Abordagem Bayesiana para estimar a biomassa das anchovas na costa do Perú

Zaida Jesús Quiroz Cornejo

Dissertação apresentada
ao
Instituto de Ciências Exatas - Departamento de Estatística
da
Universidade Federal de Minas Gerais

Programa de Pós-Graduação em Estatística

Orientador: Prof. Dr. Marcos Oliveira Prates

Coorientador: Prof. Dr. Håvard Rue

Belo Horizonte, Janeiro de 2014

Abordagem Bayesiana para estimar a biomassa das anchovas na costa do Perú

Dissertação apresentada ao Programa de
Pós-Graduação em Estatística da Universidade
Federal de Minas Gerais como requisito
parcial para a obtenção do grau de
Mestre em Estatística.

Belo Horizonte, Janeiro de 2014

A Bayesian Approach to estimate the biomass of anchovies in the coast of Perú

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

Belo Horizonte, January 2014

UNIVERSIDADE FEDERAL DE MINAS GERAIS

FOLHA DE APROVAÇÃO

Abordagem Bayesiana para estimar a biomassa das anchovas na costa do Perú

Zaida Jesús Quiroz Cornejo

Dissertação defendida e aprovada pela banca examinadora constituída por:

Prof. Dr. Marcos Oliveira Prates
Universidade Federal de Minas Gerais - UFMG

Prof. Dr. Flávio Bambirra Gonçcalves
Universidade Federal de Minas Gerais - UFMG

Profª. Drª. Thais Cristina Oliveira da Fonseca
Universidade Federal do Rio de Janeiro - UFRJ

Belo Horizonte, Janeiro de 2014

Acknowledgments

Agradeço primeiramente a Deus, por me dar forças para superar às dificuldades, pela saúde e pelas alegrias vividas ao longo desses dois anos.

A minha querida família por todo o carinho e amor, em especial a minha mãe e irmã pelo apoio incondicional, ao meu pai por sempre me acompanhar nas idas e voltas ao aeroporto e os gratos cafés de amanhã ou lanches, ao Victor pela compreensão, por estar ao meu lado se fazendo sempre presente mesmo eu estando tão longe, ao meu cunhado Juan Carlos pelas bem-vindas sempre a casa e ao meu sobrinho Luquitas pelas alegrias e beijinhos pelo skype.

Ao meu orientador, o Professor Marcos Oliveira Prates, pela ajuda na realização desse trabalho, pela paciência em cada uma das nossas reuniões, por me motivar a gostar da pesquisa em todo momento e por me incentivar a continuar meus estudos de pós-graduação.

Gostaria de agradecer também ao meu co-orientador, o Professor Håvard Rue, por aceitar em participar dessa dissertação e pelas valiosas sugestões dadas.

A todos os meus professores do programa de pós graduação em Estatística na UFMG, pelo conhecimento transmitido.

À Professora Glaura, pelas sugestões e grande apoio dado tanto antes como depois da seleção do mestrado.

À Professora Lourdes, pela sua amizade e apoio, e acreditar em mim desde o comencinho.

Gostaria de agradecer também ao Professor Renato Assunção, pelos comentários valiosos para melhoria desse trabalho.

Aos Professores Flávio Bambirra Gonçcalves e Thais Cristina Oliveira da Fonseca, por aceitar em participar da banca dessa dissertação.

Às secretárias do programa de pós graduação, Rogéria e Rose, por prestativamente responder os infinitos e-mails que eu enviei, e pela amizade e ajuda ao longo do mestrado.

Á Arnaud Bertrand e Sophie Bertrand, mercie beaucoup pour mon initiation scientifique et leur motivation et l'intérêt à poursuivre mes études.

Aos meus amigos do IMARPE-IRD, que mesmo longe sempre lembraram de mim, à Chio, à Ross, e em especial ao Danielinho, você me ajudou muito mesmo.

Aos amigos na UFMG, em especial a minha querida amiga Márcia, pela grata companhia no LESTE, pelo bate-papo, carinho e sincera amizade, valeu por todo guria!.

A minha avozinha brasileira Carmen, por todo o apoio, carinho e amizade ao longo desses dois anos, você foi meu anjo, não tenho palavras para lhe-agradecer, e a minha irmãzinha Cecilia pela amizade e brincadeiras, muito obrigada!

Resumo

O Sistema da Corrente de Humboldt do norte (NHCS) é um dos mais produtivos ecossistemas em termos de peixes do mundo. Em particular, a anchova peruana (Engraulis ringens) é a maior presa dos predadores superiores, como mamíferos, aves, peixes e pescadores. Nesse contexto, é importante compreender a dinâmica da distribuição de anchova para preservá-la, bem como para explorar sua capacidade econômica. Usando os dados recolhidos pelo "Instituto del Mar del Perú" (IMARPE), durante uma pesquisa científica em 2005, apresenta-se uma análise estatística que tem como objetivos principais: (i) se adaptar às características dos dados amostrados como: dependência espacial, altas proporções de zeros e grandes tamanhos de amostras, (ii) fornecer informações importantes da dinâmica da população de anchovas e propor um modelo para estimação e previsão da biomassa da anchova no NHCS do Perú. Os dados são analisados em um contexto Bayesiano usando a metodologia Integrated Nested Laplace Approximation (INLA). Finalmente, usa-se critérios de comparações entre modelos para selecionar o modelo proposto de melhor ajuste. Também é feito um estudo do poder preditivo de cada modelo. Além disso, é realizado um diagnóstico de influência Bayesiana para o modelo preferido.

Palavras-chave: Inferência Bayesiana Aproximada ; Geoestatística; Integrated Nested Laplace Approximation; Modelo Gaussiano Latente; Ecologia Marinha.

Abstract

The Northern Humboldt Current System (NHCS) is the world most productive ecosystem in terms of fish. In particular, Peruvian anchovy (Engraulis ringens) is the major prey of the principal top predators, like mammals, seabirds, fish and fishers. In this context, it is important to understand the dynamics of the anchovy distribution to preserve it as well as to explore its economical capacities. Using the data collected by the "Instituto del Mar del Perú" (IMARPE), during a scientific survey in 2005, we present a statistical analysis that has as main goals: (i) adapt to the characteristics of the sampled data, such as spatial dependence, high proportions of zeros and big samples size, (ii) provide important insights on the dynamics of the anchovy population and propose a model for estimation and prediction of anchovy biomass in the NHCS of Perú. These data are analyzed in a Bayesian framework using the Integrated Nested Laplace Approximation (INLA) methodology. Finally, model comparison is performed to select the best model and predictive checks to study the predictive power of each model. Moreover, a Bayesian spatial influence diagnostic is performed for the preferred model.

Keywords: Approximate Bayesian inference; Geostatistics; Integrated Nested Laplace Approximation; Latent Gaussian model; Marine Ecology.

Resumo Estendido

O ecossistema pelágico peruano é dominado pela anchova peruana. Atualmente, a anchova é responsável pela maior pesca de peixes no mundo. Devido a sua importância econômica e ecológica o "Instituto del Mar del Perú" (IMARPE) realiza pesquisas todos os anos sobre o ecossistema para orientar as decisões de gestão do país. Assim, a bordo de um cruzeiro de pesquisa científico-acústico obtem-se a massa da anchova por milha náutica que será chamada de biomassa da anchova ao longo desse trabalho. A distribuição da anchova é caracterizada por estruturas de agregação, tal padrão é devido ao seu comportamento de defesa para enfrentar a predação. Logo, os dados da biomassa da anchova caracterizam-se por uma elevada proporção de valores zeros e dependência espacial.

A motivação desse trabalho é propor um modelo estatístico capaz de estimar a biomassa de anchova, o qual deve considerar os valores zero e não zeros usando uma abordagem integrada como os modelos zero-inflacionados ou os modelos Hurdle. Nesse sentido, nos últimos anos um grande esforço tem se dedicado a lidar com esse tipo de modelos. Porém, muitas vezes esses desenvolvimentos têm sido focados principalmente na modelagem de dados discretos (Mullahy, 1986; Cameron and Trivedi, 1998; Agarwal et al, 2002). A ideia subjacente nesse trabalho é estender a definição desses modelos para dados discretos à modelagem de dados contínuos, e, ao mesmo tempo, acomodar a dependência espacial. Para isso, propomos uma modelagem Hierárquica Bayesiana. Os modelos hierárquicos, diferentemente dos modelos tradicionais baseados na modelagem multivariada da variável resposta, através de uma estrutura de covariância permitem modelar de forma simplificada respostas discretas, contínuas e/ou misturas, impondo à estrutura de auto-correlação espacial em um nível inferior na hierarquia.

Geralmente, inferência Bayesiana de modelos complexos pode ser realizada utilizando métodos tais como o Markov Chain Monte Carlo (MCMC). No entanto, sabe-se que para modelos que incluem vários efeitos fixos e aleatórios com dependência espacial e/ou grandes conjuntos de dados, obter as distribuições posteriores é um desafio, pois estas raramente possuem solução analítica, tornando assim a inferência através de métodos MCMC computacionalmente muito cara. Uma nova abordagem, chamada Intagrated Nested Laplace Approximation (INLA), foi proposta por Rue et al. (2009) para executar de forma eficiente inferência Bayesiana em modelos hierárquicos Gaussianos latentes. O método baseia-se em aproximações precisas e determinísticas ao invés de simulações aleatórias, e portanto, não precisa de diagnósticos de convergência necessários nos métodos MCMC.

Esse trabalho está organizado da seguinte maneira: No Capítulo 1 falamos brevemente da motivação do trabalho e definimos os objetivos e as contribuições do mesmo. No Capítulo 2 é feita uma revisão da literatura utilizada nos próximos capítulos. Introduzimos os conceitos de Modelos Gaussianos Latentes e a relação deles com os modelos Bayesianos hierárquicos. Em seguida, apresentamos os Campos aleatórios Gaussianos e campos aleatórios Markovianos Gaussianos (CAMG), assim como as vantagens em termos computacionais dos CAMG. Logo, apresentamos a aproximação Gaussiana no caso univariado. Finalizamos o capítulo apresentando o método de aproximação determinística INLA usado para a obtenção das distribuições marginais a posteriori na inferência Bayesiana. No Capítulo 3 é apresentada uma descrição detalhada dos dados que motivam a nossa modelagem. Nesse capítulo é descrita de forma resumida como são obtidos os dados, em particular, a variável de interesse, a biomassa da anchova. Além disso, descrevemos outras variáveis disponíveis que podem contribuir para explicar à variabilidade da variável resposta. O Capítulo 4 descreve a estrutura da modelagem para o ajuste e previsão da biomassa de anchova proposta pela autora. Além disso, apresentamos como a inferência Bayesiana é realizada usando o INLA. O capítulo termina apresentando uma variedade de critérios de seleção dos modelos, critérios de previsão dos modelos, e finalizamos apresentando um possível diagnóstico de influência. O Capítulo 5 apresenta à aplicação da modelagem proposta. Primeiramente, é apresentada uma análise exploratória das covariáveis, e exploramos possíveis distribuições contínuas para compor o modelo misto. Em seguida, apresentamos diversos modelos e os seus resultados utilizando a metodologia descrita no capítulo anterior. Os resultados da seleção de modelos mostram que a inclusão de efeitos espaciais estruturados nas duas componentes do modelo Hurdle são realmente necessários para um melhor ajuste. Assim o modelo preferido incluindo ambos efeitos espaciais é o de melhor ajuste para a biomassa da anchova e, é ainda, capaz de explicar melhor a distribuição espacial dos dados. Além disso, o poder de previsão dos modelos é estudado através de um estudo de simulação. Nesse estudo, separamos o banco em treino e validação com diferentes porcentagens de corte e rodamos 100 iterações para determinar o modelo com melhor previsão. Os resultados mostraram que o modelo com melhor previsão é também o modelo selecionado como o de melhor ajuste. Finalizamos o capítulo diagnosticando as regiões de influência através da divergência de Kullback-Liber para o modelo selecionado. No Capítulo 6 apresentamos algumas discussões a respeito das vantagens do modelo escolhido para estimar a biomassa da anchova. Finalmente, no Capítulo 7 apresentamos ainda algumas possíveis extensões que podem ser feitas como trabalhos futuros.

Contents

List of Figures viii

List of Tables x

1 Introduction 1
- 1.1 Motivation . . . 1
- 1.2 Objetives . . . 1
- 1.3 Contributions . . . 2
- 1.4 Organization of manuscript . . . 2

2 Literature Review 3
- 2.1 Latent Models and Hierarchical models . . . 3
- 2.2 Gaussian Fields and Gaussian Markov Random Fields . . . 4
  - 2.2.1 Gaussian Fields . . . 4
  - 2.2.2 Gaussian Markov Random Fields . . . 5
- 2.3 Gaussian approximation . . . 7
- 2.4 Integrated Nested Laplace Approach . . . 8
  - 2.4.1 Approximating $\pi(\theta|y)$ . . . 9
  - 2.4.2 Approximating $\pi(x_i|\theta, y)$ . . . 10
  - 2.4.3 Approximating $\pi(\theta_j|y)$ . . . 10

3 Description of Data 12

4 Model Structure 14
- 4.1 Bayesian Inference . . . 16
- 4.2 Model Assessment . . . 17
  - 4.2.1 Model Comparison . . . 17
  - 4.2.2 Model Predictive checks . . . 18
  - 4.2.3 Influence Diagnostics . . . 19

5 Application 20
- 5.1 Exploratory Analysis . . . 20
- 5.2 Data Analysis . . . 22

6 Discussion **28**

7 Future works **30**

7.1 Introduction 30

7.2 The Stochastic Partial Differential equation (SPDE) approach 30

A Proof of result 4.7 **34**

Bibliography **35**

List of Figures

2.1 Simulation of a GRMF using sparse Precision matrix. 6
2.2 Correlation function for the fitted GMRF (red line) and the Matern CF (blue line) with range 15 and 5×5 neighbourhood. 7
2.3 Original posterior density (continuous black line) and Gaussian approximation of posterior density (dashed grey line) for each x_0=0,0.5,1,1.5. The value of x_0 is represented by the small dot point in each plot. 8
2.4 Location of the integration points in a two dimensional θ-space using the (a) grid and the (b) CCD strategy. 10

3.1 Left: The observed data, where the trajectory of survey tracks is represented by parallel cross-shore transects (red and gray dots). Furthermore, the size of red dots correspond to the biomass of anchovy (higher than zero) and gray dots correspond to the biomass of anchovy equal to zero. Right: Exploratory analysis. Histogram for all anchovy biomass observations and Histogram for non-zero anchovy biomass observations. 13

5.1 The black grids dots represent the translated regular lattice, here red grid dots are samples of anchovy biomass traslated too. And green grids dots represent the rotated Regular lattice, here blue grid dots are samples of anchovy biomass rotated too. 21
5.2 Left: Regular lattice for distance to the coast. Right: Regular lattice for ocean depth. 21
5.3 Results. Left: Observed anchovy absence/presence. Right: Mean posterior probability of anchovy absence under model I. 26
5.4 Results. Left: Observed anchovy biomass (on the logarithmic scale). Right: Mean posterior anchovy biomass under model I (on the logarithmic scale). . 27
5.5 Results. Left: Probability q_i of influence regions. Right: Mean posterior of anchovy biomass influence regions (on the logarithmic scale). 27

7.1 Representation of piewise-linear approximationof a function in two dimensions over a triangulated mesh. 31

7.2 Meshs constructed using Constrained refined Delaunay triangulation. Red grid dots are samples of anchovy biomass samples. The region defined by the sky-blue line is the edge boundary which defines the priority area for estimation. Outside this region the boundary effects are higher. Right: Mesh have less resolution than left plots. Top: Mesh for all data. Down: Mesh for northern region of top panel. 33

List of Tables

5.1 Selection criteria for the different positive Distributions 21
5.2 Linear predictors and hyperparameters for each proposed model 23
5.3 The selection criteria for the models proposed with different linear predictors 24
5.4 Predictive model checks. Mean of RMSPE (MRMSPE) out 100 validation samples. 24
5.5 Summary statistics (point, standard deviation and 95% credible interval (CI)) for Fixed effects and Hyperparameters estimation. 25

7.1 Summary of functions ϕ_k and precision matrices Q_α for each α 32

Chapter 1

Introduction

1.1 Motivation

The Peruvian pelagic ecosystem (Northern HCS) is highly dominated by the Peruvian anchovy. Anchovy is a small pelagic fish characterized by a fast growth, an early maturity (1 year), a short life span (4 years), fast response to environmental variability, plasticity in terms of the prey it consumes and foraging behaviour (Bertrand et al, 2008). At the present, anchovy species sustain the largest single-species fishery in the world with average landings over 6.5 millions tons per year during the last decade (Bertrand et al, 2008). Because of its economic and ecological importance "Instituto del Mar del Perú" (IMARPE) conducts research on the ecosystem and fisheries to guide management decisions. In Perú, annual acoustic surveys of fish population distribution and abundance have been conducted since 1983 by IMARPE. In particular, acoustic data are collected onboard the research vessel "José Olaya Balandra" from the IMARPE during a scientific acoustic survey "Pelagic 2005". Using these data it is obtained the mass of anchovy per nautical mile which will be called anchovy biomass along the current paper.

Another relevant feature of anchovy involves its distribution which within suitable habitat is characterized by nested aggregation structures (Fréon and Misund, 1999). Such pattern results from its foraging behavior and defense behavior to face predation. For this reason, anchovy biomass data are characterized by a high proportions of zero values and spatial dependence.

Such characteristics have to be considered to develop an appropriate statistical model to estimate anchovy biomass. The excessive zero observations is the main reason to assume that anchovy distribution can not be modeled by only one distribution. In this context, Woillez et al. (2009) developed an integrated model approach involving non-parametric transformation processes to investigate schooling fish estimates, while Boyd (2012) developed another integrated model to simulate the spatial distribution of anchovy biomass for a small region using likelihood-based Geostatistics approach (Diggle et al, 1998). Therefore, an appropriate model needs to account for the zero and for the non-zero values as an integrated approach, like zero-inflated models or Hurdle models. In particular, Hurdle models are attractive because they do not assume that zero values represent error measures, may be due to a poor sample design, or are false zeros, in fact, here anchovy biomass is equally collected over all studied area, and most important, anchovy absence is not unusual within unsuitable habitat. Furthermore, to accommodate spatial dependence, we need to develop a modeling framework that allow for spatial autocorrelation with excessive zeros in the observations.

1.2 Objetives

The main objectives of this study are: (i) Adapt to the characteristics of the sampled data to predict anchovy biomass at unsampled locations, (ii) Provide important spatial insights

on the anchovy distribution and biomass. In order to achieve our objectives a Hurdle model for continuous data is developed. In particular, we use a Bayesian Hierarchical Model which is very flexible and powerful allowing for the model to have all characteristics presented.

1.3 Contributions

A fair amount of statistical effort has been devoted to dealing with zero-inflated data sets. However, often these developments have been focused mainly on the modelling of discrete data (Mullahy (1986), Agarwal et al. (2002)). The idea behind a Hurdle model for discrete data is to separate the zero structure from the non zero structure with a finite mixture of a point mass at zero with a truncated-at-zero distribution (Cameron and Trivedi 1998), such definition will be extended for continuous data.

Generally, Bayesian inference for complex models can be performed using simulation methods such as Markov Chain Monte Carlo Method (MCMC). However, it is known that for models which include several fixed and random effects with spatial dependence and/or large datasets, like in our framework, obtain posterior distributions is seldom analitically available, making inference using MCMC methods computationally expensive. A new approach, called integrated nested Laplace approximation (INLA), was proposed by Rue et al. (2009) to perform fast Bayesian inference. The method relies on accurate and deterministic approximations instead of sthocastic simulations and thus, it does not require convergence diagnostics. Recently, Muñoz et al. (2013) studied the presence/absence of *Trachurus mediterraneus* in the Western Mediterrian under the context of Bayesian Hierarchical Model using the R-INLA software.

1.4 Organization of manuscript

The dissertation is organized as follows: Chapter 2 presents some literature review needed for next sections. Chapter 3 presents some description of data. Chapter 4 describes the proposed model structure to fit and to predict anchovy biomass. Also, it is presented a summary of how Bayesian inference is achieved using INLA. The chapter ends presenting a variety of model assesment criteria. Chapter 5 presents the application of the proposed model and the results obtained. Chapter 6 discusses meaningful results and finally Chapter 7 discusses future works.

Chapter 2

Literature Review

Integrated Nested Laplace Approximation (INLA) is a relatively new approach to implement fast Bayesian inference for Latent Gaussian Models (LGM). Many well known models like Geostatistical models, spatial and spatio-temporal models, among many others, are LGM's. These models are usually complex because they usually include several fixed and random effects. As a result their posterior distributions are rarely analitically available and inference becomes very difficult. In these cases model fitting is usually based on simulation methods like Markov Chain Monte Carlo (MCMC), which are very accurate if the convergence is achieved, but on the other hand in term of computational time are very expensive. In that sense, INLA with accurate, deterministic approximations to posterior marginal distributions is an attractive alternative to MCMC simulations.
In the next sections LGM's and their main features are described, then some definitions about Gaussian Fields and Gaussian Random Markov Fields are given, then it is discussed the relationship between those ones and the computational advantages of GRMF over Gaussian Fields; and finally Gaussian approximations used to introduce the INLA approach are defined.

2.1 Latent Models and Hierarchical models

Latent Models are a subclass of structured additive models which can also be seen as a representation of a Hierarchical model. First of all, let us assume that for $i = 1, ..., n$ we have n observed (or response) variables y_i with a distribution usually but not necessarily from the exponential family. The latent variable η_i defined by Equation (2.1) is a linear predictor which enters the likelihood through some link function $g(.) = \eta_i$. Thus, η_i is modeled additively on different effects of various covariates,

$$\eta_i = \beta_0 + \sum_{j=1}^{\eta_f} w_{ij} f^{(k)}(u_{ij}) + \sum_{k=1}^{\eta_{\beta_k}} \beta_k z_{ki} + \epsilon_i. \tag{2.1}$$

Here, $\beta_k' s$ are coefficients for linear effects on a vector of covariates z, which capture the variability in data caused by explanatory variables; $f^{(k)}$'s represent unknown functions on a set of covariates u, useful to incorporate dependence between observations which can be of various kind like spatial, temporal or spatiotemporal; w_{ij} are known as weights; and ϵ represents unstructured random effects.
The latent field x is composed by a vector: $x = \{\{\eta_i\}, \{\beta_0\}, \{\beta_k\}, \{f^{(k)}\}\}$. If the distribution of latent field is set as Gaussian such model becomes a Latent Gaussian Model (LGM). If, in addition, this latent field is Gaussian and admits conditional independence properties it is called Gaussian Random Markov Field.

Hierarchical Models are a generalization of Linear and Generalized Linear Models. They are

specified by several stages of observations and parameters. 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 composed by all the unknown parameters called hyperparameters, here 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 variable 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:

$$x|\theta \sim N(\mu, \Sigma(\theta))$$

(iii) And hyperparameters θ:

$$\theta \sim \pi(\theta).$$

In many LGM's and Hierarchical models the latent Gaussian field is also a Gaussian Markov Random Field (GRMF), or can be approximated by GRMF's, an overview of this topic is presented in the following section.

2.2 Gaussian Fields and Gaussian Markov Random Fields

2.2.1 Gaussian Fields

Informally a random field, also called as spatial process in spatial statistics, is a collection of random variables that exist exclusively in the d-dimensional space domain D and these variables are indexed by some set $D \subset \Re^d$ containing spatial coordinates $s_1, s_2, ..., s_k \in D$. Furthermore, if all these random variables follow a jointly Gaussian distribution the random field is called "Gaussian random field".

In geostatistics, it is usually used a spatial random field which is assumed to be normally distributed and is known as "Gaussian field" (GF). A large reference about Gaussian fields can be found for example in Cressie (1993) or Diggle et al. (1998).

Definition 1. *Let $\{z(s), s \in D\}$ be a stochastic process where $D \subset \Re^d$ and $s \in D$ represents the location. The process $\{z(s), s \in D\}$ is a Gaussian Field (GF) if for any $k \geq 1$ and for any location $s_1, s_2, \ldots, s_k \in D$, $\{(z(s_1), ...z(s_k))\}^t$ follows a multivariate Gaussian distribuion. The mean function and covariance function of z are:*

$$\mu(s) = E(z(s)); s = (s_1, s_2, \ldots, s_k)^t,$$

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

which are assumed to exist for all s_i and s_j.
The Gaussian field is Weakly stationary if $\mu(s) = \mu$ for all $s \in D$ and if the covariance function only depends on $s_i - s_j$. The Gaussian field 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 $h = \parallel s_i - s_j \parallel$.

Thus, the covariance matrix of the Gaussian field *z(s)* is defined to be the $k \times k$ matrix with *ij* element $C(s_i, s_j)$, then the covariance structure reflects the strengths of relationship

between random variables $z(s_i)$ and $z(s_j)$.

One of the most used correlation functions 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 K_ν is the modified Bessel function of order $\nu > 0$, this last one is a shape parameter and determines the smoothness of the process and s_v is a scale parameter. Such correlation function can be re-defined depending on the range ($r = \frac{\sqrt{(8\nu)}}{s_\nu}$) by

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

This last useful parameter introduced in the correlation function called range (r) is interpreted as the minimum distance for which two locations are not more correlated. It means that if two locations are separated for more than r distance, these locations are nearly independent.

2.2.2 Gaussian Markov Random Fields

Previously to define a Gaussian Markov Random Field (GMRF) will be introduced some basic theory about graphs.

Definition 2. *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 ($i \sim j$).*

If all edges have no direction this graph is called *undirected graph*. Furthermore, from this definition it is implicity 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 example in the spatial context, a regular or irregular lattice can represent a graph.

Theorem. *Let a random vector $x = (x_1, x_2, \ldots, x_n)^t$ be normal distributed with mean μ and precision matrix $Q > 0$. Then for $i \neq j$,*

$$x_i \perp x_j | x_{-ij} \Longleftrightarrow Q_{ij} = 0.$$

In other words, this theorem 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. Now follows a formal definition of GMRF.

Definition 3. *A random vector x ($\in R^n$) is a Gaussian Markov Random Field (GRMF) with respect to a graph $G=(V,E)$ with mean μ and precision matrix $Q > 0$ (positive definite), if and only if, a joint distribution of x is given by*

$$f_X(x) = (2\pi)^{(-n/2)}|Q^{1/2}|exp(-\frac{1}{2}(x-\mu)^T Q(x-\mu))$$

where

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

Here the vertex set V corresponds to the nodes (indices) $\{1, \ldots, n\}$ and the edge set E specifies the dependencies between the random variables $x_1, x_2, \ldots, x_n$. Furthermore, if Q is

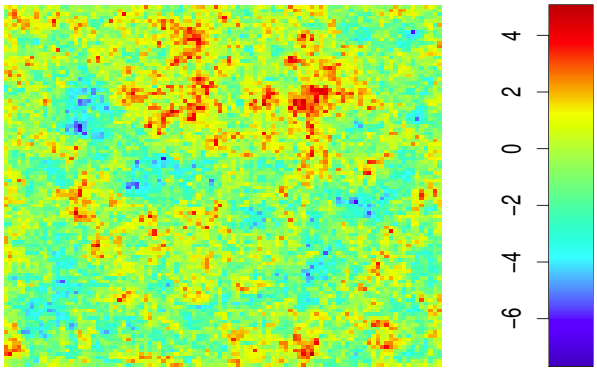

Figure 2.1: *Simulation of a GRMF using sparse Precision matrix.*

a symmetric and positive definite matrix n x 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 $i \neq j$,

$$x_i \perp x_j | x_{-ij} \Longleftrightarrow Q_{ij} = 0,$$

which implies that x_i and x_j are conditionally independent and it means that the conditional distribution of observed variable at some node only depends on its neighbors. In addition, any multivariate normal distribution with symmetric positive definite precision matrix which admits conditional independence properties it is also a Gaussian Markov Random Field. Then, any "Gaussian Field" which admits conditional independence properties it is also a GRMF with respect to some neighboor graph.
Another important feature about GMRF's it 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 sparness of the precision matrix, because there are many numerical methods which use this feature for fast computing.

In order to understand better the idea behind a GRMF let us simulate the GRMF $x \sim N(0, Q^{-1})$, where Q is a sparse precision matrix. Specifically, let the *(i,j)-th* element of Q to be 0 if and only if $(i, j) \notin E$, to be equal to -0.25 if $(i, j) \in E$ and $i \neq j$, and to be 1 if $i = j$. A consequence of this construction it is that the conditional distribution of each random variable x_i given all other random variables is equal to the conditional distribution of x_i given only its neighbors. The algorithm to simulate a GRMF from Q is very simple (see Rue and Held 2005). Given that Q is sparse, its Cholesky decomposition $Q = LL^t$, where L is a lower triangular matrix, can be computed very efficiently. Then, it is easy to show that $x = \mu + L^{-t}z$ is a sample from the GMRF $x \sim N(\mu, Q^{-1})$, where $z \sim N(0, I)$. Figure 2.1 shows the simulation for this GRMF.

It is important to notice that Gaussian fields can be well "approximated" by GMRFs. Figure 2.2 shows the approximated correlation function provided by GMRFs (red line) and compare it with the true Matérn correlation function (blue line). For further details on how to perform such approximations see Rue and Tjelmeland (2002); Rue and Held (2005); Lindgren et al. (2011).

Definition 4. *Let Λ be a regular lattice of size n_r x n_c (for a two dimensional lattice) and x_{ij} denote the value of x at site ij. Then a Gaussian field on Λ is a GMRF if it satisfies the*

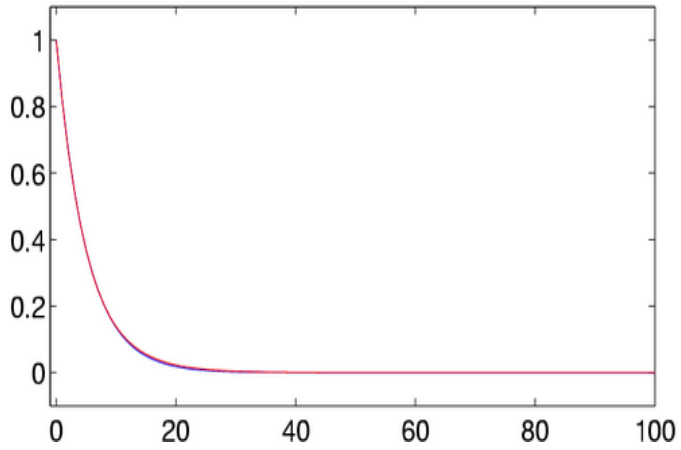

Figure 2.2: *Correlation function for the fitted GMRF (red line) and the Matern CF (blue line) with range 15 and* 5×5 *neighbourhood.*

Markov property

$$\pi(x_{ij}|x_{kl} \in \Lambda \backslash (i,j)) = \pi(x_{ij}|x_{kl}, (k,l) \in \partial_{ij})$$

where ∂_{ij} *is the neighbourhood of (i,j)*

This neigbourhood $\partial_{ij}(i,j) \in \Lambda$ implies the Markov property. Then the precision matrix Q associated to this regular lattice will have many zeros because there is a finite number of neighbors for each cell and consequently Q will be an sparse matrix. Again the benefit of using a GMRF instead of GF it is purely computational.

2.3 Gaussian approximation

Let $\pi(x|y)$ be a posterior density distribution of the form

$$\pi(x|y) \propto \pi(x)\pi(y|x) = \exp(f(x)),$$

such function *f(x)* can be approximated using a quadratic Taylor expansion around the value x_0. That is,

$$\begin{aligned} f(x) &\approx f(x_0) + f^{(1)}(x_0)(x - x_0) + \frac{1}{2}f^{(2)}(x_0)(x - x_0)^2 \\ &= a + bx - \frac{1}{2}cx^2, \end{aligned}$$

where $b = f^{(1)}(x_0) - f^{(2)}(x_0)x$ and $c = -f^{(2)}(x_0)$. Thus, the Gaussian approximation of $\pi(x|y)$ is given by

$$\widetilde{\pi}_G(x|y) \propto exp\left(-\frac{1}{2}cx^2 + bx\right),$$

then $\widetilde{\pi}_G(x|y)$ is normally distributed with mean b/c and variance $1/c$. In order to illustrate this approximation suppose that y follows a Poisson distribution with mean λ and the prior of x follows a normal distribution with mean μ=0 and variance equal to k^{-1}, define $x = \log(\lambda)$. Then,

$$\pi(x|y) \propto \pi(x)\pi(y|x) = \exp\left(-\frac{k^2}{2}(x - \mu)^2 + yx - \exp(x)\right), \text{and}$$

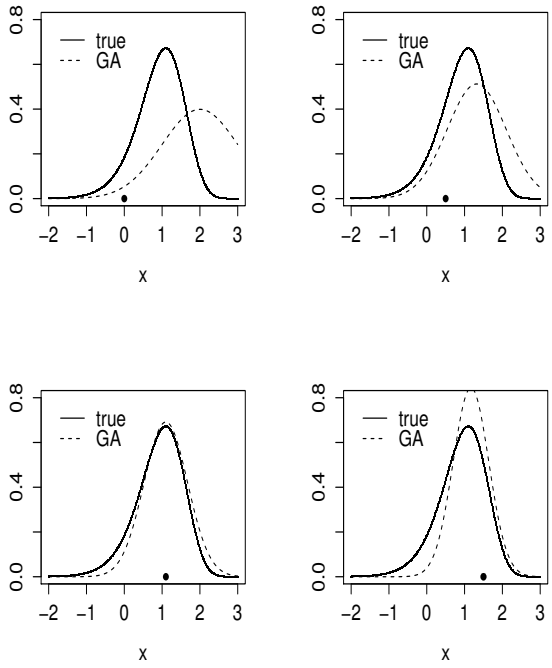

Figure 2.3: *Original posterior density (continuous black line) and Gaussian approximation of posterior density (dashed grey line) for each x_0=0,0.5,1,1.5. The value of x_0 is represented by the small dot point in each plot.*

$$\widetilde{\pi}_G(x|y) \propto exp\left(-\frac{1}{2}cx^2 + bx\right),$$

where $b = -k^2x_0 + k^2\mu + y - \exp(x_0) + cx_0$ and $c = -k^2 - \exp(x_0)$. Figure 2.3 shows $\pi(x|y)$ and $\widetilde{\pi}_G(x|y)$ for different values of x_0=0,0.5,1,1.5; y=3, μ=0 and k=0.001. Note that the normal approximation for the density $\pi(x|y)$ improves when x_0 is closer to the mode of $\pi(x|y)$. The Gaussian approximation from this univariate case can easily be generalized to the multivariate case (Rue and Held, 2005).

2.4 Integrated Nested Laplace Approach

Although inference for LGM's is usually performed through MCMC methods, it is also known that such methods are computational expensive, specially when we are dealing with complex models. The main reasons are that the components of the latent field x are strongly dependent on each other and that θ and x are strongly dependent, specially when n is large. On the other hand INLA (Rue et al, 2009) works out with LGM's that satisfy two properties: (i) The latent field x admits conditional independence properties, as a result the latent field is a GMRF; (ii) The number of hyparameters m is small (m$\leq$15). And these properties make it possible to obtain fast Bayesian inference.
The join posterior of LGM can be calculated using the likelihood distribution, latent Gaussian distribution and the distribution of hyperparameters,

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

Let

$$x|\theta \sim N(0, \Sigma(\theta))$$

here $Q^{-1} = \Sigma(\theta)$ is the precision matrix, then

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

The posterior marginals of the latent variables $\pi(x_i|y)$ and the posterior marginal of hyperparameters $\pi(\theta_j|y)$ are defined by:

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

$$\pi(\theta_j|y) = \int \pi(\theta|y)d\theta_{-j}.$$

These posterior marginals are not easy to calculate, and that is the main aim of INLA. Thus, approximations to the posterior marginals of the latent variables and hyperparameters are given by Equation (2.2) and Equation (2.3), which are both very accurate and extremely fast to compute,

$$\widetilde{\pi}(x_i|y) = \int \widetilde{\pi}(x_i|\theta, y)\widetilde{\pi}(\theta|y)d\theta \tag{2.2}$$

$$\widetilde{\pi}(\theta_j|y) = \int \widetilde{\pi}(\theta|y)d\theta_{-j}, \tag{2.3}$$

where $\widetilde{\pi}$ denotes an approximation to a probability density function (pdf).

In summary the main idea of INLA is divided into the next tasks:

- First, it provides an approximation of $\widetilde{\pi}(\theta|\mathrm{y})$ to the join posterior of hyperparameters given the data $\pi(\theta|\mathrm{y})$,
- Second, it provides an approximation of $\widetilde{\pi}(x_i|\theta,\mathrm{y})$ to the marginals of the conditional distribution of the latent field given the data and the hyperparameters $\pi(x_i|\theta,\mathrm{y})$,
- And third, it explores $\widetilde{\pi}(\theta|\mathrm{y})$ on a grid and use it to integrate out θ in Equation (2.2) and θ_{-j} in Equation (2.3).

2.4.1 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:

$$\widetilde{\pi}(\theta|y) \propto \frac{\pi(x, \theta, y)}{\widetilde{\pi}_G(x|\theta, y)}\Big|_{x=x^*(\theta)} \tag{2.4}$$

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 (Section 2.3). 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 (2.4) is also called Laplace approximation.

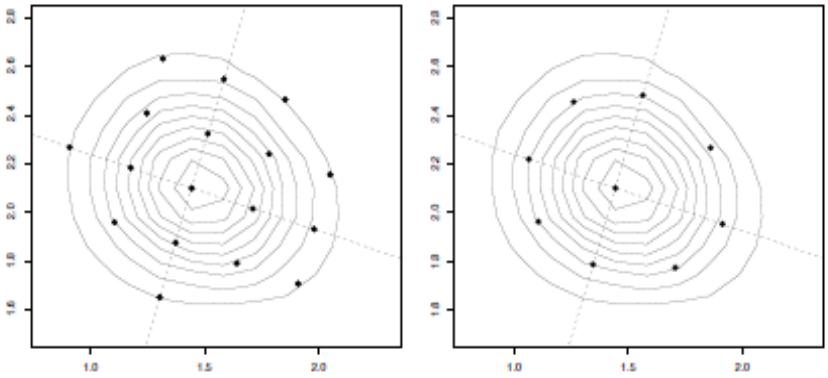

Figure 2.4: *Location of the integration points in a two dimensional θ-space using the (a) grid and the (b) CCD strategy.*

2.4.2 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 (see Rue and martino, 2007). Then, the second option is to do again a Laplace approximation, this approximation is more accurate and it is denoted by $\widetilde{\pi}_{LA}(x_i|\theta, y)$:

$$\widetilde{\pi}_{LA}(x_i|\theta, y) \propto \frac{\pi(x, \theta, y)}{\widetilde{\pi}_{GG}(x_{-i}|x_i, \theta, y)}|_{x_{-i}=x^*_{-i}(x_i,\theta)}, \tag{2.5}$$

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 (2.5). It thus correct the Gaussian approximation for location and skewness with a moderate extra cost when compared to the Laplace approximation.

2.4.3 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 $\widetilde{\pi}(\theta|y)$ (Figure 2.4). 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:

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

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

Finally to conclude this section, we have to add that when the model is too complex it is recommended to use the Simplified Laplace approximation and the Central composit design (CCD) strategy; both options are used by default via the R-INLA-package.

Chapter 3

Description of Data

The data used in this study were collected onboard the research vessel "José Olaya Balandra" from the "Instituto del Mar del Perú" (IMARPE) during a scientific acoustic survey "Pelagic 2005", between February 20 and April 4, 2005. The idea behind acoustic surveys performed by marine researchers, it is to do sea "travels" at some area of interest onboard a research vessel which has an echosounder. An echosounder has a transducer that emits sound waves which are spreaded at sea and when they find any "target" (fish, zooplankton, bottom, etc.) a sound is reflected, so a certain energy is returned to the transducer, this back scattered energy (echo) is detected again by the transducer and converted into electrical signal. After a while the transducer emits again a pulse and repeats this process (for more details see Simmonds and MacLennan (2005)). Finally the sounds measured in decibels (dB) are known as back-scattered strength (Sv). This technology allows to study the composition of the sea, spatial distributions of marine populations and their change over time, among others. How it is designed the trajectory of survey (survey tracks) depends on areas of study, but commomnly parallel cross-shore transects are performed, as was done in this study. Another feature in acoustic surveys is that the data can be collected in different frequencies depending on the echosounders used, this is important because for each frequency the Sv of marine organisms are different depending on their size, volume, among other features. Thus, using the catches from associated fishing trawls and frequencies it is possible to classify better those organisms. In this study, acoustic data were collected using a scientific echo-sounder EK500 working at frequencies 38 and 120 kHz. Selection and classification (including anchovy data) of acoustic data were carried out by IMARPE.
The variables used in this study are: (i) *Biomass of anchovy* (NASC in m^2/nm^2), back-scattered strength (Sv) of anchovy are recorded along survey tracks in each geo-referenced elementary sampling distance unit (ESDU), in this case, equal to one nautical mile. Then each Sv is trasnformed into the Nautical area scattering coefficient (NASC in m^2/nm^2, MacLennan et al. (2002) for acoustic units), NASC $= 4\pi(1852)^2 Sv$. This variable is an indicator of fish biomass. In particular, the mass of anchovy in each ESDU. (ii) *Distance to the coast* (km), is computed as the minimum orthodromic distance to the Peruvian coast (km). The orthodromic distance is the shortest distance between two points on the surface of a sphere. Let Φ_s and Λ_s be some position (longitude, latitude) at the shoreline and let Φ_{ij} and Λ_{ij} be some position (longitude, latitude) in the sea, then the orthodromic distance (od) between them is computed as follows, $od(s, ij) = 60 \times 1.852 \times \frac{180 \times \arccos(A+B)}{\pi}$ where, $A = \sin(\frac{\Lambda_s \pi}{180}) \sin(\frac{\Lambda_{ij}}{180})$ and $B = \cos(\frac{\Lambda_s \pi}{180}) \cos(\frac{\Lambda_{ij}}{180}) \cos(\frac{\Phi_s \pi}{180} - \frac{\Phi_{ij}}{180})$. (iii) *Depth (meters* < 0*)*, the gridded bathymetric data sets for Peruvian's ocean are provided by the General Bathymetric Chart of the Oceans (GEBCO, http://www.gebco.net/). (iv) *Latitude (degrees* < 0*) and Longitude (degrees* < 0*)* are also used to incorporate the spatial effect.

An appropriate model structure to estimate anchovy biomass depends on the understanding of the processes that structure their distribution. In fact, (Figure 3.1) shows that

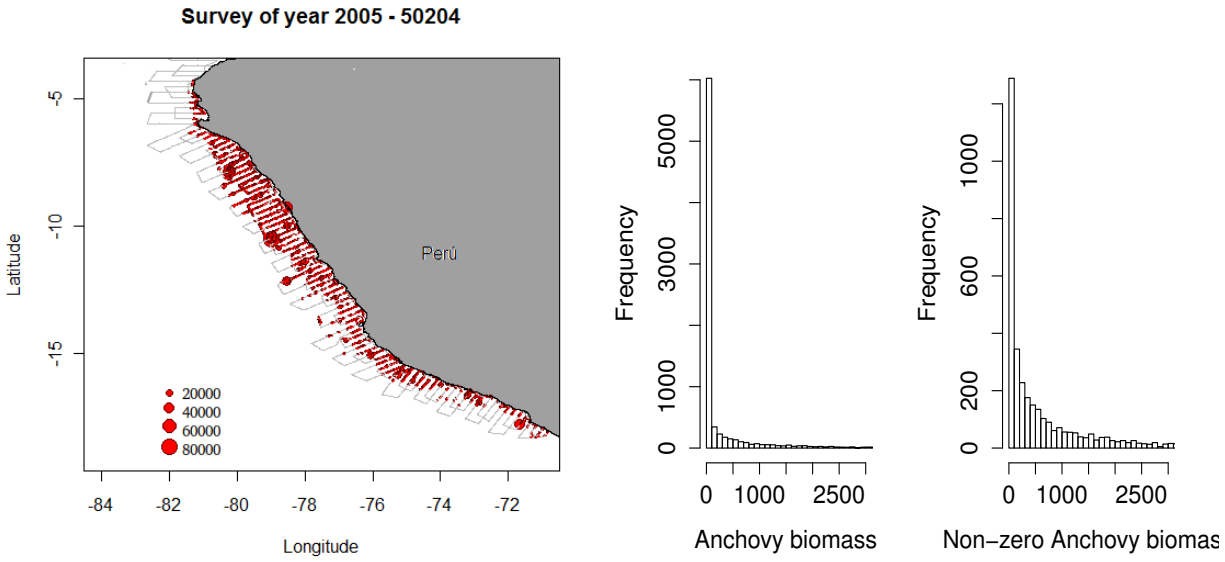

Figure 3.1: *Left: The observed data, where the trajectory of survey tracks is represented by parallel cross-shore transects (red and gray dots). Furthermore, the size of red dots correspond to the biomass of anchovy (higher than zero) and gray dots correspond to the biomass of anchovy equal to zero. Right: Exploratory analysis. Histogram for all anchovy biomass observations and Histogram for non-zero anchovy biomass observations.*

anchovy presence was fairly broadly distributed from near shore to the shelfbreak, with some medium random aggregations over this area. One of the reasons for this spatial distribution of anchovy may be attributable to the aggregative behaviors of anchovy within suitable habitat. Furthermore, the distribution of anchovy biomass was characterized by high proportions of zero values, in particular offshore, reaching approximately 57% of the 8308 observations. This may suggests that a trend surface term in the model might be appropriate and also could be necessary to lead with these zero and non-zero values (absence/presence). In addition, the left-hand histogram in Figure 3.1 shows a high frequency of zero values for all anchovy biomass while the right-hand histogram also shows a strongly right-skewed distribution for non-zero anchovy biomass. Those last results confirm that an appropriate model structure would model zero and non-zero values as an integrated process.

Chapter 4

Model Structure

Let us first define a regular lattice composed by n grid points, with $n = nrow \times ncol$, $i = 1, \ldots, nrow$; $j = 1, \ldots, ncol$. Then, let Y_{ij} be the observational (response) variable and y_{ij} be the observed values computed as the mean of anchovy biomass at each sample location that belong to the discrete n set of sampling grids. Note that this is a lattice approximation of the observed anchovy biomass. This approximation can be as good as the lattice resolution, where there is a trade off between better resolution and computational time.

As presented in Chapter 1 the idea behind a Hurdle model is to separate the zero structure from the non zero structure. Generally such Hurdle model is defined as a finite mixture of a point mass at zero with a truncated-at-zero distribution. However, in our case y_{ij}'s are positive values (≥ 0), then our Hurdle model has to be defined as a finite mixture of a a degenerate distribution with point mass at zero and a distribution with support on $\Re^+$. Bayesian approaches of this kind of mixture have been developed mainly for longitudinal data in biomedical applications using for instance logarithmic transformations of data (Ghosh and Albert, 2009) or a Log-Normal distribution (Neelon et al, 2011). In the spatial context, recently Dreassi et al. (2014) adopted a mixture with a Gamma distribution to get small area estimates of grape wine production. Hence, we proposed to use the next continuous distributions with support on $\Re^+$: Gamma, the Log-Normal and the Log-Logistic distributions, well-known distributions used quite effectively in analyzing skewed positive data. The Gamma has more of a tail on the left, and less of a tail on the right; while the far right tail of the Log-Normal is heavier and its left tail lighter. Moreover, the Log-Logistic is similar in shape to the Log-normal distribution but it has heavier tails.

Let's suppose that anchovy absence ocurrs with probability p_{ij}. Therefore, presence ocurrs with probability $1 - p_{ij}$. Define h as a probability density function (pdf) for some parametric unknown distribution with support on $\Re^+$, thus, the associated distribution for Y_{ij} has the following mixture density

$$\pi(y_{ij}|p_{ij}, \mu_{ij}, \psi) = p_{ij}\delta_0 + (1 - p_{ij})h(y_{ij}|\mu_{ij}, \psi)I_{[y_{ij}>0]}$$

where δ_0 is the Dirac measure at zero, μ_{ij} and ψ parameters corresponding to the distribution with pdf h. That is, Y_{ij} might assume a zero value with probability p_{ij} while with probability $1-p_{ij}$, $Y_{ij} > 0$ follows an unkonwn distribution with pdf h. Such notation can be a bit confusing, to avoid missunderstanding and understand how this model is obtained, we introduce a "latent" indicator variable T_{ij}, that marginally follows a Bernoulli distribution with success probability p_{ij}, while conditionally on Y is defined by

$$T_{ij} = \begin{cases} 1 \text{ if } y_{ij} = 0, \\ 0 \text{ if } y_{ij} > 0. \end{cases}$$

In particular, $\pi(y_{ij}|t_{ij} = 1, \mu_{ij}, \psi) = \delta_0$ if $y_{ij} = 0$, while $\pi(y_{ij}|t_{ij} = 0, \mu_{ij}, \psi) = h(y_{ij}|\mu_{ij}, \psi)$ if

$y_{ij} > 0$. Then, we have that

$$\pi(y_{ij}|t_{ij}, \mu_{ij}, \psi) = t_{ij}\delta_0 + (1 - t_{ij})h(y_{ij}|\mu_{ij}, \psi); t_{ij} = 0, 1. \tag{4.1}$$

Another important result is the relation between the mean posterior probability of success and the posterior predictive probability of success,

$$\begin{aligned} P(T_{ij} = 1|Y = y) &= \int P(T_{ij}|p_{ij}, Y = y)p(p_{ij}|Y = y)dp_{ij}, \\ &= \int p_{ij}p(p_{ij}|Y = y)dp_{ij}, \\ &= E[p_{ij}|Y = y]. \end{aligned}$$

This result is intuitive and it says that the future probability to have anchovy absence conditioned to the observed data is the mean posterior probability of success. A consequent and trivial result is that $P(T_{ij} = 0|Y = y) = 1 - E[p_{ij}|Y = y]$.

Using the fact that T_{ij} follows a Bernoulli distribution with succes probability p_{ij}, $\pi(y_{ij}|t_{ij}, \mu_{ij}, \psi) = \pi(y_{ij}|t_{ij}, p_{ij}, \mu_{ij}, \psi)$, then the marginal density of Y_{ij} can be calculated as follows

$$\begin{aligned} \pi(y_{ij}|p_{ij}, \mu_{ij}, \psi) &= \sum_{t_{ij}} \pi(y_{ij}|t_{ij}, p_{ij}, \mu_{ij}, \psi)\pi(t_{ij}|p_{ij}), \\ &= \sum_{t_{ij}} \pi(y_{ij}|t_{ij}, p_{ij}, \mu_{ij}, \psi)p_{ij}^{t_{ij}}(1 - p_{ij})^{1-t_{ij}}. \end{aligned}$$

and using Equation (4.1),

$$\pi(y_{ij}|p_{ij}, \mu_{ij}, \psi) = \sum_{t_{ij}} \left[t_{ij}\delta_0 + (1 - t_{ij})h(y_{ij}|\mu_{ij}, \psi)\right] p_{ij}^{t_{ij}}(1 - p_{ij})^{1-t_{ij}}, t_{ij} = 0, 1.$$

Finally, the marginal density of Y_{ij} is defined by the next Hurdle model,

$$\pi(y_{ij}|x, \theta) = p_{ij}\delta_0 + (1 - p_{ij})h(y_{ij}|\mu_{ij}, \psi)I_{[y_{ij}>0]}, \tag{4.2}$$

where p_{ij} and μ_{ij} are components of the Latent Gaussian process x, and ψ is a component of the hyperparameters θ.
The marginal likelihood function is given by

$$L(y|x, \theta) = \prod_{ij}^{n} \{p_{ij}\delta_0 + (1 - p_{ij})h(y_{ij}|\mu_{ij}, \psi)I_{[y_{ij}>0]}\}, \tag{4.3}$$

where Y_{ij}'s given a Latent Gaussian process x and the hyperparameters θ are conditionally independent.

In order to accommodate the spatial dependence and covariates we can use Equation (4.2) to define our hierarchical model as:

$$\begin{aligned} \pi(y_{ij}|x, \theta) &= p_{ij}\delta_0 + (1 - p_{ij})h(y_{ij}|\mu_{ij}, \psi)I_{[y_{ij}>0]}, \\ \text{logit}(p_{ij}) &= \eta_{ij}^{(1)} = \mathbf{Z}^{(1)}\beta^{(1)} + f_s(s_{ij})^{(1)}, \\ g(\mu_{ij}) &= \eta_{ij}^{(2)} = \mathbf{Z}^{(2)}\beta^{(2)} + f_s(s_{ij})^{(2)}, \end{aligned} \tag{4.4}$$

where logit is a cannonical link function connecting the linear predictor $\eta_{ij}^{(1)}$ with the probability of zeros p_{ij} and g is an appropriate link function which connects the linear predictor $\eta_{ij}^{(2)}$ to the parameter μ_{ij}, and it could be an identity link, a log-link, among others, depending on the unknown distribution pdf h. For each linear predictor we have that $\mathbf{Z}^{(k)}$ is the covariate matrix, $\beta^{(k)}$ are the fixed effects and $f_s(s_{ij})^{(k)}$ are the structured random effects for $k = 1, 2$. Here, $\mathbf{Z}^{(1)}$ and $\mathbf{Z}^{(2)}$ may share some common covariates but they do not need to be the same.

On the other hand, to account for the spatial random dependence we represent $f_s(s_{ij})^{(k)}$ with a Gaussian field. More specifically, with a Gaussian field with a Matérn covariance function defined by $\frac{1}{\tau_s}\rho(h)$, where $\rho(h) = \frac{1}{\Gamma(\nu)2^{\nu-1}}(\frac{\sqrt{(8\nu)}h}{r})^{\nu} K_\nu(\frac{\sqrt{(8\nu)}h}{r})$ is a Matérn correlation function, K_ν the modified Bessel function of fixed order $\nu > 0$, ν is a shape parameter and determines the smoothness of the process, h is the Euclidean distance between two locations, and the last parameter introduced called range (r) is interpreted as the minimum distance for which two locations are nearly independent. Rue and Tjelmeland (2002) showed that for a regular lattice, the Matérn correlation function can be well approximated by a Gaussian Markov random field (GMRF) (Rue and Held, 2005) which joined with analytical results given in Lindgren et al. (2011) can improved computational perfomance dramatically. Finally, it would be worth to mention that if p_{ij} and μ_{ij} are not related, it is reasonable to assume that $f_s(s_{ij})^{(1)}$ and $f_s(s_{ij})^{(2)}$ are also independent of each other but if they are related one of them may depend from the other one.

4.1 Bayesian Inference

The posterior estimates of parameters and hyperparameters are computed using Integrated Nested Laplace Approximation (INLA) (Rue et al, 2009). INLA works out with Latent Gaussian Models (LGM's), a subclass of structured additive models which can be seen as a representation of hierarchical models. In order for INLA to work properly it is necessary that the LGM's satisfy: (i) The latent field x admits conditional independence properties, thus the latent field is a GMRF; (ii) The number of hyparameters is small. In our model proposed (Equation 4.4), the spatial Gaussian fields can not be exactly GMRF's but they can be approximated to GMRF's (Rue and Held, 2005; Lindgren et al, 2011), and the number of hyperparameters which we might include is reasonable small $(\dim(\theta) \leq 5)$. These properties make it possible to obtain fast Bayesian inference.

The join posterior of LGM can be computed using the likelihood distribution of Y (Equation (4.3)), the latent Gaussian field x and the distribution of hyperparameters θ,

$$\pi(x, \theta|y) \propto \pi(\theta)\pi(x|\theta) \prod_{ij \in I} \{p_{ij}\delta_0 + (1 - p_{ij})h(y_{ij}|\mu_{ij}, \psi)I_{[y_{ij}>0]}\}.$$

From Equation (4.4), p_{ij} and μ_{ij} are connected to the likelihood through the two linear predictors defined into the latent field x. Let $x|\theta \sim N(U, Q(\theta)^{-1})$, then

$$\pi(x, \theta|Y) \propto \pi(\theta)|Q^{1/2}| \exp\left(-\frac{1}{2}(x - U)^t Q(x - U) + \sum_{ij \in I} \log\{\pi(y_{ij}|x_{ij}, \theta)\}\right),$$

where U is a mean vector which dependes on $\mathbf{Z}^{(1)}\beta^{(1)}$ and $\mathbf{Z}^{(2)}\beta^{(2)}$ and Q is a precision matrix which depend on hyperparameters $(\tau_\epsilon^{(1)}, \tau_s^{(k)}, r_s^{(k)}, \psi)$, $k = 1, 2$. Furthermore, the fixed effects $\beta^{(k)}$ have independent Gaussian priors and the priors for the hyperparameters model are chosen accordingly.

The posterior marginals of the latent variables $\pi(x_{ij}|y)$ and the posterior marginal of

hyperparameters $\pi(\theta_p|Y)$ are defined by:

$$\pi(x_{ij}|y) = \int \pi(x_{ij}|\theta, y)\pi(\theta|y)d\theta$$

$$\pi(\theta_p|y) = \int \pi(\theta|y)d\theta_{-p}.$$

These posterior marginals are not easy to compute, and that is the main aim of INLA, providing approximations to the posterior marginals of the latent variables and hyperparameters, given by Equation (4.5) and Equation (4.6),

$$\widetilde{\pi}(x_{ij}|y) = \int \widetilde{\pi}(x_{ij}|\theta, y)\widetilde{\pi}(\theta|y)d\theta, \tag{4.5}$$

$$\widetilde{\pi}(\theta_p|y) = \int \widetilde{\pi}(\theta|y)d\theta_{-p}, \tag{4.6}$$

which are both very accurate and extremely fast to compute. Here $\widetilde{\pi}$ denotes an approximation to a probability density function (pdf). In summary the main idea of INLA is divided into the following tasks: First, it provides a Gaussian approximation of $\widetilde{\pi}(\theta|y)$ to the join posterior of hyperparameters given the data $\pi(\theta|y)$. Then, it provides an approximation of $\widetilde{\pi}(x_{ij}|\theta, y)$ to the marginals of the conditional distribution of the latent field given the data and the hyperparameters $\pi(x_{ij}|\theta, y)$. And finally, it explores $\widetilde{\pi}(\theta|y)$ on a grid and use it to integrate out θ and θ_{-p} in Equation (4.5) and Equation (4.6) respectively. For more details on INLA calculations we refer to Rue et al. (2009).

4.2 Model Assessment

4.2.1 Model Comparison

In order to study the goodness of fit of the studied models we use the Deviance Information Criterion (DIC), the logarithm of the Pseudo Marginal Likelihood (LPML), accuracy rate and the mean squared estimation error (MSEE).

The DIC (Spiegelhalter et al, 2002) is a common Bayesian criterion also computed by INLA as the posterior mean of the deviance $E^{x,\theta}(D(\theta, x))$ plus the effective number of parameters p_D. For the model proposed the deviance is given by,

$$D(\theta, x) = -2\sum_{k=1}^{N} log(\pi(y_k|x, \theta)) = -2\sum_{k=1}^{N} log(p_k\delta_0 + (1 - p_k)h(y_k|\cdot)I_{[y_k>0]}),$$

where N is the number of observations (total number of grids with $y_{ij} \geq 0$). Then using INLA approximations, the posterior mean of the deviance is computed by

$$E^{x,\theta}(D(\theta, x)) = \int_{\theta,x} D(\theta, x)\pi(\theta|y)\pi(x|\theta, y)\partial\theta\partial x.$$

And finally the effective number of parameters is approximated by

$$p_D \approx N_x - \text{trace}\{Q(\theta^{me})Q^{\star}(\theta^{me})^{-1}\},$$

where N_x is the dimension of x, θ^{me} denotes the posterior median, Q denotes the prior precision matrix and $Q^{\star}$ denotes the posterior covariance matrix of the Gaussian approximation

$\widetilde{\pi}(\theta|y)$ (Rue et al, 2009).

Another alternative Bayesian model choice criterion is the conditional predictive ordinate (Geisser and Eddy, 1979; Gelfand et al, 1992) defined as $CPO_k = \pi(y_k|y_{-k}) = 1/\int \frac{\pi(x_k|y)}{\pi(y_{-k}|x_k)} dx_k$, where y_{-k} is given by y without the k-th component. Although INLA can also compute CPO_k values, it was not providing the correct values for our Hurdle model. Therefore, we decided to compute them explicitly. The Monte Carlo estimation for the CPO_k (Dey et al, 1997; Held et al, 2010) is defined as the harmonic mean of the conditional density $\pi(y_k|x_w, \theta)$,

$$\widehat{CPO}_k = \left\{ \frac{1}{W} \sum_{w=1}^{W} \frac{1}{\pi(y_k|x_w, \theta)} \right\}^{-1} ; k = 1, ..., N;$$

evaluated at samples $x_1, ..., x_W$ from $\pi(x_k|y)$. Furthermore, since the CPO_k is a goodness of fit measure for each observation, it can be summarized for all the data via a single value called the Logarithm of the Pseudo Marginal Likelihood (LPML), so comparison between models can be made using, $LPML = \sum_{k=1}^{N} \log \pi(y_k|y_{-k}) \approx \sum_{k}^{N} \log(\widehat{CPO}_k)$, that is, the higher value of LPML better the model.

On the other hand, we also calculate the accuracy rate, that is, which observations are estimated as presence when actually they are presence, and which observations are estimated as absence when actually they are absence.

Finally, to assess the closeness between the mean posterior estimation of anchovy biomass and the observed anchovy biomass it is computed the root of mean squared estimation error (RMSEE). Here, the mean posterior estimation of anchovy biomass is computed using the mean posterior estimated parameters which are computed using all observations. The root of mean squared estimation error (RMSEE) is computed as follows

$$MSEE = \sqrt{\frac{1}{N} \sum_{k}^{N} d_k^2}; \quad d_k = y_k - E(Y_k|x, \theta).$$

4.2.2 Model Predictive checks

In this section, to evaluate the predictive power of the proposed models is performed further comparison. In particular, we define three different scenarios in which each one of them follows the next four steps:

1. First it is selected a random validation sample y_V^* whose size is $V^*\%$ of the total observed values ($y_{ij} \geq 0$), and these values are not considered in the model fitting.

2. Then we fit all proposed models with the training values (data with the validation sample removed).

3. Then using the mean posterior estimated parameters by each fitted model the root of the mean squared prediction error (RMSPE) is computed using the validation random sample y_V^*,

$$RMSPE = \sqrt{\frac{1}{\mathrm{V}} \sum_{\mathrm{v}}^{\mathrm{V}} d_{\mathrm{V}}^2}; d_{\mathrm{V}} = y_{\mathrm{V}}^* - E(Y_{\mathrm{V}}|Y_{-\mathrm{V}}); \mathrm{v} = 1, \cdots, \mathrm{V}; \mathrm{V} = \mathrm{N} \times \mathrm{V}^*\%$$

4. Then steps 1-3 are repeated M times, where M is the number of simulations. Finally, the mean RMSPE is computed for each model, MRMSPE=$Mean(RMSPE_{m,(model)})$, where $m = 1, \ldots, M$.

The three scenarios are created selecting V^* to be 5, 10 and 20, repectively.

4.2.3 Influence Diagnostics

In order to locate any globally influential observations, we use Bayesian influence diagnostics. A common measure to assess the influence from one observation into the posterior estimations is the Kullback-Leibler(KL) divergence measure. The KL is defined by

$$\mathrm{KL}(\pi(x|y_{(-k)}), \pi(x|y)) = E_{x|Y}\left[-\log\left(\frac{\pi(x|y_{(-k)})}{\pi(x|y)}\right)\right].$$

A simplified expression of KL for general Bayesian models is derived in Lachos et al. (2013), and after some algebra, as shown in Appendix A the KL for a Latent Gaussian model is

$$\mathrm{KL}(\pi(x|y_{(-k)}), \pi(x|y)) = -\log(CPO_k) + E_{x|Y}[\log(\pi(y_k|x,\theta))], \tag{4.7}$$

where $E_{x|Y}[.]$ denotes the expectation with respect to the posterior distribution $\pi(x|y)$. Therefore, a Monte Carlo estimation for KL measure, is given by

$$\mathrm{KL}(\pi(x|y_{(-k)}), \pi(x|y)) = -\log(\widehat{CPO}_k) + \frac{1}{W}\sum_{w=1}^{W}\log[\pi(y_k|x_w,\theta)]. \tag{4.8}$$

Note that the KL measure does not directly define when some observation is influential or not. In order to define an inlfluential observation it is necessary to define a cutoff point. McCulloch (1989) proposes a calibration method to determine which observations are influential. The calibration is done by comparing the density of unbiased coin π_1 with the density of a biased coin π_2. The divergence of these densities can be calculated as a function of the KL measure. The measure is zero only when $q_k = 0.5$ and it is increasing when $|q_k - 0.5|$ increases. Therefore, this calibration can be done by solving for q_k such that $\mathrm{KL}(\pi(x|y_{(-k)}), \pi(x|y)) = \mathrm{KL}(Ber(0.5); Ber(q_k)) = -\log[4q_k(1-q_k)]/2$, where $Ber(q_k)$ denotes the Bernoulli distribution with success probability q_k. This implies that, $q_k = 0.5[1 + \sqrt{1 - \exp(-2\mathrm{KL}(\pi(x|y_{(-k)}), \pi(x|y)))}]$ and the ith observation is called influential if $q_k \gg 0.5$.

Finally, to end this chapter it would be worth to mention the importance of Monte Carlo methods to estimate the conditional predictive ordinate and the KL measure. Because even when INLA is used, such methods are helpful when we are dealing with complex models where these kind of measures have to be computed explicitly.

Chapter 5

Application

The shape of the coast of Perú is a bit northwest diagonal, using this fact we define a regular lattice for the spatial domain of interest as a diagonal regular lattice. This strategy reduces significatly computational time requirements, in particular for Bayesian inference. Although diagonal regular lattices can not be directly treated as matrices into their original space, applying suitable translations and rotations to the coordinates it is possible to work with the diagonal regular lattice as a matrix into another space. Thus, original grid coordinates (Lon_{ij}, Lat_{ij}) are translated to the origin (0,0) and rotated α radians, that is,

$$\widehat{Lon_{ij}} = (Lon_{ij} + a)cos(\alpha) + (Lat_{ij} + b)sin(\alpha),$$

$$\widehat{Lat_{ij}} = -(Lon_{ij} + a)sin(\alpha) + (Lat_{ij} + b)cos(\alpha),$$

where $a = max(Lon_{ij})$ and $b = min(Lat_{ij})$.
The translated and rotated regular lattice is composed by 6400 (n=6400) grid dots (Figure 5.1).

It should also be noted that although the model structure is implemented using this translated and rotated regular lattice, our final results are re-transformed into the original spatial coordinates, to avoid misleading interpretations.

5.1 Exploratory Analysis

The relationship between anchovy presence and biomass against covariates may involve the existence of spatial trends which need to be included in the model. In Figure 3.1 it is shown that the higher region of anchovy presence ranged from latitudes between 6° and 12°S, while anchovy biomass is dispersed all over the coast. Furthermore, a higher distance to the coast seems to involve higher anchovy absence and higher anchovy biomass when it is present (Figure 5.2, left panel) . Also bigger depth seems to involve higher anchovy absence while anchovy biomass seems to be concentrated within depths less than 500 meters (Figure 5.2, right panel). Until this moment, the model proposed has not been defined completely, it is necessary to specify the pdf h of the mixture. In order to find out which distribution better adjust to the positive biomass observations $(y^{\star}_{ij} = y_{ij} \geq 0)$ we fit a simple generalized linear regression model of the form

$$g(\mu_{ij}) \quad = \quad \mathbf{Z_{ij}}\beta,$$

where g is the appropriate link function for the unknown distribution, and $\mathbf{Z_{ij}}$ is the covariate matrix defined by an intercept, distance to the coast, latitude, latitude^2 and depth.

Finally, the three models are adjusted with different distribution choices: Gamma, Log-Logistic and Log-Normal. Table 5.1 shows that the DIC and LPML statistics agreed that the Gamma distribution is the preferred model. From now on the paper we set the unknown

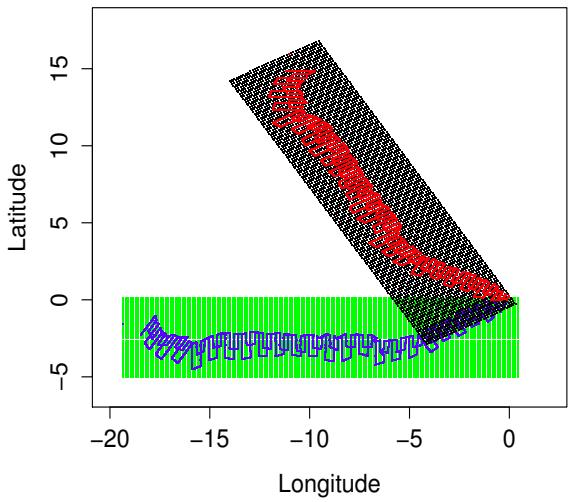

Figure 5.1: *The black grids dots represent the translated regular lattice, here red grid dots are samples of anchovy biomass traslated too. And green grids dots represent the rotated Regular lattice, here blue grid dots are samples of anchovy biomass rotated too.*

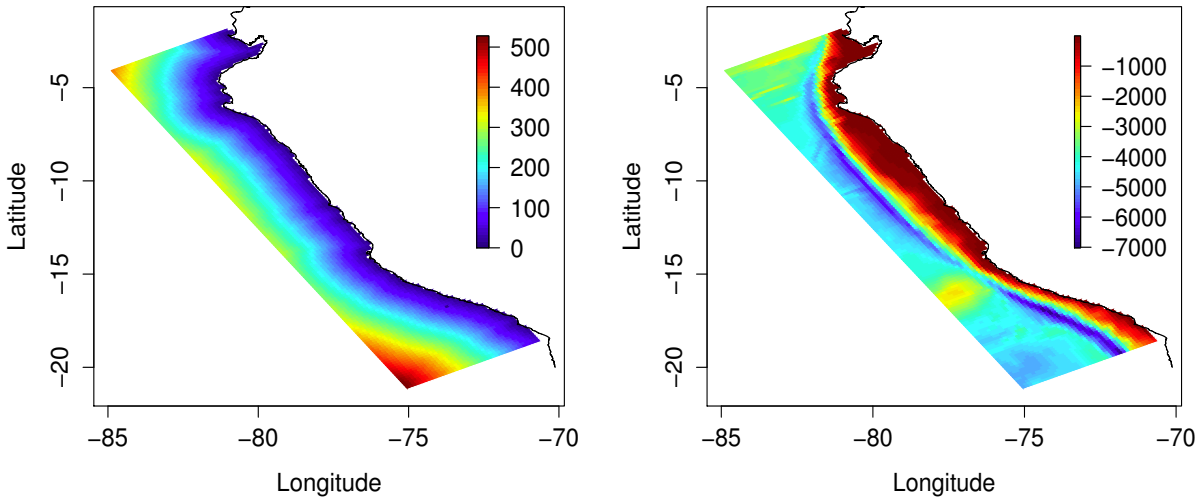

Figure 5.2: *Left: Regular lattice for distance to the coast. Right: Regular lattice for ocean depth.*

Table 5.1: *Selection criteria for the different positive Distributions*

	DIC	*LPML*
Gamma	**8923.844**	**-4458.199**
Log-Logistic	8991.108	-4492.816
Log-Normal	9020.682	-4507.793

distribution to be a Gamma$(\phi, \phi/\mu)$ having the following density

$$h(y_{ij}^{\star}) = \frac{1}{\Gamma(\phi)} \left(\frac{(\phi)}{\mu_{ij}}\right)^{(\phi)} (y_{ij}^{\star})^{\phi-1} \exp\left(-(\phi)\frac{y_{ij}^{\star}}{\mu_{ij}}\right)$$

then $\mathrm{E}(Y_{ij}^{\star}) = \mu_{ij}$ and $\mathrm{Var}(Y_{ij}^{\star}) = \mu_{ij}^2/(\phi)$, where μ_{ij} is the mean and ϕ is the precision parameter. And the linear predictor $\eta_{ij}^{(2)}$ is linked to the mean μ_{ij} using a log-link function, then $\mu_{ij} = \exp(\eta_{ij}^{(2)})$.

Thus, the model proposed in Equation (4.2) is defined by

$$\begin{aligned} \pi(y_{ij}|x,\theta) &= p_{ij}\delta_0 + (1-p_{ij})h(y_{ij}|\mu_{ij},\psi)I_{[y_{ij}>0]}, \\ \mathrm{logit}(p_{ij}) &= \eta_{ij}^{(1)} = \mathbf{Z}^{(1)}\beta^{(1)} + f_s(s_{ij})^{(1)}, \\ \log(\mu_{ij}) &= \eta_{ij}^{(2)} = \mathbf{Z}^{(2)}\beta^{(2)} + f_s(s_{ij})^{(2)}. \end{aligned}$$

Furthermore, only the distance to the coast and the depth result significant covariates. Then, all explanatory variables will be included in $\mathbf{Z}^{(1)}$ but only the distance to the coast and the depth will be included in $\mathbf{Z}^{(2)}$.

It would be worth to mention that this criteria to choose the unknown distribution might not be the best choice because it does not assure that if we run the proposed models (Table 5.2) including the structured spatial terms for the three distributions the distribution chosen would not be changed. But it is a reasonable criteria in order to reduce the computational time requirements.

5.2 Data Analysis

After selecting the distribution for positive anchovy biomass as Gamma and in order to verify the necessity of the full model presented in Equation (4.4), we introduce a variety of submodels that will be used for model comparison. All submodels are presented in Table 5.2. The "full model", model I, is the model with all possible components. Model II have a shared spatial component instead of two separate spatial components, this model incorporate another hyperparameter called δ, an unknown scale parameter that explain the degree of relation from the structured spatial term $f_s(s_{ij})^{(1)}$ to the linear predictor $\eta_{ij}^{(2)}$. Models III and IV have only one spatial component in the linear predictor of the anchovy biomass or in the linear predictor of the probability of zero, respectively. Finally, Model V has no spatial effect.

Table 5.3 presents the selection criteria of the fitted models with different choices of the smooth parameter $\nu = 1, 2, 3$ as presented in Section 4.2.1. Overall model I is the prefered one among all criteria and all scenarios. Specifically, the best accuracy rate of classification (anchovy absence/presence) is for model I and model IV (97.61%) with $\nu = 1$. Although model IV classifies fairly good anchovy presence, the RMSEE for the anchovy biomass is not good for this model due to the lack of a spatial effect to specifically predict anchovy biomass. For this reason DIC and LPML values indicate that model I with $\nu = 1, 2$ and 3 have a better goodness of fit than the rest of models. On the other hand, RMSEE is by far in favour of model I for all choices of ν, being better for $\nu = 1$. We can conclude, if the model classifies correctly anchovy presence/absence, the global estimation of anchovy biomass would be better too. This means that it is really necessary first to classify anchovy presence and then estimate anchovy biomass, using the knownledge that anchovy is present with high probability, like model I does. Furthermore, models with $\nu = 1$ have a better perfomance than its similar with other choices of ν.

Table 5.2: *Linear predictors and hyperparameters for each proposed model*

Models	Linear predictor	Hyperparameters
Model I	$\eta_{ij}^{(1)} = \mathbf{Z}^{(1)}\beta^{(1)} + f_s(s_{ij})^{(1)}$ $\eta_{ij}^{(2)} = \mathbf{Z}^{(2)}\beta^{(2)} + f_s(s_{ij})^{(2)}$	$\tau_s^{(1)}, r_s^{(1)}$ $\phi, \tau_s^{(2)}, r_s^{(2)}$
Model II	$\eta_{ij}^{(1)} = \mathbf{Z}^{(1)}\beta^{(1)} + f_s(s_{ij})^{(1)}$ $\eta_{ij}^{(2)} = \mathbf{Z}^{(2)}\beta^{(2)} + \delta f_s(s_{ij})^{(1)}$	$\tau_s^{(1)}, r_s^{(1)}$ ϕ, δ
Model III	$\eta_{ij}^{(1)} = \mathbf{Z}^{(1)}\beta^{(1)}$ $\eta_{ij}^{(2)} = \mathbf{Z}^{(2)}\beta^{(2)} + f_s(s_{ij})^{(2)}$	 $\phi, \tau_s^{(2)}, r_s^{(2)}$
Model IV	$\eta_{ij}^{(1)} = \mathbf{Z}^{(1)}\beta^{(1)} + f_s(s_{ij})^{(1)}$ $\eta_{ij}^{(2)} = \mathbf{Z}^{(2)}\beta^{(2)}$	$\tau_s^{(1)}, r_s^{(1)}$ ϕ
Model V	$\eta_{ij}^{(1)} = \mathbf{Z}^{(1)}\beta^{(1)}$ $\eta_{ij}^{(2)} = \mathbf{Z}^{(2)}\beta^{(2)}$	 ϕ

The last column of Table 5.3 presents the total running time for each model. From the computational time presented, we can see that even the full model runs in a resonable time. Recall, that the grid has 6400 sites and model I has fixed effects and two different latents spatial fields. This observation allows to emphasize the computational adavantage of using INLA methodology when comparing to traditional MCMC methods for spatial data analysis.

Since $\nu = 1$ provides the best overall fit for all models, we now focus on $\nu = 1$ to investigate the models with best predictive performance (section 4.2.2). To do so, we used a hundred validation random samples ($M = 100$) of size 57 ($V = 5\%$), 113 ($V = 10\%$) and 226 ($V = 20\%$) for each model. From results reported in Table 5.4, the lower mean RMSPE values (MRMSPE) for each case are again in favour of model I. This means that model I with $\nu = 1$ not only have the better goodness of fit, it is also better for predicting anchovy biomass among all models.

After selecting model I as the preferred one for both fitting and predicting, we investigate its posterior parameters estimates. The posterior parameters estimates are reported in Table 5.5. From Table 5.5 we can see that the mean posterior fixed effect for the distance to the coast in the absence/presence probability part (0.053) indicates that the higher distance to the coast higher the probability of anchovy absence. The contribution of this parameter to the probability anchovy absence/presence (when all other parameters are kept fixed) increases really fast as distance to the coast increases. Bertrand et al. (2011) argues that there is a permanent pattern of reduction in anchovy presence with an increasing distance to the coast, where the slope of this reduction depends on the environmental conditions. On the other hand, the mean posterior of the distance to the coast effect in the positive anchovy biomass (0.015) indicates that a higher distance to the coast implies a higher anchovy biomass. When the distance to the coast increases the model estimates that the anchovy biomass increases too. These two considerations may appear a contradiction at a first sight. However, based on the posterior effect we can see that the anchovy biomass increases much slower than the probability of anchovy absence increases. For instance, by each 1km incremented of distance to the coast, the probability of anchovy absence is increased 50% while anchovy biomass is increased 15%. Thus, for large distances to the coast the model classifies the observation as an anchovy absence, as it was observed in the real data.

The mean posterior fixed effect for the latitude and latitude squared in the probability

Table 5.3: *The selection criteria for the models proposed with different linear predictors*

	LPML	DIC	Accurate rate	RMSEE	Total run time (min)
$\nu = 1$					
Model I	-3209.41	7114.95	97.61%	165.94	16.05
Model II	-3390.66	7283.56	89.46%	1069.71	7.94
Model III	-3384.17	7244.60	86.71%	499.55	3.22
Model IV	-4632.66	9523.28	97.61%	1507.39	4.41
Model V	-4843.13	9696.66	86.71%	1539.13	0.13
$\nu = 2$					
Model I	-3236.78	7073.40	96.46%	251.83	22.70
Model II	-3378.76	7224.86	89.01%	1052.14	12.23
Model III	-3405.80	7253.58	86.71%	499.42	5.70
Model IV	-4647.40	9517.15	96.46%	1509.88	5.62
Model V	-4843.13	9696.66	86.71%	1539.88	0.13
$\nu = 3$					
Model I	-3254.41	7114.28	96.10%	262.04	31.20
Model II	-4686.77	9526.58	94.06%	1528.39	132.81
Model III	-3388.50	7273.37	86.71%	499.32	7.73
Model IV	-4653.28	9515.72	96.10%	1510.47	7.89
Model V	-4843.13	9696.66	86.71%	1539.88	0.13

Table 5.4: *Predictive model checks. Mean of RMSPE (MRMSPE) out 100 validation samples.*

	MRMSPE		
$\nu = 1$	5%	10%	20%
Model I	1341	1515	1786
Model II	1541	1886	2355
Model III	1387	1581	1827
Model IV	1863	2386	3147
Model V	2106	2745	3568

Table 5.5: *Summary statistics (point, standard deviation and 95% credible interval (CI)) for Fixed effects and Hyperparameters estimation.*

	Mean	sd	95% CI
Probability of anchovy absence/presence			
Intercept	16.530	7.215	(4.641,33.138)
Distance to the coast	0.053	0.022	(0.019,0.104)
Latitude	4.384	1.608	(1.819,8.127)
$Latitude^2$	0.184	0.069	(0.072,0.346)
Depth	-0.002	0.001	(-0.003,-0.001)
$\tau_s^{(1)}$	0.056	0.027	(0.019,0.124)
$r_s^{(1)}$	7.089	1.291	(4.853,9.903)
Positive anchovy biomass			
Intercept	5.056	0.208	(4.650,5.467)
Distance to the coast	0.015	0.007	(0.015,0.015)
Depth	0.001	0.000	(0.000,0.001)
ϕ	148.601	111.494	(25.170,438.440)
$\tau_s^{(2)}$	0.214	0.015	(0.186,0.245)
$r_s^{(2)}$	1.370	0.159	(1.085,1.708)

anchovy absence/presence (4.384 and 0.184, respectively) are evidence that there is a higher probability of anchovy absence when latitudes are near to the extremes. When looking at the mean posterior fixed effect for the depth in the absence/presence probability part (-0.002) there is an indication that for deeper ocean parts there is a higher probability of anchovy absence. On the other hand, the fixed effect for the depth in the positive anchovy biomass (0.001) suggests that a lower ocean depth size implies a higher anchovy biomass.

The mean posterior range $r_s^{(1)}$ for the structured spatial effect of the probability of anchovy absence/presence is approximately 7.089 "units", where units here means number of cells. Here the lattice have cells of size aproximately 14x14km (width x height). Thus, the model states that the probability of absence of anchovy for some site are depedent of neighbors observations until a distance of 100 km. The mean posterior range $r_s^{(2)}$ for the structured spatial effect of the positive biomass is approximately 1.37 "units". Therefore, by model I anchovy biomass (when anchovy is present) depends on neighbors observations until a distance of 20 km.

Finally we can see that the spatial dependence is captured by the random spatial effect $f_s^{(1)}$, thus, we can conclude that the spatial model is capable of accomodating the variability in the anchovy distribution. On the otherhand, the variability of positive anchovy biomass not explained by the structured spatial term $f_s^{(2)}$ depends on μ_{ij} and $/\phi$, that is, if $\phi \gg \mu_{ij}^2$ then there is very little unexplained variability in the positive anchovy biomass, otherwise, there is high variability of anchovy biomass not explained by the covariates and/or the structured spatial effect.

The mean posterior probability of anchovy absence/presence (Figure 5.3, right panel) is computed using the posterior mean linear predictor $\eta_{ij}^{(1)}$. On the left panel we can see the original anchovy absence/presence values. Comparing both sides we can see that the right panel classifies very closely to the observed absence/presence of anchovy in the left side, supporting the good classification performance under model I.

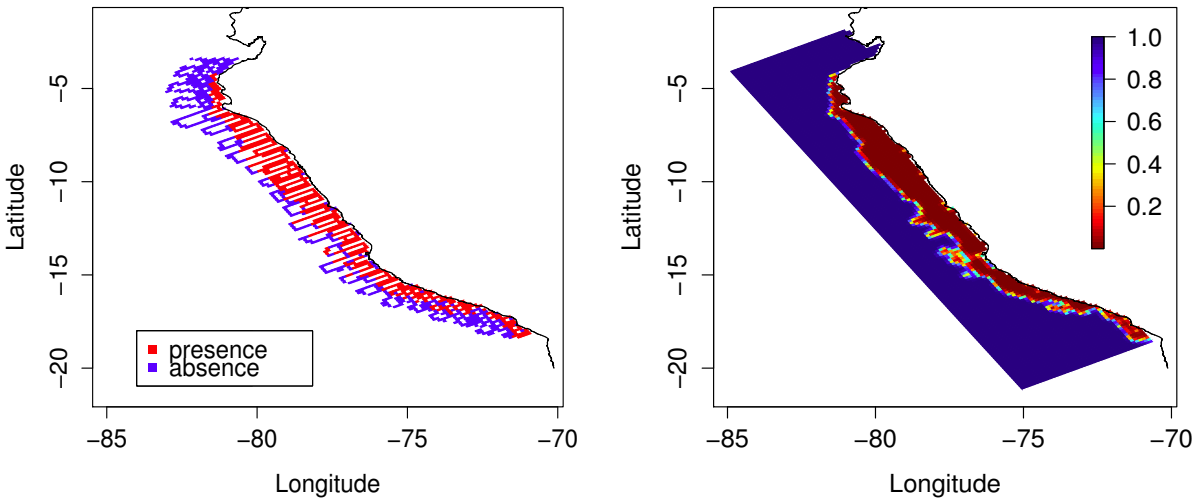

Figure 5.3: *Results. Left: Observed anchovy absence/presence. Right: Mean posterior probability of anchovy absence under model I.*

Then the estimated mean posterior anchovy biomass is computed using the above result and the model proposed in Equation (5.1). Thus, for each grid point it is defined the following prediction rule: If the mean posterior of anchovy absence probability is higher than 0.5 the site is classified like anchovy absence, while if the mean posterior anchovy absence probability is lower than 0.5 the site is defined like anchovy presence (anchovy biomass(> 0)) with estimation computed using the mean posterior linear predictor $\exp(\eta_{ij}^{(2)})$.

The posterior mean estimated anchovy biomass map is presented in Figure 5.4. On the left panel we can see the original observation values, while on the right panel the estimated pattern is presented (both cases on the logarithmic scale for anchovy biomass). Comparing both sides we can see that the right panel agrees very closely to the observed values in the left side. This is an indication of good model fitting, and thus, model I is used for prediction on the unobserved sites.

Furthermore, probabilities q_i from section 4.2.3 are computed using Kullback-Leibler(KL) measure presented in Equation(4.8). Then a threshold of 0.9 is set for q_i. After detecting the influential observations we define an influential region at those sites i where $q_i > 0.9$ and at least five neighboor(from eight) have $q_j > 0.9$ for site j neighboor from site i. Thus, if a grid point have more than one neighboor with $q_j > 0.90$, we call it an influential region. Figure 5.5 shows the probability q_i value and mean posterior biomass anchovy for all influential regions. Such regions probably represent shoal clusters with higher anchovy biomass values.

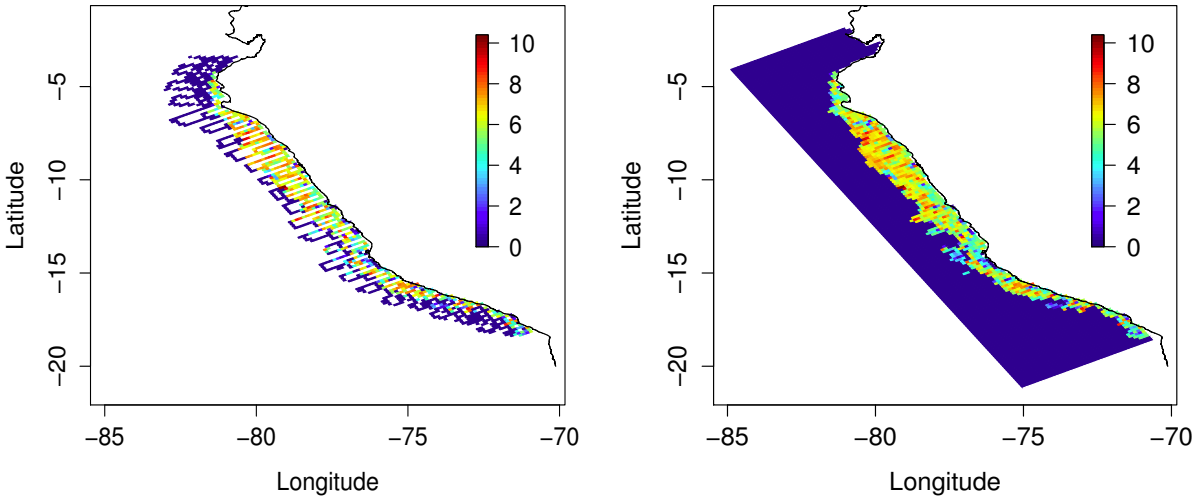

Figure 5.4: *Results. Left: Observed anchovy biomass (on the logarithmic scale). Right: Mean posterior anchovy biomass under model I (on the logarithmic scale).*

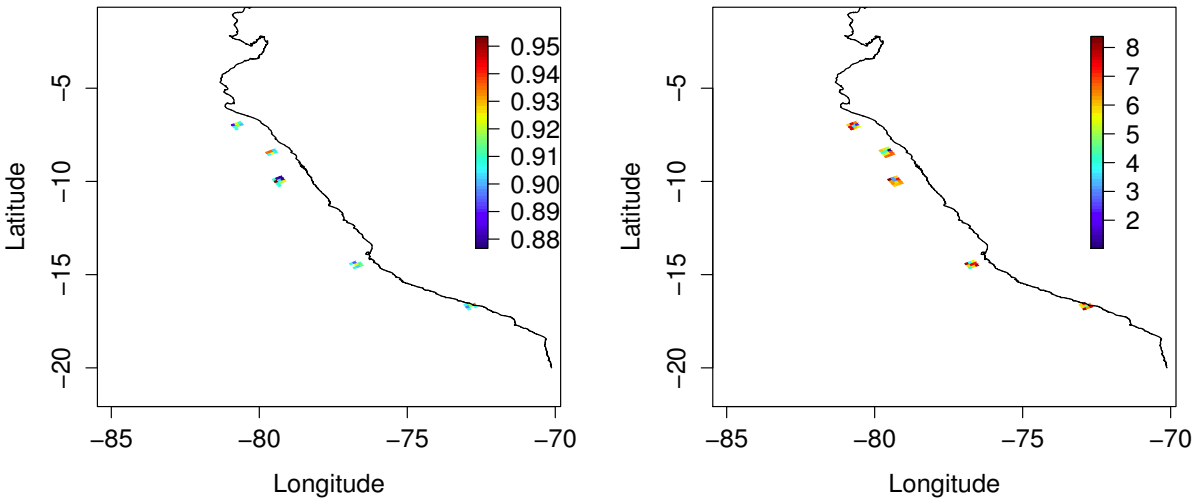

Figure 5.5: *Results. Left: Probability q_i of influence regions. Right: Mean posterior of anchovy biomass influence regions (on the logarithmic scale).*

Chapter 6

Discussion

The Peruvian anchovy is a dominant fish at the Peruvian pelagic ecosystem. For this reason, understanding the anchovy biomass distribution is very important for either an economic perspective as well as an ecological perspective. Because of the nature of anchovy biomass there is a need for a spatial model that can accommodate its main characteristics. The anchovy biomass has three main characteristics: (1) high proportions of zero, (2) strong spatial dependence and (3) very right skewed distribution. In this study we proposed a Bayesian Hurdle type model which showed its capability of adapting to the characteristics of the anchovy biomass in the coast of Perú. In fact, including a component to capture the excess of zeros was necessary and it was an indication that the source of overdispersion present in the data might comes essentially from the excess of zeros. The Gamma distribution for fitting anchovy positive biomass data was very effectively for analyzing our skewed non-negative data.

The main advantage of the Bayesian formulation, in particular using INLA, is computational. The complex model chosen including two spatial components runs in slightly over sixteen minutes. Moreover, Bayesian inference and model comparison suggest that such model provides a plausible description of the anchovy biomass, being capable of successfully address the data challenging characteristics. The effectiveness of such model for identifying the absence/presence of anchovy is fairly good. And the mean posterior estimations verify that this model has potential not only to identify absence/presence of anchovy but also to estimate anchovy biomass given that anchovy is present.

In addition, we verify the prediction performance of the proposed models. The dataset were separated into two parts: a validation part and a fitting part. This procedure was repeated one hundred times and the prediction power of the models were estimated. It is worth noting that again the computational cost to run all this simulations using INLA was very low. From the results we conclude that the full model has the best performance in both fitting and predicting, being the one selected from the majority of the model assesment measures.

From our results we observe that anchovy absence is higher when distance to the coast increases, that is, in offshore waters (>150km from the coast) where ocean depth is also deeper. And anchovy presence is higher in the central region, such area is characterized by lower oxygen, cold and fresh water in coastal surface layers which is an ideal habitat for anchovy (Bertrand et al, 2011). Furthermore, anchovy biomass slightly increases when distance to the coast increases within 150 km from the coast, reflecting the existence of dense schools not too near to the coast but where ocean depth is not too deeper. Moreover, we showed that anchovy absence/presence at some location is not independent of absence/presence at neighboring locations, as well as biomass anchovy at some location is not independent of biomass anchovy at nearest neighboring locations, supporting the inclusion of spatial effects. Bayesian influence diagnostic also suggests that classifying absence/presence suffers

less influence than correctly estimating the current anchovy positive biomass.

Finally, the current model is based on data from only one year, but it can be extended in several ways. The most inmediate one, is to a spatio-temporal model including data from more years, seasons, among others.

All code implementing our data analysis are available, please contact to the authors for providing them.

Chapter 7

Future works

7.1 Introduction

In this study the GRMF was a discretely indexed Gaussian field, thus our approach was restricted to graphs like regular lattices (Rue and Held, 2005). The problem of doing this is the necessity to approximate the position of the locations. Thus a lot of effort going into collecting locations data with a high degree of precision is lost. Clearly, the finer the lattice, better estimations, in particular, better predictions at unsampled locations. Therefore the quality of our estimations primarily depends on the size of the grid. As a result, we are required to compute on a much finer grid than it is necessary for better results, then lattice based approaches might be computationally expensive.

Throughout the next section, it is discussed a recent work of Lindgren et al. (2011) that has broken down the barrier between GMRFs and spatially continuous Gaussian random field models. They suggest a link between Gaussian Random fields and Gaussian Markov Random Fields (GRMF) through the stochastic partial differential Equations (SPDE) for a Gaussian field with Matérn covariance function. In particular, for irregular grids they use the finite element method (FEM) to discretize complex geometries, even irregular geometric areas, and at the same time they get an approximation to the solution of the SPDE using basis functions. That allows us to hold on to the continuous interpretation of space, while the computational algorithms only see discrete structures with Markov properties. A great variety of applications using SPDE approach for geostatistical data can be found in Bolin (2012), Simpson et al. (2012), Blangiardo et al. (2013) and Cameletti et al. (2013).

7.2 The Stochastic Partial Differential equation (SPDE) approach

A differential equation can be defined informally as any equation which contains derivatives. In particular, a differential which has differential derivatives is called partial differential. Generally, a Stochastic Partial Differential equation (SPDE) is a equation which allows randomness in the coefficients of the differential equation, i.e., involves one or more stochastic processes. Then any solution of a stochastic differential equation must involve some randomness, that is, we are only able to say something about the probability distribution of the solutions.

In this context, a Gaussian field $Z(s)$ with the Matérn Covariance is a solution to the linear fractional stochastic partial differential equation (SPDE)

$$(\kappa^2 - \triangle)^{\alpha/2} Z(s) = W(s), s \in R^n, \alpha = \nu + d/2, \tag{7.1}$$

where $(\kappa^2 - \triangle)^{\alpha/2}$ is the pseudo - differential operator of Laplace and W is spatial Gaussian White noise. This explicit formulation is the strong version of the SPDE specification, an

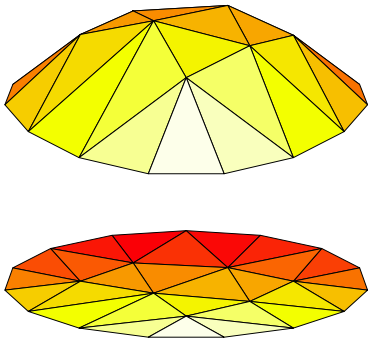

Figure 7.1: *Representation of piewise-linear approximationof a function in two dimensions over a triangulated mesh.*

alternative is the weak formulation given by

$$[< \phi_j, (\kappa^2 - \triangle)^{\alpha/2} Z >_\Omega] \xrightarrow{\text{d}} [< \phi_j, W >_\Omega]; j = 1, ..., m \tag{7.2}$$

where ϕ_j is an arbitrary well-behaved function, $< f, g >$ is the inner product of functions f and g , and $\xrightarrow{\text{d}}$ denotes convergence in distribution. This definition implies that a process is said to achieve weakly convergence if the first and second moments exist.

If the spatial locations are on some irregular grid Lindgren et al. (2011) proposed to construct a representation of the finite element as solution to the SPDE, thus using the finite element method (FEM) observations are interpolated to the nearest grid point. In order to do this, suppose that $\Re^2$ is subdivided into a set of non-intersecting triangles, where any two triangles meet in at most a common edge. The suggestion is to start with the locations of the observed points and add some triangles with some restriction to maximize the allowed length edge and to minimize the desired angles (Figure 7.1). Therefore, this approach need to construct a finite dimensional representation of solutions to the Equation (7.1), for some chosen basis functions ψ_k, weights w_k normally distributed and n vertices in the triangulation. This approximation is of the form,

$$Z(s) = \sum_{k=1}^{n} \psi_k(s) w_k. \tag{7.3}$$

In particular, the basic functions ψ_k used by Lindgren et al. (2011) are piecewise linear in each triangle, where ψ_k is 1 at vertex k and 0 at all other vertices. Therefore, the weights determine the field values at the vertices, and the values inside the triangles are calculated by linear interpolation. Hence , the finite dimensional solution is obtained by finding the distribution of weights w_k in Equation (7.3) that fulfils the weak SPDE specification (Equation (7.2)) for a specific set of functions ϕ_k and for $m = n$. Then $w_k \sim N(0, Q_\alpha^{-1})$ and the finite dimensional representation of solutions to the Equation (7.1) has precision matrix Q_α defined as a function of κ^2 e α. Functions ϕ_k and precision matrices Q_α for each α are defined in Table 7.1. Here C, G and K are matrices defined by,

$$C_{i,j} = \langle \psi_i, \psi_j \rangle, \qquad G_{i,j} = \langle \nabla\psi_i, \nabla\psi_j \rangle, \qquad (K_{\kappa^2})_{i,j} = \kappa^2 C_{i,j} + G_{i,j} \tag{7.4}$$

Table 7.1: *Summary of functions ϕ_k and precision matrices Q_α for each α*

α	ϕ_k	Q_α
1	$(\kappa^2 - \triangle)^{1/2}\psi_k$	K_{κ^2}
2	ψ_k	$K_{\kappa^2}C^{-1}K_{\kappa^2}$
3,4,...	recursive Garleky formulation	$K_{\kappa^2}C^{-1}Q_{\alpha-2,\kappa^2}C^{-1}K_{\kappa^2}$

In particular, these matrices are calculated by using triangulation and geometry of any arbitrary triangle defined by their vertices, edges and angles (Lindgren and Rue, 2007; Lindgren et al, 2011). Finally, the fact that C^{-1} is a dense matrix implies that the precision matrix will also be dense, for this reason C is approximated by a diagonal sparse matrix $\widetilde{C}$, thus we get the precision matrix of a GRMF.

Then we are able to re-write the linear predictor of interest using such GRMF, for example:

$$\eta(s) = \beta_0 + \sum_{k=1}^{\eta_{\beta_k}} \beta_k x_k(s) + Z(s) + \epsilon(s),$$

where Z is a Gaussian field, can be re-written as

$$\eta(s) = \beta_0 + \sum_{k=1}^{\eta_{\beta_k}} \beta_k x_k(s) + \widetilde{Z}(s) + \epsilon(s),$$

where the $\widetilde{Z}(s)$ is a GRMF with precision matrix Q_α.

Coming back to our data analysis, as future work is feasible to extend our study using irregular grids to hold on the natural continuous interpretation of space. Thus, we might extend the model structure defined in Chapter 4 but instead of using a regular lattice and compute the observed response as the mean of anchovy biomass we could use a irregular grid and construct a mesh using Constrained Refined Delaunay Triangulation. Thus, each sample location of anchovy biomass can be defined at each vertice or if we want to reduce computational requirements we might interpolate sample locations given some maximum triangle edge length. Although this step is easy to implement using the R-INLA package, the mesh size influences computational time needed to fit the model. More nodes on the mesh need more computational time. Moreover, spatial dependence is also influenced by scales. Thus, when using irregular grids the size of triangles has to be carefully thought out. Figure 7.2 displays examples of meshs constructed for anchovy biomass data. Then making use of the link between Gaussian fields and GRMF through SPDE, the proposed model defined in Equation (4.4) can be re-defined. Then we are able to focus on estimation and prediction procedures.

Finally, although this work was focus on a particular hierarchical spatial Hurdle-model, the SPDE approach can also be extended to spatio-temporal models. For instance, Cameletti et al. (2013) consider a spatio-temporal model with separable covariance function and use also INLA for fast Bayesian inference. Furthermore, it is feasible to consider models with more complex structures like non-separable covariance functions. With this modeling would be plausible to interpret changes on especies behavior due to changes on environmental conditions, and develop efficient and realistic simulation tools enabling to anticipate better the efficacy of different management strategies and conservation of peruvian anchovy.

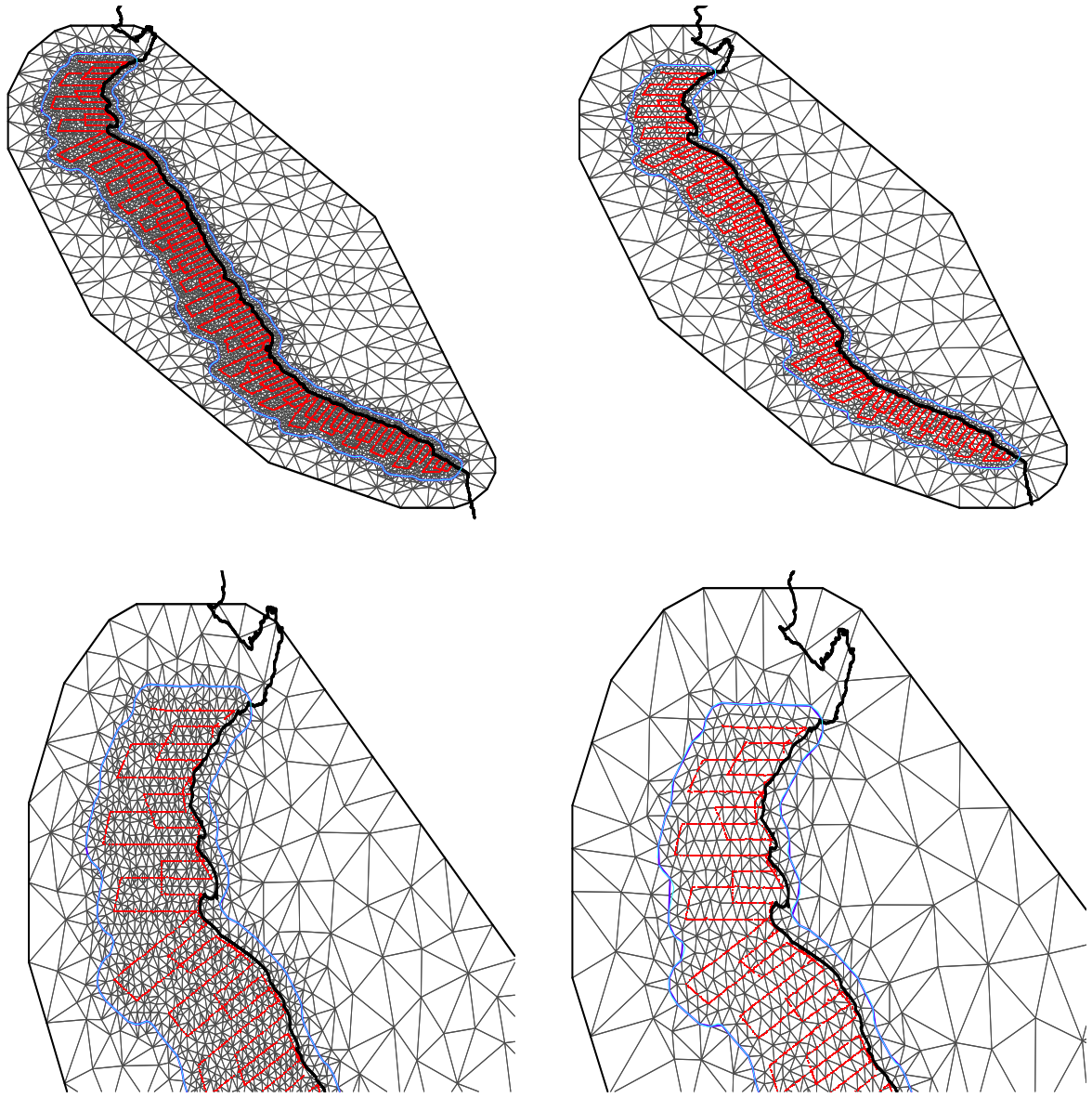

Figure 7.2: *Meshs constructed using Constrained refined Delaunay triangulation. Red grid dots are samples of anchovy biomass samples. The region defined by the sky-blue line is the edge boundary which defines the priority area for estimation. Outside this region the boundary effects are higher. Right: Mesh have less resolution than left plots. Top: Mesh for all data. Down: Mesh for northern region of top panel.*

Appendix A

Proof of result 4.7

Proof of Kullback-Leibler (KL) divergence measure:

$$\begin{aligned}
\mathrm{KL}(\pi(x|Y_{-k}), \pi(x|Y)) &= E_{x|Y}\left[-\log\frac{\pi(x|Y_{-k})}{\pi(x|Y)}\right] = E_{x|Y}\left[\log\frac{\pi(x|Y)}{\pi(x|Y_{-k})}\right], \\
&\text{by definition of expectation,} \\
\mathrm{KL}(\pi(x|Y_{-k}), \pi(x|Y)) &= \int \pi(x|Y)\log\left[\frac{\pi(x|Y)}{\pi(x|Y_{-k})}\right]dx, \\
&\text{by Bayes Theorem} \\
\mathrm{KL}(\pi(x|Y_{-k}), \pi(x|Y)) &= \int \pi(x|Y)\log\left[\frac{\pi(Y|x)\pi(x)}{\pi(Y)}\frac{\pi(Y_{-k})}{\pi(Y_{-k}|x)\pi(x)}\right]dx \\
&= \int \pi(x|Y)\log\left[\frac{\pi(Y|x)}{\pi(Y_{-k}|x)}\frac{\pi(Y_{-k})}{\pi(Y)}\right]dx \\
&= \int \pi(x|Y)\log\left[\frac{\pi(Y|x)}{\pi(Y_{-k}|x)}\right]dx + \int \pi(x|Y)\log\left[\frac{\pi(Y_{-k})}{\pi(Y)}\right]dx \\
&= \int \pi(x|Y)\log\left[\frac{\pi(Y|x)\pi(x|\theta)}{\pi(Y_{-k}|x)\pi(x|\theta)}\right]dx + \int \pi(x|Y)dx\log\left[\frac{\pi(Y_{-k})}{\pi(Y)}\right] \\
&\pi(x|Y) \text{ is a density, then} \\
\mathrm{KL}(\pi(x|Y_{-k}), \pi(x|Y)) &= \int \pi(x|Y)\log\left[\frac{\pi(Y|x)\pi(x|\theta)}{\pi(Y_{-k}|x)\pi(x|\theta)}\right]dx + 1\times\log\left[\frac{\pi(Y_{-k})}{\pi(Y)}\right] \\
&\text{by conditional probability} \\
\mathrm{KL}(\pi(x|Y_{-k}), \pi(x|Y)) &= \int \pi(x|Y)\log\left[\frac{\pi(Y|x,\theta)}{\pi(Y_{-k}|x,\theta)}\right]dx + \log\left[\frac{\pi(Y_{-k})}{\pi(Y)}\right] \\
&= \int \pi(x|Y)\log\left[\frac{\pi(Y|x,\theta)}{\pi(Y_{-k}|x,\theta)}\right]dx - \log\left[\frac{\pi(Y)}{\pi(Y_{-k})}\right] \\
&= \int \pi(x|Y)\log\left[\frac{\pi(Y|x,\theta)}{\pi(Y_{-k}|x,\theta)}\right]dx - \log\left[\frac{\pi(Y_k, Y_{-k})}{\pi(Y_{-k})}\right] \\
&= \int \pi(x|Y)\log\left[\frac{\pi(Y|x,\theta)}{\pi(Y_{-k}|x,\theta)}\right]dx - \log[\pi(Y_k|Y_{-k})] \\
&Y\text{'s are independent given } x,\theta, \text{ then} \\
\mathrm{KL}(\pi(x|Y_{-k}), \pi(x|Y)) &= \int \pi(x|Y)\log\left[\pi(Y_k|x,\theta)\right]dx - \log[\pi(Y_k|Y_{-k})] \\
\mathrm{KL}(\pi(x|Y_{-k}), \pi(x|Y)) &= E_{x|Y}[\log[\pi(Y_k|x,\theta)] - \log[CPO_k].
\end{aligned}$$

Bibliography

Agarwal, D. K., Gelfand, A. E., and Citron-Pousty, S. (2002). Zero-inflated models with application to spatial count data. *Environmental and Ecological Statistics* **9,** 341–355. iv, 2

Bertrand, A., Chaigneau, A., Peraltilla, S., Ledesma, J., Graco, M., Monetti, F., and Chavez, F. (2011). Oxygen: A fundamental property regulating pelagic ecosystem structure in the coastal southeastern tropical pacific. *PLoS One* **6,** 12. 23, 28

Bertrand, A., Gerlotto, F., Bertrand, S., Gutiérrez, M., Alza, L., and Chipollini, A. (2008). Schooling behaviour and environmental forcing in relation to anchoveta distribution: An analysis across multiple spatial scales. *Progress in Oceanography* **79,** 264–277. 1

Bertrand, S., DÃaz, E., and Lengaigne, M. (2008). Patterns in the spatial distribution of Peruvian anchovy (Engraulis ringens) revealed by spatially explicit fishing data. *Progress in Oceanography* **79,** 379–389. 1

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. 30

Bolin, D. (2012). *Models and methods for random fields in spatial statistics with computational efficiency from Markov properties.* PhD thesis, Lund University. 30

Boyd, C. (2012). *The Predator Dilemma: Investigating the responses of seabirds to changes in the abundance and distribution of small pelagic prey.* PhD thesis, University of Washington. 1

Cameletti, M., Lindgren, F., Simpson, D., and Rue, H. (2013). Spatio-temporal modeling of particulate matter concentration through the spde approach. *AStA Advances in Statistical Analysis* **97(2),** 109–131. 30, 32

Cameron, A. and Trivedi, P. (1998). *Regression Analysis of Count Data.* University Press, Cambridge. iv, 2

Cressie, N. A. C. (1993). *Statistics for Spatial Data.* Wiley. 4

Dey, D. K., Chen, M.-H., and Chang, H. (1997). Bayesian Approach for Nonlinear Random Effects Models. *Biometrics* **53(4),** 1239–1252. 18

Diggle, P., Tawn, J., and Moyeed, R. A. (1998). Model based geostatistics (with discussion). *Applied Statistics* **47,** 299–350. 1, 4

Dreassi, E., Petrucci, A., and Rocco, E. (2014). Small area estimation for semicontinuous skewed spatial data: An application to the grape wine production in Tuscany. *Biometrical Journal* **1,** 141–156. 14

Fréon, P. and Misund, O. A. (1999). *Dynamics of Pelagic Fish Distribution and Behaviour: Effects on Fisheries and Stock Assessment*. Wiley-Blackwell. 1

Geisser, S. and Eddy, W. F. (1979). A Predictive Approach to Model Selection (Corr: V75 p765). *Journal of the American Statistical Association* **74,** 153–160. 18

Gelfand, A. E., Dey, D. K., and Chang, H. (1992). Model Determination Using Predictive Distributions, with Implementation Via Sampling-based Methods (Disc: P160-167). In Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M., editors, *Bayesian Statistics 4. Proceedings of the Fourth Valencia International Meeting*, pages 147–159. Clarendon Press [Oxford University Press]. 18

Ghosh, P. and Albert, P. S. (2009). A Bayesian analysis for longitudinal semicontinuous data with an application to an acupuncture clinical trial. *Computational Statistical Data Analysis* **53,** 699–706. 14

Held, L., Schordle, B., and Rue, H. (2010). Posterior and cross-validatory predictive checks: a comparison of MCMC and INLA. *Statististical Modelling and Regression Structures* pages 91–110. 18

Lachos, V. H., Barbosa, C. R., and Medina, A. W. (2013). *Modelos Não Lineares Assimétricos*. Instituto de Matemática e Estatística-USP and ABE Associação Brasileira de Estatística. 19

Lindgren, F. and Rue, H. (2007). Explicit construction of GRMF approximations to generalised matérn fields on irregular grids. Technical report, Centre for Mathematical Sciences, Lund University. 32

Lindgren, F., Rue, H., and Lindstrom, 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(1),** 423–498. 6, 16, 30, 31, 32

MacLennan, D. N., Fernandes, P. G., and Dalen, J. (2002). A consistent approach to definitions and symbols in fisheries acoustics. *ICES Journal of Marine Science* **59,** 365–369. 12

McCulloch, R. E. (1989). Local model influence. *Journal of the American Statistical Association* **84,** 473–478. 19

Mullahy, J. (1986). Specifications and testing of some modified count data model. *Journal of Econometrics* **33,** 341–365. iv, 2

Muñoz, F., Pennino, M., Conesa, D., López-Quélez, A., and Bellido, J. (2013). Estimation and prediction of the spatial ocurrence of fish species using Bayesian Latent Gaussian models. *Sthocastic Environmental Research and Risk Assessment* **27,** 1171–1180. 2

Neelon, B., O'Malley, A. J., and Normand, S.-L. T. (2011). A Bayesian two-part latent class model for Longitudinal Medical expenditure data: Assesing the impact of mentral health and substance abuse parity. *Biometrics* **67,** 280–289. 14

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. Chapman and Hall-CRC Press. 6, 8, 16, 30

Rue, H. and martino, S. (2007). Approximate Bayesian Inference for Hierarchical Gaussian Markov Random fields models. *Journal of Statistical Planning and Inference* **137,** 3177–3192. 10

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. iv, 2, 8, 10, 16, 17, 18

Rue, H. and Tjelmeland, H. (2002). Fitting Gaussian Markov Random Fields to Gaussian Fields. *Scandinavian Journal of Statistics* **29(1),** 31–49. 6, 16

Simmonds, J. and MacLennan, D. (2005). *Fisheries Acoustics: Theory and Practice*. Wiley-Blackwell. 12

Simpson, D., Lindgren, F., and Rue, H. (2012). Think continuous: Markovian Gaussian models in spatial statistics. *Spatial Statistics* **1,** 16–29. 30

Spiegelhalter, D. J., Best, N., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B* **64,** 583–639. 17

Woillez, M., Rivoirard, J., and Fernandes, P. G. (2009). Evaluating the uncertainty of abundance estimates from acoustic surveys using geostatistical simulations. *ICES Journal of Marine Science* **66,** 1377–1383. 1