Block Nearest Neighbor Gaussian processes

Zaida C. Quiroz , Marcos O. Prates and Dipak K. Dey

Block Nearest Neighbor Gaussian processes

Z. C. Quiroz, M. O. Prates and Dipak K. Dey

Abstract

This work develops a valid spatial block-Nearest Neighbor Gaussian process (block-NNGP) for estimation and prediction of location-referenced large spatial datasets. The key idea behind our approach is to subdivide the spatial domain into several blocks which are dependent under some constraints. As consequence, the cross-blocks should mainly capture the large-scale spatial variation, while each block should capture the small-scale dependence. Of course, the optimal blocking depends on the sampled spatial locations, and the number of blocks represents a trade-off between computational and statistical efficiency. The block-NNGP is included as prior in the hierarchical modeling framework and efficient Markov chain Monte Carlo (MCMC) algorithms exploit the sparsity of the block precision matrix, which can be computed by distributing the operations using parallel computing. The performance of the block-NNGP is illustrated using simulation studies and applications with massive data.

Keywords: Bayesian hierarchical models, block-NNGP, large datasets, MCMC, parallel computing.

1 Introduction

New technologies such as GPS and remote sensing enable the collection of massive amounts of high-resolution geographically referenced observations over large spatial regions. These data are analyzed through spatial random fields, usually based on Gaussian processes (GP). However, it is well-known that computations can be prohibitive for a spatial random field where the number of locations is large because calculations over a GP depend on the covariance and precision matrix, which are usually dense.

One approach to model large spatial datasets proceeds inducing sparsity in the precision matrix through Gaussian Markov random fields (GMRF), assuming that the spatial correlation between pairs of distantly located observations is nearly zero (Rue and Tjelmeland, 2002). In particular, this sparsity can be achieved either

through stochastic partial differential equations (SPDE - Lindgren et al. 2011) when the covariance function is Matérn, or the Nearest Neighbor Gaussian process (NNGP- Datta et al. 2016) which is less restrictive, working for any valid covariance function. In the SPDE approach, the Gaussian field with Matérn covariance function is approximated to a GMRF, through the solution of a SPDE using finite element methods. On the other hand, the NNGP is a well-defined spatial GMRF, built from lower-dimensional conditional distributions which depends on the nearest neighbor observations, providing a unified fully process-based framework for estimation and prediction.

Another approach to deal with computationally intractable large matrices of spatial random fields is the spatial blocking, that is, the partition of the spatial domain into blocks. This approach was often restricted to covariance matrices, ignoring the dependence between different blocks. For instance, see Stein (2013) or Bolin and Wallin (2016), where they showed that this simple approach is better than covariance tapering and spatially adaptive covariance tapering, methods that set "distant" observations of the covariance matrix into zero to get its sparsity. Kim et al. (2005) presented a similar approach but their method automatically decomposes the spatial domain into disjoint blocks. Otherwise, Stein et al. (2004) and Caragea and Smith (2007) proposed composite-likelihood methods to achieve computational feasibility by treating blocks of observations as independent and/or conditionally independent, but it is not clear how to obtain proper joint predictive distributions for locations in different blocks. To allow for some dependence between blocks Eidsvik et al. (2014) used block composite-likelihood methods to propose a unified framework for both parameter estimation and prediction, however, it is restricted to fit Gaussian response variables through classic inference.

One drawback of the NNGP approach is that we need to predetermine a collection of the "past" neighbors to construct the model, but in spatial settings, the locations are not naturally ordered. In addition, the information of "non-past" nearest neighbors is not considered, and thereby some small-scale spatial dependence may be lost. While Eidsvik et al. (2014) assumed negligible dependence between distant blocks, they only used this feature to approximate the likelihood, affecting directly the covariance matrix instead of the precision matrix. Here, we merge both approaches, first, we assume that pairs of blocks are conditionally independent given some of the "past" blocks, and then we extend the NNGP theory to get a new valid GMRF called block-NNGP.

The main goal of block-NNGP is to capture much of the spatial dependence, because the cross-blocks should mainly capture the large-scale spatial variation while each block captures the small-scale dependence. This new process enables a consistent

way to combine parameter estimation and spatial prediction. Of course, the optimal blocking depends on the sampled spatial locations, and the number of blocks represents a trade-off between computational and statistical efficiency. The higher the number of blocks, more sparse will be the precision matrix, but we have to be careful to avoid loosing important spatial dependences. Finally, to perform inference we adopt a Bayesian framework to demonstrate the full inferential capabilities in terms of estimation, prediction and goodness of fit, of the block-NNGP hierarchical models and parameters therein. In particular, the parameters were estimated through the collapsed MCMC method (Finley et al., 2017) to improve convergence and run time. This algorithm enjoys the frugality of a low-dimensional MCMC chain but allows for full recovery of the latent random effects.

The paper is organized as follows. Section 2 gives the details of the proposed block-NNGP process. In Section 4, simulations are assessed for the predictive performance of the proposed process. The example of mining and precipitation data are used in Section 5 to illustrate the use of the proposed process when the data size is large. Some discussions are given in Section 6.

2 Block NNGP process

Assume that $w(s) \sim \text{GP}(0, C(\theta))$ defined for all $s \in D \subset \Re^2$, where $C(\theta)$ is any valid covariance function. Let $S = \{s_1, \ldots, s_n\}$ be a fixed set of locations in D. Then the joint density of $w_S = (w(s_1), \ldots, w(s_n))'$ can be written as

$$p(w_S) = p(w(s_1)) \prod_{i=2}^{n} p(w(s_i)|w(s_1), \ldots, w(s_{i-1})). \tag{5.1}$$

Vecchia (1988) proposed to replace the conditioning sets on the right-hand side of Equation (5.1) with conditioning sets of size at most m, where $m \ll n$. In particular, Datta et al. (2016) propose to use some fixed number of nearest neighbors observations from the "past", then Equation (5.1) is approximated by $\widetilde{p}(w_S) = p(w(s_1)) \prod_{i=2}^{n} p(w(s_i)|w(s_{i_m}))$, where $w(s_{i_m})$ are the neighbor observations of $w(s_i)$. This approach seems very reasonable, since correlations between pairs of distant locations are nearly zero, and little information might be lost when taking them to be conditionally independent given intermediate locations. They also proved that $\widetilde{p}(w_S)$ is a valid joint distribution for w_S, which is used to built up a valid spatial process called NNGP, thus the traditional GP is replaced by the NNGP.

Stein et al. (2004) proposed a generalization of the Vecchia approximation, a restricted version of the conditional probability approximation, where the joint density of

Equation (5.1) is approximated by assuming a partition of w_S in vectors of not uniform lengths and some conditioning vector sets of each vector. Here we extend the NNGP introducing another valid spatial process through such approximation built on blocks of data. In particular, we consider a partition of the region D into M blocks $b_1, \ldots, b_M$, with $U_{k=1}^{M} b_k = D, b_k \cap b_l = \phi$, for all pairs of blocks b_k and b_l. The vector $w_{b_k} = \{w(s_i); s_i \in b_k\}$ where $\dim(w_{b_k}) = n_k$ such that $\sum_{k=1}^{M} n_k = n$. Then, we assume that the w_{b_l} and w_{b_j}, for $l \neq j$, are conditionally independent given some "past" blocks, and the joint density of w_S is approximated by

$$\widetilde{p}(w_S) = p(w_{b1}) \prod_{k=2}^{M} p(w_{b_k} | w_{N(b_k)}), \tag{5.2}$$

where $N(b_k) \subset S \setminus [s_i \in b_k]$ is the set of nb neighbor blocks of b_k.

Proposition 1. *Let $G = \{S, \xi\}$ be a chain graph, where $S = \{s_1, \ldots, s_n\}$ is the set of nodes, and ξ is comprised by: (i) the set of directed edges from every node in the set $s_{b_k} = \{s_i \in b_k, \forall i = 1, \ldots, n\}$, to all nodes in $N(b_k)$, $\forall k = 1, \ldots, M$, and (ii) the set of undirected edges between every pair of nodes in b_k. Let G_b be a subgraph of G composed by M nodes, such that each node is one node of the set s_{b_k}. If G_b is acyclic and $p(w_S)$ is a valid multivariate joint density, then $\widetilde{p}(w_S)$ in Equation (5.2) is also a valid multivariate joint density.*

The proof of this proposition and subsequent proofs are found in Appendix A1. A chain graph G, also called partially directed acyclic graphs, is defined by a set of nodes disjointly partitioned into several chain components, edges between nodes in chain are undirected and edge between nodes in different chains are directed. If we take one node per chain they form a directed graph which we call G_b. Proposition 1 states that Equation (5.2) is a proper multivariate joint density when $p(w_S)$ is a valid multivariate joint density and G is a chain graph which has a directed acyclic graph (DAG) G_b. In particular, if $N(b_k)$ is any subset of $\{N(b_1), \ldots, N(b_{k-1})\}$ then G_b is acyclic (Figure (1)). This choice of neighbor sets do not unvalidate the acyclic property between blocks and also produce valid densities. With this choice, we are assuming that $\forall w(s_i); s_i \in b_k$, they are dependent between them, but also that each one depends on $w(s_j) \in w_{N(b_k)}$, that is, depends on the neighbor blocks of b_k. Hence, $w(s_i)$ is explained by all of its nearest neighbors in the block b_k and some nearest neighbor blocks from the past, which avoids loss of information at small scale while preserving information at large scale, respectively. In fact, sometimes when the spatial dependence is strong relative to the spatial domain of observation, it can be advantageous to include some observations in $N(b_k)$ that were rather distant from $s_{b_k} = \{s_i \in b_k, \forall i = 1, \ldots, n\}$ (Stein et al., 2004). This situation was not presented

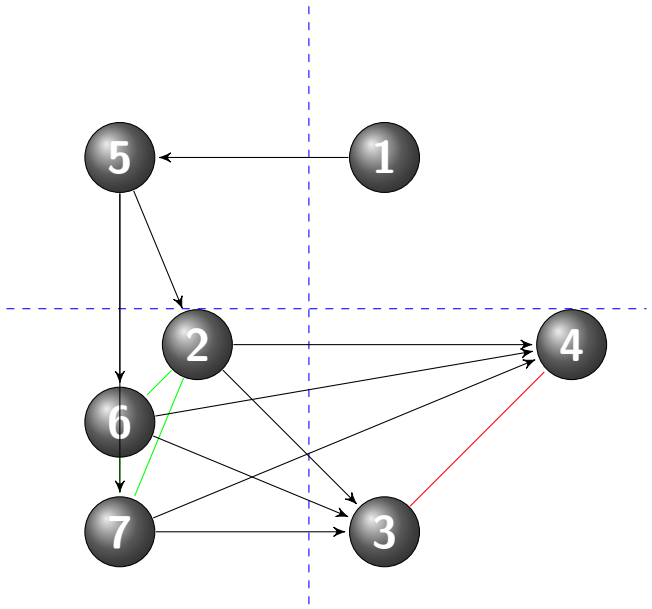

Figure 1: Illustration of a chain graph with $n = 7$ nodes and $M = 4$ blocks: $b_1 = \{1\}$, $b_2 = \{5\}$, $b_3 = \{2, 6, 7\}$, $b_4 = \{3, 4\}$.

in any of Datta et al. (2016) examples. Each $w(s_i)$ depends on $n_{bk} = n_k - 1 + N_{b_k}$ neighbors, where N_{b_k} is the number of locations in the neighbor blocks of b_k. And G is sufficiently sparse if n_k and N_{b_k} are sufficiently small. Note that for the NNGP process, each $w(s_i)$ only depends on at most m nearest neighbors from the past such that $m \ll n$. In particular, the NNGP will be a special case of our proposed spatial process (see corollary 3).

Let w_S be a realization of a GP over S with covariance function $C(\theta)$, therefore $p(w_S)$ is the probability density (pdf) of a n-variate normal distribution with mean zero and covariance matrix C_S. From the proposition 1 holds the next corollary.

Corollary 1. *$p(w_S)$ is the pdf of a n-variate normal distribution with mean zero and covariance matrix C_S. If G is a chain graph and G_b is a DAG, as we specified in proposition 1, then $\widetilde{p}(w_S)$ is a proper density.*

From basic properties of normal distributions, $w_{b_k}|w_{N(b_k)} \sim N_n(B_{b_k} w_{N(b_k)}, F_{b_k})$, $B_{b_k} = C_{b_k,N(b_k)} C^{-1}_{N(b_k)}$ and $F_{b_k} = C_{b_k} - C_{b_k,N(b_k)} C^{-1}_{b_k} C_{N(b_k),b_k}$, where $C_{i,j}$ and C_i are elements of C_s. Thens if f is the pdf of a normal distribution, Equation (5.2) is defined by

$$\widetilde{p}(w_S) = \prod_{k=1}^{M} f(w_{b_k}|B_{b_k} w_{N(b_k)}, F_{b_k}). \tag{5.3}$$

Proposition 2. *If $p(w_S)$ is a proper pdf of a n-variate normal distribution with mean zero and covariance matrix C_S, G is a chain graph and G_b is a DAG, as we*

specified in proposition 1, then

(i) $\widetilde{p}(w_S)$ *is also the pdf of a n-variate normal distribution with mean zero and covariance matrix* $\widetilde{C}_S = (B_S^T F_S^{-1} B_S)^{-1}$,

(ii) B_S *is a block matrix and a lower triangular matrix,*

(iii) F_S *is block diagonal,*

(iv) $\widetilde{C}_S$ *is positive definite, and*

(v) If $n_{bk} \ll n; \forall k = 1, \ldots, M$ *then* $\widetilde{C}_S^{-1}$ *is sparse.*

The multivariate normal distribution is completely specified by its expectation which is assumed to be zero, and its covariance function which is valid since it is positive definite from Proposition 2 (iv). In addition, Proposition 2 also states that B_S is a block matrix and F_S is block diagonal, due to these features we are able to implement our algorithm using parallel processing. The sparsity of the precision matrix in fact represents that distant pair of observations, as well as, distant block of observations are independent. Note that if we assume more blocks, the precision matrix will be more sparse. The reduction in computational complexity is achieved through such sparsity of the precision matrices and we also can parallelize many computations for blocks of data. Then, $\widetilde{p}(w_S)$ is a proper multivariate joint density with a sparse precision matrix which enjoys great features, as a result, it is easier to work with $\widetilde{p}(w_S)$ than with $p(w_S)$. We remark that $\widetilde{p}(w_S)$ is a valid pdf, and we could perform inference directly from a likelihood function not a composite or pseudo-likelihood. For instance, Eidsvik et al. (2014) achieved inference through block-composite likelihood, but their approach ignores information about components of the covariance structure, as a consequence, there is loss of statistical efficiency.

To build a general valid spatial process, we need to provide a pdf consistent with some well-defined random field. Hence, following the NNGP approach, we use $\widetilde{p}(w_S)$ to provide such pdf. We also assume that S is a set of fixed and observed locations. And, we define $U = \{u_1, \ldots, u_l\}$ as any finite set of locations such that $S \cap U = \emptyset$ and $V = S' \cup U$, $S' \subset S$. Using the conditional distribution properties and corollary 1, we have that the approximated conditional pdf of $p(w_U|w_S)p(w_S)$ defined by $\widetilde{p}(w_U|w_S)\widetilde{p}(w_S)$ is also a proper density if $\widetilde{p}(w_U|w_S)$ is proper. Notice that $p(w_U|w_S)$ is proper since $\{w_U, w_S\}$ is a realization of the $GP(0, C(\theta))$. For simplicity, if we assume that w_{u_i} is independent of w_{u_j} given w_S, then we define $\widetilde{p}(w_U|w_S) = \prod_{i=1}^{l} p(w_{u_i}|w_S)$. Further, if we also assume that w_{u_i} only depends on some observations of w_S, $N(u_i)$, which is the set of neighbors of u in S, then $\widetilde{p}(w_U|w_S) = \prod_{i=1}^{l} p(w_{u_i}|w_{N(u_i)})$ which is proper. Now, we assume that $w_U|w_S$ follows a multivariate normal distribution with the

following pdf, $\widetilde{p}(w_U|w_S) = \prod_{i=1}^{l} f(w_{u_i}|B_{u_i}w_{N(u_i)}, F_{u_i})$, where $B_{u_i} = C_{u_i,N(u_i)}C_{N(u_i)}^{-1}$ and $F_{u_i} = C_{u_i} - C_{u_i,N(u_i)}C_{u_i}^{-1}C_{N(u_i),u_i}$, $C_{i,j}$ and C_i are elements of C_s.
Then we can define an approximation of the pdf $p(w_V)$ as follows,

$$\widetilde{p}(w_V) = \int \widetilde{p}(w_U|w_S)\widetilde{p}(w_S) \prod_{s_i \in (S')^c} d(w(s_i)), \tag{5.4}$$

where $(S')^c$ is the complement of S' and $\widetilde{p}(w_V)$ is a proper density for any choice of $N(u_i)$. Katzfuss and Guinness (2017) proposed a general Vecchia approximation, which is very similar in form to the pdf $\widetilde{p}(w_V)$, if we assume $S = S'$, they proved that such approximation yields a joint multivariate distribution. Their most similar case assumes a similar $\widetilde{p}(w_S)$, using vectors of observations, but we define $\widetilde{p}(w_U|w_S)$ different from their approach to build a valid spatial process. We prove that the joint distribution of w_V is consistent with some well-defined stochastic process, in the sense that the Kolmogorov's consistency conditions are verified, that is, if the symmetry and compatibility conditions hold for the process defined through the finite-dimensional distributions in Equation (5.4). For this reason we need to be careful when defining $\widetilde{p}(w_U|w_S)$ to ensure that it will be the same under reordering of the sites.

Lemma 1. *Let $\widetilde{p}(w_V)$ in Equation (5.4) be a pdf, where S is fixed, w_{u_i} given $w_{N(u_i)}$ is independent of w_{u_j} given $w_{N(u_j)}$, for $N(u_i) = \{s_{b_j} \in S, u_i \in b_j\}$ $\forall i = 1, \ldots, l$ and proper normal densities $\widetilde{p}(w_{u_i}|w_{N(u_i)})$. Then the finite-dimensional distributions with pdf $\widetilde{p}(w_V)$ support a valid random field w_V for all $V \subset \Re^2$, that is, they satisfy the Kolmogorov's conditions of symmetry and consistency.*

Following the NNGP we could have chosen $N(u_i)$ to be the m nearest neighbors of u_i in S. Nevertheless, henceforth $N(u_i)$ comprises the observed locations in the block where u_i belongs in the spatial domain D, therefore, $N(u_i)$ depends on the same observations in S for any order of U. Hence, lemma 1 defines a new valid spatial process and the next theorem proves that such spatial process derived from a GP is also a GP.

Theorem 1. *For any finite set $V \in D$, $\widetilde{p}(w_V)$ in Equation (5.4) is the finite dimensional density of a Gaussian process, called block-NNGP, with cross covariance funcion*

$$\widetilde{C}_{v_i,v_j} = \begin{cases} \widetilde{C}_{s_i,s_j} & \text{if } (v_1 = s_i, v_2 = s_j) \in S \\ B_{v_1}\widetilde{C}_{N(v_1),s_j} & \text{if } v_1 \notin S \text{ and } v_2 = s_j \in S \\ \delta_{(v_1=v_2)}F_{v_1} + B_{v_1}\widetilde{C}_{N(v_1),N(v_2)}B_{v_2}^T & \text{if } (v_1, v_2) \notin S, \end{cases}$$

where $\widetilde{C}_{m,n}$ is the covariance matrix of $\widetilde{C}_S$.
The block-NNGP contains existing processes as special cases. If we consider one observation per block and nb is the number of "past" nearest neighbor observations, the NNGP with S being the set of all observed locations is a particular case of block-NNGP. Also when $N(b_k) = \emptyset, \forall k$ each block $w_{N(b_k)}$ is independent from the other blocks, that is, $w_{N(b_k)} \perp w_{N(b_j)}, \forall k \neq j$, and we say that the spatial process is composed by independent blocks (Stein, 2013).

Corollary 2. *The block-NNGP with $M = n$ and $nb = m$ recovers the NNGP when S is the set of all observed locations.*

Corollary 3. *The block-NNGP with M blocks and $nb = 0$ recovers the independent blocks approach.*

Following previous blocking strategies (Kim et al., 2005; Eidsvik et al., 2014), the spatial domain can be partitioned into several regions, either using a regular block design (Figure (2)a) or an irregular block design (Figure (2)b). If the observed locations are approximately uniformly distributed over the domain D, the partitions can simply be obtained by splitting the spatial domain into M subregions of approximately equal area. If the observation locations are far from uniform, more complicated partitioning schemes might be necessary to achieve fast inference. In our approach, for the regular block design we fixed the number of blocks, and each block can have different number of observations. While for the irregular block design, we have fixed the number of observations per block. Of course, different block designs can also be implemented, for instance Voronoi/Delaunay designs (Eidsvik et al., 2014).

3 Bayesian estimation for block-NNGP

Let $Y = (Y(s_1), \ldots, Y(s_n))$ be a realization of a spatial stochastic process defined for all $s_i \in D \subset \Re^2$, $i = 1, \ldots, n$. The basic geostatistical Gaussian regression model is of the form

$$Y(s_i) = X'(s_i)\beta + w(s_i) + \epsilon(s_i),$$

where β is a coefficient vector (or regression parameter), X is a a vector of covariates, $w(s)$ is a spatial structured effect, it captures the spatial association, and $\epsilon(s_i) \sim N(0, \tau^2)$ models the measurement error. Thus, $y|\beta, w, \tau^2 \sim N(X\beta + w, D(\tau^2))$, where D is a diagonal matrix with entries τ^2. Full Bayesian specification is available if we assing priors to β, w, τ, and hyperparameters. Hence, instead of the Gaussian process prior for w, we assume that $w \sim$ block-NNGP$(0, \widetilde{C}(\theta))$, and also we assume

$\beta \sim N(\mu_\beta, V_\beta)$ and $\theta^\star = (\phi, \sigma^2, \tau^2) \sim \pi(\theta^\star)$. So, the joint posterior distribution is given by

$$p(\theta^\star, \beta, w|y) \propto p(\theta^\star) \times p(\beta|\mu_\beta, \Sigma_\beta) \times p(w|0, \widetilde{C}(\theta)) \times p(y|X\beta + w, D(\tau^2)). \quad (5.5)$$

In particular, we assume that $S = \{s_1, \ldots, s_n\}$ is the set of locations where the outcomes have been observed and $S' = S$, then for estimation $w = w_S$ and $\widetilde{C}(\theta) = \widetilde{C}_S(\theta)$ in Equation (5.5).

The Markov Chain Monte Carlo (MCMC) implementation usually requires updating the n latent spatial effects w sequentially, in addition to the regression and covariance parameters (for instance, see Datta et al. (2016)). Finley et al. (2017) studied the convergence for very large spatial datasets using NNGP to prove that such sequential updating of the random effects often leads to very poor mixing in the MCMC. To overcome this issue they proposed the Collapsed MCMC NNGP, which in summary performs Gibbs Sampling and random walk Metropolis steps to update β and θ, respectively, and then recover w and predictions y_0 using composition sampling.

The Collapsed MCMC for block-NNGP follows the steps: (i) update $\theta^\star$ through Random walk Metropolis-Hastings (MH). The target log-density is $p(\theta^\star|y) \propto p(\theta^\star) \times N(y|X\beta, \Sigma_{y|\beta,\theta})$; where $\Sigma_{y|\beta,\theta} = \widetilde{C}_S + D$; (ii) Gibb's sampler update for β, from the full conditional $\beta|y \sim N(Bb, B)$ where $B = (\Sigma_\beta^{-1} + X^T\Sigma_{y|\beta,\theta}X)^{-1}$ and $b = \Sigma_\beta^{-1}\mu_\beta + X^T\Sigma_{y|\beta,\theta}^{-1}y$; (iii) Recover $w_S|\theta^\star, \beta$ for each post-burn in MCMC sample; $w_S|\beta, \theta^\star, y \sim N(Ff, F)$, where $F = (\widetilde{C}_s^{-1} + D^{-1})^{-1}$ and $f = D^{-1}(y - X\beta)$.

Spatial prediction can be carried out after parameter inference. Conditioning on a particular estimated value of the parameters (θ, β), spatial prediction amounts to finding the posterior predictive distribution at a set of prediction locations u_i, that is, $p(y(u_i)|y)$. Note that we consider all observed data for estimation, thus S comprises the observed locations, while the new location points for predictions belong to the finite set U. Furthermore, since the components of $w_U|w_S$ are independent, we can update $w(u_i)$ for each $i = 1, \ldots, l$, from $p(w(u_i)|w_S, \beta, \theta^\star, y \sim N(m, v)$, where $m = C_{u_i,N(u_i)}^T C_{N(u_i),N(u_i)}^{-1} w(N(s_0))$ and $v = \sigma^2 - C_{u_i,N(u_i)}^T C_{N(u_i),N(u_i)}^{-1} C_{u_i,N(u_i)}$. Block NNGP are especially useful here as posterior sampling for w_U is cheap because their components are independent and each $w(u_i)$ is only based on the observations that lie in the block that it belongs. Now using the posterior samples of $w(u_i)$, the posterior predictive sampling $y(u_i)|w_U, w_S, \beta, \theta^\star, y \sim N(X(u_i)^T\beta + w(u_i), \tau^2)$.

Our approach does not need to store $n \times n$ dense distance matrices, it stores M "small" dense matrices. It is scalable to massive datasets, we can compute the precision matrix from the block-NNGP using faster (parallel) computation for the defined blocks. For shared Memory, good parallel libraries are available, such as the

multi-threaded BLAS/LAPACK libraries included in Microsoft R Open and parallel Packages in R like the doMC Package (Calaway et al., 2017).

4 Simulation Studies

To assess the performance of the block-NNGP models, we present the next simulation experiments. We generate a spatial process with n = 2500 observation sites on a spatial domain $(0,1) \times (0,1)$. The covariates are $X(s_i) = (1, x_i)$, $x_i \sim (N(0,1))$ with true regression parameters $\beta = (1,5)^T$. We use an exponential covariance function $C(h) = \sigma^2 \exp(\|s_i - s_j\|)$ with $\sigma^2 = 1$. The so-called effective range (r), the distance at which the correlation decays to 0.1, is studied using simulation scenarios, (i) SIM I: $r = 0.16$ ($\phi = 12$) (ii) SIM II: $r = 0.33$ ($\phi = 6$), and (iii) SIM III: $r = 0.67$ ($\phi = 3$), where $\phi = \sqrt{(8 \times \nu)}/range$ (with $\nu = 0.5$) is called the spatial decay. For all locations we considered $\tau^2 = 0.1$.

Let S be the set of $n = 2000$ observed locations and U the set of the remaining 500 observations used to assess predictive performance. We fit the models: (i) full Gaussian process (full GP), (ii) block-NNGP models with $M = n$ for $nb = 10$ and $n = 20$, which by Corollary 4.1 is equivalent to the NNGP model with 10 and 20 neighbors respectively, (iii) regular (R) block-NNGP models and (iv) irregular (I) block models. We vary the number of spatial blocks to investigate the way blocking schemes influences the estimation and prediction capabilities. We use regular blocks and irregular blocks (Figure (S1)). The regular blocks have the same of size. The number of blocks $M = n_m \times n_m$, for instance, $3^2, 5^2, 7^2$, and 10^2. A similar configuration was also used in Eidsvik et al. (2014). Our irregular blocks design requires grouping approximately n/M observations per block, so the region D is subdivided into M irregular regions. In the regular case, we are also able to know the number of observations per block (n_k), but our main concern comes when the observed locations are not uniformly distributed over the domain D because the (n_k) will be very different for each block k, and for some blocks it will be expensive to perform matrices operations. On the other hand, with irregular blocks we can control the approximated number of observations per block and the sparsity of $\widetilde{C}_S^{-1}$ (Figure (S2)). In both cases the maximum number of blocks should be constrained by some prior information about the range of the process. Although there might not be an explicit number of blocks and neighbor blocks for optimal blocking, we will determine them by the computational speed as well as statistical efficiency, maximizing the number of blocks.

The parameters of the models are estimated from a Bayesian point of view, so we run

the MCMC for a small number of iterations (1000) to determine the "best" number of blocks in terms of less time. Figure (2) shows that for this configuration and different values of ϕ, the time does not significatly decrease for $M > 9^2$. We also test the WAIC and LPML to study the goodness of fit for different number of blocks and neighbor blocks, but we did not get any pattern. Then full posterior inference for subsequent analysis was based upon one chain of 25000 iterations (with a burn-in of 5000 iterations). In particular, the collapsed MCMC method (Finley et al., 2017) was adapted to the block-NNGP. We use flat prior distributions for β, for σ^2 we assigned inverse Gamma $IG(2, 1)$ prior, for τ^2 we assigned $IG(2, 0.1)$ prior, and for the spatial decay ϕ we assigned a uniform prior $U(2, 30)$ which is equivalent to a range between approximately 0.067 and 1 units. We also used a parameterization on the real line, with log variance, log precision and log range parameters.

Parameter estimates and performance metrics for the models proposed when $\phi = 12$ are provided in Table (1). In all cases, the mean posterior estimates for block-NNGP are very close to the full-GP mean posterior estimates. The goodness of fit and predictive performance for all models are very similar. The number of neighbors of block-NNGP models with irregular blocks ($n_{bk} = 40 - 1$ for $M = 100, nb = 1$ and $n_{bk} = 30 - 1$ for $M = 200, nb = 2$) is higher than the number of neighbors of NNGP models (10 and 20). Nevertheless these block-NNGP models are faster and they also show a slightly better performance to fit the data, thus it has more information about the process without increaing the computational cost. In fact, Figure (3) shows the similarity of estimations of w_S, interpolated over the domain, between all block-NNGP models and Full GP. We conclude that for this simulated data we detect no differences between the block-NNGP models, and they fit the data very well when the range is very small.

To study the statistical efficiency when the effective correlation length increases, while keeping the domain fixed, we evaluate the performance of the proposed models when $\phi = 6$. In Table (2), it is observed that estimations of the block-NNGP models closely approximate to the ones of full GP model. Figure (4) shows the posterior mean estimates of the spatial random effects interpolated over the domain. As illustrated in Figure (4), the block-NNGP models can result in considerably better approximations, specially for M=225 and $nb = 2$. The LPML and WAIC values suggest that the block-NNGP models are the best to fit the data. Computing times requirements for NNGP and block-NNGP models are similar, but as we expected lower than the full GP model time.

Further comparisons show that the mean posterior estimates of σ^2, ϕ and β_0 for the NNGP model with 20 neighbors and full GP are a little different (Table (2)). We might think that if we increase the number of neighbors, the estimation of

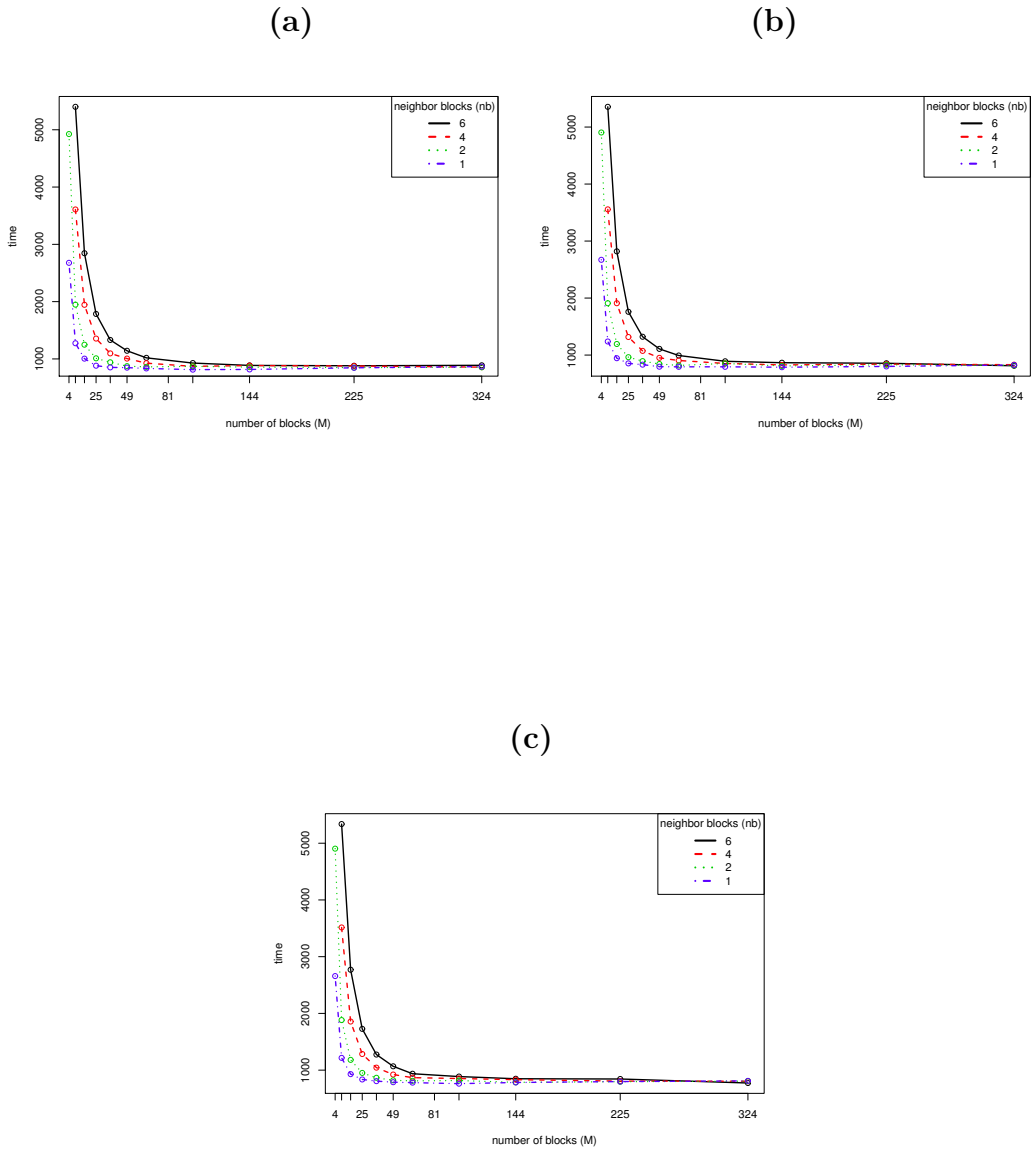

Figure 2: MCMC time for block-NNGP models running 1000 iterations, for regular blocks. (a) SIM I ($\phi = 12$), (b) SIM II ($\phi = 6$) and (c) SIM III ($\phi = 3$).

Table 1: SIM I ($\phi = 12$) Summary of mean parameter estimates. Parameter posterior summary (2.5, 97.5) percentiles.

Model		**Full GP**	**NNGP (20)**	**NNGP (10)**	(R)**M=100 nb=1**	(R)**M=64 nb=1**	(I)**M=100 nb=1**	(I)**M=200 nb=2**
σ^2	1	0.99 (0.79, 1.36)	1.07 (0.91, 1.47)	1.04 (0.89, 1.32)	1 (0.83, 1.31)	0.99 (0.81, 1.27)	0.92 (0.78,1.12)	0.94 (0.77,1.18)
τ^2	0.1	0.1 (0.08, 0.12)	0.09 (0.07, 0.11)	0.09 (0.07, 0.11)	0.1 (0.08, 0.12)	0.1 (0.08, 0.12)	0.1 (0.08,0.12)	0.1 (0.08,0.12)
ϕ	12	13.74 (9.48, 17.74)	13.09 (9.08, 14.93)	13.64 (10.2, 14.91)	13.52 (9.87, 16.95)	13.6 (10.15, 17.38)	14.79 (11.67, 17.98)	14.22 (10.56, 17.86)
β_0	1	1.09 (0.79, 1.49)	1.18 (0.85, 1.61)	0.98 (0.67, 1.29)	1.12 (0.91, 1.37)	1.12 (0.89, 1.39)	0.72 (0.52,0.91)	0.92 (0.69,1.19)
β_1	5	5.01 (4.99, 5.03)	5.01 (4.99, 5.03)	5.01 (4.99, 5.03)	5.01 (4.99, 5.04)	5.01 (4.98, 5.03)	5.01 (4.99,5.04)	5.01 (4.99,5.03)
LPML		-31084.36	-35783.45	-36204.53	-30747.6	-30260.92	**-30245.45**	**-29973.36**
WAIC2		184256.5	228406.8	232569.2	181016	176769.5	**176363.1**	**174723.1**
G		66.46808	58.47473	57.72994	65.74294	67.48258	66.19069	69.10049
P		329.3224	304.9878	303.4614	335.6308	338.47	340.2987	337.6664
D		395.7905	**363.4625**	**361.1913**	401.3738	405.9526	406.4894	406.7669
RMSPE		–	0.562189	**0.5506445**	0.5674386	0.5875377	0.5636098	**0.5569855**
Accep		23.73333	34.45	35.74	23.13333	23.63	23.36333	23.69333
time (sec)		31637.95	23915.9	23357.57	23758.02	24683.79	**22990.89**	**22915.74**

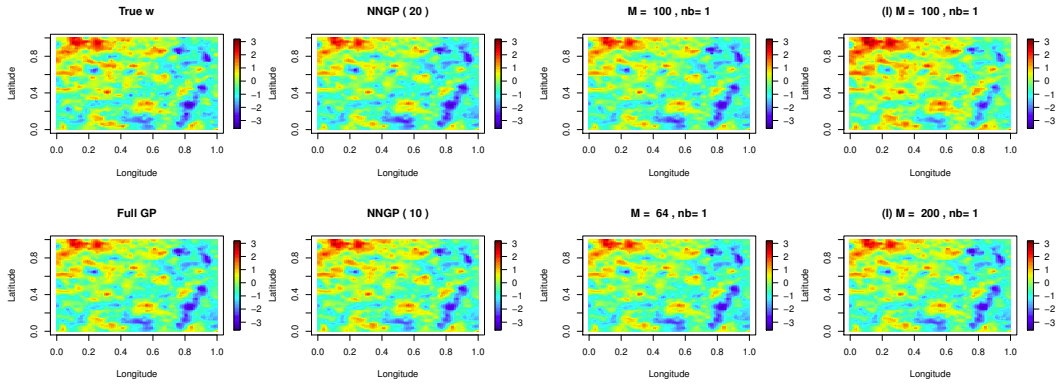

Figure 3: SIM I ($\phi = 12$). True spatial random effects and posterior mean estimates for different models.

parameters using the NNGP model should be better, but this is not guaranteed as we can see from this simulation. In fact, Figure (4) also shows that the NNGP model with 20 neighbors did not approximate well the spatial field of the full GP model. The patterns differ greatly from the original spatial random field and the one estimated using the full-GP. Otherwise the block-NNGP model with $M = 64$ and $nb = 1$ has bigger blocks but the estimation is improved without increasing the computing time requirements drastically. So, although the NNGP has proven to be successful in capturing local/small-scale variation of spatial processes, it might have one disadvantage: inaccuracy in representing global/large scale dependence. This might happen because the NNGP built the DAG based on observations, where the locations are ordered by one of the coordinates. Adversely, the block-NNGP chain graph is based on blocks of observations, which captures both small and large dependence.

Table 2: SIM II ($\phi = 6$) Summary of mean parameter estimates. Parameter posterior summary (2.5, 97.5) percentiles, n = 2000.

Model		**Full GP**	**NNGP (20)**	**NNGP (10)**	(R)**M=64 nb=1**	(R)**M=144 nb=6**	(I)**M=100 nb=2**	(I)**M=100 nb=1**
σ^2	1	1.35 (0.83, 2.12)	1.72 (1.01, 2.31)	1.01 (0.74, 1.66)	1.38 (0.95, 2.05)	1.08 (0.75, 1.93)	1.06 (0.78,1.12)	0.96 (0.75,1.31)
τ^2	0.1	0.14 (0.09, 0.12)	0.1 (0.08, 0.12)	0.1 (0.08, 0.11)	0.11 (0.09, 0.12)	0.1 (0.09, 0.12)	0.11 (0.09,0.12)	0.11 (0.09,0.12)
ϕ	6	4.93 (3.14, 8.17)	3.88 (3.03, 6.96)	6.91 (4.03, 9.87)	4.67 (3.11, 7.04)	6.26 (3.35, 9.28)	6.2 (3.75, 8.75)	6.85 (4.76,9.07)
β_0	1	1.5 (0.77, 2.67)	1.97 (1.05, 3.23)	0.87 (0.36, 1.42)	1.43 (0.93, 2.16)	1.11 (0.55, 1.91)	1.03 (0.57,1.7)	0.61 (0.31,0.94)
β_1	5	5.01 (4.99, 5.03)	5.01 (4.99, 5.03)	5.01 (4.99, 5.03)	5.01 (4.99, 5.03)	5.01 (4.99, 5.03)	5.01 (4.99, 5.03)	5.01 (4.99,5.03)
LPML		-26101.92	-27891.44	-29292.24	**-24944.26**	-26345.72	**-25364.79**	-25435.02
WAIC2		146979.3	161696.3	173963.3	**137423.1**	148797.4	140935.4	**140785.8**
G		98.60967	94.5057	89.47476	101.9886	97.27277	100.5962	98.87815
P		313.3962	303.9089	298.1956	324.1011	312.4023	319.8328	324.3391
D		412.0059	398.4146	**387.6703**	426.0897	409.6751	420.429	423.2173
RMSPE		–	0.7678724	**0.4926092**	0.5830725	0.5996594	0.5310345	**0.4934768**
Accep		28.69	32.61333	28.59	29.14	26.50333	26.19	25.17333
time (sec)		31677.58	23896.63	**23423.89**	23840.58	24166.43	23867.53	**22746.49**

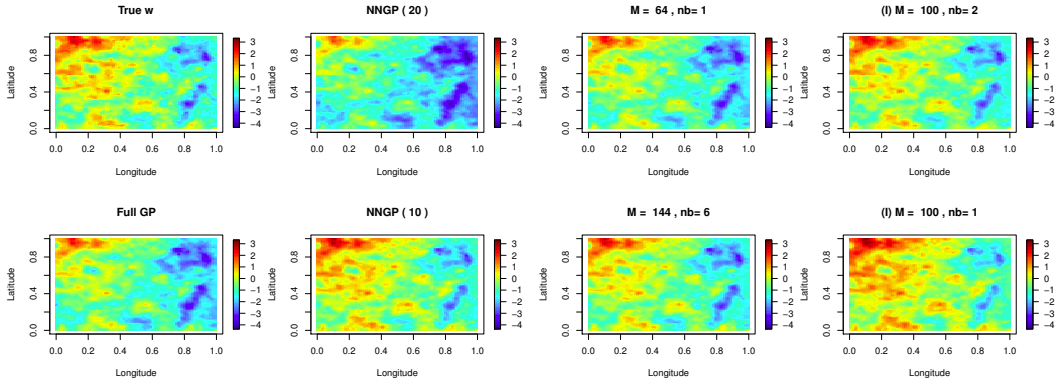

Figure 4: SIM II ($\phi = 6$). True spatial random effects and posterior mean estimates for different models.

Table (3) provides parameter estimates and performance metrics for all models when $\phi = 3$. It is observed that estimations of the block-NNGP models closely approximate to the ones of full GP model, except the block-NNGP model with $M = 100$ and $nb = 1$. Figure (5) shows the posterior mean estimates of the spatial random effects interpolated over the domain. We can see that the block-NNGP models result in considerably better approximations, specially for $M = 225$ and $nb = 2$. The LPML and WAIC values support also this statement. Computing times requirements for NNGP and block-NNGP models are similar.

Further comparisons show that mean posterior estimates of β_0 for NNGP and full GP are a little different (Table (3)). Figure (5) shows that the spatial random effects with 20 neighbors is too smooth. Also we can see that the map for the NNGP-model and the block-NNGP model with $M = 100$ and $nb = 1$ are very similar to the true process, but different of the full GP model. In NNGP and the block-NNGP models

the number of neighbors is small, so we might think that if we increase the number of neighbors, the estimation of parameters using the NNGP model should be better, but this is not guaranteed as we can saw from simulation with $\phi = 6$. And if we use more neighbors than the "necessary" the model oversmooth the spatial process. In general, if the block-NNGP models has more neighbors per obervation, that is more neighbor blocks, the block-NNGP process is more similar to the GP Full process without increasing the computing time requirements drastically.

Table 3: SIM III ($\phi = 3$): Summary of mean parameter estimates. Parameter posterior summary (2.5, 97.5) percentiles, n = 2000.

Model		**Full GP**	**NNGP (20)**	**NNGP (10)**	(R)**M=225 nb=2**	(R)**M=324 nb=2**	(I)**M=100 nb=1**	(I)**M=200 nb=2**
σ^2	1	1.37 (0.86, 1.79)	2.65 (1.36, 3.55)	1.12 (0.66,2.68)	2.03 (0.97, 3.07)	2.37 (1.38, 3.21)	0.97 (0.69, 1.65)	1.98 (0.96, 1.04)
τ^2	0.1	0.1 (0.09, 0.12)	0.1 (0.09, 0.12)	0.1 (0.09,0.12)	0.11 (0.1, 0.12)	0.11 (0.1, 0.12)	0.11 (0.1, 0.12)	0.11 (0.1, 0.12)
ϕ	3	2.41 (2.01,3.97)	1.25 (1.01, 2.48)	3.03 (1.22,5.46)	1.48 (1.03, 3.17)	1.26 (1.01, 2.22)	3.22 (1.76, 4.7)	1.54 (1.04, 3.33)
β_0	1	1.95 (0.85, 3.27)	3.09 (1.29, 5.33)	0.8 (-0.18,1.83)	1.55 (0.35,3.16)	1.81 (0.44, 3.39)	0.67 (0.2, 1.25)	1.73 (0.68, 3.29)
β_1	5	5.01 (4.99,5.03)	5.01 (4.99,5.03)	5.01 (4.99,5.03)	5.01 (4.99, 5.03)	5.01 (4.99, 5.03)	5.01 (4.99, 5.03)	5.01 (4.99, 5.03)
LPML		-21883.68	-22780.12	-23219.55	-20010.68	-20096.75	-20784.85	-21050.7
WAIC2		115927.4	122610.6	126136.8	102370.9	102763.3	107298.9	110547.8
G		127.6543	126.3956	123.6112	138.2599	137.6801	132.412	132.6907
P		291.0201	285.967	285.0954	299.7904	300.6322	303.5624	295.7998
D		418.6744	412.3627	408.7066	438.0503	438.3123	435.9745	428.4904
RMSPE		–	1.077448	0.5020536	0.8271246	0.8799603	0.4356712	0.8119286
Accep		32.72667	32.31667	24.49	30.09667	32.02	23.82	29.16667
time (sec)		32814.86	23061.93	23760.52	23284.51	24529.85	23973.31	23912.96

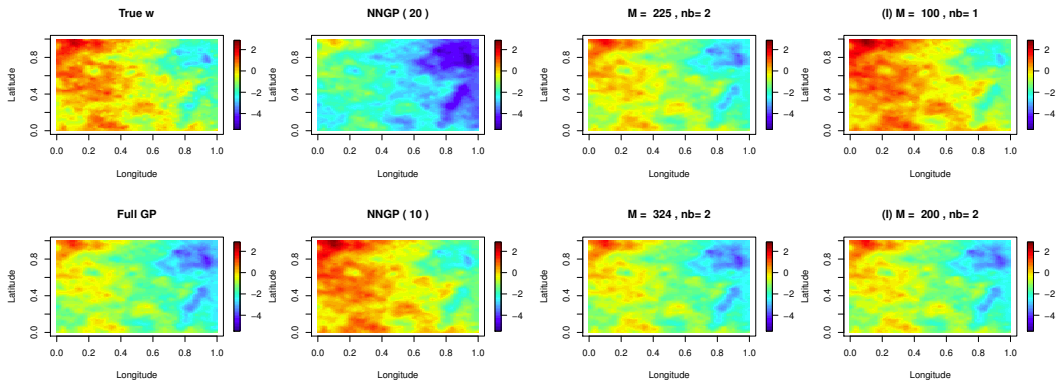

Figure 5: SIM III ($\phi = 3$). True spatial random effects and posterior mean estimates for different models.

5 Application

In this section, we illustrate the application of block-NNGP to large spatial data from the mining industry. In the process of extracting ore, stability is crucial because

it is one of the key characteristics that influence the success of underground mining work. If it is not possible to produce ore above cut-off at stable conditions, the ore is made inaccessible, which often results in lost production. To assess the stability of the rock mass it is studied the spatial joint frequency distribution in a mine because a joint is a planar or semiplanar discontinuity in a rock mass and represents zones of weakness in the rock mass (Ellefmo and Eidsvik, 2009).

Here we study joint-frequency data in an iron mine in the northern part of Norway to estimate the most probable joint frequency at unsampled locations. Eidsvik et al. (2014) aggregated the raw joint data along the boreholes, thus we have the total number of 11,701 measurements. Then they transformed the data, the logarithm of the joint-frequency observations are standardized. In Figure (6), we display locations of the measurements (east, north) of the joint-frequency data. The depth of boreholes is used as covariate, along with an intercept. More references about these data can be found in (Ellefmo and Eidsvik, 2009) and Eidsvik et al. (2014).

We first divide the joint-frequency data in two subsets, the set S composed by a random subset of 11000 observed locations and the remaining 701 observations were withheld to assess predictive performance, so they belong to the set U. We fit the block-NNGP models with different number of regular blocks and different neighbor blocks. We only run the MCMC for 1000 iterations to choose between these models, thus we choose the model with $M = 289$ blocks and $nb = 1$ block (Figure (6)). Then full Bayesian inference and posterior inference were based upon 10000 iterations. We use flat prior distributions for β, for σ^2 we assigned inverse Gamma $IG(2, 1)$ prior, for τ^2 we assigned $IG(2, 0.1)$ prior, and for the spatial decay ϕ we assigned a uniform prior $U(0.001, 2)$ which is equivalent to a range between approximately 1 and 2000 m.

From the parameter estimates, the mean effective spatial range is approximately 29m ($\phi = 0.07$), the nugget effect equal to 0.1 and the marginal variance equal to 0,16. These results are very similar to the parameter estimates of block composite likelihood proposed by Eidsvik et al. (2014) using a Matérn covariance function with $\nu = 3/2$. Figure (7) shows a maps of posterior estimates for the spatial random effect and interpolated posterior predictive mean of joint-frequency data. Comparing to Figure (6), it is easy to see that our estimations are rather accurate.

6 Discussion

We have presented the block-NNGP, a new GMRF for approximating Gaussian processes with any covariance function. The precision matrix of the block-NNGP has

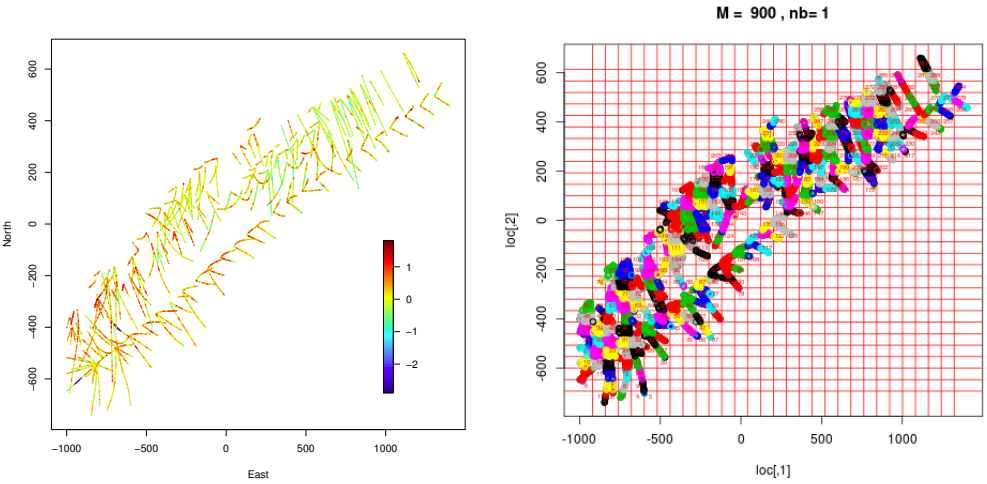

Figure 6: Left: Joint-frequency data, $n = 10701$ locations. Right: Regular blocks for these data.

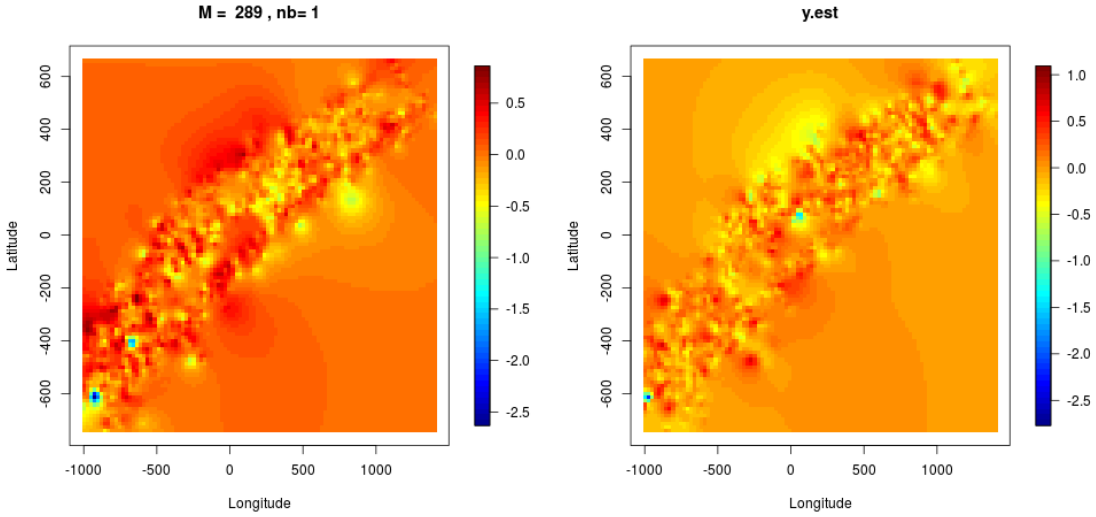

Figure 7: Left: Mean Posterior of w_S. Right: Mean posterior of joint-frequency data.

a block-sparse structure, which allows scalable inference and distributed computations. It is one of the methods in the state-of-the-art for large spatial data and can be viewed as a general case of the NNGP (with M = n) of Datta et al. (2016). The results for block-NNGP and NNGP are very similar for small ranges of the spatial random field. In addition, it improves the NNGP when the range is not too small. Using theoretical results, a toy example, large simulated datasets, and a real-data application, we have shown that the block-NNGP can provide a better approximation at the same or lower computational complexity and computation time. It should also be noted that our inference results for $M \neq n$ provide an algorithm for parallel blocks and distributed computations for inference. The block-NNGP not only approximates the data precision matrix to a sparse precision matrix, but it is also a valid Gaussian process in its own right. Extensions to more complicated scenarios are therefore possible by assuming different sets S and U, or chain graphs. Finally, we remark that a more sophisticated implementation would allow more speed-up for the block-NNGP model, using a parallel for-loop and running matrix decompositions in parallel. This is future work.

Acknowledgments

The first author would like to thank ProUNI, and the second author would like to thank FAPEMIG and CNPq, for partial financial support.

Bibliography

Bolin, D. and Wallin, J. (2016). Spatially adaptive covariance tapering. *Spatial Statistics*, 18(Part A):163 – 178.

Calaway, R., Analytics, R., and Weston, S. (2017). domc: Foreach parallel adaptor for "parallel". Technical report.

Caragea, P. C. and Smith, R. L. (2007). Approximate likelihoods for spatial processes. *Journal of Multivariate Analysis*, 98 (7):1417–1440.

Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E. (2016). Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*, 111(514):800–812.

Eidsvik, J., Shaby, B. A., Reich, B. J., Wheeler, M., and Niemi, J. (2014). Estimation and prediction in spatial models with block composite likelihoods. *Journal of Computational and Graphical Statistics*, 23(2):295–315.

Ellefmo, S. and Eidsvik, J. (2009). Local and spatial joint frequency uncertainty and its application to rock mass characterisation. *Rock mechanics and rock engineering*, 42(4):667–688.

Finley, A. O., Datta, A., Cook, B. C., Morton, D. C., Andersen, H. E., and Banerjee, S. (2017). Applying nearest neighbor Gaussian processes to massive spatial data sets: Forest canopy height prediction across Tanana Valley Alaska. (Under review).

Katzfuss, M. and Guinness, J. (2017). A general frameprocessesVecchia approximations of gaussian processes. (Under review).

Kim, H.-M., Mallick, B. K., and Holmes, C. C. (2005). Analyzing nonstationary spatial data using piecewise gaussian processes. *Journal of the American Statistical Association*, 100(470):653–668.

Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: The SPDE approach. *Journal of the Royal Statistical Society, Series B*, 73(4):423–498.

Rue, H. and Tjelmeland, H. (2002). Fitting Gaussian Markov random fields to Gaussian fields. *Scandinavian Journal of Statistics*, 29(1):31–50.

Stein, M. L. (2013). Statistical properties of covariance tapers. *Journal of Computational and Graphical Statistics*, 22(4):866–885.

Stein, M. L., Chi, Z., and J.Welty, L. (2004). Approximating likelihoods for large spatial data sets. *Journal of the Royal Statistical Society, Series B*, 66(2):275–296.

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

Appendix

A: Proofs of main results

Proof of Proposition 1. If $p(w_S)$ is a valid multivariate joint density, $p(w_{b_k}|w_{N(b_k)})$ is also proper, and we have that $\int p(w_{b_k}|w_{N(b_k)})dw_{b_k} = 1, \forall k = 1, \ldots, M$.
From the definitions of G and G_b there exists a set of nodes $s_{\pi(b_1)}$ in G, such that "the last node" from a DAG G_b belongs to $s_{\pi(b_1)}$. Then the nodes in $s_{\pi(b_1)}$ do not have any directed edge originating from them. As consequence, any node in block $\pi(b_1)$ can not belong to the set of nodes of any other block. So the term in Equation (5.2) where all locations of $\pi(b_1)$ appear is $p(w_{\pi(b_1)}|w_{N(\pi(b_1))})$. From Fubini's theorem, we can interchange the product and integral, thus

$$\begin{aligned}\int p(\widetilde{p}(w_S))dw_S &= \int \ldots \int \prod_{i=1}^{M} p(w_{\pi(b_i)}|w_{N(\pi(b_i))})dw_{\pi(i)} \\ &= \int \ldots \int \prod_{i=2}^{M} p(w_{\pi(b_i)}|w_{N(\pi(b_i))})dw_{\pi(i)}.\end{aligned}$$

Then, removing any node of $\pi(b_1)$ from G and G_b, we have the chain graph G' and DAG G'_b, respectively. There exists another set of nodes $s_{\pi(b_2)}$ in G', such that "the last node" from a DAG G'_b belongs to $s_{\pi(b_2)}$. Then the nodes $s_{\pi(b_2)}$ do not have any directed edge originating from them. As consequence, any node in block $\pi(b_2)$ can not belong to the set of nodes of any other block. So the term in Equation (5.2) where all locations of $\pi(b_2)$ appear is $p(w_{\pi(b_2)}|w_{N(\pi(b_2))})$. Applying the Fubini's theorem again,

$$\int p(\widetilde{p}(w_S))dw_S = \int \ldots \int \prod_{i=3}^{M} p(w_{\pi(b_i)}|w_{N(\pi(b_i))})dw_{\pi(i)}.$$

In a similar way, we find $s_{\pi(b_3)}, \ldots, s_{\pi(M)}$, such that,
$\int p(\widetilde{p}(w_S))dw_S = \int \prod_{i=1}^{M} p(w_{\pi(b_i)}|w_{N(\pi(b_i))})dw_{\pi(i)} = 1.$ □

Matrix Analysis Background
Theorem A1: A matrix $B \in \Re^{m\times n}$ is full column rank if and only if B^TB is invertible
Theorem A2: The determinant of an $n \times n$ matrix B is 0 if and only if the matrix B is not invertible.
Theorem A3: Let T_n be a triangular matrix (either upper or lower) of order n. Let $\det(T_n)$ *be the determinant of T_n. Then* $\det(T_n)$ *is equal to the product of all the diagonal elements of T_n, that is,* $\det(T_n) = \prod_{k=1}^{n}(a_{kk})$.
Proposition A1: If B is positive definite (p.d.), then if S has full column rank, then

$S^T BS$ is positive definite.
Corollary A1: If B is positive definite, then B^{-1} is positive definite.

Proof of Proposition 2. Without loss of generality, assume that the data were re-ordered by blocks. From known properties of Gaussian distributions, $w_{b_k}|w_{N(b_k)} \sim N(B_{b_k} w_{N(b_k)}, F_{b_k})$, where $B_{b_k} = C_{b_k,N(b_k)} C_{N(b_k)}^{-1}$ and $F_{b_k} = C_{b_k} - C_{b_k,N(b_k)} C_{N(b_k)}^{-1} C_{N(b_k),b_k}$. Hence,

$$\begin{aligned}
\widetilde{p}(w) &= \prod_{k=1}^{M} p(w_{b_k}|w_{N(b_k)}) \\
&\propto \prod_{k=1}^{M} \frac{1}{|F_{b_k}|^{1/2}} \exp\left\{-\frac{1}{2}(w_{b_k} - B_{b_k} w_{N(b_k)})^T F_{b_k}^{-1} (w_{b_k} - B_{b_k} w_{N(b_k)})\right\} \\
&\propto \frac{1}{\prod_{k=1}^{M} |F_{b_k}|^{1/2}} \exp\left\{-\frac{1}{2}\sum_{k=1}^{M}(w_{b_k} - B_{b_k} w_{N(b_k)})^T F_{b_k}^{-1} (w_{b_k} - B_{b_k} w_{N(b_k)})\right\}.
\end{aligned}$$

Let $w_{b_k} - B_{b_k} w_{N(b_k)} = B_{b_k}^{\star} w_S$, and j be the $j-th$ observation of block b_k, then $\forall k = 1, \ldots, M$, $i = 1, \ldots, n$ and $j = 1, \ldots, n_{bk}$:

$$B_{b_k}^{\star}(j,i) = \begin{cases} 1 & \text{if } s_i \in s_{b_k} \\ B_{b_i}[j,l] & \text{if } s_i \in s_{b_k}; s_i = s_{N(b_k)}[l]; l = 1, \ldots, N_{bk} \\ 0 & \text{otherwise,} \end{cases}$$

and

$$B_{b_k}^{\star} = \begin{bmatrix} B_{b_k}^{\star}(1) \\ \vdots \\ B_{b_k}^{\star}(j) \\ \vdots \\ B_{b_k}^{\star}(n_{b_k}) \end{bmatrix}_{n_{b_k} \times n}.$$

From these definitions, $B_{b_k}^{\star}$ is a matrix with i-th column full of zeros if $s_i \notin s_{b_k}$ or $s_i \notin N(s_{b_k})$. Since the data were reordered by blocks and the neighbor blocks are from the past, $B_{b_k}^{\star}$ has the next form:

$$B_{b_k}^{\star} = [\ R_k\ \underline{A_k}\ 0\ \ \ldots\ \ 0\ ],$$

where A_k is a $n_{bk} \times n_{bk}$ matrix and R_k is a $n_{bk} \times \sum_{r=1}^{k-1} n_{br}$ matrix with at least one column with none null-element if $nb \neq 0$.

Then,

$$\begin{aligned}
\widetilde{p}(w) &\propto \frac{1}{\prod_{k=1}^{M}|F_{b_k}|^{1/2}}\exp\left\{-\frac{1}{2}\sum_{k=1}^{M}(B^{\star}_{b_k}w_S)^T F^{-1}_{b_k}(B^{\star}_{b_k}w_S)\right\}\\
&\propto \frac{1}{\prod_{k=1}^{M}|F_{b_k}|^{1/2}}\exp\left\{-\frac{1}{2}\sum_{k=1}^{M}w_S^T(B^{\star}_{b_k})^T F^{-1}_{b_k}(B^{\star}_{b_k}w_S)\right\}\\
&\propto \frac{1}{\prod_{k=1}^{M}|F_{b_k}|^{1/2}}\exp\left\{-\frac{1}{2}\sum_{k=1}^{M}w_S^T((B^{\star}_{b_k})^T F^{-1}_{b_k}B^{\star}_{b_k})w_S\right\}\\
&\propto \frac{1}{\prod_{k=1}^{M}|F_{b_k}|^{1/2}}\exp\left\{-\frac{1}{2}w_S^T(\sum_{k=1}^{M}(B^{\star}_{b_k})^T F^{-1}_{b_k}B^{\star}_{b_k})w_S\right\}.
\end{aligned}$$

Let $\sum_{k=1}^{M}(B^{\star}_{b_k})^T F^{-1}_{b_k}B^{\star}_{b_k} = (B^{\star}_s)^T F_s^{-1}B^{\star}_s$, where $B_s = [\ B^{\star}_{b1}\ \ \ldots\ \ \ldots\ \ \ldots\ \ B^{\star}_{bM}\ ]$ and $F_s^{-1} = \mathrm{diag}(F^{-1}_{b_k})$. F_s^{-1} is a block diagonal matrix and (iii) is proved. And given that $B^{\star}_{b_k}$ is a matrix with i-th column full of zeros for $i > \sum_{r=1}^{k} n_{br}$, then B_s is a block matrix and lower triangular, and (ii) is proved.
Finally, $\widetilde{p}(w) \propto \frac{1}{\prod_{k=1}^{M}|F_{b_k}|^{1/2}}\exp\left\{-\frac{1}{2}w_S^T(B_s^T F_s^{-1}B_s)w_S\right\}$ and $\widetilde{C}_s^{-1} = B_s^T F_s^{-1}B_s$.

$\widetilde{C}_s$ is positive definite
From properties of the Normal distribution, the covariance of the conditional distribution of $w_{b_k}|w_{N(b_k)}$ is also p.d. (by Schur complement conditions), then $F_{bi} = C_{b_i} - C_{b_i,N(b_i)}C^{-1}_{N(b_i)}C_{N(b_i),b_i}$, is p.d. Moreover, $F_s = diag(F_{bi})$ and a block diagonal matrix is p.d. iff each diagonal block is positive definite, so given that F_{bi} is p.d. and F_s is block diagonal with blocks F_{bi} p.d then F_s is p.d. By Corollary A1, F_s is p.d. then F_s^{-1} is p.d. By Theorem A1, B_s has full column rank iff $R_s = B_s^T B_s$ is invertible. By Theorem A2, the inverse of R_s exists iff $\det(R_s) \neq 0$. Using the well-known matrix theorems (Henderson and Searle 1981), we can prove the following: $\det(R_s) = \det(B_s^T B_s) = \det(B_s^T)\det(B_s) \neq 0$ if $\det(B_s^T) = \det(B_s) \neq 0$. Given that B_s is a lower triangular matrix, by Theorem A3, $\det(B_s) = \prod_{k=1}^{n}(b_{kk})$. And, $b_{kk} = 1, \forall k$, then $\det(B_s) \neq 0$. So, the R_s is invertible and B_s has full column rank. By Proposition A1, given that B_s has full column rank, and F_s^{-1} is p.d. then $\widetilde{C}_s^{-1} = B_s^T F_s^{-1}B_s$ is p.d. And by corollary A1, $\widetilde{C}_s^{-1}$ is p.d. then $\widetilde{C}_s$ is p.d. and (iv) is proved.

Since $\widetilde{p}(w_S) \propto \frac{1}{\prod_{k=1}^{M}|F_{b_k}|^{1/2}}\exp\left\{-\frac{1}{2}w_S^T(\widetilde{C}_s^{-1})w_S\right\}$, $\widetilde{C}_s^{-1} = B_s^T F_s^{-1}B_s$, and $\widetilde{C}_s$ is p.d., then $\widetilde{p}(w_S)$ is a pdf of a multivariate normal distribution and (i) is proved.

If $n_{bk} \ll n$ then $i > \sum_{r=1}^{k} n_{br}$ and $B^{\star}_{b_k}$ will be more sparse. Also, if n_k is small, the block diagonal matrix F_s^{-1} will be more sparse. As result, $\widetilde{C}_s^{-1} = B_s^T F_s^{-1}B_s$, will

still be sparse. □

Proof of Lemma 1. We need to prove that the finite dimensional distributions in Equation (5.4) are consistent with a stochastic process. The Kolmogorov consistency conditions are checked as follows:

Symmetry under permutation: Let $\pi_1, \ldots, \pi_n$ be any permutation of $1, \ldots, n$, note that S is fixed, then it is clear that $\widetilde{p}(w(v_1), \ldots, w(v_n)) = \widetilde{p}(w(v_{\pi_1}), \ldots, w(v_{\pi_n}))$ if and only if the same holds for the distribution of $u_i|N(u_i)$. Since $w_U|w_S$ follows a l-multivariate normal distribution, then the symmetry condition is satisfied by $p(w_U|w_S)$, and it holds that the next condition $\widetilde{p}(w(u_1), \ldots, w(u_l)|w_S) = \widetilde{p}(w(u_{\pi_1}), \ldots, w(u_{\pi_l})|w_S)$ is necessary and sufficient to prove the symmetry condition of $\widetilde{p}(w_V)$. To prove this we define the next pdfs,

$$\begin{aligned}\widetilde{p}(w(u_1), \ldots, w(u_l)|w_S) &= |2\pi F_U|^{-1/2} \exp\left\{-\frac{1}{2}(w_U - B_U w_S)^T F_U^{-1}(w_U - B_U w_S)\right\} \\ &= |2\pi F_U|^{-1/2} \exp\left\{Q(w_U)\right\},\end{aligned}$$

and

$$\begin{aligned}\widetilde{p}(w(u_{\pi_1}), \ldots, w(u_{\pi_l})|w_S) &= |2\pi \Sigma'|^{-1/2} \exp\left\{-\frac{1}{2}(w_{U\pi} - m')^T \Sigma'^{-1}(w_{U\pi} - m')\right\} \\ &= |2\pi \Sigma'|^{-1/2} \exp\left\{Q(w_{U\pi})\right\}.\end{aligned}$$

We also define a permutation matrix P such that $(\pi_1, \ldots, \pi_l)^T = P(1, \ldots, l)^T$. Then $Pw_U = P(w(u_1), \ldots, w(u_l))^T = (w(u_{\pi_1}), \ldots, w(u_{\pi_l}))^T = w_{U\pi}$. And the mean and covariance matrix of $w_U\pi|w_S$ are $m' = PB_U w_S$ and $\Sigma' = PF_U P'$. Since $P^{-1} = P^T$ it follows that $|P| = \pm 1$ which implies that $|\Sigma'| = |F_U|$. Using this we have,
$Q(w_{U\pi}) = (Pw_U - m')^T \Sigma'^{-1}(Pw_U - m') = (Pw_U - PB_U w_S)^T (PF_U P')^{-1}(Pw_U - PB_U w_S) =$
$(w_U - B_U w_S)^T P^T (P^T F_U^{-1} P^T) P(w_U - B_U w_S) = (w_U - B_U w_S)^T P^T \Sigma'^{-1} P(w_U - B_U w_S) =$
$(w_U - B_U w_S)^T F_U^{-1}(w_U - B_U w_S) = Q(w_U)$.
Since both $|F_U|$ and $Q(w_U)$ are invariant under permutations, $\widetilde{p}(w(u_1), \ldots, w(u_l)|w_S) = \widetilde{p}(w(u_{\pi_1}), \ldots, w(u_{\pi_l})|w_S)$ and hence the symmetry condition is satisfied.

Dimensional consistency: We also assume that S is fixed, so, this proof does not differ from the one found in (Datta et al., 2016) although $\widetilde{p}(w_S)$ has a different definition.
Let $V_1 = V \cup \{v_0\}$ then $V_1 = S' \cup \{v_0\} \cup U$. We need to verify $\widetilde{p}(w_V) = \int \widetilde{p}(w_{V_1}) d(w(v_0))$. So, we have two cases:

Case 1: If $v_0 \in S$. By definition $\widetilde{p}(w_{V_1}) = \int \widetilde{p}(w_{V_1|S}|w_S)\widetilde{p}(w_S)\prod_{si\in S|V_1} d(w_{s_i})$, then

$$\int \widetilde{p}(w_{V_1})d(w(v_0)) = \int \widetilde{p}(w_{V_1|S}|w_S)\widetilde{p}(w_S)\prod_{si\in S|V_1} d(w(s_i))d(w(v_0)).$$

If $v_0 \in S$, and $V = S' \cup U$ then $v_0 \in (S')^c$, and $\prod_{si\in S|V_1} d(w(s_i)d(w(v_0)) = \prod_{si\in(S')^c} d(w(s_i)$, and

$$\int \widetilde{p}(w_{V_1})d(w(v_0)) = \int \widetilde{p}(w_{V_1|S}|w_S)\widetilde{p}(w_S)\prod_{si\in(S')^c} d(w_{s_i}).$$

Also, $V_1|S = U$ since $v_0 \in S$, then

$$\int \widetilde{p}(w_{V_1})d(w(v_0)) = \int \widetilde{p}(w_U|w_S)\widetilde{p}(w_S)\prod_{si\in(S')^c} d(w_{s_i}) = \widetilde{p}(w_V).$$

Case 2: If $v_0 \notin S$, then $V_1|S = U \cup \{v_0\}$, $\widetilde{p}(w_{V_1|S}|w_S) = \widetilde{p}(w_{U|S}|w_S)\widetilde{p}(w(v_0)|w_S)$ and $S|V_1 = (S')^c$. Now,

$$\begin{aligned}\widetilde{p}(w_{V_1}) &= \int \widetilde{p}(w_{V_1|S}|w_S)\widetilde{p}(w_S)\prod_{si\in S|V_1} d(w_{s_i}) \\ &= \int \widetilde{p}(w_U|w_S)\widetilde{p}(w(v_0)|w_S)\widetilde{p}(w_S)\prod_{si\in(S')^c} d(w_{s_i}).\end{aligned}$$

Hence,

$$\begin{aligned}\int \widetilde{p}(w_{V_1})d(w(v_0)) &= \int \widetilde{p}(w_U|w_S)\widetilde{p}(w(v_0)|w_S)\widetilde{p}(w_S)\prod_{si\in(S')^c} d(w_{s_i})d(w(v_0)) \\ &= \int \widetilde{p}(w_S)\widetilde{p}(w_U|w_S)[\widetilde{p}(w(v_0)|w_S)d(w(v_0))]\prod_{si\in(S')^c} d(w_{s_i}),\end{aligned}$$

where $\int \widetilde{p}(w(v_0)|w_S)d(w(v_0)) = 1$, since $w(v_0)$ does not appear in any other term. Finally,

$$\int \widetilde{p}(w_{V_1})d(w(v_0)) = \int \widetilde{p}(w_S)\widetilde{p}(w_U|w_S)\prod_{si\in(S')^c} d(w_{s_i}) = \widetilde{p}(w_V).$$

□

Proof of Theorem 1. To verify that $\widetilde{p}(w_V)$ is the pdf of finite dimensional distribution of a Gaussian process, only rests to prove that $\widetilde{p}(w_V)$ is the pdf of a multivariate normal distribution. Since $w_U|w_S$ follows a l-multivariate normal distribution and w_S follows a n-multivariate normal distribution, the product of these densities is also

a multivariate normal distribution.
Let $\widetilde{C}_{m,n}$ is the covariance matrix of $\widetilde{C}_S$. The cross-covariance is computed for the next possible cases:
Case 1: If $v_1 \in S$ and $v_2 \in S$, that is, $v_1 = s_i$ and $v_2 = s_j$, then $\text{cov}(w(v_1), w(v_2)|\theta)) = \widetilde{C}_{s_i,s_j}$.
Case 2: If $v_1 \in U$ and $v_2 \in S$, we may suppose also that $v_2 \in b_l$. Using the law of total covariance,

$$\text{cov}(w(v_1), w(v_2)|\theta)) = \text{E}(\text{cov}(w(v_1), w(v_2)|w_S)|\theta) + \text{cov}(\text{E}(w(v_1)|w_S), \text{E}(w(v_2)|w_S)|\theta).$$

From our definition $w(v_1)|w_S \perp w(b_l)|w_S$ and $v_2 \in b_l$, then we have that $w(v_1)|w_S \perp w(v_2)|w_S$ and $\text{cov}(w(v_1)|w_S, w(v_2)|w_S) = 0$. Further, $\text{E}(w(v_1)|w_S) = B_{v_1} w_{N(v_1)}$ and using the next property, $\text{E}(g(X)|X) = g(X)$, $\text{E}(w(v_2)|w_S) = w(v_2)$. It follows that,

$$\text{cov}(w(v_1), w(v_2)|\theta)) = \text{E}(0|\theta) + \text{cov}(B_{v_1} w_{N(v_1)}, w(v_2)|\theta) = B_{v_1} \widetilde{C}_{N(v_1),w(v_2)} = B_{v_1} \widetilde{C}_{N(v_1),w(s_j)}.$$

Case 3: If $v_1 \in U$ and $v_2 \in U$. This part of the proof is the same for the NNGP, found in (Datta et al., 2016). We have $\text{E}(w(v_1)|w_S) = B_{v_1} w_{N(v_1)}$ and $\text{E}(w(v_2)|w_S) = B_{v_2} w_{N(v_2)}$. Then,

$$\begin{aligned} \text{cov}(\text{E}(w(v_1)|w_S), \text{E}(w(v_2)|w_S)|\theta) &= \text{cov}(B_{v_1} w_{N(v_1)}, B_{v_2} w_{N(v_2)}) \\ &= B_{v_1} \text{cov}(w_{N(v_1)}, w_{N(v_2)}) B_{v_2}^T. \end{aligned}$$

Observe that if $v_1 \neq v_2$, then $w(v_1)|w_S \perp w(v_2)|w_S$ and $\text{cov}(w(v_1), w(v_2)|w_S) = 0$. Conversely, if $v_1 = v_2$ now $\text{cov}(w(v_1), w(v_2)|w_S) = \text{var}(w(v_1)|w_S) = F_{v_1}$. Then, $\text{cov}(w(v_1), w(v_2)|w_S) = \delta(v_1 = v_2) F_{v_1}$, and $\text{E}(\delta(v_1 = v_2) F_{v_1}|\theta) = \delta(v_1 = v_2) F_{v_1}$. Hence,

$$\text{cov}(w(v_1), w(v_2)|\theta)) = \delta(v_1 = v_2) F_{v_1} + B_{v_1} \widetilde{C}_{N(v_1),N(v_2)} B_{v_2}^T.$$

□

B: Supplementary Material

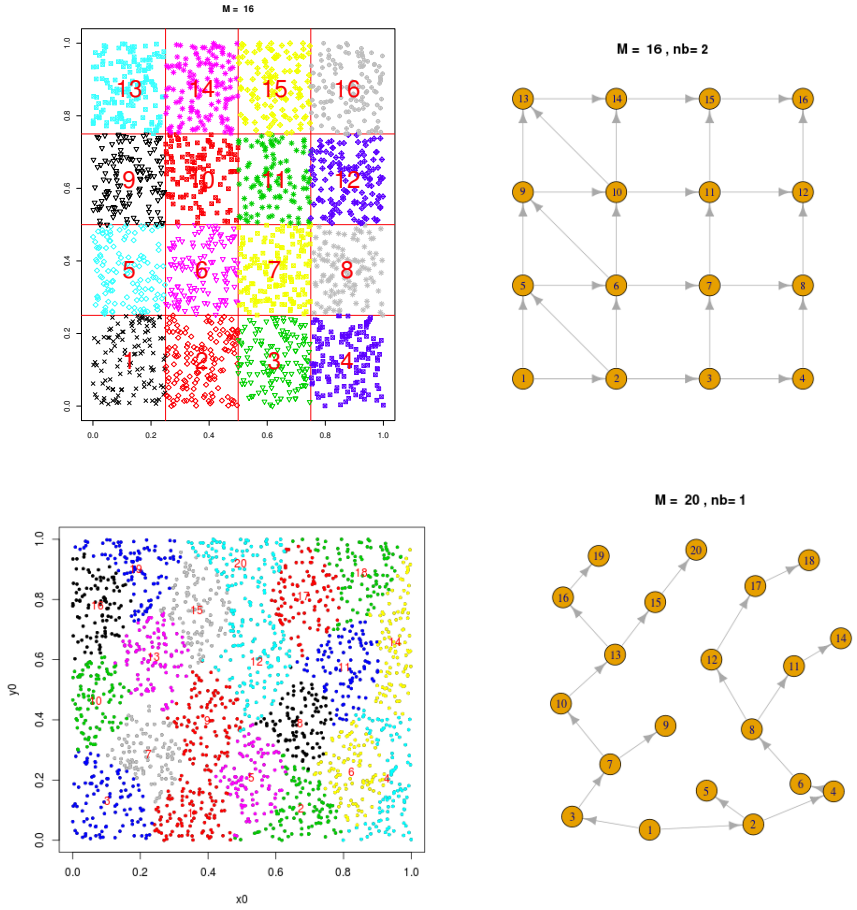

Figure S1: First row: Regular block. Second row: Irregular block. Left: Block design. Right: DAG of blocks.

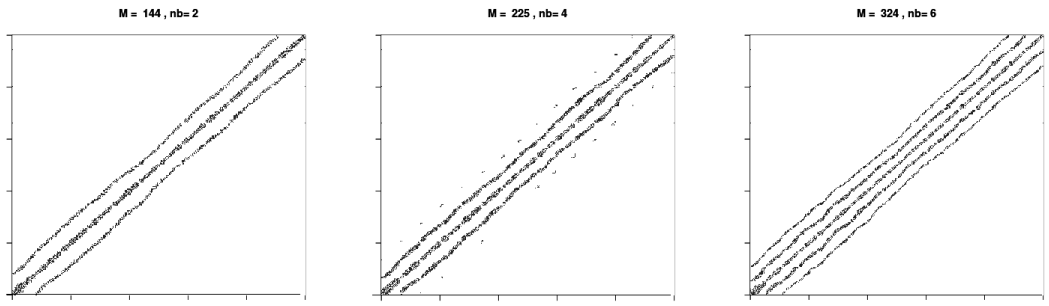

Figure S2: Sparse pattern of precision matrices $\widetilde{C}_S^{-1}$ of block-NNGP, with different number of blocks (M) and differente number of neighbor blocks (nb). Only the nonzero terms are shown and those are indicated by a dot.

Part III

Conclusions and Future work

Chapter 6

Conclusions

This thesis makes original contributions. The first paper is my extension of modeling the spatial distribution of anchovy abundance off the coast of Peru, to study its spatio-temporal distribution. There is a need for efficient and fast methods for estimating abundance of species data. Here the anchovy distribution is modeled taking into account the features of anchovy (many locations with non anchovy, patchiness) as well as the inherent challenges of the data (irregular samples across years, big dataset). The distribution we used is a mixture of a discrete probability mass at zero and a Gamma distribution for nonzero values. We study many possible spatial, temporal and spatio-temporal dependencies, these spatio-temporal models shows great promise for understanding the spatial dependencies of anchovy distribution across years. Although biologists knows that there are different kinds of clusters of anchovy agregations (large, medium, small patches with high or low densities) depending on seasons, it is new that this features can be obtained through spatio-temporal fields changing in time, this is a substantial gain compared with previous visual analysis of anchovy data. By using the SPDE approach and estimating the model parameters using INLA, we also obtain a substantial gain in computational cost compared with a full MCMC-based approach. Faster estimates can help to understand and identify underlying reasons for the detected changes in anchovy behaviour.

Because of the increasing number of large data sets, there is a need for computationally efficient statistical models. In the sencond paper, we generalize the NNGP to build a new valid spatial process called block-NNGP. To be useful for a broad range of practical applications, this process contains a wide family of covariance functions. We can use the block-NNGP model for parameter estimation and prediction in large Gaussian spatial models. We show through a simulation study that the block-NNGP performs well. Using the divide and conquer strategy inherent in the block-NNGP

model, the required computation time is reduced relative to GP model calculations. The block-NNGP model approach requires the selection of blocks. We recommend testing results with different choices of block sizes and number of neighbor blocks. Our choice must depend on the spatial correlation (small or high spatial dependence) and the design of data points. We made use of parallel computing environments (shared memory) for block-NNGP model, the doMC package in R and C++ for some part of the code. A topic for future work is to implement all code in C++ and to extend the code to the distributed memory for maximum reduction of the computing time.

Chapter 7

Future works

In addition to the papers previously mentioned, there is a general idea of potential future work involving block-NNGP models that we are already working on and may be fruitful. Gaussian process models have been widely used in spatial statistics but still face modeling and computational challenges for large spatial datasets. Most often, the random field is specified to have a stationary isotropic correlation function, assuming that the variability of many spatial processes is the same throughout the domain. In such cases, non-stationary Gaussian random fields are used to model non-stationary data. Nevertheless, non-stationary Gaussian random fields are not always necessary to model non-stationary spatial data. In this context, to model nonstationary spatial data we develop a modeling approach using a valid covariance function based on selected partitions that allows one to knit together local covariance parameters. Thus, the local covariance parameters are allowed to be estimated within each partition to reduce computational time requirements. Finally, to facilitate the computations in local covariance estimation, we use the block-NNGP approach for the Bayesian inference of our model.