NOVA SCHOOL OF SCIENCE & TECHNOLOGY

DEPARTAMENTO DE MATEMÁTICA

Modelação Geoestatística

BELCHIOR COELHO MIGUEL

Mestre em Estatística Computacional

DOUTORAMENTO EM ESTATÍSTICA E GESTÃO DO RISCO

Universidade NOVA de Lisboa
Maio, 2022

NOVA SCHOOL OF SCIENCE & TECHNOLOGY

DEPARTAMENTO DE MATEMÁTICA

Modelação Geoestatística

BELCHIOR COELHO MIGUEL

Mestre em Estatística Computacional

Orientadora: Doutora Isabel Cristina Maciel Natário
Professora Associada da Faculdade de Ciências e Tecnologia da Universidade NOVA de Lisboa

Coorientadora: Doutora Paula Cristina Pires Simões
Professora Auxiliar da Academia Militar, Instituto Universitário Militar

Júri:

Presidente: Doutor Carlos Manuel Agra Coelho
Professor Catedrático da Faculdade de Ciências e Tecnologia da Universidade NOVA de Lisboa

Arguentes: Doutora Maria Lucília Salema Carvalho
Professora Associada Aposentada da Faculdade de Ciências da Universidade de Lisboa

Doutora Raquel Menezes da Mota Leite
Professora Associada com Agregação do Departamento de Matemática da Universidade do Minho

Orientador: Doutora Isabel Cristina Maciel Natário
Professora Associada da Faculdade de Ciências e Tecnologia da Universidade NOVA de Lisboa

Vogais: Doutor Carlos Manuel Agra Coelho
Professor Catedrático da Faculdade de Ciências e Tecnologia da Universidade NOVA de Lisboa

Doutor Manuel Leote Tavares Inglês Esquível
Professor Associado com Agregação da Faculdade de Ciências e Tecnologia da Universidade NOVA de Lisboa

DOUTORAMENTO EM ESTATÍSTICA E GESTÃO DO RISCO

Universidade NOVA de Lisboa
Maio, 2022

Modelação Geoestatística

Aos meus pais João Baptista Abílio Miguel e Carolina João Rodrigues Coelho, em memória.

Agradecimentos

Em primeiro lugar, gostaria de agradecer à minha orientadora Professora Doutora Isabel Natário. Foi um grande prazer aprender, trabalhar e discutir com ela. Obrigado por me apoiar, encorajar e confiar em mim ao longo desses anos. Em segundo lugar, gostaria de agradecer também à minha co-orientadora, Professora Doutora Paula Simões pelas discussões úteis e sugestões valiosas. À Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa (FCT-UNL) pelo acolhimento e ao Instituto Superior Politécnico de Tete (ISPT) pela Bolsa de estudos concedida para frequentar o programa doutoral, estou-lhe grato. À FCT - Fundação para a Ciência e a Tecnologia, I.P., no âmbito dos projetos PREFERENTIAL, PTDC/MAT-STA/28243/2017; UIDB/00297/2020 e UIDP/00297/2020 (Centro de Matemática e Aplicações) pela bolsa de investigação. Ao Centro de Matemática e Aplicações (CMA) e ao Departamento de Matemática (DM) da FCT-UNL por todo o apoio e recursos constantes como investigador. À Marinha Portuguesa em geral e em particular à Direção de Análise e Gestão da Informação e ao Comando Naval, na pessoa do Capitão-Tenente Rui Pedro Gonçalves de Deus pelos dados e pela colaboração. Aos meus colegas do ISPT e do CMA e, ainda, ao staff por toda a ajuda. Aos meus amigos residentes em Lisboa e em Moçambique por me ouvirem e fornecerem perspetiva. Por último, mas não menos importante, gostaria de agradecera minha família toda, especialmente aos meus irmãos pelo apoio total e muita paciência, cujo incentivo e interesse me ajudaram a seguir em frente. Esta tese é dedicada aos meus pais Miguel e Carolina, que me ensinaram a trabalhar duro e acreditar que sempre é possível alcançar aquilo que almejamos.

Resumo

Nos últimos anos, a geoestatística tem sido muito utilizada em vários estudos de diferentes áreas de pesquisa, a atividade pesqueira é uma dessas áreas de interesse, principalmente quando se pretende estudar fenómenos relacionados com a distribuição espacial da quantidade do pescado das diferentes espécies e as infrações associadas a pesca. A modelação geoestatistica tradicional baseada em técnicas de krigagem é frequentemente utilizada como uma abordagem não paramétrica na modelação de dados georreferenciados Gaussianos. Em relação a dados binários são utilizados modelos de regressão logística como padrão. Têm-se usado modelos espaciais Bayesianos para analisar dados geoestatísticos binários, incorporando a componente espacial de forma hierárquica como um campo aleatório latente. Os tradicionais métodos de Monte Carlo via Cadeia de Markov (MCMC) para estimação destes modelos, podem ser substituídos pela abordagem *Integrated Nested Laplace Approximation* (INLA), computacionalmente menos exigente sem problemas de convergência, permitindo fazer inferência Bayesiana aproximada em modelos gaussianos latentes, como os modelos lineares generalizados mistos. Conjugada com esta metodologia, a recente abordagem *Stochastic Partial Differential Equation* (SPDE), estima facilmente o campo aleatório, resolvendo problemas de grande dimensão. Em relação ao delineamento de amostragem para dados geoestatísticos binários são propostos dois critérios de seleção de delineamentos de amostragem, o de maximização do risco estimado e maximização da variabilidade associada ao risco estimado, que depois foram comparados ao delineamento aleatório simples. O principal objetivo deste trabalho é apresentar a modelação geoestatística baseada em técnicas de krigagem e a modelação de dados geoestatísticos com resposta binária usando a combinação das abordagens INLA-SPDE, e propor critérios de delineamento de amostragem baseados em geoestatística para dados binários. Posteriormente, fez-se uma aplicação desta metodologia para estimar a quantidade do pescado de lulas, inspecionado em ações de fiscalização da Marinha Portuguesa no ano 2015 na região costeira do Algarve. Os modelos baseados em técnicas de krigagem foram usados para calcular a variância da quantidade do pescado de lulas, inspecionado em ações de fiscalização da Marinha Portuguesa no ano 2015 na região costeira do Algarve. Adicionalmente, aplicou-se os modelos espaciais Bayesianos para analisar dados

geoestatísticos binários a um conjunto de dados reais de fiscalização marítima da costa Portuguesa, para produzir mapas de médias e erro padrão do efeito espacial subjacente e dos mapas de riscos. As análises foram feitas com recurso ao pacote geoR do *software* R e o pacote R-INLA, e fez-se a comparação dos delineamentos de amostragem em diferentes tamanhos de amostra e escolheu-se o melhor delineamento para cada um destes tamanhos.

Palavras-chave: Modelação geoestatística, semivariograma e krigagem, Presumíveis infrações pesqueiras, INLA-SPDE, Dados espaciais binários, Delineamento amostral, Métodos de otimização, Delineamentos baseados em modelo.

Abstract

In recent years, geostatistics has been widely used in several studies of different research areas, the fishing activity is one of these areas of interest, especially when one wants to study phenomena related to the spatial distribution of the amount of fish of different species and the infractions associated with fishing. Traditional geostatistical modelling based on kriging techniques is often used as a non-parametric approach in modelling Gaussian georeferenced data. For binary data, logistic regression models are used as standard. Bayesian spatial models have been used to analyse binary geostatistical data, incorporating the spatial component in a hierarchical way as a latent random field. Traditional Markov Chain Monte Carlo (MCMC) methods for estimating these models can be replaced by the Integrated Nested Laplace Approximation (INLA) approach, which is computationally less demanding without convergence problems, allowing approximate Bayesian inference in latent Gaussian models, such as generalized linear mixed models. Conjugated with this methodology, the recent Stochastic Partial Differential Equation (SPDE) approach, easily estimates the random field, solving high dimensional problems. Regarding the sampling design for binary geostatistical data, two criteria for selecting sampling designs are proposed, maximizing the estimated risk and maximizing the variability associated with the estimated risk, which were then compared to the simple random design. The main objective of this work is to present geostatistical modeling based on kriging techniques, modeling geostatistical data with binary response using the combination of INLA-SPDE approaches and propose geostatistics-based sampling design criteria for binary data. Subsequently, an application of this methodology was made to estimate the quantity of squid fish, inspected in enforcement actions of the Portuguese Navy in the year 2015 in the coastal region of Algarve. The models based on kriging techniques were used to calculate the variance of the quantity of squid fish, inspected in surveillance actions of the Portuguese Navy in the year 2015 in the coastal region of the Algarve. Additionally, Bayesian spatial models were applied to analyse binary geostatistical data to a real marine surveillance dataset of the Portuguese coast, to produce mean and standard error maps of the underlying spatial effect and risk maps. The analyses were done using the geoR package of the R software and the R-INLA package, and the

sampling designs were compared at different sample sizes and the best design was chosen for each of these sizes.

Keywords: Geostatistical modelling, semivariogram and kriging, Presumed fishing infringements, INLA-SPDE, Binary spatial data, Sample design, Optimization methods, Model-based designs.

Índice

Índice de Figuras

Índice de Tabelas

Siglas

BLUE Best Linear Umbiased Estimator 29

CAG Campo Aleatório Gaussiano 57

DIC Deviance Information Criterion 66

GLMM Gaussian Linear Mixed Models 46

INLA Integrated Nested Laplace Approximation 53

KO Krigagem Ordinária 29

MaxRSD Maximum risk sampling design 81
MaxVRSD Maximum variance risk sampling design 82
MCMC Monte Carlo via Cadeias de Markov 53
MLG Modelos Lineares Generalizados 44
MLGM Modelos Lineares Generalizados Mistos 43
MMSD Minimization of mean shortest distance 76
MVR Maxima verosimilhança Restrita 26

RMSE Root mean square error 76

SPDE Stochastic Partial Differential Equation 53

ZEE Zona Económica Exclusiva 57

1

Introdução

1.1 Enquadramento

A atividade pesqueira é importante para qualquer nação costeira com recursos haliêuticos nos espaços marítimos onde detém soberania e jurisdição. Portugal, como estado costeiro, possui soberania, jurisdição e responsabilidade de 14.069 quilómetros quadrados de águas interiores, de 50.957 quilómetros quadrados de mar territorial e de 1.660.456 quilómetros quadrados de Zona Económica Exclusiva (ZEE), totalizando 1.725.482 quilómetros quadrados de espaço marítimo da responsabilidade de Portugal. Esta atividade constitui uma fonte de emprego ou de sustentabilidade de certos grupos sociais, principalmente de comunidades residentes ao longo da costa, assim como fonte de receitas de algumas grandes empresas ligadas a este ramo. Portanto, para a sua sustentabilidade é necessário estabelecer regras e princípios, com objetivo de controlar a exploração dos recursos piscícolas e a exploração de algumas áreas e espécies que, quando feitas descontroladamente, podem comprometer a sobrevivência do ecossistema marinho. Para uma atividade pesqueira mais justa, regrada e sustentável com vista a assegurar a conservação, gestão e desenvolvimento dos recursos aquáticos respeitando o ecossistema e a biodiversidade, foi criado em Portugal o sistema SIFICAP, do qual a Marinha Portuguesa é entidade participante, e que pretende assegurar a vigilância, fiscalização e controlo da atividade da pesca. Neste seguimento, o presente trabalho vai aplicar as técnicas da geoestatística para estudar a distribuição espacial do pescado e de presumíveis infrações marítimas na costa portuguesa de interesse neste contexto.

A Geoestatística é uma área de conhecimento que vem sendo muito utilizada atualmente como ferramenta para analisar variáveis distribuídas no espaço e/ou no tempo, onde se assume que os valores destas variáveis estejam correlacionados. Apoia-se em dados que são medições relacionadas com um fenómeno subjacente espacialmente contínuo $\{U(s), s \in D \subseteq \mathbb{R}^2\}$. De acordo com Yamamoto e Ladim [97], a Geoestatística teve origem nos trabalhos de um engenheiro de minas sul africano, trabalhador das minas do Rand, de nome Daniel Krige (Krige, 1951), que nos anos de 1960, apresentou uma série de publicações de elevada importância para o estudo e formalização da teoria de

variáveis regionalizadas, por isso, é distinguido pelo Professor George Matheron como sendo o pai da Geoestatística. Posteriormente, dois ex-alunos do Professor Matheron de nomes André G. Journel e Michel David, entre outras obras, publicaram dois livros importantes intitulados de: *Geostatistical ore reserve estimation* (David, 1977) [78] e *Mining Geostatistical* (Journel, Huijbregts, 1979) [39], que contribuíram bastante para a divulgação da Geoestatística. Por volta dos anos 1951, a Geoestatística era aplicada na área de mineração, atualmente tem sido aplicada em muitas outras áreas científicas, tais como a saúde, agricultura, ciências do solo, entre outras, e é considerada uma ferramenta muito poderosa para o mapeamento e quantificação de incertezas, pois, possui capacidade de caracterizar a variabilidade espacial por meio de um modelo probabilístico consistente [52]. Esta estrutura espacial pode ser identificada pelo variograma, uma função que mede a associação espacial entre valores de variáveis indexadas no espaço. De acordo com Diggle e Ribeiro [31], a Geoestatística tem dois grandes objetivos científicos, a estimação e a predição.

Tradicionalmente, a geoestatística não paramétrica baseada em krigagem nas suas diferentes formas foi muito utilizada, mas com o tempo, introduziu-se a abordagem de geoestatística baseada em modelos [30]. Atualmente, nas mais diversas áreas de estudo são utilizados modelos de regressão para dados em que a variável resposta é binária, para os quais os modelos lineares clássicos não são apropriados porque dificilmente são alcançados os pressupostos dos modelos (normalidade e variâncias constantes). Nestes casos, o modelo de regressão logística tem sido o mais empregue para se trabalhar com dados binários [38]. Os modelos de regressão para dados binários, em particular o modelo logístico, podem ser enquadrados na classe dos modelos lineares generalizados (MLG) [55]. Dados geoestatísticos binários são dados espaciais em que a variável resposta toma um de dois valores possíveis, dividindo a região em estudo em duas sub-regiões disjuntas de acordo com esses valores [26]. Para se analisar este tipo de dados têm-se usado modelos espaciais Bayesianos, incorporando a componente espacial de forma hierárquica como um campo aleatório latente. Tradicionalmente são usados métodos Monte Carlo via Cadeia de Markov (MCMC) de ajuste destes modelos, que podem ser substituídos pela abordagem *Integrated Nested Laplace Approximation* (INLA), computacionalmente menos exigente e sem problemas de convergência, permitindo fazer inferência Bayesiana aproximada em modelos gaussianos latentes, tais como modelos lineares generalizados mistos. Conjugada com esta metodologia, a recente abordagem *Stochastic Partial Differential Equation* (SPDE), estima com facilidade o campo aleatório, resolvendo problemas de dimensionalidade associados às outras metodologias de estimação.

Delinear as amostragens é muito importante em vários campos de pesquisas, pois tem a função orientadora de como o pesquisador deve selecionar cuidadosamente os elementos da população que vão fazer parte da amostra que posteriormente serão usados para fazer inferência sobre uma população. [19, 70, 82]. Existem critérios de seleção que são usados para construir delineamentos de amostragem, neste trabalho foram propostos dois critérios de seleção de delineamentos de amostragem, o de maximização do risco

estimado e maximização da variabilidade associada ao risco estimado. Estes critérios foram utilizados para construir dois esquemas de amostragem que depois foram comparados ao delineamento aleatório simples de amostragem das infrações pesqueiras na costa Portuguesa usando valores reais obtidos em ações de fiscalização.

1.2 Objetivos

1.2.1 Objetivo geral

O principal objetivo deste trabalho é apresentar a modelação geoestatística tradicional baseada em técnicas de krigagem, a modelação de dados geoestatísticos baseada em modelos, para resposta binária usando a combinação das abordagens INLA-SPDE, e propor critérios de delineamento de amostragem baseados em geoestatística para dados binários.

1.2.2 Objetivos específicos

- Apresentar a modelação geoestatística baseada em técnicas de krigagem;
- Apresentar técnicas de modelação de dados geoestatísticos com resposta binária usando a combinação das abordagens INLA-SPDE;
- Apresentar critérios de seleção de delineamento de amostragem;
- Aplicar a estimação por krigagem aos dados do pescado de lulas obtidos em ações de fiscalização, na região costeira de Algarve em 2015, com recurso ao *software* R; e
- Aplicar os modelos bayesianos geoestatísticos na modelação de presumíveis infrações marítimas relacionadas a atividade pesqueira, com recurso ao *software* R.

1.3 Dados

Neste trabalho foram utilizados dados georeferenciados relativos a ações de fiscalização efetuadas pela Marinha Portuguesa em toda costa Portuguesa, com mais de duzentos e onze mil (211000) pontos fiscalizados em 22 anos (de 1998 a 2019), com uma média de 9620 pontos fiscalizados por ano.

1.4 Estrutura do trabalho

Este trabalho encontra-se dividido em 5 capítulos. No capítulo 2, apresentam-se os conceitos básicos da geoestatística, os vários modelos teóricos do variograma, a estimação paramétrica da estrutura de covariância espacial, o método de estimação krigagem, resultados da aplicação dos modelos teóricos de variogramas e da krigagem numa aplicação a dados de pescado de lulas na região do Algarve obtidos em ações de fiscalização da

Marinha Portuguesa. No capítulo 3, apresentam-se métodos e modelos relacionados a dados geoestatísticos binários, com principal destaque para as suas abordagens Bayesianas INLA e SPDE, os materiais usados no trabalho, isto é, a aquisição, descrição e apresentação dos dados, os resultados da aplicação dos modelos sobre as presumíveis infrações pesqueiras dos dados de fiscalização da costa Portuguesa. No capítulo 4, apresentam-se delineamento de amostragem de dados geoestatísticos binários, os critérios de seleção de delineamento de amostragem e de validação, os resultados da aplicação e suas conclusões e por fim o capítulo 5 apresenta as conclusões e recomendações para trabalhos futuros.

2

Modelos Geoestatísticos para dados Gaussianos

2.1 Introdução

A Geoestatística atualmente tem sido uma ferramenta muito utilizada para analisar variáveis distribuídas no espaço e/ou no tempo, onde se assume que os valores destas variáveis estejam correlacionados. Apoia-se em dados que são medições relacionadas com um fenómeno subjacente espacialmente contínuo $\{U(s), s \in D \subseteq \mathbb{R}^d\}$. No princípio, por volta dos anos 1951, a Geoestatística era aplicada na área de mineração, atualmente tem sido aplicada em muitas outras áreas científicas, tais como a saúde, agricultura e ciências do solo, entre outras.

A Geoestatística é considerada uma ferramenta muito poderosa para o mapeamento e quantificação de incertezas, pois, possui uma capacidade de caracterizar a variabilidade espacial por meio de um modelo probabilístico consistente [52]. Esta estrutura espacial pode ser identificada pelo variograma, uma função que mede a associação espacial entre valores de variáveis indexadas no espaço.

De acordo com Diggle e Ribeiro [31], a Geoestatística tem dois grandes objetivos científicos, a estimação e a predição.

Num contexto geral, de acordo com Assis *et al.* [5] estimação é a determinação do valor de um parâmetro populacional ou dos parâmetros de um modelo probabilístico, com base num conjunto de observações ou dados, e a predição, refere-se a um processo para determinar a magnitude de variáveis estatísticas em alguns pontos onde não foram observadas.

Em Geoestatística, de acordo com Diggle e Ribeiro [31], estimação, refere-se à inferência sobre os parâmetros de um modelo estocástico para os dados, por exemplo, os parâmetros que definem a estrutura de covariância de um modelo para $U(s)$, e a predição refere-se à inferência sobre a realização do processo não observado. Geralmente a predição é apresentada como mapa dos valores preditos de $U(s)$, ou como predição de algumas propriedades da realização completa de $U(s)$ que é de particular relevância para

um problema específico.

E ainda, Zhang [98], diz que a predição geralmente no contexto geoestatístico assume a forma de um mapa ou de uma série de mapas. De acordo com o mesmo autor, distinguem-se duas formas de predição, a saber: estimação e simulação. Estimação é baseada nos dados da amostra e num modelo (variograma) precisamente representando a correlação espacial dos dados da amostra, sendo esta estimação (mapa) produzida por um processo não paramétrico designado por krigagem, enquanto que na simulação, são produzidos muitos mapas de igual probabilidade (também chamadas de imagens) da distribuição de uma variável de interesse, usando o mesmo modelo de correlação espacial necessário para a krigagem.

2.1.1 Origem da Geoestatística

De acordo com Yamamoto e Ladim [97], a Geoestatística teve origem nos trabalhos de um engenheiro de minas Sul Africano que trabalhava nas minas do Rand, de nome Daniel Krige (Krige, 1951), nos anos de 1960, que apresentou uma série de publicações de elevada importância para o estudo e formalização da teoria de variáveis regionalizadas, por isso, é distinguido pelo Professor George Matheron como sendo o pai da Geoestatística.

Mais tarde, em 1968 Matheron criou o *Centre de Morphologie Mathématique*, que depois foi dividido em dois centros de pesquisas: Morfologia Matemática e Geoestatística. Estes dois centros de pesquisa, tiveram uma grande importância para o estudo, difusão e formação dos pesquisadores. A seguir, dois ex-alunos do Professor Matheron de nomes André G. Journel e Michel David, entre outras obras, publicaram dois livros importantes intitulados de: *Geostatistical ore reserve estimation* (David, 1977) e *Mining Geostatistical* (Journel, Huijbregts, 1979), que contribuíram bastante para a divulgação da Geoestatística.

Existem várias definições de Geoestatística, a seguir vai-se apresentar algumas delas.

Matheron [51] e Olea [63], consideram que a Geoestatística é um conjunto de técnicas estatísticas que modelam a variação espacial dos dados e usam esses modelos para estimar ou classificar outros dados.

Journel e Huilbregts [39], definem Geoestatística como sendo um estudo dum certo fenómeno natural, caracterizado pela distribuição no espaço de uma ou mais variáveis. Estas variáveis são designadas variáveis regionalizadas.

Cressie [21] também define a Geoestatística como sendo uma classe de métodos estatísticos para estimar ou prever o valor de um processo espacial contínuo em locais não observados, dado o valor do processo num conjunto de locais conhecidos.

Diggle e Ribeiro [31] definem Geoestatística como um sub-ramo da estatística espacial em que os dados consistem em uma amostra finita de valores medidos relacionados com fenómeno espacialmente contínuo subjacente.

Este capítulo encontra-se dividido em 6 secções. Na secção 2, apresentam-se os conceitos básicos que sustentam a parte teórica do capítulo, como a amostragem, variáveis

regionalizadas, estacionaridade e autocorrelação espacial. Na secção 3, apresentam-se os vários modelos teóricos do variograma e suas principais características, e a estimação paramétrica da estrutura de covariância espacial. Na secção 4, apresentam-se o método de estimação krigagem, subdividida em krigagem linear, krigagem não linear e cokrigagem. Na secção 5, apresentam-se resultados da aplicação dos modelos teóricos de variogramas e da krigagem numa aplicação a dados de pescado de lulas (quilogramas) na região do Algarve obtidos em ações de fiscalização da Marinha Portuguesa, e por fim, a secção 6 apresenta as conclusões e recomendações para trabalhos futuros.

2.2 Conceitos básicos da Geoestatística

Neste capítulo, apresentam-se algumas definições importantes dos conceitos Geoestatísticos.

Fenómeno espacial é o processo que define a distribuição e variabilidade espaciais de uma variável de interesse dentro dum domínio. Amostra é um subconjunto do fenómeno espacial, que se for representativa, deve espelhar a distribuição e a variabilidade espaciais tanto em tamanho, como em termos da distribuição espacial subjacente [97].

De acordo com o mesmo autor, em Geoestatística destacam-se três tipos de amostragem, a saber: amostragem aleatória simples, amostragem aleatória estratificada ou amostragem aleatória sistemática.

2.2.1 Amostragem

Amostragem Aleatória Simples é uma amostragem probabilística em que todos os elementos da população têm a mesma probabilidade (probabilidade conhecida) de fazer parte da amostra escolhida. Em Geoestatística, as observações são feitas nos pontos de amostragem localizados dentro do domínio (região de estudo) e, assim sendo, a componente aleatória serão as coordenadas geográficas a serem escolhidas aleatoriamente.

Amostragem Aleatória Estratificada é uma amostragem feita em estratos. Implica subdividir a região em estudo em células de dimensões fixas nas direções Leste-Oeste e Norte-Sul. Dentro de cada célula, as coordenadas geográficas de um ponto são escolhidas aleatoriamente e o ponto é selecionado. No final o número de unidades selecionadas será igual ao número de células.

Amostragem Sistemática é uma amostragem feita sobre os nós de uma malha regular definida com base em uma origem escolhida aleatoriamente. A malha regular é definida inicialmente pelo responsável da amostragem para otimizar a coleta das unidades dentro da região de estudo.

Ainda de acordo com Yamamoto e Ladim [97], o método de amostragem que fornece melhor resultados é a amostragem sistemática, contudo, nem sempre é possível aplicar este tipo de amostragem visto que ela depende dum conjunto de fatores, como por exemplo acesso, acidentes geográficos (rios, topografia), vegetação, entre outros.

O maior objetivo da amostragem em Geoestatística é de extrair o máximo de informação disponível na amostra recolhida.

2.2.2 Dados georreferenciados

Dados georreferenciados, de uma forma simples, podemos dizer que são observações referenciadas com base na sua localização geográfica.

Os dados georreferenciados (também chamados de dados geoespaciais) podem ser de três tipos diferentes: dados em pontos, dados em áreas ou dados referentes a processos pontuais.
Os dados georreferenciados em pontos são observações do mesmo fenómeno registadas num número de localizações específicas (como por exemplo a poluição do ar). Dados georreferenciados por área são observações das mesmas características registadas em subregiões do espaço (como por exemplo: um bairro, um concelho). E dados referentes a processos pontuais são constituídos pelas localizações aleatórias da ocorrência do fenómeno de interesse.
Uma análise adequada para dados georreferenciados depende muito do modelo estatístico considerado para o processo espacial subjacente dos dados disponíveis.

2.2.3 Variáveis regionalizadas

Segundo Matheron [52], quando um fenómeno se estende no espaço e exibe uma certa estrutura espacial então ele é regionalizado. Um valor $y(s_i)$, observado espacialmente em s_i (denominação genérica das coordenadas geográficas de um certo local) é então interpretado como uma realização da variável aleatória $U(s_i)$ que representa o referido fenómeno. No espaço D, no qual se dispersa um conjunto de valores amostrados, temos as realizações das n variáveis aleatórias $U(s_1), U(s_2), ..., U(s_n)$, correlacionadas entre si.[79].

Pode-se calcular o valor esperado e a variância (os dois primeiros momentos) de cada uma das variáveis aleatórias $U(s_i)$ da seguinte forma:

$$E[U(s_i)] = m(s_i) = \int_{-\infty}^{+\infty} u dF_{s_i}(u) = \int_{-\infty}^{+\infty} u f_{s_i}(u) du \tag{2.1}$$

$$V(U(s_i)) = \int_{-\infty}^{+\infty} [u - m(s_i)]^2 dF_{s_i}(u), \tag{2.2}$$

onde que $f_{s_i}(u)$ e $F_{s_i}(u)$ são respetivamente as funções de densidade de probabilidade e de distribuição de probabilidade da variável $U(s_i)$.

Consideremos agora duas variáveis aleatórias, $U(s_1)$ e $U(s_2)$, pode ser definida a covariância:

$$Cov[U(s_1), U(s_2)] = E[U(s_1)U(s_2)] - m(s_1)m(s_2)$$

$$E[U(s_1)U(s_2)] = \int_{\infty}^{+\infty}\int_{-\infty}^{+\infty} xyd^2F_{s_1,s_2}(x,y)$$

$$E[U(s_1)U(s_2)] = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xyf_{s_1,s_2}(x,y)dxdy, \tag{2.3}$$

sendo $F_{s_1,s_2}(x,y)$ a função de distribuição bivariada:

$$F_{s_1,s_2}(x,y) = P\{U(s_1) \leq x\ ,\ U(s_2) \leq y\}.$$

O coeficiente de correlação é dado por:

$$\rho[U(s_1),U(s_2)] = \frac{Cov[U(s_1),U(s_2)]}{\sqrt{V(U(s_1)).V(U(s_2))}} \tag{2.4}$$

e o semivariograma é dado por:

$$\gamma[U(s_1),U(s_2)] = \frac{1}{2}E[(U(s_1)-U(s_2))^2]. \tag{2.5}$$

O conjunto de variáveis aleatórias, $U(s_1),\ldots,U(s_n),\ldots$, constituem um conjunto infinito das variáveis aleatórias [3], um processo aleatório ou um processo estocástico. O conjunto de valores reais de U que compõem a realização da função aleatória é conhecido como variável regionalizada [95].

A teoria da variáveis aleatórias foi defendida por Matheron, e destacam-se dois objetivos principais das variáveis regionalizadas, a saber: um objetivo teórico, que é de descrever a correlação espacial e o outro prático que é de resolver problemas de estimativa de uma variável regionalizada com base numa amostra.

2.2.4 Estacionaridade

Estacionaridade é uma característica da variável regionalizada, que se manifesta pela variabilidade da diferença da variável regionalizada a partir duma certa distância ser sempre a mesma (constante), refletindo a independência entre os valores da variável a partir dessa distância. A estacionaridade é identificada no variograma, quando este apresenta um patamar ou tendência a estabilização a partir de uma certa distância [5].

A figura 2.1 mostra o comportamento de uma função variograma estacionária e outra não estacionária.

O conjunto de variáveis aleatórias $(U(s_i),\ i = 1,2,...,n)$ correlacionados entre si, constituem uma função aleatória da qual se conhece uma realização $y(s_i)$, o conjunto dos dados experimentais. Por consequência disto, com uma só realização de cada variável aleatória é impossível determinar as estatísticas das variáveis individuais $U(s_i)$ [79]. Assim sendo, o autor sugere que para se resolver este problema, deve-se assumir diversos graus de estacionaridade da função aleatória, tal como passamos a descrever a seguir.

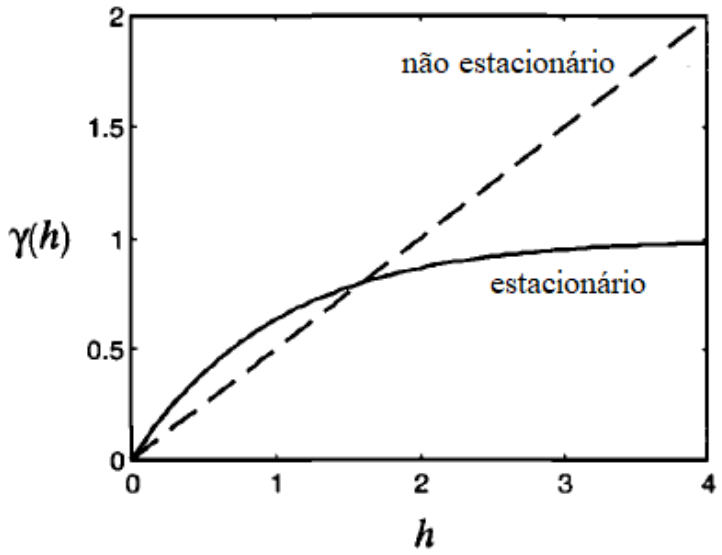

Figura 2.1: Variograma indicativo de estacionaridade e não estacionaridade, fonte: [41]

2.2.4.1 Estacionaridade espacial forte

Estacionaridade espacial forte, também conhecida como estacionaridade no sentido estrito.
A função aleatória $\left\{U(s),\ s \in D \subseteq \mathbb{R}^d\right\}$ é dita estritamente estacionária, se as variáveis aleatórias $U(s_1), U(s_2), ..., U(s_k)$ e $U(s_1+\mathbf{h}), U(s_2+\mathbf{h}), ..., U(s_k+\mathbf{h})$ têm a mesma função de distribuição conjunta para todo o k, e para quaisquer pontos dados $s_1, s_2, ..., s_k$ e qualquer translação de vector $\mathbf{h} \in \mathbb{R}^d$, fornecendo os valores $s_1, s_2, ..., s_k, s_1+\mathbf{h}, s_2+\mathbf{h}, ..., s_k+\mathbf{h}$, contidos em D [57].

Duma forma geral, esta condição é muito restritiva, por isso essa hipótese geralmente é relaxada para a estacionaridade da 2ª ordem, que limita a estacionaridade ao primeiro e ao segundo momento da função aleatória. Sabendo que em geoestatística o que nos interessa são os primeiros dois momentos.

2.2.4.2 Estacionaridade de 2ª ordem

A função aleatória $\left\{U(s),\ s \in D \subseteq \mathbb{R}^d\right\}$ é chamada de estacionária de 2ª ordem, estacionaridade fraca ou estacionária no sentido amplo, se tiver momentos finitos de 2ª ordem (ou a covariância existe) e se verifica:

- O valor esperado existe e é constante e, portanto, não depende da localização:

$$E[U(s)] = \mu(s) = \mu. \tag{2.6}$$

- A covariância existe para cada par de variáveis aleatórias $U(s)$ e $U(s+\mathbf{h})$ e depende

apenas do vector **h** que une os locais s e $(s+\mathbf{h})$, mas não especificamente deles:

$$C(U(s), U(s+\mathbf{h})) = C(\mathbf{h}), \forall\ s \in D,\ \mathbf{h} \in \mathbb{R}^d \tag{2.7}$$

Como se pode ver que a função covariância ou covariograma de uma função aleatória estacionária de 2ª ordem é apenas uma função de **h**, a variância da função aleatória existe e é finita:

$$V(U(s)) = C(\mathbf{0}) = \sigma^2, \tag{2.8}$$

e o covariograma é simétrico, $C(\boldsymbol{h}) = C(-\boldsymbol{h})$.

Por vezes, quando se pretende fazer a modelação da dependência espacial de funções aleatórias estacionárias de 2ª ordem, ao invés de usar-se o covariograma é usado o correlograma ou função de correlação, que é dado por:

$$corr(U(s), U(s+\boldsymbol{h})) = \frac{C(\boldsymbol{h})}{C(\mathbf{0})} = \rho(\boldsymbol{h}), \tag{2.9}$$

onde $C(\mathbf{0}) \neq 0$. Verifica-se que $\rho(\boldsymbol{h}) = \rho(-\boldsymbol{h})$ e $|\rho(\boldsymbol{h})| \leq 1$.

Pode-se ainda considerar no caso de estacionaridade de 2ª ordem, que a função covariância e o semivariograma são equivalentes quando se trata de definir a estrutura de dependência espacial apresentada pelo fenómeno, pois verificam a seguinte relação de correspondência entre elas:

$$\begin{aligned}
\gamma(\mathbf{h}) &= \frac{1}{2} V(U(s+\mathbf{h}) - U(s)) \\
&= \frac{1}{2}\left(V(U(s+\mathbf{h})) + V(U(s)) - 2C(U(s+\mathbf{h}), U(s))\right) && (2.10)\\
&= \frac{1}{2}(C(\mathbf{0}) + C(\mathbf{0}) - 2C(\mathbf{h})) = \frac{1}{2}(2(C(\mathbf{0}) - C(\mathbf{h}))) && (2.11)\\
&= C(\mathbf{0}) - C(\mathbf{h}) && (2.12)
\end{aligned}$$

Onde sabe-se da equação 2.7 e 2.8

$$C(U(s), U(s+\mathbf{h})) = C(\mathbf{h})$$

e

$$V(U(s)) = C(\mathbf{0}) = \sigma^2$$

Diz-se que uma função aleatória é quase estacionária quando a hipótese estacionária correspondente (geralmente a hipótese da 2ª ordem) é válida apenas para a distância $\|\mathbf{h}\| < d$, onde o d é uma distância limite. Ou seja, no caso quase estacionário da 2ª ordem (geralmente referido como quase estacionário) $\mu(s+\mathbf{h}) \approx \mu(s)$ se $\|\mathbf{h}\| < d$ e $C(U(s+\mathbf{h}), U(s)) = C(\mathbf{h})$ se $\|\mathbf{h}\| < d$.

A seguir, apresenta-se na figura 2.2 o gráfico de um processo estacionário em $\mathbb{R}$. Este processo, é regulado em torno de sua média constante.

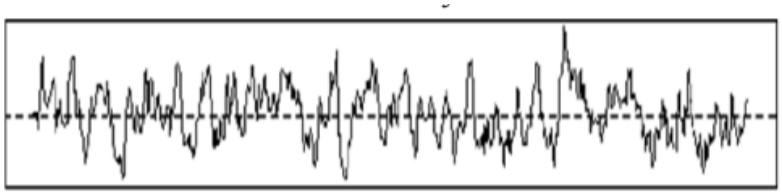

Figura 2.2: Exemplo de um processo estacionário em $\mathbb{R}$, fonte: Petitgas [68].

2.2.4.3 Estacionaridade intrínseca

A função aleatória $\{U(s),\ s \in D\}$ é chamada de estacionária intrínseca se para qualquer dado vector de translação $\mathbf{h}$, os incrementos da 1ª ordem, $U(s+\mathbf{h}) - U(s)$ são estacionários da 2ª ordem, apesar de não se exigir esta condição à própria função, ou seja

$$E\left[U(s+\mathbf{h}) - U(s)\right] = \mu(s) \tag{2.13}$$

onde $\mu(s)$ é a tendência e, é necessariamente linear em $\mathbf{h}$, e

$$C\Big(U(s+\mathbf{h}) - U(s)), U(s+\mathbf{h}+\mathbf{h}^{'}) - U(s+\mathbf{h}^{'})\Big) = C(\mathbf{h},\mathbf{h}^{'}) \tag{2.14}$$

(É suficiente definir que $\mathbf{h}^{'} = \mathbf{0}$)
que é equivalente a:

$$\frac{1}{2}V(U(s+\mathbf{h}) - U(s)) = \gamma(\mathbf{h}) \tag{2.15}$$

que é só uma função de $\mathbf{h}$.
Se a tendência linear é zero então

$$E[U(s+\mathbf{h}) - U(s)] = 0 \tag{2.16}$$

e

$$E\left[U(s+\mathbf{h}) - U(s)\right]^2 = \gamma(\mathbf{h}) \tag{2.17}$$

Se uma função aleatória for estacionária de 2ª ordem, ela também será intrinsecamente estacionária, no entanto, o recíproco não é verdadeiro.
As funções aleatórias intrínsecas que não são da 2ª ordem, são chamadas funções aleatórias estritamente intrínsecas.
A seguir, a figura 2.3 apresenta o gráfico de um processo intrínseco em $\mathbb{R}$. Este processo é mais flexível, embora neste exemplo não inclua nenhum desvio ou tendência.

2.2.4.4 Funções aleatórias não estacionárias

A função aleatória $\{U(s),\ s \in D\}$, para a qual a média e/ou função covariância dependem da localização (não são invariantes à localização) é considerada uma função aleatória não estacionária.

Quando a função aleatória $\{U(s),\ s \in D\}$ tem uma tendência, ou seja, a sua média não é constante e varia com a localização e os seus incrementos de 1ª ordem U(s+$\mathbf{h}$)-U(s) são não estacionários, diz-se que a função aleatória é não intrínseca (alguns autores chamam de função aleatória intrínseca de ordem $k > 0$).

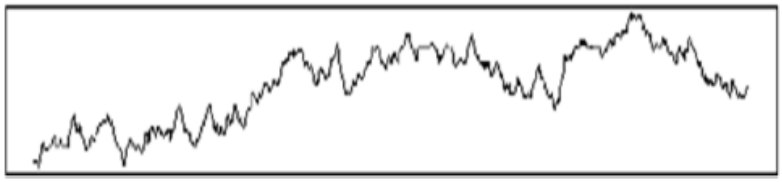

Figura 2.3: Exemplo de processo intrínseco em $\mathbb{R}$, fonte: Petitgas [68].

2.2.5 Autocorrelação espacial

Quando se fala de análise espacial ou de dados espaciais, um dos conceitos mais importantes é a dependência espacial. Alguns autores, como Waldo Tobler [86], Wikle *et al.* [96], defendem que as características das coisas mais próximas tendem a ser mais parecidas em relação as coisas mais distantes.
A dependência espacial é medida através da autocorrelação espacial, que expressa numericamente o conceito de dependência espacial. O termo autocorrelação espacial surge do conceito correlação, que em estatística serve para medir o nível de relacionamento entre duas variáveis aleatórias. Segundo Monteiro, *et al.* [56] a preposição *auto* indica que a medida de correlação é realizada com a mesma variável aleatória, medida em diferentes locais do espaço.

A autocorrelação espacial pode ser positiva ou negativa. A autocorrelação positiva ocorre quando valores semelhantes ocorrem próximos uns dos outros e, a negativa ocorre quando valores diferentes ocorrem próximos uns dos outros.

As seguintes medidas, baseadas nas medições de $U(s)$ feitas nos pontos $(s_1,...,s_n)$, $y(s) = (y_1,\ldots,y_n)$, onde $y_i = y(s_i)$, podem ser utilizadas para avaliar a correlação espacial: Índice de Moran, Índice de Geary's e variograma.

2.2.5.1 Índice de Moran

A estatística mais utilizada para medir a autocorrelação espacial é o índice global de Moran I e pode ser aplicada directamente à variável dependente ou aos resíduos de um modelo adaptado [77]. O índice global de Moran I é dado por

$$I = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2.18}$$

onde:
n é o número de pontos dos dados, y_i é o valor do atributo considerado no local s_i, $\bar{y}$ o valor médio do atributo na região de estudo e w_{ij} são os elementos de uma matriz $\boldsymbol{W}$ que correspondem a pesos de proximidade espacial entre os locais s_i e s_j. Para este caso, a correlação será calculada usando a matriz de pesos de distâncias inversas, onde as entradas para os pares de pontos que estão próximos são maiores do que os pares de pontos

distantes. Na matriz criada, os elementos da diagonal principal são zero e os elementos fora da diagonal principal w_{ij} são dados pelo inverso da distância entre os locais s_i e s_j.

O índice de Moran é utilizado num teste de permutações para as hipóteses formuladas da seguinte maneira:
H_0: não há autocorrelação espacial
H_1: há autocorrelação espacial.
A média e a variância da distribuição amostral de I são calculadas sobre a hipótese nula da seguinte maneira:

$$I = \frac{-1}{n-1} \qquad e \qquad V(I) = E[I^2] - (E[I])^2 \tag{2.19}$$

A interpretação dos resultados do índice de Moran é feita a partir da escala que varia entre -1 e 1, os valores positivos indicam uma correlação direta e valores negativos indicam uma correlação inversa. Depois de calcular o índice, é de extrema importância que se estabeleça a sua validade estatística para a avaliação da autocorrelação espacial.

2.2.5.2 Índice de Geary

O índice de Geary é uma medida de covariância, que emprega o quadrado das diferenças entre pares de valores dos atributos entre pontos e pode ser calculado da seguinte maneira:

$$C = \frac{(n-1)\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - y_j)^2}{\sum_{i=1}^{n}\sum_{i=1}^{n} w_{ij}\sum_{j=1}^{n}(y_i)^2} \tag{2.20}$$

onde n é o número de pontos dos dados, y_i é o valor do atributo considerado no local s_i e w_{ij} são os elementos de uma matriz $\boldsymbol{W}$ que correspondem a pesos de proximidade espacial entre os locais s_i e s_j. O índice C de Geary varia entre 0 e 2. Valores mais baixos (entre 0 e 1) mostram autocorrelação espacial positiva e valores mais altos (entre 1 e 2) indicam autocorrelação espacial negativa [53].

A diferença que existe entre o índice C de Geary e I de Moran é que o primeiro utiliza a diferença entre os pares de valores dos pontos, enquanto que o segundo utiliza a diferença entre cada ponto e a média global. Deste modo, o indicador C de Geary assemelha-se ao variograma e o I de Moran assemelha-se ao correlograma [15].

2.2.5.3 Variograma

Variograma é a ferramenta básica da Geoestatística que permite descrever quantitativamente a variação no espaço de um fenómeno regionalizado [5]. Geralmente, o variograma é definido como a representação gráfica da função variograma.

De acordo com Andriotti [3], em termos matemáticos, o variograma é considerado como sendo o valor esperado do quadrado dos acréscimos da variável regionalizada em

estudo numa determinada direção definida pelo vetor $\mathbf{h}$, ou seja, o valor médio do quadrado das diferenças entre todos os pares de pontos presentes na área de estudo, que têm a distância $\|\boldsymbol{h}\|$ uns dos outros.
Considere-se duas variáveis regionalizadas $U(s)$ e $U(s+\mathbf{h})$, referentes ao mesmo atributo, o variograma $2\gamma(\boldsymbol{h})$ é dado por:

$$2\gamma(\boldsymbol{h}) = E[(U(s) - U(s+\mathbf{h}))^2] \tag{2.21}$$

A função $2\gamma(\mathbf{h})$ chama-se de variograma e a função $\gamma(\mathbf{h})$ chama-se semivariograma. Portanto, o semivariograma é a metade da função variograma e é dado por:

$$\gamma(\mathbf{h}) = \frac{1}{2}E[(U(s) - U(s+\mathbf{h}))^2] = \frac{1}{2}V(U(s) - U(s+\mathbf{h})) \tag{2.22}$$

desde que a média e variância existam.

Há três características que são importantes num variograma, que são:

1) Alcance, também chamado por alguns autores de amplitude serve para medir a distância a partir da qual os valores da variável U deixam de estar correlacionados.

2) Patamar é o valor do semivariograma que corresponde ao seu alcance. A partir desse ponto em diante, considera-se que não existe mais dependência espacial entre valores amostrados, porque a variância dada pela equação 2.22 torna-se aproximadamente constante.

3) Efeito pepita mede fundamentalmente duas parcelas da variabilidade total do fenómeno em estudo:

   i) Variação espacial numa escala inferior à distância mínima entre os pontos amostrados.

   ii) Variabilidade devido a erros de medição.

Propriedades de semivariograma

Na descrição das propriedades de semivariograma que seguem, assume-se que o processo é estacionário da segunda ordem.

a) De acordo com a definição da estacionaridade, deduz-se que $\gamma(-\boldsymbol{h}) = \gamma(\boldsymbol{h})$, $\gamma(\mathbf{0}) = 0$.

   Pode acontecer que

$$\lim_{\|\boldsymbol{h}\| \longrightarrow 0} \gamma(\boldsymbol{h}) = \tau^2 > 0,$$

   em que τ^2 é o efeito pepita.

b) O semivariograma $\gamma(\boldsymbol{h})$ é finito. Semivariograma de uma função aleatória intrinsecamente estacionária sem tendência poderia crescer até o infinito, mas não incontrolavelmente, pois cresce mais lentamente do que $\|\boldsymbol{h}\|^2$ quando $\|\boldsymbol{h}\| \longrightarrow \infty$, isto é,

$$\lim_{\|\boldsymbol{h}\|\longrightarrow\infty} \frac{\gamma(\boldsymbol{h})}{\|\boldsymbol{h}\|^2} = 0$$

Um semivariograma que tenha, a grandes distâncias, um crescimento mais acentuado que o crescimento de $\|h\|^2$ é incompatível com hipótese intrínseca. Nesta situação ele apresentará um valor esperado não constante;

c) É válida a relação

$$\gamma(\boldsymbol{h}) = C(\boldsymbol{0}) - C(\boldsymbol{h})$$

Esta relação entre semivariograma e covariância, quando ambos existem, pode ser expressa pela figura 2.4.

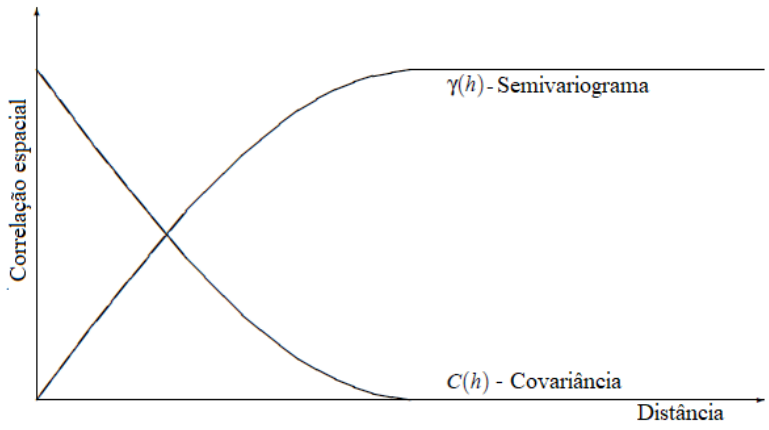

Figura 2.4: Relação de semivariograma e covariância, fonte: adaptado de Bárdossy [9].

d) A teoria mostra que o valor de $\gamma(\boldsymbol{h})$ que limita o crescimento do semivariograma em função de $\boldsymbol{h}$, quando o variograma é limitado, é igual à variância pressuposta da população estudada (σ^2). Se o variograma for limitado, pode-se estudar a variável regionalizada como se fosse uma função aleatória estacionária de segunda ordem.

e) O semivariograma assume sempre valores positivos, mas o semivariograma cruzado [1], pode assumir valores negativos.

No gráfico cartesiano da função variograma cruzado designado por variograma cruzado, um valor negativo indica que o crescimento positivo de uma variável corresponde

[1] É a metade do variograma cruzado, dado pela equação 2.24.

em média o decrescimento da outra variável aleatória. E pode-se calcular o variograma cruzado a partir da seguinte fórmula:

$$2\gamma_{12}(\mathbf{h}) = E[(U_1(s+\mathbf{h}) - U_1(s))(U_2(s+\mathbf{h}) - U_2(s))] \tag{2.23}$$

Esta fórmula geralmente é utilizada para estimar o teor de um mesmo elemento utilizando dois métodos de amostragem diferentes. Considerando a hipótese de estacionaridade de 2ª ordem, se existir covariâncias cruzadas, então existe também variogramas cruzados, cumprindo a seguinte relação:

$$2\gamma_{12}(\mathbf{h}) = 2C_{12}(0) - C_{12}(\mathbf{h}) - C_{21}(\mathbf{h}) \tag{2.24}$$

onde $C_{ij}(\mathbf{h}) = C[U_i(s), U_j(s+\boldsymbol{h})]$. Sabe-se que $\gamma_{12}(\mathbf{h}) = \gamma_{21}(\mathbf{h})$ e $\gamma_{12}(\mathbf{h}) = \gamma_{12}(-\mathbf{h})$, e que $C_{12}(\mathbf{h}) = C_{21}(-\mathbf{h})$, mas $C_{12}(\mathbf{h}) \neq C_{12}(\text{-}\mathbf{h})$.

2.3 Modelos teóricos de semivariogramas

Segundo Sturaro [85], a função semivariograma é utilizada em Geoestatística para expressar a variabilidade espacial numa direção predefinida.

2.3.1 Isotropia e anisotropia

Se o semivariograma é estacionário e tem o mesmo comportamento em todas as direções diz-se que é isotrópico, o que quer dizer que $\gamma(\mathbf{h})$ só depende de $\mathbf{h}$ apenas através do seu comprimento $\|\mathbf{h}\|$. Caso contrário, quando a estrutura de correlação varia na direção assim como na distância de $\mathbf{h}$, o fenómeno é chamado de anisotrópico [15, 3, 88].

De acordo com os mesmos autores, destacam-se dois tipos diferentes de anisotropia: anisotropia geométrica e anisotropia zonal.

2.3.1.1 Anisotropia geométrica

Se apenas uma transformação linear de coordenadas for suficiente para explicar a isotropia, então, diz-se que se está diante de uma anisotropia geométrica em termos de semivariograma.

Na anisotropia geométrica, o alcance do semivariograma varia de uma direção para outra, enquanto o patamar permanece constante. Este tipo de anistropia também é conhecido por anisotropia elíptica, pode ter zonas de influências elípticas, com os alcances distribuídos como uma elipse.
A figura 2.5, apresenta uma ilustração da anisotropia geométrica em termos de semivariograma.

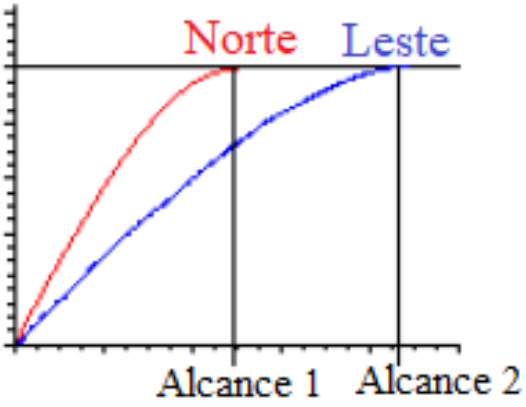

Figura 2.5: Semivariograma ilustrando a anisotropia geométrica, fonte: adaptado de Deutsch e Pyrcz [69].

2.3.1.2 Anisotropia zonal

Para este tipo de anisotropia, o patamar muda com a direção enquanto que alcance permanece constante. No caso de anisotropia zonal, um modelo complexo deve ser ajustado e esta não se consegue remover por uma simples transformação linear das coordenadas.

A anisotropia zonal pode ser provocada por mistura de populações, pois teremos variâncias diferentes conforme as populações.
A figura 2.6, apresenta uma ilustração da anisotropia zonal através de uma representação de um semivariograma.

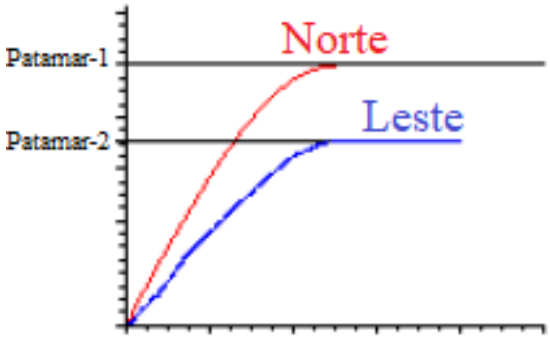

Figura 2.6: Semivariograma ilustrando a anisotropia zonal, fonte: adaptado de Deutsch e Pyrcz [69].

Ainda pode acontecer em simultâneo os dois tipos de anisotropia, o que é representado pelo semivariograma da figura 2.7.

Existem dois tipos de semivariogramas teóricos, a saber: semivariograma com patamar e sem patamar.

Nas próximas duas secções, considera-se $h = \|\boldsymbol{h}\|$, porque assume-se o pressuposto da

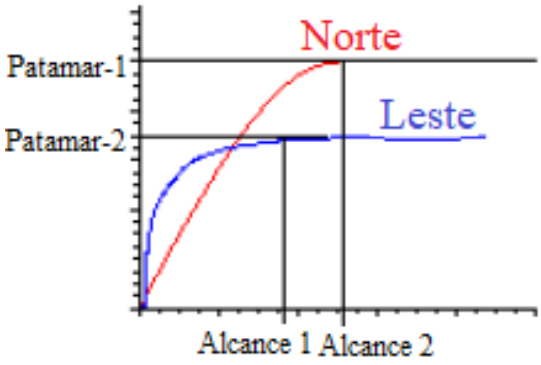

Figura 2.7: Semivariograma ilustrando a anisotropia geométrica e zonal, fonte: adaptado de Deutsch e Pyrcz [69].

isotropia.

2.3.2 Semivariograma com patamar

De acordo com Oleas [63], existem 7 modelos de semivariograma com patamar a saber: modelo esférico, modelo exponencial, modelo gaussiano, modelo cúbico, modelo penta-esférico, modelo de efeito furo e o modelo de Matérn. Para estes modelos, o alcance e o patamar C devem ser maiores que zero. São fatores que alteram a escala, mas não alteram a sua forma e, por isso são chamados de modelos transitivos.

A seguir vai-se apresentar a descrição dos quatro modelos de semivariogramas com patamar mais utilizados.

2.3.2.1 Modelo esférico

Seja h um *lag* entre quaisquer duas posições no espaço, então o modelo esférico de semivariograma é dado por:

$$\gamma(h) = \begin{cases} C\left(\frac{3}{2}\frac{h}{a} - \frac{1}{2}\left(\frac{h}{a}\right)^3\right), & 0 \leq h < a \\ C, & a \leq h, \end{cases} \tag{2.25}$$

onde C é o patamar e a é a amplitude.
Este modelo é considerado transitivo porque atinge um patamar finito numa amplitude também finita. O modelo esférico só é aplicado para um espaço com dimensões menor ou igual a 3.

2.3.2.2 Modelo exponencial

Seja h um *lag* entre quaisquer duas posições no espaço, então o modelo exponencial de semivariograma é dado por:

$$\gamma(h) = C\left(1 - e^{-\frac{3h}{a}}\right) \tag{2.26}$$

onde C é o patamar e a é a amplitude.
O modelo aproxima-se do patamar assintoticamente.

2.3.2.3 Modelo gaussiano

Seja h um *lag* entre quaisquer duas posições no espaço, então o modelo gaussiano de semivariograma é dado por:

$$\gamma(h) = C\left(1 - e^{-3\left(\frac{h}{a}\right)^2}\right) \tag{2.27}$$

O patamar é atingido assintoticamente e o gráfico desse modelo tem a forma parabólica junto à origem.

2.3.2.4 Modelo de Matérn

Seja h um *lag* entre quaisquer duas posições no espaço, então o modelo de Matern de semivariograma é dado por:

$$\gamma(h) = \begin{cases} 0, & h = 0 \\ C\left(1 - \dfrac{\frac{h}{\phi}}{2^{k-1}\Gamma(k)} K_k\left(\frac{h}{\phi}\right)\right), & h > 0 \end{cases} \tag{2.28}$$

Onde: Γ refere-se a função gama, K_k é a função Bessel modificada do segundo tipo de parâmetro k, $\phi > 0$ é um parâmetro de escala com dimensões da distância, k é um parâmetro de forma que determina a suavidade do processo subjacente.

2.3.3 Semivariograma sem patamar

Este é um tipo de semivariogramas que podem ser definidos, mas não se estabilizam em nenhum patamar. É particularmente adequado para casos de amostragem insuficiente ou incompleta ou ainda quando os dados apresentam tendências.

De acordo com Olea [63], Webster [95], os semivariogramas sem patamar são representados por um semivariograma de potências dado por:

$$\gamma(h) = \alpha h^{\beta}, \qquad 0 < \beta < 2 \tag{2.29}$$

onde α representa a intensidade da variação, o β descreve a curvatura. Os limites 0 e 2 são excluídos porque quando $\beta = 0$ indica variância constante para todo $h > 0$ e quando $\beta = 2$ a função é parabólica com gradiente zero na origem. O último significa que o processo não é aleatório.

Este é um modelo de semivariograma não transitivo. Se $\beta = 1$, teremos um caso especial de um semivariograma linear.

Sturaro [85] e Clark [18] apresentam dois modelos de semivariograma para esse grupo de semivariogramas sem patamar, que são: modelos lineares e modelos logaritmos.

2.3.3.1 Modelos Lineares

O modelo linear é dado por:

$$\gamma(h) = \alpha h \tag{2.30}$$

onde α é a inclinação da reta.
Uma extensão do modelo linear é a seguinte generalização dada por:

$$\gamma(h) = \alpha h^{\beta} \tag{2.31}$$

onde β varia de 0 a 2 (mas não pode ser 2).

2.3.3.2 Modelos Logaritmo (Esquema de Wings)

O modelo logaritmo [85] também chamado de modelo Wijsiano [18] é dado por:

$$\gamma(h) = 3\alpha log(h) \tag{2.32}$$

em que o semivariograma é linear se for traçado contra o logaritmo da distância. α é uma constante que representa a dispersão absoluta.

Pode-se ter ainda o modelo de efeito pepita puro, que é um caso especial degenerado de semivariograma transitivos com um alcance infinitesimal, o semivariograma surge diretamente de zero com um valor constante.

O modelo de efeito pepita puro é dado por:

$$\gamma(h) = C(1 - H(0)) \tag{2.33}$$

onde C é o efeito pepita e $H(0)$ é a função *Heaviside* que é 1 no *lag* 0 e zero caso contrário. De acordo com Oleas [63], existe um modelo de semivariogramas chamado de semivariogramas aninhados, que é utilizado para ajustar estruturas complexas que podem ser causadas por uma combinação de processos. Este tipo de modelo é dado por:

$$\gamma(h)_{total} = \gamma_1(h) + \gamma_2(h) + ... + \gamma_n(h) \tag{2.34}$$

Os modelos de semivariogramas aninhados podem ser criados e usados para qualquer combinação linear de modelos admissíveis.

A seguir abordar-se a análise do comportamento dum semivariograma perto da origem.

2.3.4 Comportamento dos semivariogramas perto da origem

É muito importante estudar o comportamento dos semivariogramas para valores pequenos de h, porque isso está relacionado a continuidade espacial da variável. De acordo com Amestrong [4], pode-se distinguir 4 tipos de comportamentos:

1) Comportamento quadrático indica que a variável regionalizada é altamente contínua e diferenciável. A forma quadrática também pode ser associada à presença de uma tendência;

2) Comportamento Linear indica que a variável regionalizada é contínua, mas não é diferenciável.

3) Comportamento descontínuo na origem quer dizer que $\gamma(h)$ não tende a zero enquanto h tende a zero. Isso significa que a variável é altamente irregular em distâncias curtas.

4) Comportamento de efeito pepita puro ou ruído branco indica que as variáveis regionalizadas $U(s+h)$ e $U(s)$ não estão correlacionadas para todos os valores de h não importa o quão próximos estejam. Este é um caso de uma total falta de estrutura.

2.3.5 Estimação Paramétrica da Estrutura de Covariância Espacial

Na modelação da estrutura de associação espacial em Geoestatística, a escolha de um modelo paramétrico para o variograma que a descreva é usualmente precedida por uma análise exploratória dos dados, em que se constrói um variograma empírico que ajuda a escolher qual o modelo teórico de variograma mais indicado para a aplicação em mãos [15].

Se considerarmos uma amostra o variograma pode ser estimado de forma não paramétrica pelo variograma empírico:

$$2\hat{\gamma}(\mathbf{h}) = \frac{1}{N(\mathbf{h})} \sum_{i=1}^{N(\mathbf{h})} [(U(s_i) - U(s_i + \mathbf{h})]^2 \tag{2.35}$$

onde: $2\hat{\gamma}(\mathbf{h})$ é o variograma estimado; $N(\mathbf{h})$ é o número de pares de valores observados $U(s_i)$ e $U(s_i + \mathbf{h})$, separados por um vector $\mathbf{h}$; $U(s_i)$ e $U(s_i + \mathbf{h})$ são valores observados da variável regionalizada nos pontos s_i e $s_i + \mathbf{h}$, onde $i = 1, ..., n$ separados pelo vector $\mathbf{h}$.

Escolhido então qual o modelo de variograma que parece mais apropriado, vários métodos de ajuste de semivariogramas paramétricos são sugeridos, tais como: métodos baseados em mínimos quadrados ordinários ou ponderados, métodos de máxima verosimilhança e métodos Bayesianos. A seguir vai descrever-se de acordo com Diggle e Ribeiro [27], os métodos de mínimos quadrados ordinários ou ponderados, assim como o método de máxima verosimilhança por serem os mais utilizados para ajuste dos variogramas.

De forma a simplificar o processo, o variograma amostral é organizado agrupando os seus pontos em intervalos de distância. Assim, o variograma amostral é descrito por

$\{(w_k, v_k, n_k) : k = 1, ..., m\}$, em que as ordenadas v_k representam a média das ordenadas do variograma empírico original (ou experimental [2]) correspondentes ao conjunto de n_k distâncias compreendidas na k-ésima divisão ou caixa do eixo das distâncias do variograma original, relativas às observações aí pertencentes. Este grupo de distâncias é representada pela sua distância média w_k. O variograma empírico não suavizado é o caso especial em que todos $n_k = 1$.

2.3.5.1 Método dos mínimos quadrados ordinários

As ordenadas do variograma empírico v_k são estimadores não enviesados de $\gamma(w_k)$, quaisquer que sejam as propriedades distributivas do processo subjacente $U(.)$. Assim, estimadores consistentes de θ num modelo paramétrico, $\gamma(w) = \gamma(w;\theta)$, podem ser obtidos minimizando uma das quantidades seguintes:

$$MQO_1(\theta) = \sum_{k=1}^{m} \{v_k - \gamma(w_k;\theta)\}^2, \tag{2.36}$$

ou

$$MQO_2(\theta) = \sum_{k=1}^{m} n_k \{v_k - \gamma(w_k;\theta)\}^2. \tag{2.37}$$

baseadas no variograma amostral.

Por conveniência computacional, normalmente é utilizada a equação definida em (2.37).

2.3.5.2 Método dos mínimos quadrados ponderados

Uma possível objeção aos métodos dos mínimos quadrados dados pelas equações (2.36) ou (2.37), foram propostos mais melhorias, em resposta ao facto de que a variância amostral de v_k depende do valor correspondente do variograma teórico, $\gamma(w_k;\theta)$, bem como de n_k. Sob hipóteses gaussianas, cada ordenada do variograma empírico v_{ij} tem valor esperado $\gamma(w_{ij};\theta)$ e variância $2\gamma(w_{ij};\theta)^2$, então Cressie [20] propôs a seguinte estimativa de γ-mínimos quadrados ponderados

$$MQP_2(\theta) = \sum_{k=1}^{m} n_k \left[\frac{v_k - \gamma(w_k;\theta)}{\gamma(w_k;\theta)} \right]^2, \tag{2.38}$$

Barry *et al.* [10], dizem que o uso da equação (2.38) corresponde ao uso de uma estimativa tendenciosa, porque o parâmetro desconhecido θ contribui para a ponderação. Para ver isto, vamos diferenciar $MQP_2(\theta)$ com respeito a cada elemento de θ, então temos, para cada j

[2] Sendo o variograma calculado a partir das observações disponíveis, dependendo do comportamento espacial da variável regionalizada [5].

$$\frac{\partial}{\partial\theta_j}MQP_2(\theta) = \sum_{k=1}^{m} 2n_k\left\{\frac{v_k-\gamma(w_k;\theta)}{\gamma(w_k;\theta)}\times\left(\frac{-v_k}{\gamma(w_k;\theta)^2}\right)\times\frac{\partial}{\partial\theta_j}\gamma(w_k;\theta)\right\} \quad (2.39)$$
$$= \sum_{k=1}^{m} 2n_k\left\{\frac{-v_k^2+v_k\gamma(w_k;\theta)}{\gamma(w_k;\theta)^3}\frac{\partial}{\partial\theta_j}\gamma(w_k;\theta)\right\}.$$

As estimativas dos γ-mínimos quadrados ponderados satisfazem as equações de estimativa $D_j(\theta)=0$ para todos os j, onde

$$D_j(\theta)=\frac{\partial}{\partial\theta_j}MQP_2(\theta).$$

Como v_k é aproximadamente não enviesado para o $\gamma(w_k;\theta)$, segue-se que

$$E[D_j(\theta)]\approx\sum_{k=1}^{m}2n_k\left[-\frac{V(v_k)}{\gamma(w_k;\theta)^3}\frac{\partial}{\partial\theta_j}\gamma(w_k;\theta)\right]\neq 0, \quad (2.40)$$

daí que as equações de estimativa sejam tendenciosas. Uma explicação intuitiva é que a minimização de (2.38) é equivalente à maximização de uma probabilidade gaussiana, mas ignorando o determinante da matriz de variância. Contudo, $V(v_k)$ é de ordem $(n_k)^{-1}$ e para um determinado tamanho de amostra n, o número de caixas m é de ordem $(\bar{n}_k)^{-1}$ onde $\bar{n}_k$ é a média do n_k. Assim, (2.40) sugere também que, na prática, a quantidade de enviesamento irá diminuir à medida que o n_k aumentar. Este resultado fornece uma justificação teórica para as orientações práticas dadas em textos de geoestatística aplicada.

Um conjunto não enviesado de equações de estimativa poderia ser obtido de uma forma iterativa através do algoritmo dos mínimos quadrados ponderados, como utilizado na modelação linear generalizada [31]. O conjunto resultante de equações de estimativa, também descrito em [10] resolve $D_j^*(\theta)=0$ para todos os j, temos

$$E[D_j^*(\theta)]=\sum_{k=1}^{m}n_k\left[\frac{v_k-\gamma(v_k;\theta)}{\gamma(w_k;\theta)}\frac{\partial}{\partial\theta_j}\gamma(w_k;\theta)\right] \quad (2.41)$$

e $E[D_j^*(\theta)]\approx 0$, conforme necessário. Então dentre os métodos dos mínimos quadrados acima apresentados, conclui-se que os mínimos quadrados n-ponderados dados pela equação (2.37) aplicados ao variograma da amostra dão um método simples e conveniente para obter estimativas iniciais dos parâmetros do variograma.

2.3.5.3 Envelopes de variogramas

Quando o variograma empírico parece mostrar pouca ou nenhuma correlação espacial, pode ser útil avaliar formalmente se os dados são compatíveis com um modelo subjacente com a seguinte forma $Y_i=\mu(s_i)+Z_i$, onde os Z_i são resíduos não relacionados com uma média espacialmente variável $\mu(s)$. Uma forma simples de fazer é calcular os resíduos sobre uma média ajustada $\hat{\mu}(s)$ e comparar o variograma empírico residual com um

envelope de variogramas empíricos calculados a partir de permutações aleatórias dos resíduos, mantendo fixos os locais correspondentes. De acordo com Carvalho e Natário [15] o método dos envelopes simulados para o variograma amostral (também chamado de *invólucro do variograma amostral* pelo método de Monte Carlo) consiste em simular independentemente m variogramas amostrais com base nas observações. Cada simulação corresponde a uma permutação dos resíduos entre as localizações. De seguida, para cada abcissa, calcula-se máximo e o mínimo dos valores obtidos nos m variogramas amostrais simulados, obtendo-se então os limites do referido envelope.

Os envelopes simulados são utilizados para verificar a existência de dependência espacial por meio de diagnóstico gráfico. Segundo Diggle e Ribeiro [31], rejeita-se a hipótese de independência espacial entre as observações se tiver pelo menos um ponto do variograma empírico fora do envelope.

2.3.6 Estimação por máxima verosimilhança

A geoestatística tradicional é baseada em métodos não paramétricos que assentam na função variograma para a estimação da correlação espacial e posterior utilização para fazer a krigagem da superfície do fenómeno espacial. Mais modernamente, Diggle e Ribeiro [31] e outros advogaram uma geoestatística baseada em modelos para os quais é necessário estabelecer um modelo para o fenómeno em estudo. Os processos estocásticos gaussianos são comummente utilizados como modelos para dados geoestatísticos, pela facilidade que induzem na estimação.

De acordo com Diggle e Ribeiro [27], na estatística clássica, a função de verosimilhança é a base sobre o qual métodos de inferência são construídos. Estimar os parâmetros do modelo maximizando a função de verosimilhança sob um modelo assumido fornece, sob condições muito gerais, estimadores imparciais e eficientes quando aplicados a grandes amostras. Se considerarmos modelos paramétricos, normalmente usa-se a função de verosimilhança como base para estimação de parâmetros.

2.3.6.1 Máxima verosimilhança sob suposição gaussiana

Apresenta-se aqui a descrição do método de máxima verosimilhança para estimação dos parâmetros da estrutura de covariância espacial sob suposição gaussiana de acordo com Diggle e Ribeiro [27].

Consideremos $Y = (Y_1, \ldots, Y_n)$ como dados gerados a partir de um modelo linear Gaussiano,

$$Y_i = \mu(s_i) + U(s_i) + \epsilon_i \ : \ i = 1, \ldots, n. \tag{2.42}$$

onde $U(s)$ é um processo Gaussiano estacionário com variância σ^2, a função de correlação $\rho(h, \phi)$ e ϵ são mutuamente independentes, $\epsilon \sim N(0, \tau^2)$. Na equação (2.42), $\mu(s_i)$ é

determinado por um modelo de regressão linear

$$\mu(s_i) = \sum_{k=1}^{p} f_k(s_i)\beta_k \tag{2.43}$$

onde $f_k(.)$ são funções de variáveis explicativas referenciadas espacialmente observadas. Para derivar a função de verosimilhança associada, é necessário escrever o modelo em forma de matriz. Seja $E[\boldsymbol{Y}] = \boldsymbol{F\beta}$, onde $\boldsymbol{F}$ é uma matriz $n \times p$ cuja k-ésima coluna consiste nos valores $f_k(s_i),\ldots,f_k(s_n)$, e $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_p)$ é um vetor de p-elementos de parâmetros de regressão. Agora, denote $\boldsymbol{\theta} = (\tau^2,\sigma^2,\phi)$ o conjunto de parâmetros que definem a matriz de covariância dos dados observados $\boldsymbol{Y}$, e escreva-se $V(\boldsymbol{Y}) = G(\boldsymbol{\theta})$. Note que $G(\boldsymbol{\theta}) = \tau^2\boldsymbol{I} + \sigma^2\boldsymbol{R}(\phi)$, onde $\boldsymbol{I}$ é a matriz identidade $n \times n$ e $\boldsymbol{R}(\phi)$ a matriz $n \times n$ com ij-ésimo elemento $r_{ij} = \rho\left(\|s_i - s_j\|;\phi\right)$.

Seguindo esta notação, os modelos definidos por (2.42) e (2.43) implicam que $\boldsymbol{Y} \sim MVN\{\boldsymbol{F\beta}, G(\boldsymbol{\theta})\}$. Segue que a log-verosimilhança para $\boldsymbol{(\beta,\theta)}$, a menos uma constante aditiva, é

$$l(\boldsymbol{\beta},\boldsymbol{\theta}) = -\frac{1}{2}\{log|G(\boldsymbol{\theta})| + (\boldsymbol{y} - \boldsymbol{F\beta})'\{G(\boldsymbol{\theta})\}^{-1}(\boldsymbol{y} - \boldsymbol{F\beta})\} \tag{2.44}$$

Para se maximizar log-verosimilhança da equação dada em (2.44), primeiro eliminamos o $\boldsymbol{\beta}$ da maximização numérica de $l(.)$. Para isto, para dado um $\boldsymbol{\theta}$ fixo, o estimador de máxima verosimilhança de $\boldsymbol{\beta}$ é dado por

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}) = \left[\boldsymbol{F}'\{G(\boldsymbol{\theta})\}^{-1}\boldsymbol{F}\right]^{-1}\boldsymbol{F}'\{G(\boldsymbol{\theta})\}^{-1}\boldsymbol{y} \tag{2.45}$$

Substituindo o $\hat{\boldsymbol{\beta}}$ em (2.44) teremos log-verosimilhança reduzida para $\boldsymbol{\theta}$,

$$L(\boldsymbol{\theta}) = l(\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}),\boldsymbol{\theta})$$

Em casos especiais, a estrutura algébrica de $G(\boldsymbol{\theta})$ e/ou $\boldsymbol{F}$ pode permitir simplificações adicionais. Em particular, sempre podemos extrair um fator escalar de $G(\boldsymbol{\theta})$, por exemplo escrevendo

$$G(\boldsymbol{\theta}) = \sigma^2\{\nu^2 + \boldsymbol{R}(\phi)\}$$

onde $\nu^2 = \frac{\tau^2}{\sigma^2}$ é um tipo de relação ruído-sinal. Estas reduções são importantes pois permitem eliminar σ^2 algebricamente do critério a ser numericamente maximizado. Desta forma, faz com que a confiabilidade das rotinas de maximização automática das funções sejam mais eficientes quando a dimensionalidade da função a ser maximizada é menor.

2.3.6.2 Máxima verosimilhança restrita

De acordo com Patterson e Thompson [64] citado por Diggle e Ribeiro [27] uma variante do método de máxima verosimilhança que tem como origem o planeamento de experiências é chamada de Maxima verosimilhança Restrita (MVR).

O estimador de máxima verosimilhança restrita é baseado em seguinte princípio:

- sob o modelo assumido para $E[\boldsymbol{Y}] = \boldsymbol{F\beta}$, transforme os dados linearmente para $\boldsymbol{Y}^* = \boldsymbol{AY}$ tal que a distribuição de $\boldsymbol{Y}^*$ não dependa de $\boldsymbol{\beta}$;

- estimar $\boldsymbol{\theta}$ por máxima verosimilhança aplicada aos dados transformados $\boldsymbol{Y}^*$. Observe que podemos sempre encontrar uma matriz $\boldsymbol{A}$ adequada sem conhecer os verdadeiros valores de $\boldsymbol{\beta}$ ou $\boldsymbol{\theta}$, por exemplo

$$\boldsymbol{A} = \boldsymbol{I} - \boldsymbol{F}(\boldsymbol{F'F})^{-1}\boldsymbol{F'}$$

  O estimadores de máxima verosimilhança restrita para $\boldsymbol{\theta}$ podem ser calculados maximizando

$$\begin{aligned} L^*(\boldsymbol{\theta}) &= -\frac{1}{2}\{log|G(\boldsymbol{\theta})| + log|\boldsymbol{F}'\{G(\boldsymbol{\theta})\}^{-1}\boldsymbol{F}| \\ &+ (\boldsymbol{y} - \boldsymbol{F}\tilde{\boldsymbol{\beta}})'\{G(\boldsymbol{\theta})\}^{-1}(\boldsymbol{y} - \boldsymbol{F}\tilde{\boldsymbol{\beta}})\}, \end{aligned} \tag{2.46}$$

  onde $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\boldsymbol{\theta})$. Note-se que a expressão dada em (2.46) depende de $\boldsymbol{F}$ e, portanto, depende de correta especificação do modelo para $\mu(s)$, mas não depende da escolha de $\boldsymbol{A}$.

Em comparação ao método de máxima verosimilhança tradicional, este método produz menores enviesamentos para amostras pequenas e é mais sensível a especificação incorretas de $\mu(s)$ [27].

2.3.7 Validação cruzada

Como foi descrito anteriormente na secção 2.3.2, existem vários modelos de variogramas teóricos. Os dados orientam a escolha do mais adequado para modelar a correlação espacial dos mesmos. Depois de se fazer o ajuste dos variogramas teóricos usando um dos métodos de estimação, Mínimos quadrados ou Máxima verosimilhança, deve-se escolher o melhor modelo ajustado. Essa escolha é feita através de alguns métodos de validação, tais como o de Informação de Akaike, de Filliben, e de validação cruzada. Este último, é o método que será usado neste trabalho.

Segundo Isaaks e Srivastava [40], o método de validação cruzada avalia erros de estimação permitindo comparar valores preditos com os valores amostrados. De acordo com Lv *et al.* [48], a validação cruzada *leave-one-out* consiste em ir ajustando de forma sequencial o modelo a todos os pontos da amostra exceto um (conjunto de treino) e comparar a previsão para esse ponto com o valor observado e não considerado. Este procedimento repete-se para todos os pontos da amostra e uma medida global destas comparações é obtida. Assim, para se avaliar os resultados do teste do método de validação cruzada, são usados os seguintes parâmetros de decisão: Erro médio (EM) e erro médio quadrático (EMQ), dados por

$$EM = \frac{1}{n}\sum_{i=1}^{n}(U(s_i) - U^*(s_i))$$

e

$$EMQ = \frac{1}{n}\sum_{i=1}^{n}(U(s_i) - U^*(s_i))^2$$

onde n é o número de dados, $U(s_i)$ é o valor observado no local s_i e $U^*(s_i)$ é o valor predito no local s_i.

O melhor modelo ajustado será aquele que apresentar menores valores possíveis para o EM, EMQ e desvio padrão do EM (S_{EM}), e que o valor de desvio padrão do EMQ (S_{EMQ}) mais próximo de um [87].

2.3.8 Software

Os pacotes utilizados para se encontrar o ajuste dos variogramas experimentais são *gstat* e *geoR*.

Gstat é pacote do *software* R que é usado para fazer a modelação de variograma, previsão e simulação geoestatística [65].

O *gstat* é utilizado principalmente para análise multivariada, permitindo uma implementação mais geral do modelo linear multivariado com tendências modelada como uma função linear de polinómios de coordenadas ou funções de base definidas pelo usuário e resíduos independentes ou dependentes modelados geoestatisticamente. Quanto a simulação, o *gstat* compreende a simulação sequencial gaussiana, condicional ou incondicional (multi) gaussiana de valores de pontos ou médias de blocos, ou simulação sequencial (multi) indicador [66].

No *gstat*, o ajuste do variograma ou semivariograma, utiliza-se a função *semivarigram*, e pode-se fazer o diagnostico do mesmo utilizando a função *vardiag*, mais detalhes, ver [65].

E o *geoR*, é também um pacote do *software* R que é principalmente utilizado para análise univariada. Embora, este pacote não permita alguns cálculos, tais como o cálculo do variograma cruzado, ele é muito utilizado porque permite um ajuste de modelos de variograma robusto e o ajuste do alcance.

Para o ajuste desses métodos no *geoR*, podem ser utilizados as seguintes funções: *variofit* (para mínimos quadrados) e *likfit* (para máxima verosimilhança). Estas funções e códigos no *software R* podem ser vistos em [31].

Os dois pacotes permitem fazer o ajuste de modelos de variograma ou de semivariograma robusto, e também produzem ajuste de modelos exponencial quase idênticos. O *gstat* disponibiliza uma variedade de funções para a Geoestatística univariada e multivariada, incluindo um conjunto de dados maior, enquanto que o pacote *geoR* contêm funções para Geoestatística baseada em modelos.

Neste trabalho, vai-se utilizar o pacote *geoR* para fazer-se a análise do variograma, assim como a krigagem.

2.4 Krigagem

Na abordagem clássica de estimação geoestatística, a inferência espacial é feita através de um processo de reprodução das características do fenómeno espacial baseado em pontos amostrais, que é determinado por interpolação ou estimativa. A interpolação de um ponto não amostrado é feita por um ajuste de funções matemáticas locais, nos pontos mais próximos ao ponto não amostrado, ou em todos os pontos amostrais [97].

Sabe-se que o semivariograma é um instrumento muito importante em geoestatística, visto que este fornece muitas informações sobre o fenómeno em estudo. O semivariograma é utilizado principalmente para perceber e informar sobre a variabilidade espacial do fenómeno. Para se estimar o mesmo em locais não amostrados ou estimar sua média numa área específica, usa-se então a krigagem.

Duma forma geral, dizer-se que o processo de estimação por excelência em Geoestatística é não paramétrico, feito por krigagem nas suas diferentes formas.

Por definição, krigagem é uma técnica de interpolação geoestatística que considera a distância e o grau de variação entre os pontos de dados conhecidos para estimar valores em áreas desconhecidas [91].

O nome krigagem foi dado por George Matheron em homenagem a Daniel Krige. O estimador de krigagem é designado linear por ser formado por uma combinação linear dos dados e do tipo Best Linear Umbiased Estimator (BLUE).

Yamamoto e Ladim [97], destaca dois tipos de krigagem: krigagem linear e krigagem não linear.

2.4.1 Krigagem linear

Neste tipo de krigagem, os métodos utilizados para fazer as estimativas são realizados com os dados originais, sem haver necessidade de uma transformação prévia dos mesmos. Os métodos mais conhecidos para este tipo de krigagem são: krigagem simples ou estacionária, krigagem média, krigagem ordinária, krigagem de blocos e krigagem universal.

Neste trabalho, vai-se fazer apenas a descrição da krigagem ordinária, porque é o método mais utilizado e abrangente.

2.4.1.1 Krigagem Ordinária

Krigagem ordinária é o método local de estimação geoestatística mais usado, quer pela sua simplicidade, assim como pelos resultados que proporciona [57].

De acordo com o mesmo autor, o estimador de Krigagem Ordinária (KO) pode ser descrita da seguinte maneira. Seja $U = \{U(s) : s \in D\}$ uma função aleatória estacionária de 2ª ordem com média constante mas desconhecida μ e uma função de covariância conhecida $C(h)$. Sob esta suposição, as equações de krigagem ordinária, que fornecem os pesos necessários para a previsão de krigagem da função aleatória num ponto não observado, podem ser expressas em termos da função de covariância ou em termos do

semivariograma. Em qualquer um desses dois casos, $U = \{U(s) : s \in D\}$ é predito num ponto não observado s_0 usando o preditor linear

$$U^*(s_0) = \sum_{i=1}^{n} \lambda_i U(s_i), \tag{2.47}$$

impondo ao erro de previsão que sua esperança deve ser zero e sua variância mínima.

Derivando a equação da krigagem ordinária em termos das funções de covariância, com o fim de cumprir com a condição de que o preditor de krigagem seja não enviesado (ou que seja centrado), a soma dos pesos devem ser um, então temos:

$$\begin{aligned} E[U^*(s_0) - U(s_0)] &= E\left[\sum_{i=1}^{n} \lambda_i U(s_i) - U(s_0)\right] \\ &= \sum_{i=1}^{n} \lambda_i E[U(s_i)] - E[U(s_0)] = \mu \sum_{i=1}^{n} \lambda_i - \mu = 0 \\ &\Leftrightarrow \sum_{i=1}^{n} \lambda_i = 1 \end{aligned} \tag{2.48}$$

a variância é dada por:

$$\begin{aligned} V(U^*(s_0) - U(s_0)) &= E[(U^*(s_0) - U(s_0))^2] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j E[U(s_i)U(s_j)] - 2\sum_{i=1}^{n} \lambda_i E[U(s_i)U(s_0)] + \\ &+ E[(U(s_0))^2], \end{aligned} \tag{2.49}$$

pelo qual agora, tendo em mente que $E[U(s)] = \mu$, temos que:

$$E[U(s_i)U(s_j)] = C(s_i - s_j) + \mu^2, \tag{2.50}$$

$$E[U(s_i)U(s_0)] = C(s_i - s_0) + \mu^2, \tag{2.51}$$

$$E[(U(s_0))^2] = C(\mathbf{0}) + \mu^2, \tag{2.52}$$

e, portanto, a expressão (2.49) é transformada em:

$$\begin{aligned} V(U^*(s_0) - U(s_0)) &= \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j C(s_i - s_j) - 2\sum_{i=1}^{n} \lambda_i C(s_i - s_0) + \\ &+ C(\mathbf{0}) + \mu^2 \left(\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j - 2\sum_{i=1}^{n} \lambda_i + 1 \right). \end{aligned} \tag{2.53}$$

como $\sum_{i=1}^{n} \lambda_i = 1$, então a expressão é reduzida a:

$$\begin{aligned} V(U^*(s_0) - U(s_0)) &= \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j C(s_i - s_j) - 2\sum_{i=1}^{n} \lambda_i C(s_i - s_0) + \\ &+ C(\mathbf{0}). \end{aligned} \tag{2.54}$$

O problema de krigagem ordinária é encontrar os pesos $\lambda_i,\ i = 1,...,n$ minimizando $V(U^*(s_0) - U(s_0))$ sujeito à restrições lineares $\sum_{i=1}^{n} \lambda_i = 1$.
Este problema é resolvido pelo método dos multiplicadores de Lagrange, a função de Lagrange é dada por:

$$\varphi(\lambda_i, \alpha) = V(U^*(s_0) - U(s_0)) - \alpha\left(\sum_{i=1}^{n} \lambda_i - 1\right), \tag{2.55}$$

que, derivando parcialmente em relação aos pesos $\lambda_i,\ i = 1,...,n$ e ao multiplicador de Lagrange α e igualando essas derivadas parciais com o zero, resulta no seguinte sistema de $n+1$ equações com $n+1$ incógnitas:

$$\begin{cases} \sum_{j=1}^{n} \lambda_j C(s_i - s_j) - \alpha = C(s_i - s_0),\ \ i = 1,...,n \\ \sum_{i=1}^{n} \lambda_i = 1 \end{cases} \tag{2.56}$$

a variância minimizada, chamada de variância de krigagem ordinária, é obtida substituindo $\sum_{j=1}^{n} \lambda_j C(s_i - s_j)$ com $C(s_i - s_j) + \alpha$ como pode ser deduzido de (2.56), na expressão de variância (2.54), que incorpora a condição de que é centrada:

$$\sigma^2_{OK}(s_0) = C(\mathbf{0}) - \sum_{i=1}^{n} \lambda_i C(s_i - s_0) + \alpha \tag{2.57}$$

Porém, ao resolver sistema (2.56) surge um problema: embora a média constante desconhecida da função aleatória, μ, não apareça expressamente no sistema, seria necessário que ela fosse determinada para estimar a covariância.
Para se evitar esse problema deve-se expressar o sistema de krigagem ordinária em termos de semivariogramas. No caso estacionário de 2ª ordem $C(s_i - s_j) = C(\mathbf{0}) - \gamma(s_i - s_j)$, o sistema de krigagem ordinária em (2.56) e a variância da krigagem ordinária (2.57) podem ser escritas da seguinte forma:

$$\begin{cases} \sum_{j=1}^{n} \lambda_j \gamma(s_i - s_j) + \alpha = \gamma(s_i - s_0),\ \ i = 1,...,n \\ \sum_{i=1}^{n} \lambda_i = 1 \end{cases} \tag{2.58}$$

e

$$\sigma^2_{OK}(s_0) = \sum_{i=1}^{n} \lambda_i \gamma(s_i - s_0) + \alpha, \tag{2.59}$$

respetivamente.

No caso em que $U = \left\{U(s) :\ s \in D \subseteq \mathbb{R}^d\right\}$ não é estacionário de 2ª ordem, mas intrinsecamente estacionário, o sistema de krigagem ordinária só pode ser expresso em termos do semivariograma, já que, nem a variância nem a covariância são definidas. Nesse caso, o sistema de krigagem ordinária é o mesmo que para uma função aleatória estacionária de 2ª ordem, mas a condição $\sum_{i=1}^{n} \lambda_i = 1$ não é uma condição de não ser enviesado, mas uma condição de permissibilidade.

2.4.2 Krigagem não linear

A estimação pelo método de krigagem linear descrito em 2.4.1, apresentam limitações em algumas situações, como por exemplo, quando se pretende estimar a distribuição espacial em vez de estimar apenas o valor médio em algum local ou área; quando se está perante uma distribuição fortemente enviesada, a estimativa do valor médio usando a krigagem linear não é muito aconselhada, a presença de valores extremos podem tornar qualquer estimativa linear muito instável.

Por conta dessas limitações os métodos de estimação por krigagem não linear surgem como alternativa. Segundo Vann e Guibal [90], no contexto geoestatístico, a interpolação não linear é uma tentativa de estimar a esperança condicional e de promover a distribuição condicional do atributo num local, em vez de simplesmente prever o atributo em si. Ou pode-se dizer que os estimadores geoestatísticos não lineares são aqueles que utilizam modelos dos dados para obter ou aproximar a esperança condicional.

Fazem parte deste tipo de krigagem os seguintes métodos: krigagem multigaussiana, krigagem lognormal e indicativa.

A descrição dos métodos acima mencionados podem ser vistos com mais detalhes em [63, 62, 97, 22].

2.4.3 Cokrigagem

As técnicas de krigagem lineares e não lineares acima referidas são todas no contexto univariado, mas existem casos em que se precisa de trabalhar com diversas variáveis regionalizadas, estas variáveis podem apresentar-se fortemente correlacionadas no espaço entre si, portanto isto permite uma estimativa conjunta (como por exemplo coestimativa) das variáveis em estudo. Para este tipo de situação utiliza-se as técnicas de cokrigagem, que é um procedimento de estimativa multivariada para o contexto corregionalizado. A corregionalização em geoestatística é definida como sendo a regionalização de duas ou mais variáveis em estudo medidas no mesmo local de amostragem [40, 63].

Duma forma geral, pode-se dizer que cokrigagem é essencialmente uma generalização da krigagem utilizando funções aleatórias vetoriais.

2.5 Análise de dados de pescado de lulas (quilogramas), obtidos em ações de fiscalização

Neste capítulo vai-se fazer a aplicação das técnicas geoestatísticas, especificamente a análise de variograma experimental e a krigagem descrita no capítulo 2.4. Essas técnicas são aplicadas a dados de pescado de lulas (quilogramas) inspecionado em ações de fiscalização da Marinha Portuguesa na região costeira do Algarve no ano 2015.

Os métodos geoestatísticos descritos anteriormente são agora aplicados para verificar a distribuição espacial e fazer a predição por meio de krigagem da quantidade de pescado

de lulas inspecionado em ações de fiscalização da Marinha Portuguesa (em quilogramas), na região costeira do Algarve.

No âmbito desse estudo, usou-se a base de dados de pescado, com dados georeferenciados relativos a ações de fiscalização efetuadas pela Marinha Portuguesa em toda costa Portuguesa referente ao ano 2015, com 80 observações, que respeita à zona do Algarve, delineada pelas coordenadas: -10^o a -7.5^o de longitude e 36.5^o a 38^o de latitude, conforme a figura 2.8. A única variável resposta em estudo é a quantidade do pescado de lulas, inspeccionado em ações de fiscalização da Marinha Portuguesa.

Apesar das coordenadas e do mapa da figura 2.8 serem apresentadas em graus, toda a análise dos dados será feita em quilómetros. A transformação foi feita com base no datum WGS84 para o sistema Universal Transverso de Mercator (UTM). Na análise é usado o pacote *geoR* do *software* R, descrito em Diggle e Ribeiro [31].

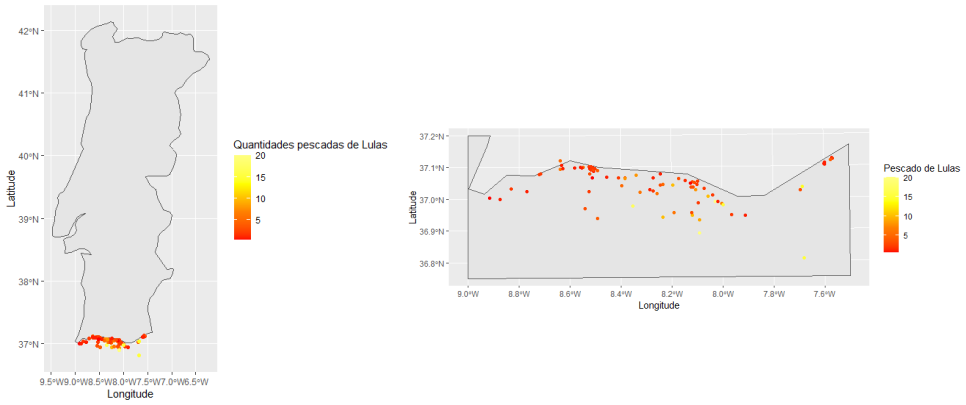

Figura 2.8: Apresentação de dados de pescado de lulas (quilogramas), obtidos em ações de fiscalização, na região costeira do Algarve

2.5.1 Análise exploratória

A figura 2.9 mostra o gráfico de círculos com diâmetro proporcional aos valores da quantidade pescada nos 80 pontos amostrados da variável em estudo. Verifica-se maiores valores da quantidade pescada para pontos mais distantes da costa na zona leste de Algarve e menores valores para os pontos mais próximos da costa com maior incidência na zona oeste de Algarve.

Dentre vários critérios existentes para gerar a matriz de pesos para calcular o índice de Moran, neste trabalho usou-se a matriz de pesos de distâncias inversas no pacote *ape* do *software* R, onde as entradas para os pares de pontos que estão próximos são maiores do que os pares de pontos distantes. Na matriz criada, os elementos da diagonal principal são zero e os elementos fora da diagonal principal w_{ij} são dados pelo inverso da distância entre os locais s_i e s_j.

Apresenta-se na tabela 2.1 o resultado do teste de hipóteses para verificar a existência de autocorrelação espacial, utilizando o Índice de Moran, calculado pela equação dada

em (2.18), para tal, consideremos as seguintes hipóteses:
H_0: Não há autocorrelação espacial;
H_1: Há autocorrelação espacial.

Tabela 2.1: Resultados do Teste de Moran.

V. observado	V.Esperado	sd	p.valor
0.09	-0.01	0.036	0.005

Com base nos resultados apresentados na tabela 2.1, rejeita-se a hipótese nula de que não há correlação espacial a um nível de significância $\alpha = 0.05$

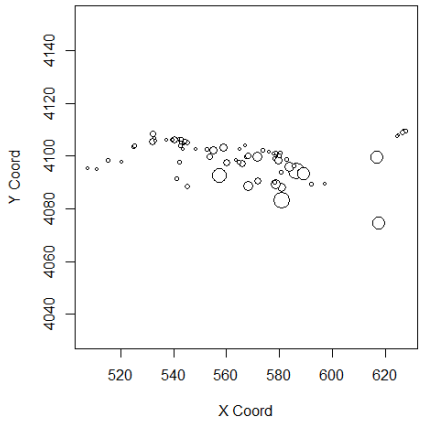

Figura 2.9: Apresentação de dados de pescado de lulas (quilogramas), obtidos em ações de fiscalização, na região costeira do Algarve

Na tabela 2.2 tem-se os valores da mediana, média e do desvio padrão da variável em estudo neste trabalho, quantidade de pescado de lulas inspecionado em ações de fiscalização em 2015 (em quilogramas).

Tabela 2.2: Tabela descritiva da variável em estudo.

Variável	Mediana	Média	Desvio padrão
Quantidade	2.0	4.18	4.47

Na figura 2.10 tem-se os gráficos descritivos da variável em estudo (quantidade de pescado de lulas em 2015 inspecionado em ações de fiscalização em quilogramas). O gráfico do canto superior esquerdo, categoriza os dados em quartis amostrais das observações, em que as cores amarela, verde, azul e vermelha, nesta ordem, indicam os quartis amostrais; os painéis superior direito e inferior esquerdo mostram os dados dispostos em relação às coordenadas y e x da correspondente localização, respetivamente, onde se procura verificar a existência de alguma tendência nos dados em função destas coordenadas. O diagrama no painel inferior direito mostra o histograma dos dados. Estes dados sugerem a existência de padrão espacial nos dados na região costeira do Algarve, concretamente

na região entre as 520 quilómetros a 590 quilómetros de longitude, uma vez que existem pequenos conglomerados das categorias nesta zona.

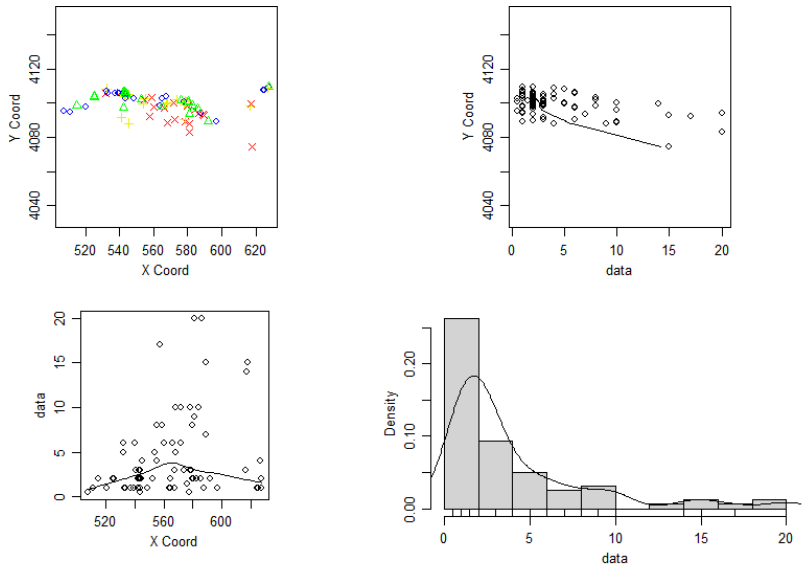

Figura 2.10: Gráficos descritivos do padrão espacial da distribuição das lulas na região costeira do Algarve-2015. No canto superior esquerdo encontra-se o gráfico de dispersão separado por quartis dos dados; no canto superior direito o gráficos dos dados contra latitude; no canto inferior esquerdo o gráfico dos dados contra longitude; e no canto inferior direito o histograma dos dados

.

Construíram-se variogramas empíricos para a quantidade de pescado de lulas, obtidos em ações de fiscalização, apresentados na figura 2.11. O painel esquerdo mostra um variograma da superfície original, enquanto que o painel direito mostra variogramas para os resíduos dos modelos de superfície de tendência linear e quadrática, indicado por linhas sólidas e tracejadas respetivamente. Pode-se notar que para a variável em estudo, os variogramas empíricos explicam muito bem a dependência espacial para as menores distâncias.

A figura 2.12 mostra envelopes dos variogramas obtidos de permutações aleatórias independentes dos resíduos de uma superfície de tendência linear e quadrática ajustada às quantidades do pescado de lulas por mínimos quadrados ordinários.

A existência de pontos fora do envelope de variograma no painel esquerdo da figura 2.12, indicam a presença de padrão espacial, enquanto que o envelope do variograma do painel direito não apresenta pontos fora. Ainda na mesma figura é também detetada a dependência espacial. Isso mostra que a tendência crescente no variograma empírico é estatisticamente significativa, o que indica a presença de correlação espacial positiva.

Ainda relativamente aos variogramas empíricos da figura 2.11, a linha contínua no painel direito mostra um comportamento típico de um processo estacionário, espacialmente correlacionado, isto é, um aumento nivelado até a estabilidade do variograma, embora nota-se uma ligeira diminuição em distâncias maiores. A forma dos variogramas

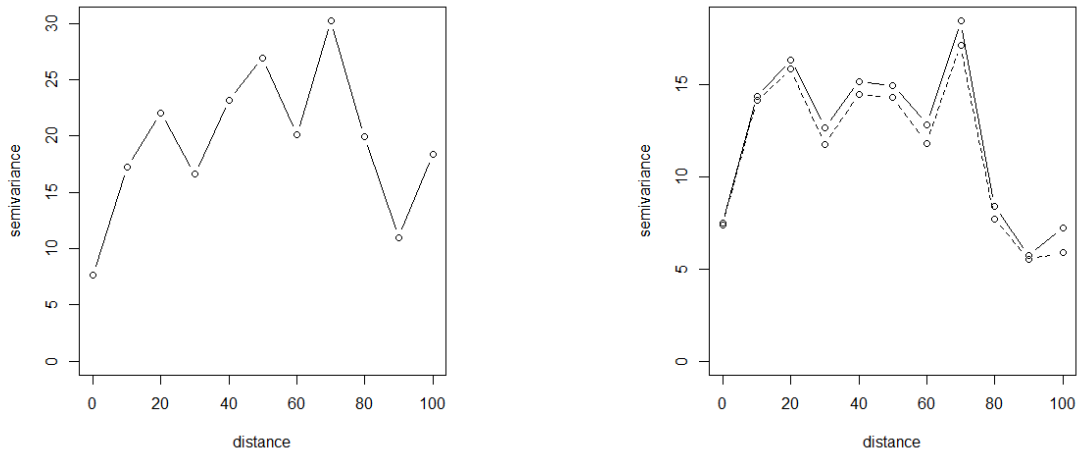

Figura 2.11: Variograma empíricos para dados originais (a esquerda) e variograma empírico dos resíduos (a direita) de uma superfície de tendência linear (linhas sólidas) ou quadrática (linhas tracejadas)

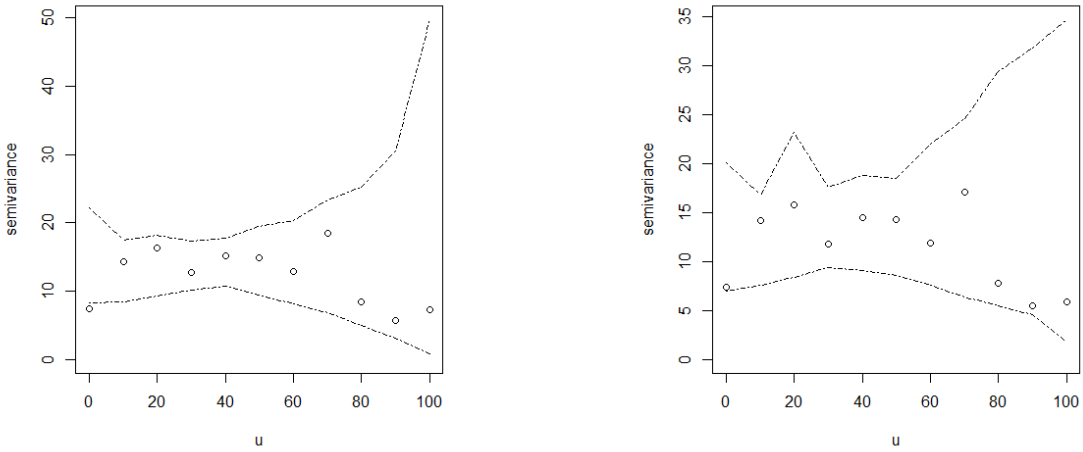

Figura 2.12: Gráficos dos envelopes de mínimos quadrados comuns de variogramas dos dados de pescado de lulas (quilogramas) após o ajuste de tendência linear (painel esquerdo) ou quadrático (painel direito), para diferentes *lags* u.

empíricos dos resíduos tanto do modelo de tendência linear e quadrática (painel direito) são muito parecidos, não evidenciando grandes diferenças.

2.5.2 Estimação de variogramas

A seguir serão ajustados os modelos teóricos de variogramas exponencial, esférico, gaussiano e de Matérn com $k = 1.5$, usando o método de estimação de máxima verosimilhança, para posterior escolha do melhor modelo ajustado.

A tabela 2.3 apresenta os parâmetros calculados pelo método de validação cruzada.

Tabela 2.3: Valores das medidas de performance pela validação cruzada (*leave one out*).

Método	Modelos	EM	EMQ	S_{EM}	S_{EMQ}
Máxima verosimilhança	Gaussiano	0.155	0.023	4.498	1.065
	Esférico	0.130	0.018	4.476	1.047
	Exponencial	0.126	0.017	4.560	1.069
	Matérn	0.101	0.012	4.186	1.043

Dos resultados apresentados na tabela 2.3, verifica-se que os valores do erro médio e erro médio padrão são menores para o modelo de variograma da família de Matérn, então, esse é o melhor modelo ajustado e será o escolhido para as análises que se seguem.

Este modelo para os dados de pescado de lulas (quilogramas), obtidos em ações de fiscalização, na região costeira do Algarve, assumem um modelo gaussiano estacionário com uma função de correlação de Matérn com valor fixo de $k = 1.5$. As estimativas usando mínimos quadrados ponderados e o método de máxima verosimilhança sob suposição gaussiana (descritos em 2.3.5.2 e 2.3.6.1) dos parâmetros restantes são dadas na tabela 2.4.

Tabela 2.4: Valores das estimativas dos parâmetros (intercepto β_0, coeficiente da coordenada latitude β_1, coeficiente da coordenada longitude β_2, efeito pepita τ^2, variância espacial σ^2, alcance ϕ e patamar $\tau^2 + \sigma^2$).

Método de ajuste	β_0	β_1	β_2	σ^2	Φ	τ^2	$\tau^2 + \sigma^2$
MQP	-	-	-	12.92	0.00	7.65	20.57
MV	5.27	-	-	14.00	5.70	11.38	25.38
MV.Tred	1394.79	0.03	-0.34	0.00	0.00	13.52	13.52

Como a superfície de tendência é responsável por parte da variação espacial, a estimativa de variância espacial (σ^2) de máxima verosimilhança com tendência (MV.Tred) é menor do que no modelo estacionário (MV) e o mesmo acontece com o parâmetro alcance (ϕ).

Usando estas estimativas do variograma podemos agora avançar para o processo de krigagem.

2.5.3 Krigagem

O painel da esquerda da figura 2.13 fornece resultados do preditor de krigagem de erro quadrático mínimo aplicado aos dados de pescado de lulas (quilogramas), obtidos em

ações de fiscalização, na região costeira do Algarve, usando como valores para os parâmetros do modelo de variograma as estimativas de máxima verosimilhança (MV) da tabela 2.4. Enquanto que o painel esquerdo da figura 2.14, mostra os correspondentes erros padrão de predição. As previsões seguem a tendência observada da quantidade de pescado de lulas, obtidos em ações de fiscalização, na região costeira do Algarve. As variâncias de predição são geralmente pequenas em locais próximos aos locais de amostragem, porque $\hat{\tau}^2$ é relativamente pequeno. Os mapas dos painéis direitos das figuras 2.13 e 2.14 são obtidos usando o modelo de superfície de tendência linear e suas estimativas de máxima verosimilhança associada, (mais detalhes em Diggle e Ribeiro [31]). Esta forma de estimação parece não ter deixado grande variabilidade espacial por explicar.

A krigagem ordinária produziu uma superfície predita que captura qualitativamente a tendência espacial aparente nos dados e que é quase idêntica às previsões obtidas usando o modelo de superfície de tendência linear mais razoável.

Verificou-se através dos mapas que o efeito espacial é alto na zona de Algarve situada entre 550 quilómetros a 592 quilómetros e valores relativamente altos ao largo do quilómetro 620.

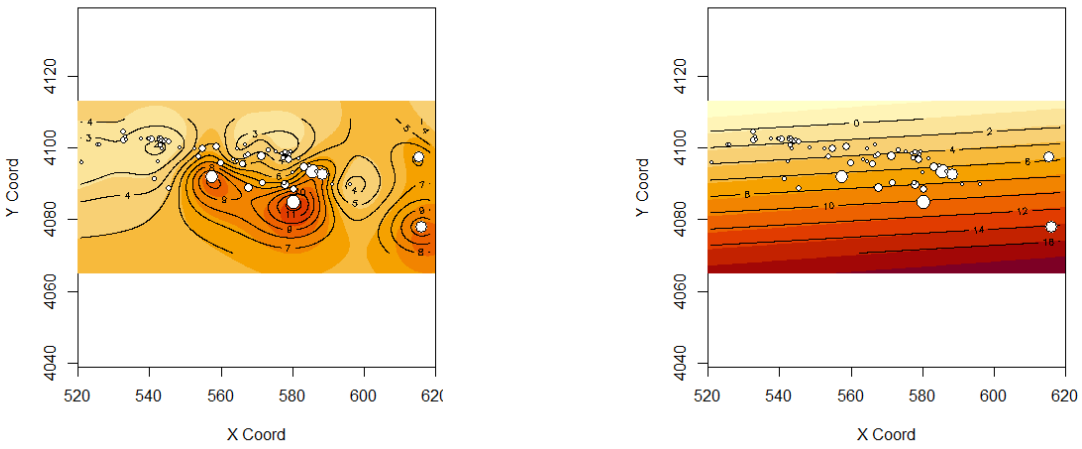

Figura 2.13: Predições por Krigagem ordinária dos dados de pescado de lulas (quilogramas), obtidos em ações de fiscalização, na região costeira do Algarve-2015. O painel esquerdo mostra o preditor de krigagem ordinária como uma imagem a cores e gráfico de contorno; os locais de amostragem são plotados como círculos com raios proporcionais às quantidades pescadas. O painel direito fornece as mesmas informações, mas com base no modelo com uma superfície de tendência linear. A coloração vermelha indica as regiões com maiores quantidades estimadas de pescado de lulas (quilogramas).

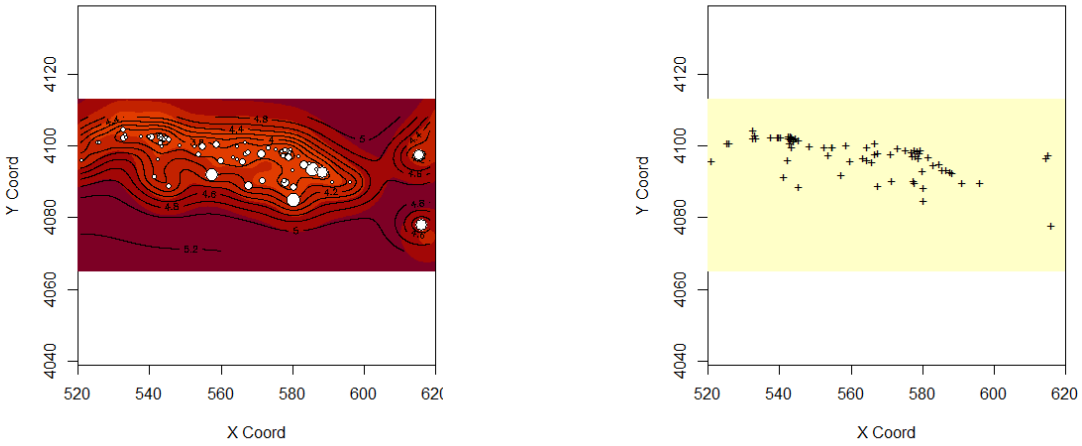

Figura 2.14: Predições por Krigagem ordinária dos dados de pescado de lulas (quilogramas), obtidos em ações de fiscalização, na região costeira do Algarve-2015, O painel esquerdo mostra os desvios padrão da previsão (valores maiores correspondem a cores mais fortes), os locais de amostragem são representados como círculos com raios proporcionais às quantidades pescadas. O painel direito fornece as mesmas informações, mas com base no modelo com uma superfície de tendência linear.

2.6 Conclusões

A krigagem é uma técnica bastante usada em geoestatística nas suas diferentes formas, principalmente a krigagem simples e a krigagem ordinária aplicadas em diferentes áreas, como por exemplo a agricultura, geologia, meio ambiente, entre outras. Neste trabalho usou-se a krigagem ordinária aplicada com recurso ao pacote *geoR* do *software* R.

Primeiramente, no desenvolvimento deste trabalho, fez-se uma pesquisa bibliográfica onde foram encontradas na literatura matérias que dizem respeito à geoestatística clássica, desde os conceitos básicos, variograma e suas propriedades, métodos de ajuste de variogramas e técnicas de interpolação.

O estudo teve como principal objetivo apresentar a modelação geoestatística baseada em técnicas de krigagem. Fez-se uma aplicação da krigagem, onde se demonstrou sua utilidade na estimação das quantidades do pescado de lulas, inspecionado em ações de fiscalização da Marinha Portuguesa no ano 2015 na região costeira do Algarve.

A metodologia tradicional da geoestatística permite o desenvolvimento de um modelo de semivariograma que captura a correlação espacial, que por sua vez permite fazer a krigagem e estimar a quantidade do pescado em pontos de locais onde as quantidades não eram conhecidas. O modelo foi desenvolvido, usando a krigagem ordinária, e através desta técnica gerou-se e apresentou-se o mapa de variâncias.

A abordagem da geoestatística com base na krigagem apresenta claras vantagens práticas, ao ser não paramétrica (como por exemplo, ser mais precisas em termos de estimação), não necessitando de pressupostos distribucionais sobre o fenómeno espacial.

Como sugestão para continuidade deste trabalho pode-se fazer estudos de outras espécies que são apanhadas durante as ações de fiscalização nesta região costeira, e pode-se fazer um estudo para tentar perceber se existe alguma relação entre a distribuição do pescado e algumas infrações associadas a pesca. Este estudo será ainda completado usando uma abordagem de modelação, sob a perspectiva clássica e também Bayesiana. Esta última abordagem tem a vantagem de não restringir o processo estocástico das observações ao modelo gaussiano.

3

Modelos Geoestatísticos para dados Binários

3.1 Introdução

Atualmente, nas mais diversas áreas de estudo são utilizados modelos de regressão para dados em que a variável resposta é binária, para os quais os modelos lineares clássicos não são apropriados porque dificilmente são alcançados os pressupostos dos modelos (normalidade e variâncias constantes). Nestes casos, o modelo de regressão logística tem sido o mais empregue para se trabalhar com dados binários [38]. Os modelos de regressão para dados binários, em particular o modelo logístico, podem ser enquadrados na classe dos modelos lineares generalizados (MLG) [55]. Para esta classe de modelos introduzida por Nelder e Wedderburn [61], a principal característica é que a variável resposta segue uma distribuição da família exponencial, e as variáveis explicativas, combinadas num preditor linear, relacionam-se com a média da variável resposta através de uma função de ligação monótona e diferenciável.

Dados geoestatísticos binários são dados espaciais em que a variável resposta toma um de dois valores possíveis, dividindo a região em estudo em duas sub-regiões disjuntas de acordo com esses valores [26]. Para se analisar este tipo de dados têm-se usado modelos espaciais Bayesianos, incorporando a componente espacial de forma hierárquica como um campo aleatório latente. Tradicionalmente são usados métodos Monte Carlo via Cadeia de Markov (MCMC) de ajuste destes modelos, que podem ser substituídos pela abordagem *Integrated Nested Laplace Approximation* (INLA), computacionalmente menos exigente e sem problemas de convergência, permitindo fazer inferência Bayesiana aproximada em modelos gaussianos latentes, tais como modelos lineares generalizados mistos. Conjugada com esta metodologia, a recente abordagem *Stochastic Partial Differential Equation* (SPDE), estima com facilidade o campo aleatório, resolvendo problemas de dimensionalidade associados às outras metodologias de estimação.

Neste contexto, este trabalho tem como principal objetivo apresentar técnicas de modelação de dados geoestatísticos com resposta binária usando a combinação das abordagens INLA-SPDE. Inicialmente é feito uma revisão sobre os dados binários, modelos lineares generalizados, modelos lineares generalizados mistos e modelos hierárquicos, INLA-SPDE. Depois é feita uma aplicação desta metodologia utilizando um conjunto de dados de fiscalização marítima da costa Portuguesa, produzindo mapas de médias e erro padrão do efeito espacial subjacente e dos mapas de riscos, usando o pacote R-INLA.

Para se desenvolver o conceito de Geoestatística para dados binários, vai-se apresentar primeiro alguns conceitos básicos sobre a geoestatística e os dados binários.

Isaaks e Srivastava [40] dizem que a **geoestatística** é um conjunto de técnicas utilizadas para descrever a continuidade espacial dos fenómenos naturais e fornece adaptações de técnicas clássicas de regressão para melhor se aproveitar dessa continuidade.

Soares [80] define **variável aleatória** em geoestatística da seguinte maneira: considere um valor localizado espacialmente em s_1, é interpretado como uma realização $y(s_1)$ da variável aleatória $Y(s_1)$. No espaço D, no qual se dispersa o conjunto de n pontos amostrais, temos realizações das n variáveis aleatórias $Y(s_1), Y(s_2), ..., Y(s_n)$ correlacionadas entre si.

Agresti [1] define **dados binários** como sendo um tipo de dados estatísticos que consiste em dados categóricos que assumem apenas dois valores possíveis, tais como, *"A"* e *"B"*, morto ou vivo, ou presença ou ausência. Os dados binários são nominais, o que quer dizer que eles representam valores qualitativamente diferentes que não se podem comparar numericamente. Frequentemente os dados binários são convertidos em dados de contagem, considerando um dos dois valores, como sendo *"sucesso"* representado por 1 (um) ou *"fracasso"* representado por 0 (zero).

Variável binária é uma variável aleatória que apresenta dois valores possíveis, isto é, valores binários. As variáveis binárias seguem uma distribuição de Bernoulli.

As contagens totais de variáveis binárias relativas à mesma experiência aleatória, obtidas de forma idêntica e independente (somas de variáveis binárias codificadas como 0 e 1), seguem uma distribuição binomial num número fixo de repetições da experiência, mas quando as variáveis binárias não são independentes e identicamente distribuídas (iid) a distribuição pode não ser binomial.

Os métodos de regressão mais destacados para os dados binários são, a regressão logística e regressão *probit*, elementos da família de Modelos Lineares Generalizados e detalhados no capítulo seguinte.

Consideremos um modelo de dados espaciais numa região D em que o fenómeno de interesse varia, a região D pode ser dividida em duas sub regiões disjuntas, $B \cup W$ (conhecida como mapa binário). Então De Oliveira [26] define **dados binários geoestatísticos** como sendo um tipo de dados espaciais que surgem quando as observações são feitas em locais de amostragem individuais, de modo que as medições determinam a qual sub-região, B ou W, cada local de amostragem pertence. Podem estar também disponíveis covariáveis que dependem da localização para ajudarem a explicar o mapa binário. A análise deste tipo de dados geralmente tem como objetivo prever qual das duas sub-regiões,

B ou *W*, uma localização não amostrada pertence (problema de predição/classificação) e estimar os efeitos das covariáveis no mapa binário (problema de regressão).

Para lidar com dados geoestatísticos, Diggle e Ribeiro [31] sistematizaram uma abordagem de modelação paramétrica assente em estimação por máxima verosimilhança. Diggle *et al.* [30] propuseram uma ampla classe de modelos hierárquicos para dados geoestatísticos, com particularidade para uma subclasse de Modelos Lineares Generalizados Mistos (MLGM), que foram utilizados para descrever dados binários. De Oliveira [26] propõe dois outros modelos além dos modelos lineares generalizados mistos que Diggle *et al.* [30] também propõe, que são Modelos de Cópula Gaussiana e modelos baseados em momentos. Uma Cópula é uma função distribuição de probabilidade multivariada onde cada variável tem uma distribuição marginal uniforme (0,1). Uma Cópula é uma ferramenta que modela a dependência entre várias variáveis aleatórias. No caso Gaussiano, utiliza-se a distribuição normal multivariada para descrever a estrutura de dependência entre algumas variáveis aleatórias. Madsen [49], Han e De Oliveira [36] e De Oliveira [26] propuseram o uso de Cópulas Gaussiana para modelar dados geoestatísticos discretos, assim como modelar dados geoestatísticos binários. Modelos baseados em momentos são ainda considerados por De Oliveira [26], como sendo os mais alinhados com os métodos geoestatísticos tradicionais e também são utilizados para análise de dados geoestatísticos binários, consistem em utilizar modelos que especificam apenas funções de média e correlação do campo aleatório binário. Alguns exemplos podem ser vistos em Gotway e Stroup [34] e Lin e Clayton [44]. Para a estimação dos parâmetros destes modelos é proposto o uso de equações de quasi verosimilhança. Esta abordagem parece atraente do ponto de vista prático, mas pode apresentar falhas em termos probabilísticos. A razão para que essas falhas aconteçam é que as funções de covariâncias dos campos aleatórios binários precisam satisfazer várias propriedades adicionais além da propriedade da simetria e de ser definida positiva.

Os autores acima usaram essencialmente metodologia de estimação assente na máxima verosimilhança, mas a natureza hierárquica destes modelos é melhor ajustada no paradigma Bayesiano, como por exemplo, Riveira *et al.* [72] que utilizaram os modelos hierárquicos Bayesianos na análise da distribuição de quatro espécies de árvores, com dados do Inventário Florestal Nacional Espanhol desenvolvido na Galiza, especificamente a presença/ausência das espécies em coordenadas espaciais especificas, utilizando a abordagem INLA-SPDE; Steinbuch *et al.* [83], no mapeamento de propriedade binária do solo na Flevolandia - Holanda através de um modelo geoestatístico linear generalizado Bayesiano usando o algoritmo MCMC; Banerjee e Gelfand [7], na previsão, interpolação e regressão para dados espacialmente desalinhados, com aplicação no estudo da espécie isópode (um tipo de crustáceo), através de dados ecológicos de uma bacia hidrográfica no oeste do deserto de Negev, Israel; Moraga [58], na previsão de prevalência de malária na Gâmbia, usando a abordagem INLA-SPDE, que é o que descreveremos neste trabalho e usaremos.

Este capítulo encontra-se dividido em 8 secções. Na secção 2 apresenta-se a sustentação teórica dos modelos relacionados a dados geoestatísticos binários, com principal

destaque para os modelos lineares generalizados. Na secção 3, apresenta-se os modelos lineares generalizados mistos. Na secção 4, apresenta-se os modelos Hierárquicos, onde se faz a descrição dos modelos hierárquicos Bayesianos, modelo geoestatístico, modelos gaussianos latentes e modelos Hierárquicos Bayesianos para dados Binários. Na secção 5, apresenta-se a descrição dos métodos utilizados neste trabalho, as abordagens Bayesianas INLA e SPDE. Na secção 6, apresenta-se os resultados da aplicação dos modelos sobre as presumíveis infrações pesqueiras ao largo de Portugal, obtidos em ações de fiscalização. Na secção 7, os resultados da aplicação dos modelos sobre as presumíveis infrações pesqueiras no comando de zona do Sul, obtidos em ações de fiscalização são apresentados, e por fim a secção 8 apresenta as conclusões e recomendações para trabalhos futuros.

3.2 Modelos Lineares Generalizados (MLG) para dados binários

De acordo com Assis *et al.* [5], Modelos Lineares Generalizados (MLG) são uma classe de modelos que foram apresentados pela primeira vez por Nelder e Wedderburn [61], com objetivo de estudar a regressão linear nos casos em que a variável dependente não é necessariamente normalmente distribuída, mas a sua distribuição ainda pertence à família exponencial. Dada uma amostra de n observações de uma variável resposta estes modelos (MLG) apresentam 3 componentes tal como se segue [61]:

i) componente aleatória - as variáveis respostas $Y_1, ..., Y_n$ são independentes e seguem uma distribuição que pertence à família exponencial na forma canónica,

$$f(y_i;\theta,\phi) = exp\left\{\frac{1}{a_i(\phi)}[y_i\theta_i - b(\theta_i)] + c(y_i,\phi)\right\} \tag{3.1}$$

onde $a(.)$, $b(.)$ e $c(.)$ são funções conhecidas, θ_i o parâmetro natural ou canónico e, em geral, $a_i(\phi) = \frac{\phi}{w_i}$, com w_i o peso a priori e $\phi > 0$, conhecido como o parâmetro de dispersão ou escala.

ii) componente sistemática - as variáveis explicativas $\boldsymbol{x}_i' = (x_{i1}, ..., x_{ip})$, $i = 1, \ldots, n$ que se conjugam num preditor linear η

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta},$$

vetor de dimensão $n \times 1$, $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)'$, $p < n$, é um vetor de p parâmetros desconhecidos a serem estimados e $\boldsymbol{X}$ a matriz dos valores das covariáveis de dimensões $n \times p$.

iii) função de ligação - faz a ligação entre a componente aleatória e a componente sistemática por meio de uma função conhecida $g(.)$, monótona e diferenciável, que liga a média μ_i em (i) ao preditor linear em (ii),

$$g(\mu_i) = \eta_i = \boldsymbol{x}_i'\boldsymbol{\beta}. \tag{3.2}$$

A seguir, apresentam-se modelos de regressão para respostas binárias como elementos dos MLG de acordo com Agresti [1].

3.2.1 Modelo de Probabilidade Linear

Para um modelo de regressão linear clássico, $\mu = E[Y]$ é uma função linear das covariáveis $\boldsymbol{x}$. Se tivermos uma variável resposta binária, um modelo análogo é

$$\pi_i = E[Y_i] = \beta_0 + \boldsymbol{x}_i'\boldsymbol{\beta} \tag{3.3}$$

que é chamado de modelo de probabilidade linear, porque a probabilidade de sucesso muda linearmente em $\boldsymbol{x}$, onde β_j representa a mudança na probabilidade por cada valor atribuído a $\boldsymbol{x}_i$, o π_i varia em função do $\boldsymbol{x}_i$, e o π foi substituído por π_i para descrever a dependência desse valor. Este é um modelo linear generalizado com componente aleatória binomial e função de ligação identidade.

Neste tipo de modelo, quando se pretende fazer a estimação utilizando métodos de regressão linear, acontece deparar-se com algumas inconsistências, tais como a possibilidade de se obter estimativas de probabilidade fora do intervalo definido entre zero e um [16, 37].

3.2.2 Modelo de Regressão Logística

A regressão logística é um modelo padrão para dados binários (por exemplo mortes); Se $(y_1,...,y_n)$ representa dados para unidades individuais (por exemplo, uma pessoa sofre ou não a morte), então a variável aleatória Y_i seguirá uma distribuição de Bernoulli, que só pode assumir valores 0 ou 1. Alternativamente, se $(y_1,...,y_n)$ representa contagem de eventos em um estudo ao longo de um número especificado de tentativas para n grupos, então Y_i seguirá uma distribuição Binomial, que pode assumir valores $0, 1, 2, ..., n_i$ onde n_i representa o número de indivíduos por unidade para i-ésimo grupo. Kleinbaum *et al.* [42] defendem que a variável binária não pode ser utilizada como uma variável resposta ou dependente numa análise de regressão linear, visto que não atende o pressuposto de normalidade, mas pertencendo à família exponencial de distribuições encaixa na moldura dos MLG.

O parâmetro de interesse é $\pi_i = P(Y_i = 1|\boldsymbol{x}_i)$, onde $\boldsymbol{x}_i$ é vetor dos preditores para o i-ésimo indivíduo ou grupo. A função de ligação é a função *logit*, definida por:

$$logit(\pi_i) = log\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{x}_i'\boldsymbol{\beta} \tag{3.4}$$

de modo que

$$\pi_i = logit^{-1}(\boldsymbol{x}_i'\boldsymbol{\beta}) = \frac{exp(\boldsymbol{x}_i'\boldsymbol{\beta})}{(1+exp(\boldsymbol{x}_i'\boldsymbol{\beta}))} \tag{3.5}$$

Ou seja, a função *antilogit* ($logit^{-1}$) transforma os valores contínuos obtidos de $\boldsymbol{x}_i'\boldsymbol{\beta}$ de volta em probabilidades, a escala para o parâmetro π_i. Em termos práticos, a relação não

linear entre $\pi(a)$ e a é monótona, onde o $\pi(a)$ aumenta continuamente à medida que o a aumenta. O modelo de regressão logística, também incluindo uma ordenada na origem (intercepto) β_0, pode ser escrito da seguinte forma:

$$y_i \sim Binomial(n_i, \pi_i)$$

$$logit(\pi_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} = \eta_i \tag{3.6}$$

3.2.3 Modelo de Regressão *Probit*

A regressão *probit*, também chamada de modelo *probit*, é usada para modelar variáveis binárias. No modelo *probit*, a função distribuição normal reduzida inversa da probabilidade π_i é modelada como uma combinação linear dos preditores. Frequentemente, os dados binários resultam duma relação não linear entre $\pi(x)$ e x. E uma mudança fixa em x, pode ter menos impacto quando $\pi(x)$ está próximo de 0 ou 1 do que quando $\pi(x)$ está próximo do meio do seu alcance; assim este método de regressão é conhecido como o modelo que possui a curva da função de ligação em forma de S, função de ligação *probit*, que relaciona a probabilidade na função distribuição de uma variável normal reduzida calculada no preditor linear:

$$probit(\pi_i) = \Phi^{-1}(\pi_i) = \beta_0 + \boldsymbol{x}_i'\boldsymbol{\beta}, \tag{3.7}$$

onde ϕ é a função distribuição inversa da normal reduzida.

Os resultados produzidos por este modelo são análogos aos resultados obtidos pelo modelo de regressão logística (*logit*), exceto para valores de π próximos de 0 ou 1. Contudo, os modelos de regressão logística têm sido amplamente aplicados devido a sua simplicidade na interpretação dos parâmetros.

3.3 Modelos Lineares Generalizados Mistos (MLGM)

Como já referido anteriormente na seção 3.2, os MLG estendem a regressão comum, permitindo que a variável resposta possa seguir uma distribuição não normal, ou seja, a variável resposta pertence a família exponencial, e que possa ter uma função de ligação entre a média e o preditor que não seja identidade. Os **modelos lineares generalizados mistos** (MLGM), do inglês *Gaussian Linear Mixed Models (GLMM)* são uma extensão natural dos MLG, que para além dos efeitos fixos, permitem também a inclusão dos efeitos aleatórios no preditor linear.

Os MLGM têm como objetivo descrever as alterações da resposta média de cada indivíduo e suas relações com as covariáveis de interesse [55, 84, 2].

Duma forma geral, pode-se considerar os MLGM da seguinte forma:

Considere-se que $\mathbf{u}$ é um vetor de efeitos aleatórios. Assume-se que as variáveis respostas em $\boldsymbol{Y}$ são condicionalmente independentes dado $\boldsymbol{u}$, que a distribuição de $\boldsymbol{Y}$ dado $\boldsymbol{u}$ pertence à família exponencial, descrita pela equação (3.1), e que se escolhe uma função de ligação dada por

$$g[E(Y_i|\boldsymbol{u})] = \boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{z}_i'\boldsymbol{u} \qquad i = 1..., n \tag{3.8}$$

onde $\boldsymbol{z}_i'$ e $\boldsymbol{x}_i'$ são vetores linha de valores conhecidos de variáveis explicativas e $\boldsymbol{\beta}$ é o vetor de efeitos fixos.

A função densidade (ou probabilidade) condicional $f(\boldsymbol{y}|\boldsymbol{u},\boldsymbol{\beta},\phi)$ para $\boldsymbol{Y}$ é dada por

$$f(\boldsymbol{y}|\boldsymbol{u},\boldsymbol{\beta},\phi) = \prod_{i=1}^{n} f(y_i|\boldsymbol{u},\boldsymbol{\beta},\phi) \tag{3.9}$$

Assume-se que a média $E[y_i|\boldsymbol{u}] = \mu_i$ depende dos efeitos fixos e dos efeitos aleatórios através da função de ligação dada em (3.8), e a variância é dada por $V(y_i|\boldsymbol{u}) = \phi\boldsymbol{v}(\mu_i)$, onde ϕ é o parâmetro de dispersão e $\boldsymbol{v}(.)$ é uma função conhecida da média condicional. Também assume-se que os efeitos aleatórios $\boldsymbol{u}$ seguem uma distribuição normal multivariada, com média $\mathbf{0}$ e a matriz de variância e covariância $\boldsymbol{\Omega_\mu}$ com forma particular imposta de acordo com o contexto prático, ou seja $\boldsymbol{u} \sim \mathcal{N}(\mathbf{0},\boldsymbol{\Omega_\mu})$. Os MLGM também podem ser modelados sem as covariáveis. Mais detalhes em [13, 30].

No contexto da modelação geoestatística, Diggle e Ribeiro [31] introduziram uma extensão de modelos lineares generalizados para acomodar respostas dependentes e introduzir efeitos aleatórios não observáveis no preditor linear para poderem capturar a dependência espacial. Desta forma, o seu modelo é um MLGM em que o preditor linear η_i é descrito por:

$$\eta_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{u} \qquad i = 1, \ldots, n \tag{3.10}$$

onde $\boldsymbol{x}_i'$ é um vetor de variáveis explicativas associadas com a resposta Y_i, georreferenciada, $\boldsymbol{\beta}$ é um vetor de parâmetros desconhecidos, $\boldsymbol{u}$ são efeitos aleatórios ou variável latente. Neste modelo $\boldsymbol{u} = (u_1, ..., u_n)$ segue uma distribuição multivariada com média $\mathbf{0}$ e com uma estrutura de covariância descrevendo a estrutura espacial do problema.

De Oliveira [26] situa nas classe dos modelos lineares generalizados mistos o modelo geoestatístico para dados binários, da seguinte forma:

Para qualquer conjunto de locais distintos $s_1, ..., s_n \in D$, observam-se as variáveis dicotómicas $Y(s_1), ..., Y(s_n)$, condicionalmente independentes, dado $\boldsymbol{u} = (u_1(s_1), ..., u_n(s_n))$, e

$$Y(\boldsymbol{s_i})|\boldsymbol{u} \sim Bernoulli(g^{-1}(p_i)), \;\; i = 1, ..., n \tag{3.11}$$

onde $p_i = \boldsymbol{\beta'x} + \boldsymbol{u}$, $g : (0,1) \to \mathbb{R}$ é uma função de ligação conhecida, assumida como estritamente crescente, as covariáveis $\boldsymbol{x}$ podem ser espacialmente dependentes, $\boldsymbol{u}$ é um campo aleatório gaussiano com média 0 e função covariância $\sigma^2\rho_\theta(\boldsymbol{s},\boldsymbol{s'})$, onde $\sigma^2 > 0$ é a

variância e $\rho_\theta(\boldsymbol{s},\boldsymbol{s'})$ é uma função de correlação em $\mathbb{R}^d$ (neste trabalho vai-se considerar $d = 2$), $\boldsymbol{u}$ representa fontes não observadas de variação espacial que afetam a probabilidade espacial da variável (risco relativo) do evento de interesse e $\rho_\theta(\boldsymbol{s},\boldsymbol{s'})$ é frequentemente (mas nem sempre) assumida como sendo contínua em todo os lugares. Este é um modelo linear generalizado misto espacial, em que os efeitos aleatórios são espacialmente estruturados, onde a sua especificação requer a escolha da função de ligação (mais comum) $g(.)$ e a função de correlação $\rho_\theta(\boldsymbol{s},\boldsymbol{s'})$.

3.4 Modelos Hierárquicos

Segundo Guo e Zao [35], os conceitos e a metodologia para análise de dados multiníveis (hierárquicos) foi desenvolvida primeiramente por Mason *et al.* [50]. Nesta secção, vai-se começar por definir o que é uma estrutura de dados hierárquica, para de seguida se falar de modelação hierárquica.

Uma **base de dados hierárquica** é um tipo de estrutura de dados que liga os seus registos (coleção de atributos ou campos que descrevem os dados) numa forma de árvore através de ligações de modo que cada tipo de registo tenha apenas uma identificação, uma ligação é uma associação entre dois registos [24].

Modelação hierárquica, também chamada de modelação estatística hierárquica, é uma forma de modelar as incertezas por níveis bem definidos de probabilidades condicionais [23], para dados possivelmente dependentes. Assis *et al.* [5], dizem que **modelos hierárquicos**, que também são chamados modelos multiníveis, modelos de efeitos aleatórios, modelos de momentos sequenciais transverso ou modelos mistos, são modelos de regressão para uma estrutura de dados hierárquicos em que o efeito aleatório para um modelo de dois níveis é a fonte de erro atribuída às unidades do segundo nível, modelando corretamente erros correlacionados.

A classe de modelos lineares generalizados mistos está contida de forma natural nesta classe abrangente dos modelos hierárquicos.

A utilização cada vez mais usual de dados observacionais de múltiplas fontes, associados a sistemas complexos requer o uso de modelos multi-nível, mais realísticos, mas com um grau de complexidade computacional grande, que só pode ser ultrapassado no âmbito da inferência Bayesiana, com os seus métodos de simulação e aproximação numéricas. O paradigma Bayesiano, o avanço tecnológico e o desenvolvimento destes métodos computacionais permitiram, desde o início do século XXI, a escolha de praticamente qualquer modelo envolvendo múltiplos níveis e incorporando efeitos aleatórios ou estruturas de dependência complicadas [8].

3.4.1 Modelos Hierárquicos Bayesianos

Os modelos hierárquicos modelam incertezas de diferentes fontes através de distribuições condicionadas. Ao incluir na modelação um modelo paramétrico para expressar a incerteza probabilística nos parâmetros temos um modelo hierárquico bayesiano [23].

Para se proceder à sua descrição, começa-se por fazer uma breve apresentação de alguns conceitos básicos de probabilidade condicional com base em [23, 60].

Consideremos a seguinte notação em que $[A]$ representa uma função densidade de A (no sentido lato, incluindo também distribuições discretas) e $[A|B]$ representa uma função densidade distribuição condicional, então

$$[A] = \int [A|B][B]dB \tag{3.12}$$

pela lei de probabilidade total, obtém o teorema de Bayes

$$[B|A] = \frac{[A|B][B]}{\int [A|B][B]dB} = \frac{[A|B][B]}{[A]} \tag{3.13}$$

onde $\int g(B)[B]dB$ é o valor esperado (no caso discreto será dado pela soma) de alguma função $g(B)$ de B.

Os modelos hierárquicos que nos interessam têm num primeiro nível o modelo dos dados, determinado pela distribuição dos dados condicionada num processo (verdadeiro) latente não observado, descrito pelos efeitos aleatórios. No segundo nível tem-se então o modelo do processo. Ambos os modelos são condicionados no modelo dos parâmetros. Sejam então as seguintes quantidades de interesse Y, u e θ no caso de modelos hierárquicos. Considere $\boldsymbol{Y}$ como dados, $\boldsymbol{u}$ como um processo (não observado), que pretendemos fazer a predição e $\boldsymbol{\theta}$ como os parâmetros desconhecidos.
Apresenta-se a descrição dos modelos hierárquicos de acordo com Cressie [23]. A sua representação básica obtém-se dividindo o modelo em 3 níveis, o *modelo de dados* que é dado por $[\boldsymbol{Y}|\boldsymbol{u},\boldsymbol{\theta}_1]$, o *modelo do processo* dado por $[\boldsymbol{u}|\boldsymbol{\theta}_2]$ e o *modelo de parâmetros* dado por $[\boldsymbol{\theta}]$.

Observe que $\boldsymbol{\theta}_1$ representa os parâmetros do modelo de dados e $\boldsymbol{\theta}_2$ representa parâmetros do modelo do processo. Então $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, e o modelo de parâmetros é $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$.

A distribuição conjunta de $\boldsymbol{Y}, \boldsymbol{u}$ *e* $\boldsymbol{\theta}$ pode ser escrita da seguinte forma

$$[\boldsymbol{Y}, \boldsymbol{u}, \boldsymbol{\theta}] = [\boldsymbol{Y}, \boldsymbol{u}|\boldsymbol{\theta}][\boldsymbol{\theta}] = [\boldsymbol{Y}|\boldsymbol{u}, \boldsymbol{\theta}][\boldsymbol{u}|\boldsymbol{\theta}][\boldsymbol{\theta}] \tag{3.14}$$

Olhando para a equação (3.14), verifica-se que temos um simples produto dos 3 modelos (modelo dos dados, modelo do processo e modelo do parâmetro).

A distribuição de $\boldsymbol{u}$ *e* $\boldsymbol{\theta}$, condicional nos dados $\boldsymbol{Y}$, obtida pelo teorema de Bayes em

(3.15), é a sua distribuição a posteriori

$$\begin{aligned}[\boldsymbol{u},\boldsymbol{\theta}|\boldsymbol{Y}] &= \frac{[\boldsymbol{Y}|\boldsymbol{u},\boldsymbol{\theta}][\boldsymbol{u}|\boldsymbol{\theta}]}{\int\int[\boldsymbol{Y}|\boldsymbol{u},\boldsymbol{\theta}][\boldsymbol{u}|\boldsymbol{\theta}]d\boldsymbol{u}\,d\boldsymbol{\theta}} \\ &= \frac{[\boldsymbol{Y}|\boldsymbol{u},\boldsymbol{\theta}][\boldsymbol{u}|\boldsymbol{\theta}][\boldsymbol{\theta}]}{\int\int[\boldsymbol{Y}|\boldsymbol{u},\boldsymbol{\theta}][\boldsymbol{u}|\boldsymbol{\theta}][\boldsymbol{\theta}]d\boldsymbol{u}\,d\boldsymbol{\theta}} \\ &= \frac{[\boldsymbol{Y}|\boldsymbol{u},\boldsymbol{\theta}][\boldsymbol{u}|\boldsymbol{\theta}][\boldsymbol{\theta}]}{[\boldsymbol{Y}]}\end{aligned} \tag{3.15}$$

Toda a inferência Bayesiana sobre $\boldsymbol{u}$ e $\boldsymbol{\theta}$ no modelo Hierárquico Bayesiano é feita com base nesta distribuição ou seja, depende desta distribuição.

De acordo com Cressie [23], existe uma abordagem alternativa, conhecida por Modelo Hierárquico Empírico (também chamado de Modelo Bayesiano Empírico), que considera o modelo dos dados e do processo dependente dos parâmetros $\boldsymbol{\theta}$, fixos mas desconhecidos. Assim, $\boldsymbol{\theta}$ é estimado separadamente, apenas com base apenas nos dados Y ou noutro estudo independente, e então a inferência é feita com base na distribuição de $[\boldsymbol{u}|\boldsymbol{Y},\boldsymbol{\theta}]$, em que $\boldsymbol{\theta}$ é substituído pelas referidas estimativas. Mais detalhes podem ser vistos em [23, 11, 33].

3.4.2 Modelo geoestatístico

O modelo geoestatística é um modelo para dados georreferenciados, portanto, vamos considerar as variáveis de interesse nas localizações s_i onde são observadas, com a seguinte notação $Y_i = Y(s_i)$, então o modelo é dado por

$$Y_i = \mu + U(s_i) + \epsilon_i = \mu + u_i + \epsilon_i, \qquad s_i \in D,\ i = 1,\ldots,n, \tag{3.16}$$

assume-se que existe um fenómeno de interesse $\left\{U(s) : s \in D \subseteq \mathbb{R}^2\right\}$ modelado por um processo estocástico contínuo espacial que é imperfeitamente observado em alguns locais $\boldsymbol{s} = (s_1,...,s_n)$ definidos de acordo com um esquema de amostragem, resultando em observações $\boldsymbol{Y} = (Y_1,...,Y_n)$, $Y_i \in \mathbb{R}$, $i = 1,...,n$. Embora a complexidade da superfície verdadeira possa variar, a abordagem mais simples e comum é modelá-la como um processo Gaussiano estacionário com média constante e função de correlação dependendo apenas da distância entre as localizações $\rho(\gamma)$. Existem várias famílias de funções de correlação que se podem usar, tais como, exponencial, esférica, gaussiana, Matérn, [31] sugeridas pelos dados.

No modelo geoestatístico dado em 3.16, $\mu \in \mathbb{R}$ é um parâmetro de média constante, U é um processo Gaussiano, com média zero, variância σ^2 e função de covariância de Matérn dependente de (ϕ, ν, σ^2), ϵ_i são erros aleatórios normais i.i.d. com $E[\epsilon_i] = 0$ e $V(\epsilon_i) = \sigma_\epsilon^2$ é a variância do efeito pepita. A inferência neste modelo é baseada na função de verosimilhança de $\boldsymbol{\theta} = (\mu, \sigma_\epsilon^2, \phi, \nu, \sigma^2)$:

$$L(\boldsymbol{\theta}) = [\boldsymbol{s},\boldsymbol{Y}] = \int [\boldsymbol{s},\boldsymbol{Y},\boldsymbol{u}]d\boldsymbol{u} = \int [\boldsymbol{Y}|\boldsymbol{u},\boldsymbol{s}][\boldsymbol{s}|\boldsymbol{u}][\boldsymbol{u}]d\boldsymbol{u}, \tag{3.17}$$

Para geoestatística padrão, o processo de seleção dos locais de medição é independente da resposta, $[\boldsymbol{s}|U] = [\boldsymbol{s}]$ muitas vezes determinístico, e consequentemente a função de verosimilhança é dada por

$$L(\boldsymbol{\theta}) = [\boldsymbol{s}, \boldsymbol{Y}] = [\boldsymbol{s}] \int [\boldsymbol{Y}|\boldsymbol{u}, \boldsymbol{s}][\boldsymbol{u}] d\boldsymbol{u} = [\boldsymbol{s}][\boldsymbol{Y}|\boldsymbol{s}]. \tag{3.18}$$

Numa perspetiva Bayesiana, o modelo geoestatístico de Diggle e Ribeiro [31] pode ter então a sua distribuição conjunta especificada de forma hierárquica por três níveis:

$$[Y, \boldsymbol{u}, \boldsymbol{\theta}] = [\boldsymbol{\theta}][\boldsymbol{u}|\boldsymbol{\theta}][Y|\boldsymbol{u}, \boldsymbol{\theta}], \tag{3.19}$$

onde:
Modelo dos dados: $[Y|\boldsymbol{u}, \boldsymbol{\theta}] : Y(s_i) = \mu + u(s_i) + \epsilon_i$.
Modelo do processo ou do sinal: $\boldsymbol{u}|\boldsymbol{\theta}$ é um campo aleatório Gaussiano com uma função de covariância específica $\rho(\gamma)$.
$[\boldsymbol{\theta}]$ são as distribuições a priori para o $\boldsymbol{\theta}$.

3.4.2.1 Função Covariância de Matérn

Para Diggle e Ribeiro [31] a função de covariância de Matérn é uma das mais utilizada para descrever a dependência espacial dos dados geoestatísticos. Esta função é considerada um modelo flexível para as dependências encontradas em observações do mundo real e é descrita da seguinte forma:

$$C_M\left(u_i, u_j\right) = \sigma_u^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(k \parallel s_i - s_j \parallel\right)^{\nu} K_{\nu}\left(k \parallel s_i - s_j \parallel\right), \tag{3.20}$$

onde $\|s_i - s_j\|$ é a distância Euclidiana entre duas localizações genéricas $s_i, s_j \in \mathbb{R}^d$, σ_u^2 é a variância marginal do processo, ν ($\nu > 0$) é o parâmetro de suavidade que se fixa em 1 [43], K_ν é a função de Bessel modificada do 2º tipo e de ordem ν, k é um parâmetro de escala dado por $k = \frac{\sqrt{8\nu}}{R}$, onde R é o alcance [12, 45, 43].

A variância marginal é dada por,

$$\sigma_u^2 = \frac{\Gamma(\nu)}{\Gamma(\alpha)(4\pi)^{\frac{d}{2}} k^{2\nu} \tau^2} \tag{3.21}$$

onde $\nu = \alpha - \dfrac{d}{2}$ e τ^2 é o parâmetro que controla a variância. Por defeito no R-INLA, o parâmetro de suavidade é $\alpha = 2$ (correspondente a $\nu = 1$), mas os valores $0 \leq \alpha < 2$ estão também disponíveis, mesmo que os valores não inteiros de α não tenham sido totalmente testados, o método de aproximação está descrito em [45].

3.4.3 Modelos Gaussianos Latentes(MGL)

Para *Rue at al.* [73] , Blangiardo e Cameletti [12] os modelos gaussianos latentes são modelos em que a variável resposta de interesse (ou variável dependente) Y_i, é assumida como uma variável que pertence a uma família de distribuições, esta família não tem obrigatoriedade de ser a família exponencial, e em que duma forma geral, alguns parâmetros μ da família de distribuições (geralmente é a média $E[Y_i]$ relacionam-se com um preditor aditivo estruturado η_i através de uma função de ligação $g(.)$ para que $g(\mu) = \eta_i$. O preditor η_i acomoda os efeitos de várias covariáveis de uma forma aditiva, sendo dado por

$$\eta_i = \beta_0 + \sum_{j=1}^{n_f} f_j(z_{ij}) + \sum_{k=1}^{n_B} \beta_k x_{ki} \qquad i = 1,\ldots,n \tag{3.22}$$

Onde β_0 é um escalar que representa o intercepto, $f_j(.)$ são funções desconhecidas das covariáveis $\boldsymbol{z}$ e β_k são os efeitos lineares das covariáveis $\boldsymbol{x}$. A escolha da distribuição Gaussiana para as quantidades não observáveis (latentes) $\boldsymbol{w} = (\beta_0, \boldsymbol{f}, \boldsymbol{\beta})$ completa a formulação destes modelos.

Note-se a estrutura hierárquica do modelo em (3.22), identifica-se como modelo de dados $[Y|\boldsymbol{w}, \boldsymbol{\theta}_1]$, modelo do processo $[\boldsymbol{w}|\boldsymbol{\theta}_2]$ e modelo dos parâmetros $[\boldsymbol{\theta}]$, onde $\boldsymbol{\theta} = (\boldsymbol{\theta_1}, \boldsymbol{\theta_2})$. Assume-se que cada ponto observado $y = \{y_1, ..., y_n\}$ tem independência condicional, dado um campo latente $\boldsymbol{w} = (\beta_0, \boldsymbol{f}, \boldsymbol{\beta})$ e hiperparâmetros $\boldsymbol{\theta}_1$, (Lindgen e Rue [45]).

Modelo dos dados: $\boldsymbol{y}|\boldsymbol{w}, \boldsymbol{\theta}_1 \sim \pi(\mathbf{y}|\boldsymbol{w}, \boldsymbol{\theta}_1) = \prod_{i=1}^{n} \pi(y_i|\boldsymbol{w}, \boldsymbol{\theta}_1)$,

Modelo do processo: $\boldsymbol{w}|\boldsymbol{\theta_2} \sim \pi(\boldsymbol{w}|\boldsymbol{\theta}_2) = \mathcal{N}(\boldsymbol{\mu(\theta_2)}, \boldsymbol{Q^{-1}(\theta_2)})$,

onde $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{Q^{-1}})$ é uma distribuição Gaussiana multivariada com média o vetor $\boldsymbol{\mu}$ e a matriz de precisão $\boldsymbol{Q}$.

Modelo dos parâmetros: $\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta})$.

3.4.4 Modelo Hierárquico Bayesiano para dados Binários

O modelo hierárquico Bayesiano para dados binários é um caso dos modelos Gaussianos latentes em que o modelo dos dados $[Y|\boldsymbol{w}, \boldsymbol{\theta}_1]$ segue uma distribuição de Bernoulli e a função de ligação é dada por

$$logit(\pi_i) = log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{u}, \qquad i = 1,\ldots,n \tag{3.23}$$

o modelo do processo é $[\boldsymbol{w}|\boldsymbol{\theta}_2]$, onde as variáveis latentes $\boldsymbol{w} = (\beta_0, \boldsymbol{u}, \boldsymbol{\beta})$ incluem um efeito aleatório espacial ou espaço-temporal $\boldsymbol{u}$ e o modelo dos parâmetros $[\boldsymbol{\theta}]$ descreve todas as distribuições a priori para os dois modelos anteriores, onde $\boldsymbol{\theta} = (\boldsymbol{\theta_1}, \boldsymbol{\theta_2})$. Neste modelo (3.23), interessa-nos que o $\boldsymbol{u}$ tenha uma estrutura espacial descrita pela função de covariância de Matérn dada pela equação (3.20).

Vários estudos têm sido desenvolvidos utilizando modelos hierárquicos para dados binários. Por exemplo: Diggle e Giorgi [29] utilizaram estes modelos no contexto clássico

para mapeamento de prevalência em estudos com poucos pontos amostrais (configurações de poucos recursos) no distrito de Chikwawa, na pesquisa dos indicadores de malária e no mapeamento da prevalência da oncocercose também conhecida por cegueira do rio em Moçambique, Malawi e Tanzânia, utilizando o método de máxima verosimilhança com recurso a Monte Carlo; ainda Giorgi e Diggle [32] apresentaram *PrevMap* como um pacote do R para mapeamento de prevalência utilizando modelos hierárquicos aplicados aos dados de um tipo de vermes chamados Loa loa dos Camarões e Nigéria, utilizando o método de máxima verosimilhança; E no contexto Bayesiano, Riveira *et al.* [72] utilizaram os modelos hierárquicos Bayesianos na análise da distribuição de quatro espécies de árvores, com dados do Inventário Florestal Nacional Espanhol desenvolvido na Galiza, especificamente a presença/ausência das espécies em coordenadas espaciais especificas, utilizando a abordagem de estimação INLA-SPDE descrita mais à frente neste trabalho; Steinbuch *et al.* [83], no mapeamento de propriedade binária do solo na Flevolandia - Holanda através de um modelo geoestatístico linear generalizado Bayesiano usando o algoritmo de estimação MCMC; Banerjee e Gelfand [7], na previsão, interpolação e regressão para dados espacialmente desalinhados, com aplicação no estudo da espécie isópode (um tipo de crustáceo), através de dados ecológicos de uma bacia hidrográfica no oeste do deserto de Negev, Israel; Moraga [58], na previsão de prevalência de malária na Gâmbia, usando a abordagem INLA-SPDE.

Para se ajustar os modelos dos dados binários no contexto dos modelos hierárquicos bayesianos, pela complexidade das distribuições de probabilidade envolvidas e do consequente processo de estimação, é muito aconselhável utilizar a metodologia *Integrated Nested Laplace Approximation (INLA)* e *Stochastic Partial Differential Equation (SPDE)*. A seguir vai-se apresentar uma breve descrição do INLA e SPDE.

3.5 INLA e SPDE

Aproximação de Laplace Aninhada Integrada (INLA) do Inglês *Integrated Nested Laplace Approximation*, é um procedimento de aproximação determinístico para inferência Bayesiana, proposto por Rue *et al.* [73], que foi desenvolvido especialmente para ser aplicado à classe de modelos gaussianos latentes (MGL). Tais modelos englobam uma gama de modelos de regressão com estrutura aditiva (mais detalhes em [73, 12, 92].
A principal diferença desta metodologia com relação às abordagens Bayesianas tradicionais é a de que esta metodologia não necessita de simulações estocásticas das distribuições marginais a posteriori como os métodos de Monte Carlo via Cadeias de Markov (MCMC). Esta abordagem substitui as simulações por aproximações determinísticas, precisas e, sobretudo, de cálculos computacionais rápidos na classe de modelos gaussianos latentes.
A qualidade do ajuste com esta abordagem é extremamente alta, tal que mesmo para, muitos casos, longas simulações de MCMC não apresentam diferenças significativas dos resultados do INLA. A principal vantagem é a sua eficiência computacional que é muito superior quando comparado com os métodos tradicionais, mesmo para campos latentes

alta dimensão. Outra vantagem é de que, devido a natureza de aproximação analítica, o INLA não sofre dos conhecidos problemas de convergência dos métodos MCMC que geralmente são discutidos.

3.5.1 INLA

A seguir vai se apresentar a descrição da metodologia INLA de acordo com [59, 77]. Considere-se o modelo da equação (3.22), descrito na subsecção 3.4.3.

A classe de modelos Gaussianos é muito flexivel, os termos $f_j(.)$ podem assumir muitas formas diferentes como efeitos não lineares de covariáveis, efeitos sazonais, efeitos temporais ou espaciais aleatórios, abrangendo modelos lineares generalizados mistos, modelos hierárquicos, modelos espaciais e modelos espaço-temporais. O vetor de efeitos latentes gaussianos $\boldsymbol{w} = (\beta_0, \boldsymbol{f}, \boldsymbol{\beta})$ assume-se ser um campo aleatório de gaussiano de Markov (CAGM) com $\boldsymbol{\pi}(\boldsymbol{w}|\boldsymbol{\theta}_1) \equiv \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q}(\boldsymbol{\theta_1}))$, onde $\boldsymbol{Q}(\boldsymbol{\theta}_1)$ é a matriz de precisão. Considere $\boldsymbol{y}$ o vetor de observações com função densidade de probabilidade $\boldsymbol{\pi}(\boldsymbol{y}|\boldsymbol{w}, \boldsymbol{\theta}_2)$ cujos elementos se assumem condicionalmente independentes, dado $\boldsymbol{w}$ e $\boldsymbol{\theta}_2$ e $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ o vetor dos hiperparâmetros com função densidade probabilidade $\boldsymbol{\pi}(\boldsymbol{\theta})$ (não necessariamente gaussiana). Neste contexto, a distribuição a posteriori das quantidades desconhecidas é dada por:

$$\begin{aligned} \pi(\boldsymbol{w}, \boldsymbol{\theta}|\boldsymbol{y}) &\propto \pi(\boldsymbol{\theta}) \times \pi(\boldsymbol{w}|\boldsymbol{\theta}) \prod_{i=1}^{n} \pi(y_i|w_i, \boldsymbol{\theta}) \\ &\propto \pi(\boldsymbol{\theta}) \times \left|\boldsymbol{Q}(\boldsymbol{\theta})\right|^{\frac{1}{2}} exp\left[-\frac{1}{2}\boldsymbol{w}^T \boldsymbol{Q}(\boldsymbol{\theta})\boldsymbol{w} + \sum_{i=1}^{n} log\left[\pi(y_i|w_i, \boldsymbol{\theta})\right]\right] \end{aligned} \tag{3.24}$$

As distribuições marginais de interesse (campo latente) podem ser definidas como se segue:

$$\pi(w_i|\boldsymbol{y}) = \int \pi(w_i|\boldsymbol{\theta}, \boldsymbol{y}) \pi(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta} \tag{3.25}$$

$$\pi(\theta_j|\boldsymbol{y}) = \int \pi(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta}_{-j} \tag{3.26}$$

Construindo agora aproximações "aninhadas"[1] para estas distribuições, temos:

$$\tilde{\pi}(w_i|\boldsymbol{y}) = \int \tilde{\pi}(w_i|\boldsymbol{\theta}, \boldsymbol{y}) \tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta} \tag{3.27}$$

$$\tilde{\pi}(\theta_j|\boldsymbol{y}) = \int \tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta}_{-j} \tag{3.28}$$

onde $\tilde{\pi}$ é uma função densidade (condicional) aproximada.
As aproximações de $\pi(w_i|\boldsymbol{y})$ são obtidas aproximando $\pi(w_i|\boldsymbol{\theta}, \boldsymbol{y})$ e $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ e então usando a integração numérica em $\boldsymbol{\theta}$ (soma finita) para o fazer desaparecer - o que é possível, devido a baixa dimensionalidade de $\boldsymbol{\theta}$. O mesmo aplica-se para $\pi(\theta_j|\boldsymbol{y})$.

[1] Obtidas pelo método de aproximação de Laplace, que permite aproximar a função densidade pelos primeiros termos da expansão em série de Taylor do logaritmo da densidade.

Deste modo, aproximação Laplace proposta para $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ é dada por:

$$\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y}) \propto \frac{\pi(\boldsymbol{w},\boldsymbol{\theta},\boldsymbol{y})}{\tilde{\pi}_G(\boldsymbol{w}|\boldsymbol{\theta},\boldsymbol{y})} \mid_{\boldsymbol{w}=\boldsymbol{w}^*(\boldsymbol{\theta})}, \tag{3.29}$$

onde $\pi(\boldsymbol{w},\boldsymbol{\theta},\boldsymbol{y}) = \pi(\boldsymbol{y}|\boldsymbol{w},\boldsymbol{\theta})\pi(\boldsymbol{w}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$, $\tilde{\pi}_G$ é aproximação de Laplace (Gaussiana) para a distribuição condicional completa de $\boldsymbol{w}$ e $\boldsymbol{w}^*(\boldsymbol{\theta})$ a sua moda, para um dado $\boldsymbol{\theta}$.

Então a aproximação INLA de $\pi(w_i|\boldsymbol{y})$ é feita em três passos:

1. Utilize-se a aproximação Laplace (3.29) para aproximar $\pi(\boldsymbol{\theta}|\boldsymbol{y})$;
2. Utilize-se a aproximação de Laplace de $\pi(w_i|\boldsymbol{y},\boldsymbol{\theta})$ para valores seleccionados de $\boldsymbol{\theta}$ (tomando partido computacional de $\boldsymbol{w}$ ser um campo aleatório gaussiano);
3. Combinação dos dois passos anteriores usando a integração numérica,

$$\tilde{\pi}(w_i|\boldsymbol{y}) = \int \tilde{\pi}(w_i|\boldsymbol{\theta},\boldsymbol{y})\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}$$

pode ser resolvido através de uma soma finita ponderada:

$$\tilde{\pi}(w_i|\boldsymbol{y}) \approx \sum_m \tilde{\pi}(w_i|\boldsymbol{\theta}_m,\boldsymbol{y})\tilde{\pi}(\boldsymbol{\theta}_m|\boldsymbol{y})\boldsymbol{\Delta}_m \tag{3.30}$$

onde $\boldsymbol{\Delta}_m$ são os pesos de quadratura e $\tilde{\pi}(w_i|\boldsymbol{\theta}_m,\boldsymbol{y})$ e $\tilde{\pi}(\boldsymbol{\theta}_m|\boldsymbol{y})$ são as aproximações de $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ e $\pi(w_i|\boldsymbol{\theta},\boldsymbol{y})$ respetivamente, nos pontos $\boldsymbol{\theta}_m$.
A precisão numérica das aproximações acima referidas depende da escolha adequada dos pontos de avaliação $\boldsymbol{\theta}_m$, para tal, é efetuada uma busca na grelha de modo que o maior número de pontos selecionados corresponda à zona com maior massa de probabilidade e se consiga obter um conjunto de pontos $\boldsymbol{\theta}_m$ considerados relevantes com um conjunto correspondente de pesos $\boldsymbol{\Delta}_m$.

A seguir, vai-se apresentar uma breve descrição das SPDE.

3.5.2 SPDE

De acordo com Lindgren *et al.* [45], Blangiardo e Cameletti [12], SPDE é uma abordagem que proporciona uma representação de todo o processo espacial que varia continuamente no domínio D. Consiste em aproximar o campo aleatório de Matérn (ver 3.4.2.1) indexado continuamente, através de um campo aleatório de Markov Gaussiano indexado discretamente, através de uma representação de uma função de base definida numa triangulação do domínio D (método dos elementos finitos).

Segundo Blangiardo e Camaletti [12], Lindgren *et al.* [45], Krainski *et al.* [43], reconhecem a importância da seguinte SPDE:

$$(K^2 - \Delta)^{\frac{\alpha}{2}}(\tau \boldsymbol{u}(s)) = W(s) \tag{3.31}$$

onde $s \in \mathbb{R}^d$, $\Delta = \sum_{i=1}^{d} \frac{\partial^2}{\partial S_i^2}$ é um Laplaciano, $\alpha = \nu + \frac{d}{2}$ controla a suavidade, $K > 0$ é o parâmetro de escala, τ controla a variância, e W(s) é um processo de ruído branco espacial gaussiano. A solução exata e estacionária para essa SPDE é um campo gaussiano estacionário $U(s)$ com uma função de covariância de Matérn dada pela equação (3.20).

A solução de uma SPDE representada pelo campo gaussiano $U(s)$ de Matérn estacionário e isotrópico que pode ser aproximada pelo métodos dos elementos finitos em uma triangulação [2] do domínio D da seguinte forma:

$$\boldsymbol{u}(s) = \sum_{g=1}^{G} \varphi_g(s)\tilde{\xi}_g \tag{3.32}$$

onde G é o número total de triângulos, φ_g é conjunto de funções de base (determinísticas), e $\tilde{\xi}_g$ são os pesos com distribuição gaussiana com média zero. Para se obter uma estrutura de Markov, as funções de base são escolhidas para terem um suporte local e serem lineares por partes em cada triângulo, ou seja, φ_g é 1 no vértice g e é 0 em todos outros vértices. Usando condições de fronteira de Neumann, segue que (para o caso em que $\alpha = 2$) a matriz de precisão $\boldsymbol{Q}$ para o vetor de pesos gaussiano $\tilde{\boldsymbol{\xi}} = \left(\tilde{\xi}_1, ..., \tilde{\xi}_G\right)$ é dado por

$$\boldsymbol{Q} = \tau^2(k^4\boldsymbol{C} + 2k^2\boldsymbol{G} + \boldsymbol{G}\boldsymbol{C}^{-1}\boldsymbol{G}), \tag{3.33}$$

onde o elemento genérico da matriz diagonal $\boldsymbol{C}$ é $C_{ii} = \int \varphi_i(s)ds$ e o da matriz esparsa $\boldsymbol{G}$ é $G_{ij} = \int \nabla\varphi_i(s)\nabla\varphi_j(s)ds$ (∇ é o gradiente). A matriz de precisão $\boldsymbol{Q}$, cujos elementos dependem de τ e k, é esparsa e $\boldsymbol{u}$ é um campo aleatório de Markov Gaussiano com distribuição Normal $(\boldsymbol{0}, \boldsymbol{Q}^{-1})$ e representa a solução aproximada para uma SPDE (em sentido estocasticamente fraco). Mais detalhes ver [45, 43, 12]).

A abordagem SPDE é baseada numa representação finita para definir o campo aleatório de Matérn como uma combinação linear de funções de base definidas em triangulação no domínio D. Esta triangulação consiste em dividir o domínio D (região de estudo) numa malha, ou seja, em conjuntos de triângulos que não se intercetam, unidos em pelo menos uma aresta ou vértice. Primeiro, os vértices dos triângulos iniciais são colocados nas localizações $s_i, \ldots, s_n$, e de seguida, os vértices adicionais são adicionados para obter uma triangulação útil para a previsão espacial desejada. A altura de cada triângulo (o valor do campo espacial em cada vértice do triângulo) é dado pelo peso $\tilde{\xi}_g$ e os valores no interior do triângulo são determinados por interpolação.

O efeito geral da triangulação é que dependendo do que se deseja, pode-se ter triângulos menores, portanto, representam maior precisão do campo, onde os locais de observação são densos, triângulos maiores, onde os dados são mais esparsos, portanto, fornecem informações menos detalhadas, e grandes triângulos onde não há dados é gastar recursos computacionais, será um desperdício.

[2] Triangulação consiste em subdividir o domínio espacial em um conjunto de triângulos sem interseção, onde quaisquer dois triângulos se encontram no máximo em uma aresta ou canto comum.

Nas triangulações deve-se ter muita atenção à fronteira da região; se a fronteira for muito irregular, então muitos triângulos serão construídos próximos da fronteira para conseguir definir bem esta área. Para se resolver este problema, pode-se "simplificar"a fronteira, isto é, definir de forma não muito detalhada, ou ainda, fazer uma expansão da triangulação para além da fronteira por forma a garantir uma transição mais suave.

O principal objectivo da abordagem SPDE é a representação finita dada pela equação (3.32) que estabelece a ligação entre o campo aleatório Gaussiano $U(s)$ e o campo aleatório de Markov Gaussiano $\boldsymbol{u}(s)$ o qual pode ser atribuído uma estrutura Markoviana como pode ser visto em Lindgren [45, 43].

De acordo com Lindgren *et al.* [45], Bakka *et al.* [6], na metodologia R-INLA, utilizando a abordagem Equação Diferencial Parcial Estocástica que vem do Inglês *Stochastic Partial Differential Equation* (SPDE), consegue-se ter a estrutura espacial esparsa desejada para a matriz de precisão do Campo Aleatório Gaussiano (CAG) continuamente indexada.

Em substituição da construção de um modelo discreto para o CAG num determinado conjunto de localizações ou células da grelha utilizando uma covariância, constrói-se uma aproximação continuamente indexada do CAG utilizando um modelo contínuo, uma SPDE que é definida em toda a área de estudo.

3.6 Análise de dados de presumíveis infrações pesqueiras ao largo de Portugal, obtidos em ações de fiscalização

A atividade pesqueira é importante para qualquer nação costeira com recursos haliêuticos nos espaços marítimos onde detém soberania e jurisdição. Portugal, como estado costeiro, possui soberania, jurisdição e responsabilidade de 14.069 quilómetros quadrados de águas interiores, de 50.957 quilómetros quadrados de mar territorial e de 1.660.456 quilómetros quadrados de Zona Económica Exclusiva (ZEE) (ver figura 3.1), totalizando 1.725.482 quilómetros quadrados de espaço marítimo da responsabilidade de Portugal. Esta atividade constitui uma fonte de emprego ou de sustentabilidade de certos grupos sociais, principalmente de comunidades residentes ao longo da costa, assim como fonte de receitas de algumas grandes empresas ligadas a este ramo. Portanto, para a sua sustentabilidade é necessário estabelecer regras e princípios, com objetivo de controlar a exploração dos recursos piscícolas e a exploração de algumas áreas e espécies protegidas que, quando feitas descontroladamente, podem comprometer a sobrevivência do ecossistema marinho. Para uma atividade pesqueira mais justa, regrada e sustentável com vista a assegurar a conservação, gestão e desenvolvimento dos recursos aquáticos respeitando o ecossistema e a biodiversidade, foi criado em Portugal o sistema SIFICAP, do qual a Marinha Portuguesa é entidade participante, e que pretende assegurar a vigilância, fiscalização e controlo da atividade da pesca.

Neste seguimento, o presente trabalho vai aplicar as técnicas da geoestatística para estudar a distribuição espacial de presumíveis infrações pesqueiras na costa portuguesa

de interesse neste contexto.

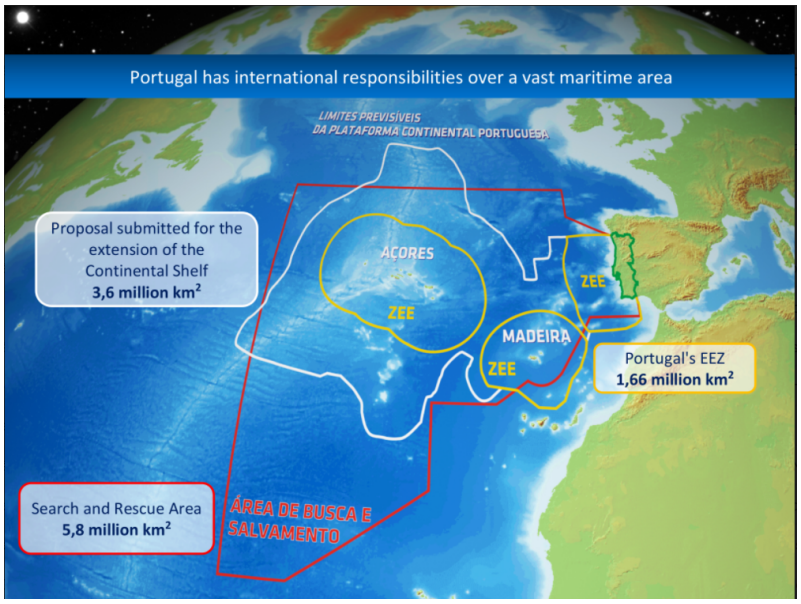

Figura 3.1: Zona Económica Exclusiva. Fonte: Marinha Portuguesa.

A informação disponível refere-se a dados georeferenciados relativos a ações de fiscalização efetuadas pela Marinha Portuguesa em toda costa Portuguesa, com mais de duzentos e onze mil (211000) pontos fiscalizados em 22 anos (de 1998 a 2019), com uma média de 9620 pontos fiscalizados por ano. O presente estudo respeita à zona costeira da extensão marítima portuguesa (mar territorial), delimitado pelas coordenadas de -11º a -5º de longitude e 36º a 42.3º de latitude.

No presente estudo, considerou-se uma coleção de pontos onde se observaram dados binários de presença/ausência de certas presumíveis infrações obtidas nas ações de fiscalização feitas pela Marinha Portuguesa, sendo os pontos amostrados $(s_1,...,s_n)$. A presença de cada presumível infração em cada ponto s é dada por $y_s = y(s)$ e a sua distribuição é especificada da seguinte maneira:

$$y_s \sim Bernoulli(\pi_s) \tag{3.34}$$

onde π_s é a probabilidade da presença da referida presumível infração. Então, especifica-se em $logit(\pi_s) = ln\left(\frac{\pi_s}{1-\pi_s}\right)$ um modelo linear incluindo diferentes covariáveis, x_{ms}, e um efeito aleatório espacial $\boldsymbol{u}$

$$logit(\pi_s) = \beta_0 + \sum_{m=1}^{M} \beta_m x_{ms} + \boldsymbol{u} \tag{3.35}$$

onde β_0 é o intercepto, β_m é o efeito da covariável x_{ms} e $\boldsymbol{u}$ um efeito aleatório que se modela como um processo gaussiano de média zero e com uma função covariância de Matérn dada pela equação (3.20).

Para a estimação dos parâmetros, vai-se usar a inferência Bayesiana através da abordagem *Integrated Nasted Laplace Approximation* (INLA) de acordo com o apresentado na subsecção 3.5.1. As distribuições a priori para os parâmetros são as que o R-INLA usa por defeito [3]. Segundo Blangiardo e Cameletti [12], para $\theta_1 = log(\tau)$ e $\theta_2 = log(k)$ são especificadas as prioris padrão com distribuição Normal $\mathcal{N}(0,1)$ e para a precisão $log(\tau) \sim logGamma(1, 0.0005)$, onde a precisão e dada por $\tau = \frac{1}{\sigma^2}$. Existem casos em que pode ser mais útil definir as prioris para o desvio padrão σ e para o alcance (*range*) R, onde $R = \frac{\sqrt{8\nu}}{k}$ ao invés de θ_1 e θ_2, nestes casos pode-se utilizar a seguinte parametrização [12]

$$log(\sigma) = log(\sigma_0) + \theta_1$$

$$log(R) = log(R_0) + \theta_2$$

onde σ_0 e R_0 são valores base para o desvio padrão e o alcance. Alguns cálculos simples produzem as seguintes equações para serem usadas na representação interna:

$$log(k) = \frac{log(8)}{2} - log(R_0) - \theta_2 = log(k_0) - \theta_2 \tag{3.36}$$

$$\begin{aligned} log(\tau) &= \frac{1}{2} log\left(\frac{1}{4\pi}\right) - log(k_0) - log(\sigma_0) - \theta_1 + \theta_2 \\ &= log(\tau_0) - \theta_1 + \theta_2 \end{aligned} \tag{3.37}$$

agora os parâmetros θ_1 e θ_2 controlam em conjunto o parâmetro τ. De acordo com Krainski *et al.* [43], para os efeitos fixos a priori padrão é uma distribuição Gaussiana com média zero, e a precisão é zero para o intercepto e 0.001 para os coeficientes das covariáveis.

3.6.1 Análise exploratória

O conjunto de dados disponível inclui presumíveis infrações relacionadas com documentação da embarcação, artes proibidas, pesca em zona proibida, pesca proibida, capturas indevidas e as condições das embarcações, entre outras. Para este trabalho escolheram-se 3 tipos de presumíveis infrações (presença/ausência) relacionadas com pesca, nomeadamente artes proibidas (infração 3) ou pesca em zona proibida ou interdita (infração 4) ou pesca proibida por potência motora ou arqueação excessiva (infração 5). Consideraram-se apenas 2 anos de dados (2014 e 2015).

A figura 3.2, apresenta a distribuição espacial dos pontos de todas as ações de fiscalização nos dois anos (2014 e 2015) no mar territorial. Neste mapa, os pontos vermelhos representam os locais fiscalizados onde foram encontradas presumíveis infrações e os pontos pretos representam os locais onde não foram encontradas alguma presumível infração das considerada.

[3] https://www.r-inla.org/

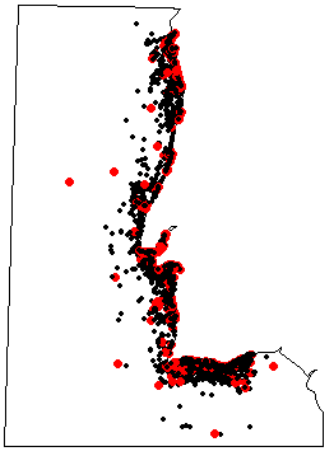

Figura 3.2: Mapa de distribuição de dados de fiscalização da costa Portuguesa.

O número total de dados analisados nos dois anos (2014 e 2015) foi de 28934, dos quais 27474 (94.95%) correspondem aos locais onde não foram encontradas presumíveis infrações, e 1460 (5.05 %) correspondem aos locais onde foram encontradas pelo menos uma das presumíveis infrações considerada, ver tabela 3.1.

Tabela 3.1: Contribuição de cada presumível infração nos dados

Infrações	total	infr.=0	infr.=0(%)	infr.=1	infr.=1(%)
infração 3	28934	28411	98.19	523	1.81
Infração 4	28934	27886	96.38	1048	3.62
Infração 5	28934	28927	99.98	7	0.02
Todas infr.	28934	27474	94.95	1460	5.05

A tabela 3.1 apresenta a contribuição de cada uma das três presumíveis infrações em análise neste estudo para os dois anos. "infr.=0"representa os locais onde não foram encontradas presumíveis infrações e "infr.=1"representa os locais onde foram encontradas presumíveis infrações. Das 1460 presumíveis infrações encontradas, 523 correspondem a infração 3, o que representa 1.81% dos dados, 1048 correspondem a infração 4, o que representa 3.62% e 7 correspondem a infração 5, o que representa 0.02%.

Existem locais onde foram encontradas mais de uma presumível infração.

Destas 3 presumíveis infrações em análise neste estudo, pode-se verificar que a infração 4 é que teve maior contribuição, ou seja, foi a mais encontradas durante a fiscalização feita pela Marinha Portuguesa e, a infração 5 foi a menos encontrada.

3.6.2 Modelação espacial das presumíveis infrações

De forma a estimar a probabilidade de ocorrência destas infrações sobre esta área, vai-se estimar o modelo 3.35 sem variáveis explicativas, em que a componente espacial do modelo será dada pelo campo aleatório, que por sua vez será aproximada pelo campo

aleatório de Markov Gaussiano, para o qual precisamos de fazer a triangularização do domínio para poder utilizar o método de aproximação pelos métodos finitos.

A figura 3.3 apresenta duas *meshs* (malhas) referente aos dois anos em análise neste trabalho, onde são feitas a triangulação do domínio espacial da área de estudo, que é necessária para estimar o modelo geoestatístico através da abordagem SPDE. A região de estudo foi subdividida em triângulos e ambas *meshs* apresentaram 995 vértices, e 9574 pontos de dados para 2014 e 9268 pontos de dados para 2015.

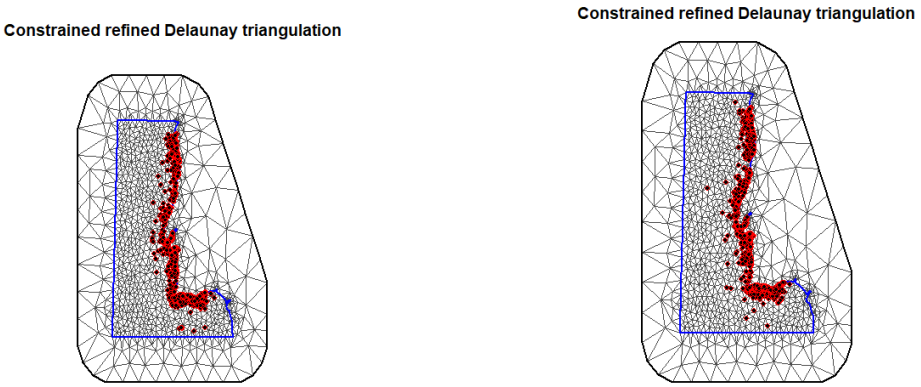

Figura 3.3: Gráficos de mesh dos anos 2014 (a esquerda) e 2015 (a direita).

Os resultados da estimação obtidos no R-INLA apresentam-se na tabela 3.2, onde se apresenta o parâmetro de intercepto β_0 para os anos 2014 e 2015. São também apresentados valores estimados dos hiperparâmetros σ_u e R da covariância de Matérn do efeito aleatório espacial para os dois anos, sendo que a variância explicada (σ_u^2) pelo variograma de Matérn, é de 1.792 e 5.324, respetivamente para 2014 e 2015, e a média a posteriori para o alcance R é de 48.62 quilómetros e 24.17 quilómetros com um intervalo de credibilidade a 95% de (22.01;92.27) e (11.21;47.04), respetivamente.

Tabela 3.2: Resumo das distribuições de probabilidade a posteriori para anos 2014 e 2015.

Parâmetro	Média	SD	Quantil 0.025	Quantil 0.5	Quantil 0.975
			2014		
β_0	-3.725	0.384	-4.516	-3.717	-2.978
σ_u^2	1.792	0.713	0.810	1.650	3.571
R	48.62	18.10	22.01	45.65	92.27
			2015		
β_0	-4.234	0.41	-5.115	-4.21	-3.489
σ_u^2	5.324	1.871	2.59	5.002	9.869
R	24.17	9.244	11.21	22.42	47.04

A figura 3.4 apresenta os mapas da média e a figura 3.5 os mapas do desvio padrão da distribuição a posteriori do campo aleatório para os anos 2014 (a esquerda) e 2015 (a direita) (Zuur, *et al.* [99]). Verifica-se que o efeito aleatório do campo em 2014 e 2015

varia entre -1.5 e 2.5 e entre -2.0 e 4.0, respetivamente. Os valores mais altos no ano 2014 estão localizados na zona centro, com maior incidência em Lisboa, e também no sul, no Algarve, enquanto que para o ano 2015 o efeito espacial apresenta valores elevados na zona centro do país, concretamente na zona de Lisboa e quase toda a extensão da zona norte, com valores altos também na região da Figueira da Foz; enquanto que a zona sul do país apresentam valores mais baixos do efeito espacial. Quanto ao desvio padrão, para 2014 varia entre 0.4 e 1.8 e para 2015 varia entre 0.5 e 2.5; o ano 2014 apresenta menor variação em relação a 2015, portanto, a incerteza de previsão parece ser menor em 2014.

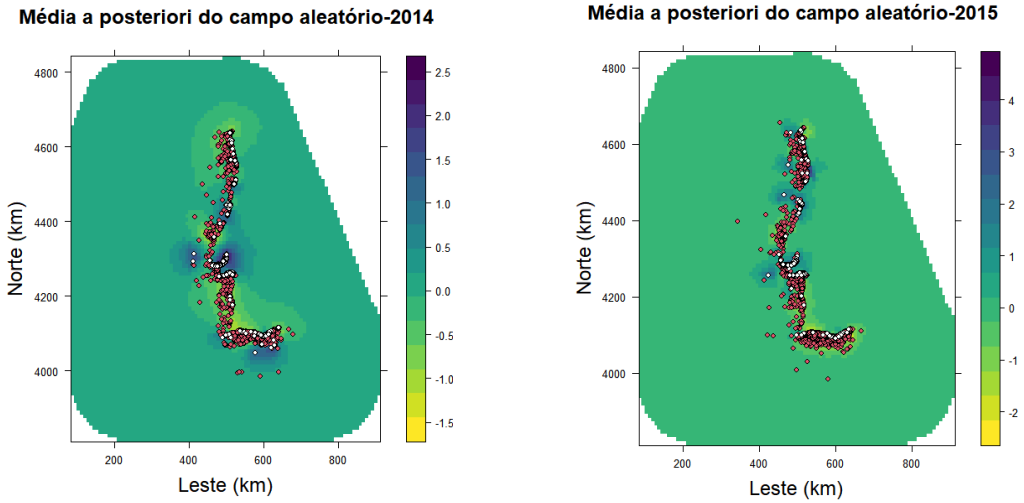

Figura 3.4: Mapas de média a posteriori do efeito espacial do campo aleatório do ano 2014 (a esquerda) e do ano 2015 (a direita).

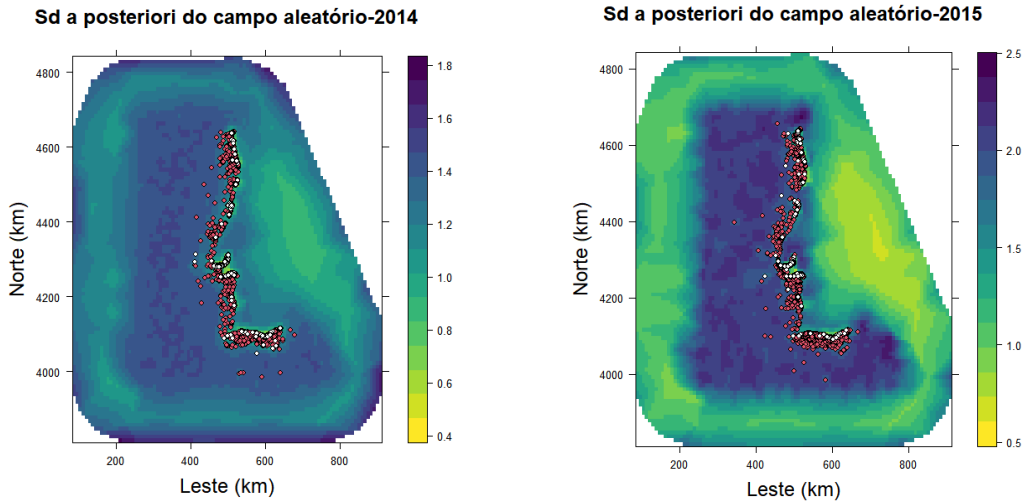

Figura 3.5: Mapas de desvio padrão a posteriori do campo aleatório dos ano 2014 (a esquerda) e do ano 2015 (a direita).

A figura 3.6 apresenta mapas de valores ajustados ou mapas de risco associado às presumíveis infrações de pesca com escalas diferentes. Verifica-se que, em ambos os anos, a probabilidade estimada de ocorrência das presumíveis infrações ao largo de Lisboa ronda os 20%, baixando para cerca de 5% ao largo da zona centro, a norte de Lisboa até ao Porto, e também ao largo de Faro (mais em 2014). No ano de 2015, a estimativa desta probabilidade eleva-se para cerca de 60% ao largo do Porto. De forma a tentar perceber melhor e fazer a comparação, apresentamos também na figura 3.7 os mapas de risco associados às presumíveis infrações de pesca com escalas iguais.

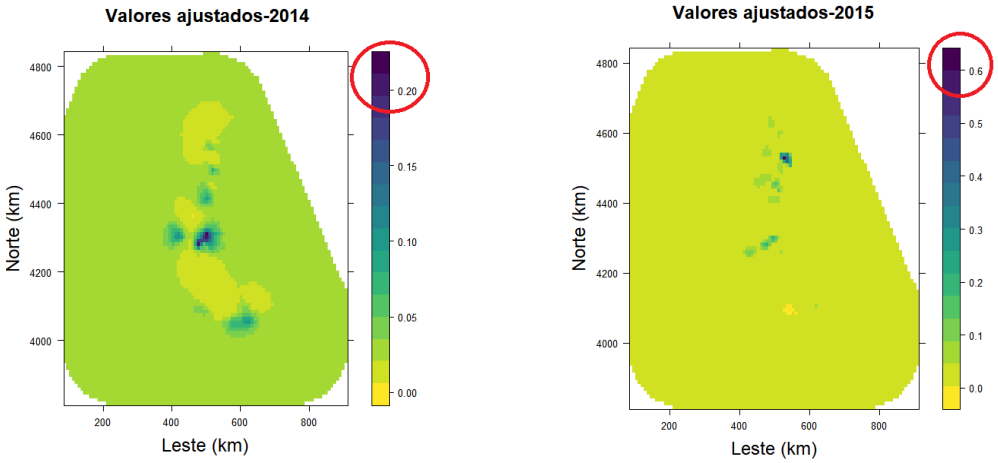

Figura 3.6: Mapas de risco com escalas diferentes do ano 2014 (a esquerda) e do ano 2015 (a direita).

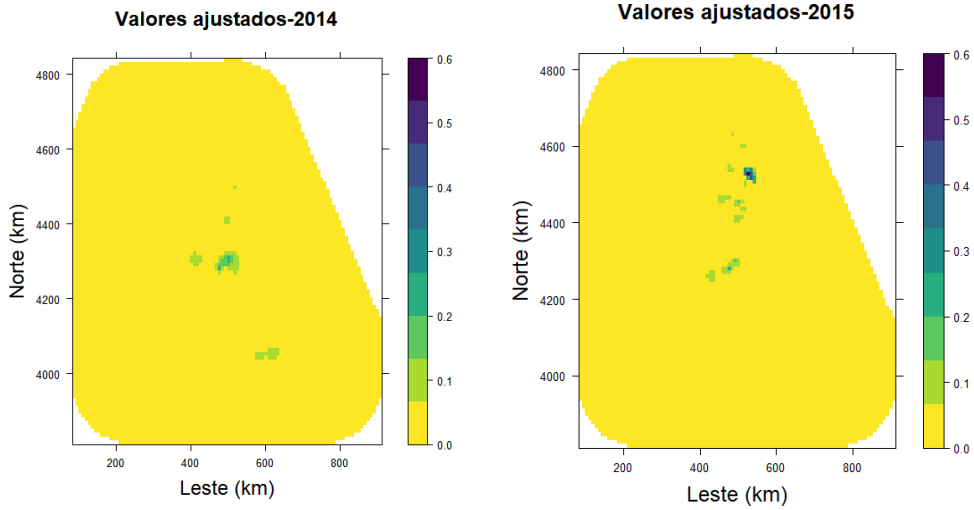

Figura 3.7: Mapas de risco com escalas iguais do ano 2014 (a esquerda) e do ano 2015 (a direita).

3.7 Análise de dados de presumíveis infrações pesqueiras no comando de zona do Sul, obtidos em ações de fiscalização

Como a Costa Portuguesa tem uma particularidade ter duas costas, uma a oeste e outra a sul do país, então vamos apresentar os resultados por comandos de zonas. Neste estudo, vamos começar por apresentar análises dos resultados obtidos no comando de zona Sul, concretamente na região costeira do Algarve, e escolheram-se 3 tipos de presumíveis infrações (presença/ausência) relacionadas com a pesca, nomeadamente artes proibidas (infração 3), pesca em zona proibida (infração 4) e pesca proibida por motorização ou tonelagem excessiva (infração 5). São considerados 5 anos de dados (de 2013 a 2017).

A figura 3.8 mostra a distribuição espacial dos pontos de todas as ações de fiscalização nos cinco anos (de 2013 a 2017) na zona costeira do Algarve. Os pontos vermelhos representam os locais inspecionados onde foram encontradas presumíveis infrações e os pontos pretos representam os locais onde não foram encontradas presumíveis infrações.

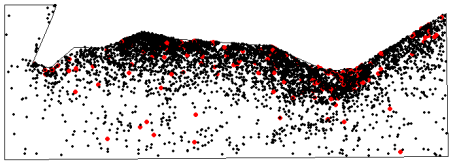

Figura 3.8: Mapa de distribuição de dados de fiscalização marítima.

O total de dados analisados nos cinco anos (de 2013 a 2017) foi de 25.335, dos quais 24.754 (97,7%) correspondem a locais onde não foram encontradas as presumíveis infrações ("infr.=0") e 581 (2,3 %) correspondem aos locais onde foram encontradas pelo menos uma das presumíveis infrações consideradas ("infr.=1"). Existem locais onde foram encontradas mais de uma presumível infração. A contribuição de cada uma das três presumíveis infrações analisadas neste estudo para os cinco anos pode ser vista na tabela 3.3. Das 581 presumíveis infrações encontradas, 202 correspondem à infração 3, que representa 0,8% dos dados, 409 correspondem à infração 4, que representa 1,6% e 7 correspondem à infração 5, que representa 0,03%. Destas 3 presumíveis infrações em análise neste estudo, verifica-se que a infração 4 foi a que mais contribuiu, ou seja, foi a mais encontrada durante a inspeção feita pela Marinha Portuguesa, e a infração 5 foi a menos encontrada.

3.7.1 Modelo espacial para presumíveis infrações pesqueiras

Para estimar a probabilidade de ocorrência dessas infrações nesta área, foi escolhido um modelo Bayesiano hierárquico com componente espacial.

Tabela 3.3: Contribuição de cada presumível infração nos dados.

Infração	total	infr.=0	infr.=0(%)	infr.=1	infr.=1(%)
Infração 3	25335	25133	99.2	202	0.8
Infração 4	25335	24926	98.4	409	1.6
Infração 5	25335	25328	99.97	7	0.03
Todas infr.	25335	24754	97.7	581	2.3

Considera-se a recolha de pontos onde foram observados dados binários de presença/ausência de determinadas presumíveis infrações, nas ações de fiscalização efetuadas pela Marinha Portuguesa, nos pontos amostrados $(s_1,...,s_n)$. A presença/ausência de cada presumível infração em cada ponto s_i é dada por $Y_i = y(s_i)$ e sua distribuição é especificada como

$$Y_i \sim Bernoulli(\pi_i) \tag{3.38}$$

onde π_i é a probabilidade da presença de presumível infração em s_i. Para $logit(\pi_i) = ln\left(\frac{\pi_i}{1-\pi_i}\right)$ um preditor linear incluindo covariáveis x_m, $m = 1,\ldots,M$, e um efeito aleatório espacial $\boldsymbol{u}$ é especificado,

$$logit(\pi_i) = \beta_0 + \sum_{m=1}^{M} \beta_m x_{m,i} + \boldsymbol{u} \tag{3.39}$$

onde β_0 é o intercepto, β_m é o efeito da covariável x_m e $\boldsymbol{u}$ é um efeito aleatório que é modelado como um processo Gaussiano de média zero com uma função de covariância de Matérn dada pela equação (3.20). Este campo aleatório deve ser aproximado por um campo aleatório Gaussiano de Markov, para o qual o domínio correspondente é triangularizado para a aplicação do método de aproximação de elementos finitos. A figura 3.9 apresenta a malha considerada onde o domínio espacial da área de estudo é triangularizado, para estimar este modelo geoestatístico através da abordagem SPDE. A região de estudo foi subdividida em triângulos e a malha possui 1770 vértices e 12232 pontos de dados.

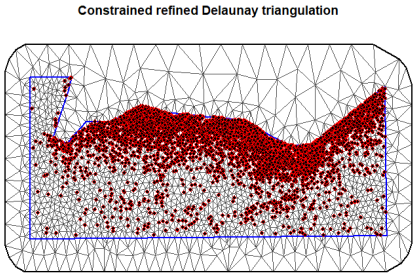

Figura 3.9: Gráfico da *mesh* (malha) para os dados dos cinco anos em análise (2013-2017).

Para a estimação dos parâmetros, utiliza-se a inferência Bayesiana através da abordagem INLA, conforme apresentado na subsecção 3.5. As distribuições a priori para os parâmetros são aquelas que o R-INLA usa como padrão [3]. De acordo com Blangiardo e Cameletti [12], para $\theta_1 = log(\tau)$ e $\theta_2 = log(k)$, as distribuições a priori gaussianas padrão $\mathcal{N}(0,1)$ são especificado e para a precisão τ um log-Gamma$(1, 0.0005)$, onde a precisão é dada por $\tau = \frac{1}{\sigma_u^2}$. Existem casos em que pode ser mais útil definir as prioris para o desvio padrão σ_u e para o alcance (*range*) R, onde $R = \frac{\sqrt{8\nu}}{k}$ em vez de θ_1 e θ_2, nestes casos pode-se usar a seguinte parametrização [12]

$$log(\sigma_u) = log(\sigma_0) + \theta_1$$

$$log(R) = log(R_0) + \theta_2$$

onde σ_0 e R_0 são valores base para o desvio padrão e o alcance. Alguns cálculos simples produzem as seguintes equações para usar na representação interna do R-INLA:

$$log(k) = \frac{log(8)}{2} - log(R_0) - \theta_2 = log(k_0) - \theta_2 \tag{3.40}$$

$$\begin{aligned} log(\tau) &= \frac{1}{2} log\left(\frac{1}{4\pi}\right) - log(k_0) - log(\sigma_0) - \theta_1 + \theta_2 \\ &= log(\tau_0) - \theta_1 + \theta_2. \end{aligned} \tag{3.41}$$

Agora, os parâmetros θ_1 e θ_2 controlam em conjunto o parâmetro τ. De acordo com Krainski *et al.* [43], para os efeitos fixos distribuições a priori padrão Gaussianas com média zero e precisão zero para o intercepto (uniforme) e de 0,001 para os coeficientes das covariáveis.

3.7.1.1 Seleção de modelo

Para selecionar o modelo espacial mais adequado, vários modelos com diferentes graus de complexidade foram ajustados, conforme tabela 3.4. *Período* representa quatro períodos do dia de seis horas cada, 0:01-6:00; 06:01-12:00; 12h01-18h00 e 18h01-24h00; $fator(ano)$ corresponde ao ano, de 2013 a 2017, e $\boldsymbol{u}$ representa o efeito aleatório espacial. O Deviance Information Criterion (DIC) proposto por Spiegehalter *et al.* [81] e o critério de informação de Watanabe-Akaike (WAIC) proposto por Watanabe [94] foram usados para fazer a seleção do modelo. O modelo mais adequado foi aquele com o menor DIC ou WAIC.

O modelo A com apenas o intercepto e a componente aleatória espacial foi utilizada como base de comparação. O modelo B também leva em consideração o efeito dos períodos do dia que foram divididos em quatro (horas), 0:01-6:00; 06:01-12:00; 12h01-18h00 e 18h01-24h00. O modelo C também leva em consideração o efeito do tempo em anos,

[3] https://www.r-inla.org/

Tabela 3.4: DIC e WAIC dos modelos ajustados.

Modelos	Preditor Linear	DIC	WAIC
A	$\beta_0 + \boldsymbol{u}$	2242.08	2175.33
B	$Periodo + \boldsymbol{u}$	2202.76	2137.16
C	$factor(ano) + \boldsymbol{u}$	2237.12	2169.46
D	$Periodo + factor(ano) + \boldsymbol{u}$	2190.62	2127.53

nesse caso entra como fator (de 2013 a 2017). E por fim, no modelo D, adicionamos ao modelo A as duas covariáveis utilizadas nos modelos B e C.

O modelo que apresenta o menor DIC é o modelo D, que considera os períodos do dia, os anos e o efeito aleatório espacial.

A seguir, detalhamos os resultados do modelo mais adequado (modelo D) obtido usando a abordagem INLA-SPDE.

3.7.1.2 Ajuste do modelo espacial

Os resultados da estimação para o modelo D são apresentados na tabela 3.5, onde é apresentado o resumo das distribuições a posteriori para os anos 2013 a 2017. Estes resultados foram obtidos em R-INLA com um tempo de execução inferior a 3 minutos. Valores estimados dos hiperparâmetros do modelo para os cinco anos também são apresentados. A variância explicada (σ_u^2) pelo variograma de Matérn é estimada em 4.883, e a média a posteriori para o alcance R é de 4.904 quilómetros, com um intervalo de credibilidade a 95% de (3.07;7.414). Percebe-se também que os efeitos dos quatro períodos do dia, e dos anos de 2015 e 2017 são significativos.

Tabela 3.5: Resumo das distribuições de probabilidade a posteriori para os anos 2013 a 2017.

Parâmetro	Média	SD	Quantil 0.025	Quantil 0.5	Quantil 0.975
0:01-6:00	-4.572	0.454	-5.528	-4.549	-3.744
06:01-12:00	-5.846	0.420	-6.750	-5.818	-5.101
12:01-18:00	-5.678	0.427	-6.592	-5.651	-4.915
18:01-24:00	-4.456	0.429	-5.370	-4.431	-3.684
Year 2014	-0.112	0.181	-0.469	-0.112	0.241
Year 2015	-0.810	0.243	-1.302	-0.805	-0.347
Year 2016	0.003	0.196	-0.385	0.004	0.383
Year 2017	-0.625	0.252	-1.136	-0.620	-0.145
σ_u^2	4.883	1.121	3.047	4.756	7.433
R	4.904	1.110	3.07	4.784	7.414

A figura 3.10 mostra os mapas da média (esquerda) e desvio padrão (direita) da distribuição a posteriori do campo aleatório para os cinco anos analisados neste estudo

(2013-2017) (Zuur , *et al.* [99]). Verifica-se que o campo médio do efeito espacial varia entre -2 e 6 e o desvio padrão varia entre 0,5 e 3,0. Os valores mais elevados localizam-se ao longo de toda a costa da região do Algarve, com maior incidência na zona junto à linha da costa, e também na costa leste do Algarve (região de Tavira e Faro).

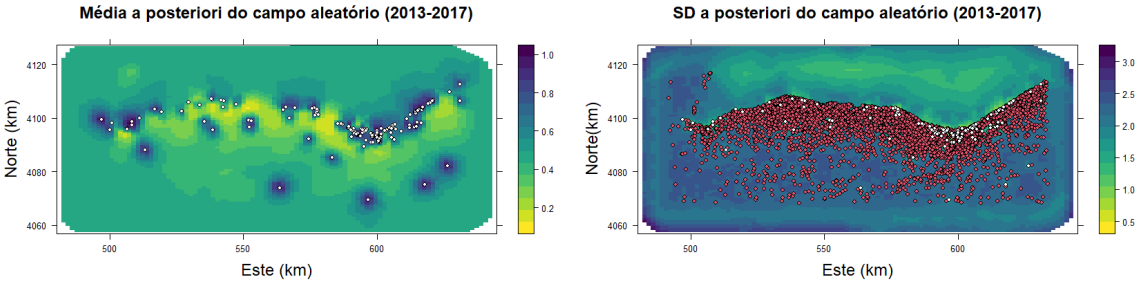

Figura 3.10: Mapa da média a posteriori (esquerda) e mapa de desvio padrão (direita) do efeito espacial do campo aleatório para os anos 2013-2017.

A figura 3.11 apresenta mapas de risco associados às presumíveis infrações de pesca por períodos do dia para o ano 2013, de modo que o 2º e 3º período correspondem ao período do dia com maior disponibilidade de recursos de vigilância (6:01-12:00 e 12:01-18:00) e o 1º e 4º período correspondem ao período noturno (menos recursos). Verifica-se que a probabilidade estimada máxima de ocorrência das presumíveis infrações ao largo da zona de Sagres e Tavira é de cerca de 50% no 2º e 3º período, baixando para cerca de 5% ao largo da zona de Faro até Albufeira. No 1º e 4º período, a estimativa desta probabilidade eleva-se para cerca de 70% ao largo de Sagres e Tavira.

A figura 3.12 apresenta mapas de risco associados às presumíveis infrações de pesca para os quatros períodos do ano 2014. Verifica-se que a probabilidade estimada máxima de ocorrência das presumíveis infrações não difere muito entre o 1º e o 4º período na ordem de 70%, o 2º período apresenta uma probabilidade estimada máxima na ordem de 40%, e o 3º período apresenta uma probabilidade estimada máxima na ordem de 50%, sendo a estimativa é alta ao largo de Sagres e Tavira nos quatros períodos do ano 2014.

A figura 3.13 apresenta mapas de risco associados às presumíveis infrações de pesca para os quatros períodos do ano 2015. Verifica-se que a probabilidade estimada máxima de ocorrência das presumíveis infrações não difere muito entre o 1º e o 4º período na ordem de 60%, o 2º período apresenta uma probabilidade estimada máxima na ordem de 30%, e o 3º período apresenta uma probabilidade estimada máxima na ordem de 35%, sendo que a estimativa é alta ao largo de Sagres e Tavira nos quatros períodos do ano 2015.

A figura 3.14 apresenta mapas de risco associados às presumíveis infrações de pesca para os quatros períodos do ano 2016. Verifica-se que a probabilidade estimada máxima de ocorrência das presumíveis infrações não difere muito entre o 1º e o 4º período na

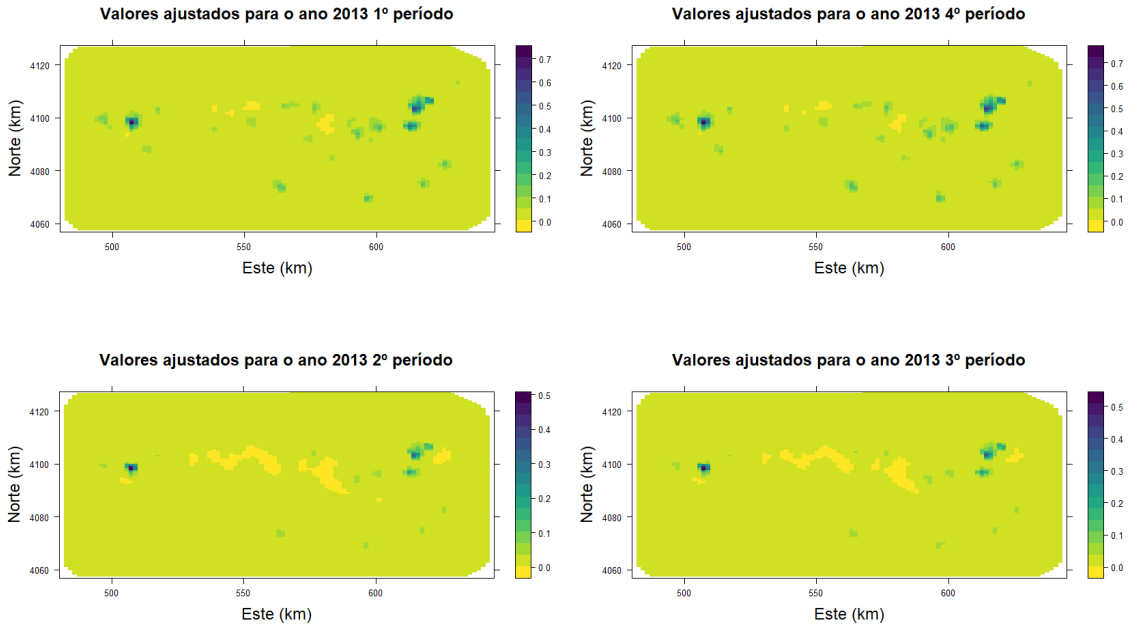

Figura 3.11: Mapas de risco para o 1º período (em cima-esquerda); 4º período (em cima-direita); 2º período (em baixo-esquerda); e 3º período (em baixo-direita) para o ano 2013.

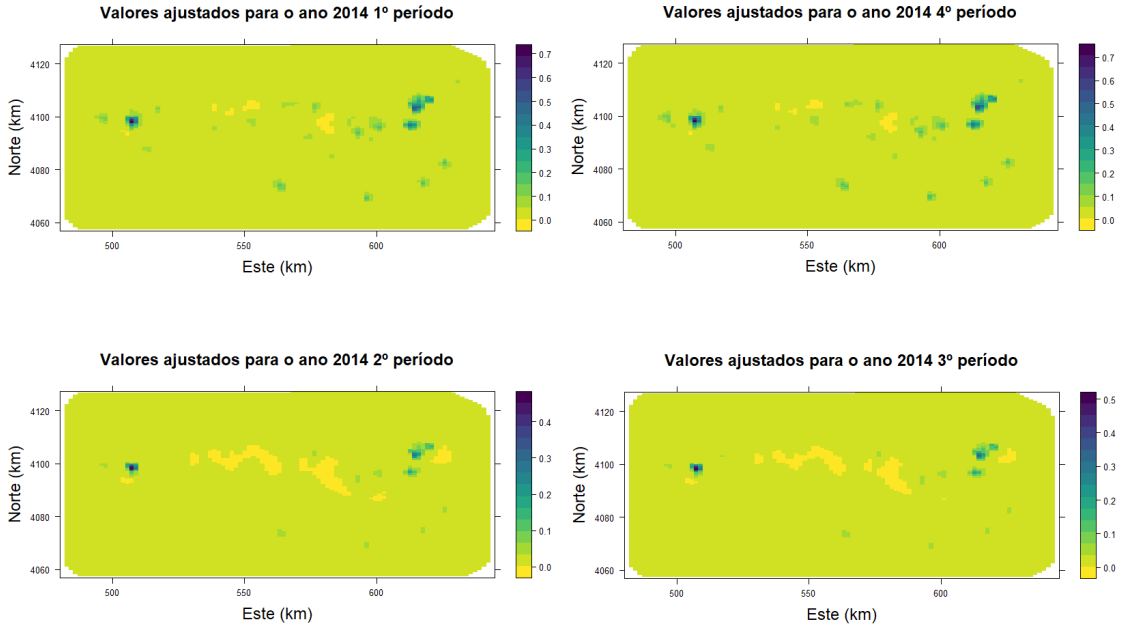

Figura 3.12: Mapas de risco para o 1º período (em cima-esquerda); 4º período (em cima-direita); 2º período (em baixo-esquerda); e 3º período (em baixo-direita) para o ano 2014.

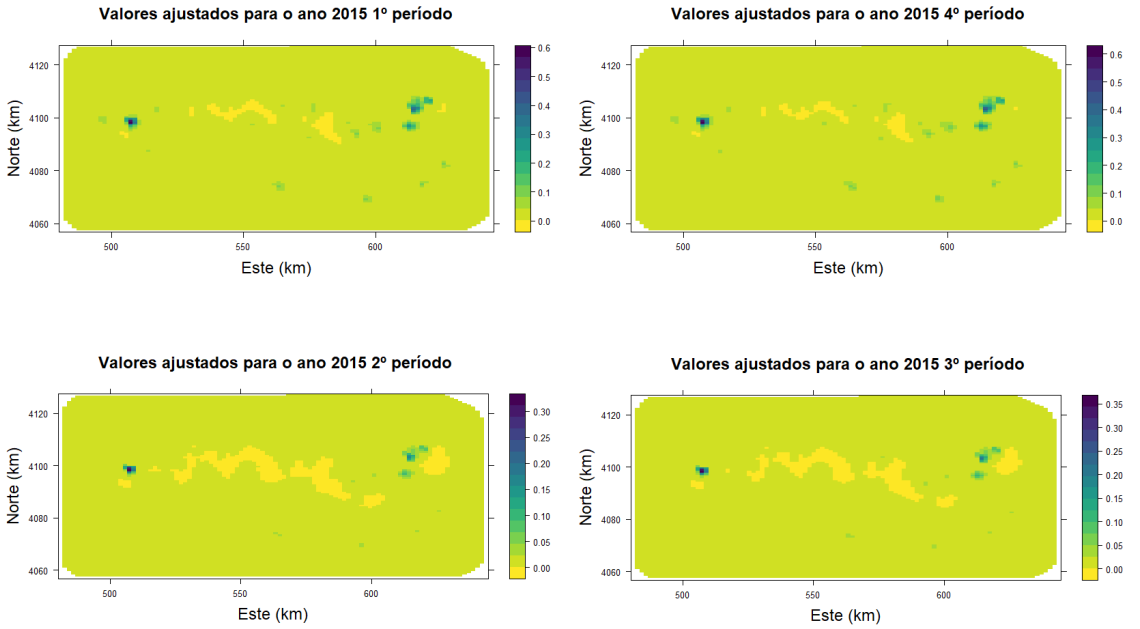

Figura 3.13: Mapas de risco para o 1º período (em cima-esquerda); 4º período (em cima-direita); 2º período (em baixo-esquerda); e 3º período (em baixo-direita) para o ano 2015.

ordem de 70%, baixando para 20% na região de Faro e Albufeira, o 2º e o 3º na ordem de 50%, sendo que a estimativa é alta ao largo de Sagres e Tavira nos quatros períodos do ano 2016.

A figura 3.15 apresenta mapas de risco associados às presumíveis infrações de pesca para os quatros períodos do ano 2017. Verifica-se que a probabilidade estimada máxima de ocorrência das presumíveis infrações não difere muito entre o 1º e o 4º período na ordem de 60%, o 2º período apresenta uma probabilidade estimada máxima na ordem de 35%, e o 3º período apresenta uma probabilidade estimada máxima na ordem de 40%, sendo que a estimativa é alta ao largo de Sagres e Tavira nos quatros períodos do ano 2017.

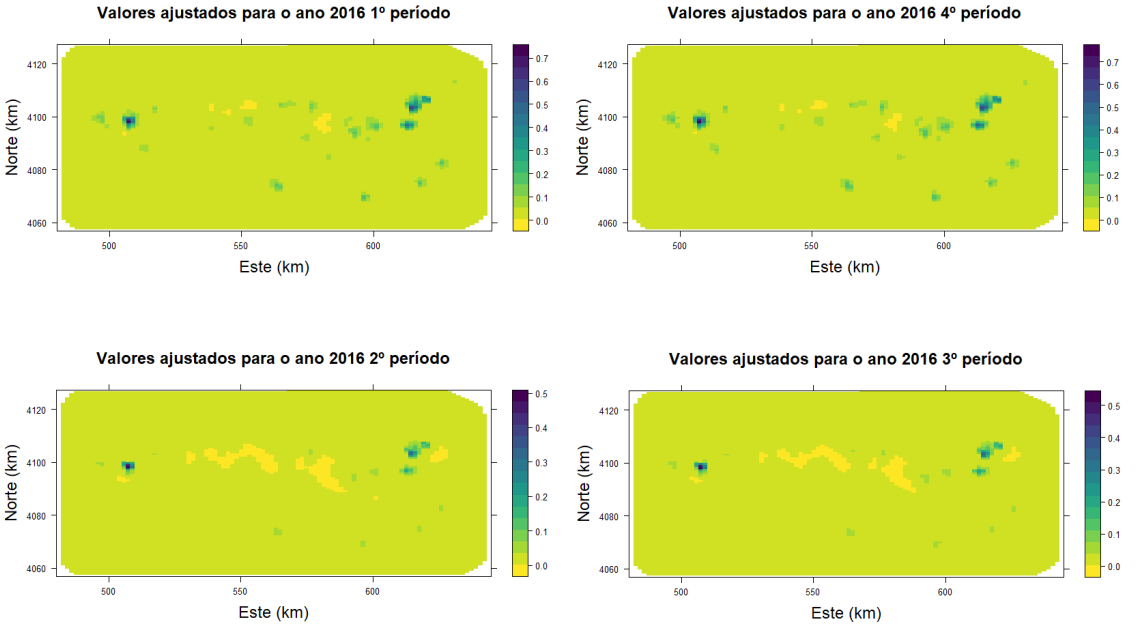

Figura 3.14: Mapas de risco para o 1º período (em cima-esquerda); 4º período (em cima-direita); 2º período (em baixo-esquerda); e 3º período (em baixo-direita) para o ano 2016.

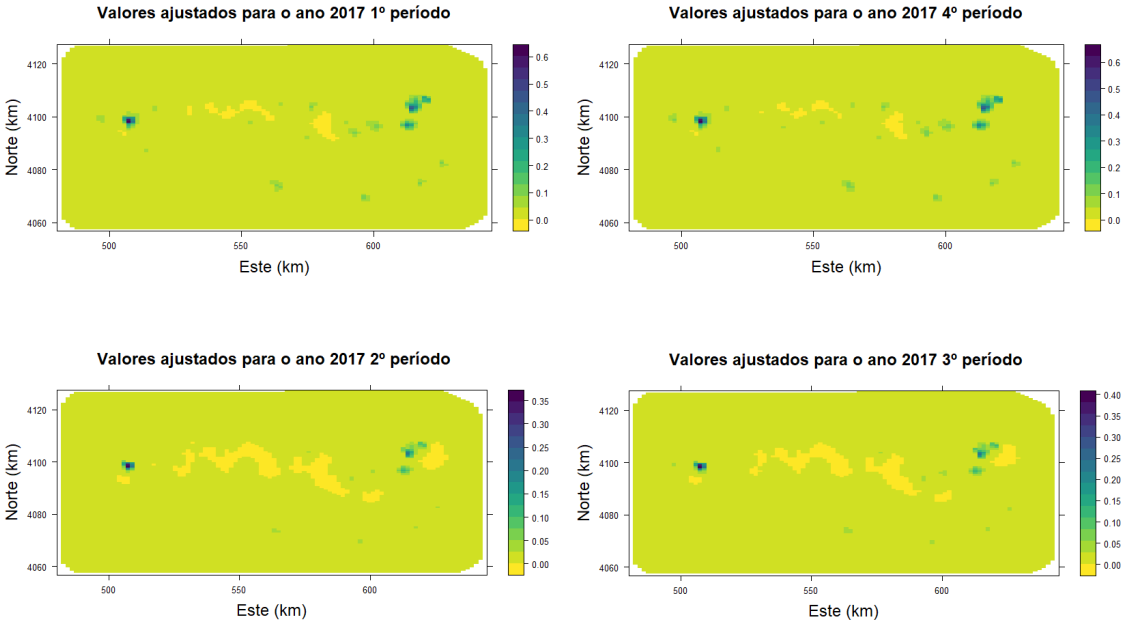

Figura 3.15: Mapas de risco para o 1º período (em cima-esquerda); 4º período (em cima-direita); 2º período (em baixo-esquerda); e 3º período (em baixo-direita) para o ano 2017.

3.7.2 Modelo espaço-temporal para presumíveis infrações pesqueiras

O modelo espaço-temporal é considerado, por exemplo, por Riveira *et al.* [72] e Blangiardo e Camaletti [12] como uma extensão do modelo espacial incluindo a componente temporal. Dados são definidos por um processo dado por

$$y(s,t) \equiv \{y(s,t),(s,t) \in D \subset \mathbb{R}^2 \times \mathbb{R}\} \tag{3.42}$$

À semelhança do modelo espacial, consideramos uma coleção de dados de pontos de presença/ausência de certas presumíveis infrações encontradas em ações de fiscalização marítima na região do Algarve, onde $(s_1,\ldots,s_n)$ são os pontos amostrados e t representa os anos. A presença de cada presumível infração em cada ponto s_i é dada por $y_{it} = y(s_i,t)$ e sua distribuição é dada por

$$y_{it} \sim Bernoulli(\pi_{it}) \tag{3.43}$$

π_{it} é a probabilidade de presença de presumível infração. Então, especifica-se em $logit\,(\pi_{it}) = ln\left(\frac{\pi_{it}}{1-\pi_{it}}\right)$ um modelo linear incluindo covariáveis x_m, $m = 1,\ldots,M$, e um termo referente ao espaço temporal w_{it}, dado por

$$logit\,(\pi_{it}) = \beta_0 + \sum_{m=1}^{M} \beta_m x_{mi} + w_{it} \tag{3.44}$$

na aplicação, x_m são as covariáveis períodos do dia, que foram divididos em quatro (horas), 0:01-6:00; 06:01-12:00; 12:01-18:00 e 18:01-24:00, w_{it} representa um processo espaço-temporal latente que muda no tempo com uma dinâmica autorregressiva de primeira ordem e inovações de correlação espacial, que modelamos da seguinte forma:

$$w_{it} = aw_{i(t-1)} + \boldsymbol{u}_t, \tag{3.45}$$

com $t = 2,\ldots,T$, $|\,a\,| < 1$, $w_{it} \sim Normal(0,\sigma^2/(1-a^2))$. $\boldsymbol{u}_t$ é um campo Gaussiano de média zero que é temporalmente independente com a seguinte função de covariância espaço-tempo:

$$Cov(u_{it},u_{jz}) = \Big\{(0,t \neq z; Cov(u_i,u_j)),t = z\Big\}, \tag{3.46}$$

para $i \neq j$, onde $Cov(u_i,u_j)$ é modelado usando a função de covariância espacial de Matérn.

3.7.2.1 Ajuste do modelo espaço-temporal

O modelo espaço-temporal foi ajustado para os cinco anos analisados neste estudo usando o modelo de correlação autoregressiva de ordem 1 (ar1).

Os resultados da estimação obtidos no R-INLA são apresentados na tabela 3.6, o resumo das distribuições a posteriori do efeito período nos anos 2013 a 2017.

Pode-se observar na figura 3.16 que não houve diferenças relevantes nos efeitos espaciais ao longo dos anos, embora que o efeito espacial nos anos 2013 e 2014 varia entre -2 a 6, enquanto nos anos 2015, 2016 e 2017 o efeito espacial varia de -2 a 5. Os valores mais elevados localizam-se ao longo de toda a costa de Algarve, com maior incidência na

Tabela 3.6: Resumo das distribuições de probabilidades a posteriori para o modelo espaço-temporal dos anos 2013 a 2017.

Parâmetro	Média	SD	Quantil 0.025	Quantil 0.5	Quantl 0.975
Intercepto	-4.415	14.146	-32.187	-4.415	23.335
0:01-6:00	-0.588	14.144	-28.357	-0.588	27.158
06:01-12:00	-1.806	14.143	-29.573	-1.806	25.938
12:01-18:00	-1.631	14.143	-29.399	-1.632	26.113
19:01-24:00	-0.387	14.143	-28.155	-0.388	27.357

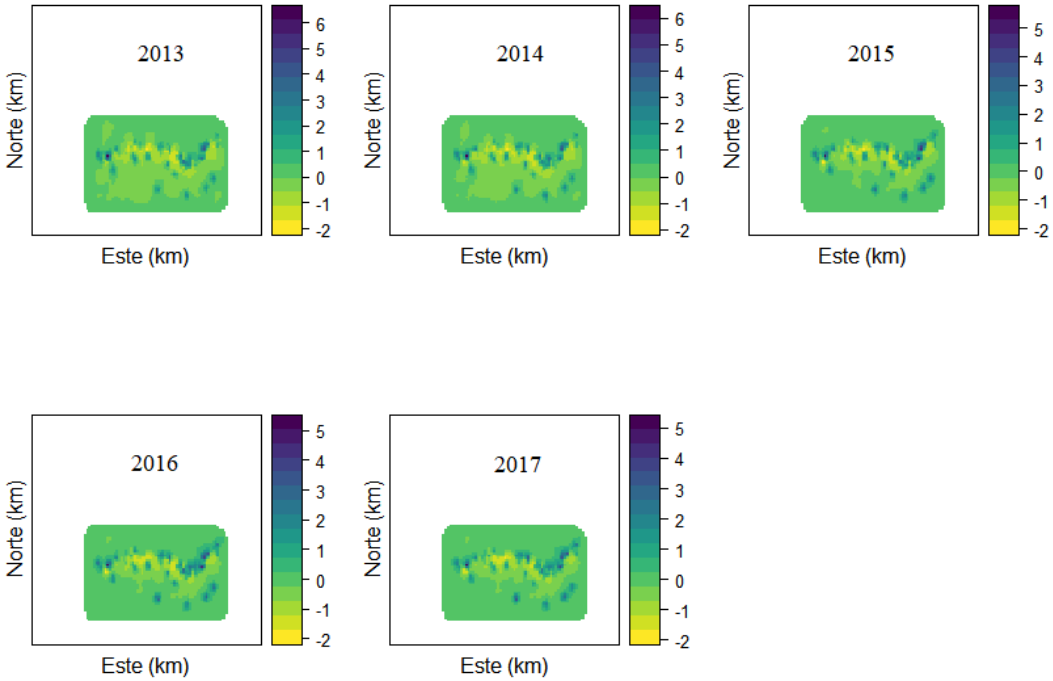

Figura 3.16: Campos aleatórios espaciais correlacionados temporalmente para cada ano. Superior esquerdo: 2013. Superior direito: 2015. Inferior esquerdo: 2016. inferior direito: 2017.

zona junto à linha da costa, principalmente na região de Sagres, Faro e Tavira. A tabela 3.6 também mostra que os quatro períodos do dia não são significativos.

A seguir, é apresentada a comparação, em termos de DIC e WAIC, dos dois modelos ajustados neste estudo, o modelo espacial e o modelo espaço-temporal (ver tabela 3.7). O modelo espacial é o que melhor se ajusta (tem o menor DIC), refletindo uma constância das estimativas ao longo do tempo.

Tabela 3.7: Comparação do modelo espacial e espaço-temporal.

Critério	Modelo espacial	Modelo espaço-temporal
DIC	2190.62	2224.43
WAIC	2127.53	2137.00

3.8 Conclusões

Nos últimos anos, dados geoestatísticos binários têm sido utilizados em diversas áreas de pesquisa. Neste estudo, um modelo hierárquico bayesiano foi ajustado a dados geoestatísticos ou referenciados por pontos usando abordagens INLA-SPDE e verificou-se que modelos hierárquicos bayesianos para dados binários podem ser usados para analisar dados de fiscalização marítima.

No desenvolvimento deste trabalho, foram encontrados na literatura matérias que dizem respeito à técnicas de modelação de dados geoestatísticos, especificamente quando a variável resposta é binária, também encontram-se trabalhos descrevendo os modelos lineares generalizados, modelos lineares generalizados mistos, modelos hierárquicos Bayesianos, sendo todos eles para dados binários, e de seguida, foram descritos os métodos INLA e SPDE. Por último, fez-se uma aplicação das técnicas descritas com dados reais de fiscalização marítima da costa Portuguesa, para construção de mapas de média e erro padrão do efeito aleatório e mapas de riscos. A inferência bayesiana foi totalmente feita utilizando a abordagem INLA, que comparado aos métodos MCMC apresenta bons resultados a nível computacional, ou seja, é muito mais rápido computacionalmente [73].

Na aplicação aos dados de fiscalização marítima, o modelo conseguiu estimar muito bem os dados, mapeando as regiões onde ocorrem as presumíveis infrações obtidas nas ações de fiscalização. A aplicação foi feita em duas partes, na primeira parte, analisou-se dados de presumíveis infrações pesqueiras ao largo de toda a costa Portuguesa, obtidos em ação de fiscalização nos anos 2014 e 2015; e a segunda parte, analisou-se dados de presumíveis infrações no comando de zona do Sul, obtidos em ações de fiscalização nos anos de 2013 a 2017.

Da aplicação destas técnicas feitas na primeira parte, obtiveram-se as seguintes conclusões:

Os mapas de média a posteriori do efeito aleatório do campo para os dois anos (2014 e 2015) descreveram a distribuição espacial das infrações em toda a costa Portuguesa, mostrando os locais com maior e com menor efeito aleatório.

Os mapas de desvio padrão estimaram a incerteza de previsão para cada ano em estudo.

Os mapas de riscos, apresentaram uma distribuição do risco associado as presumíveis infrações de pesca. E verificou-se que o risco associado a essas presumíveis infrações nos dois anos é maior na zona de Lisboa e algumas regiões da zona Norte de Portugal, mas

em 2014 a região de Algarve apresentou um risco elevado.

Na segunda parte, dois modelos foram ajustados, um espacial e outro espaço-temporal [99]. Das análises realizadas pelo critério de comparação DIC, o modelo espacial foi o que melhor se ajusta aos dados de fiscalização marítima ao largo da costa sul de Portugal. Foram construídos mapas de médias onde não foram observadas diferenças significativas durante os 5 anos.

Ainda na segunda parte, outro objetivo deste estudo foi construir mapas de risco para os quatro períodos do dia em todos os anos em análise nesse estudo. Os mapas da média a posteriori e de erro padrão do campo aleatório foram apresentados. Os mapas de efeito aleatório de campo médio a posteriori para os cinco anos (de 2013 a 2017) descreveram a distribuição espacial das infrações na região costeira do Algarve, mostrando os locais com os maiores e os com menores efeitos aleatórios. Os mapas de erro padrão estimaram a incerteza de previsão para os cinco anos em estudo [43, 58] .

Em seguida foram construídos mapas de valores ajustados por períodos do dia que estimaram a probabilidade de encontrar as presumíveis infrações de pesca, onde foram na ordem de 50% no 2º e 3º período do dia e 70% no 1º e 4º período do dia no ano 2013. Construiu-se também mapas de riscos para os quatro períodos do dia para os anos 2014 a 2017 e constatou-se que o risco associado a estas presumíveis infrações é maior no 1º e no 4º período do dia (correspondente ao período noturno), na zona de Sagres e Tavira e, no 2º e 3º período, as mesmas regiões (Sagres e Tavira) apresentaram um risco relativamente alto. As probabilidades estimadas podem contribuir para definir futuras rotas para as ações de fiscalização.

Como trabalho futuro, o estudo desenvolvido será aplicado/desenvolvido para outras zonas costeiras de Portugal (Centro e Norte) e outro grupo/tipo de infrações, articulando-se com a futura definição dos percursos por zonas e por grupo de infrações mas com o objetivo de estar disponível a nível nacional. Pretende-se também fazer um desenho amostral onde as rotas com os pontos a serem inspecionados sejam definidas de acordo com as probabilidades estimadas deste estudo.

4

Delineamento de amostragem para dados geoestatísticos binários

4.1 Introdução

A fiscalização da atividade pesqueira é muito importante para a preservação das espécies que são pescadas, assim como para assegurar a continuidade dessas espécies [46]. Estas ações de fiscalização podem implicar custos muito altos, pelo que para as levar a cabo é necessário um bom planeamento dos delineamentos de amostragem utilizados, de forma a maximizar a eficiência na obtenção da informação a partir dos dados das ações desenvolvidas sobre a área em estudo.

Delineamento de amostragem (do inglês *Sampling Design*) tem a função orientadora de como selecionar cuidadosamente os elementos da população que vão fazer parte da amostra que posteriormente será usada para fazer inferência sobre a população [19, 70, 82].

O principal objetivo deste estudo é propor critérios de delineamento de amostragem baseados em modelos de geoestatística, em contexto de dados binários, que demonstrem ser vantajosos no seu objetivo de otimizar as ações de fiscalização em termos do esforço empregue na sua execução. Adicionalmente, queremos comparar estas propostas com delineamentos de amostragem convencionais.

Com base num modelo geoestatístico semelhante aos ajustados no capítulo 3, para modelar a prevalência de infrações pesqueiras na costa Portuguesa a partir de valores reais obtidos em ação de fiscalização, pretende-se otimizar esquemas de amostragem para futuras fiscalizações pesqueiras com base em dois critérios: maximização do risco de infração e exploração de áreas onde a variabilidade associada ao risco é elevada. Estes critérios foram utilizados para construir dois esquemas de amostragem que se comparam ao delineamento de amostragem aleatória simples através da raíz do erro quadrático médio, Root mean square error (RMSE). Alternativas propostas na literatura de outros esquemas de amostragem incluem a minimização da média da distância mais curta entre os pontos amostrados Minimization of mean shortest distance (MMSD) e otimização do

ajuste a uma distribuição ideal, definida a priori, da distribuição dos pares de pontos amostrados para a estimação do variograma experimental (WM) [44, 76, 89, 74].

Este capítulo encontra-se dividido em 4 secções. Na secção 2, apresenta-se alguns delineamentos de amostragem, tanto baseados em desenhos amostrais como baseados em modelos, alguns delineamentos de amostragem convencionais, delineamentos de amostragem para dados geoestatísticos e a abordagem proposta para novos delineamentos de amostragem. Detalha-se ainda o critério de seleção e validação do delineamento de amostragem. Na secção 3, apresenta-se os resultados da aplicação dos delineamentos de amostragem para os dados geoestatísticos binários propostos neste trabalho e, por fim, a secção 4 apresenta as conclusões e recomendações para trabalhos futuros.

4.2 Delineamento de Amostragem

Dos estudos já realizados, existem duas abordagens diferentes para os delineamentos de amostragem: abordagem baseada em desenhos e abordagem baseada em modelos [67, 14]. A seguir vai-se apresentar a descrição das duas abordagens de delineamentos de amostragem referidas.

4.2.1 Abordagens baseadas em desenhos

Segundo Brus [14] e Liu [46], a abordagem baseada em desenhos é frequentemente referida como teoria de amostragem clássica, onde a estocasticidade é introduzida pela amostragem, e os locais da amostra são selecionados por um critério de seleção aleatória pré-determinado. Na amostragem clássica, várias amostras podem ser selecionadas numa determinada área usando o mesmo delineamento de amostragem.

Este tipo de amostragem, incluí amostragem aleatória simples, estratificada e sistemática. [19, 71].

Existem vários delineamentos de amostragem convencionais, mas para este trabalho vai-se descrever basicamente três tipos amplamente utilizados em pesquisas: amostragem aleatória simples, amostragem aleatória Sistemática e Estratificada.

4.2.1.1 Amostragem aleatória simples (SSD)

O delineamento de amostragem aleatória simples é a forma mais básica de amostragem probabilística e fornece a base teórica para as formas mais complexas. Existem duas formas de se obter uma amostra aleatória simples: com reposição, em que a mesma unidade pode ser incluída mais de uma vez na amostra, e sem reposição, em que todas as unidades da amostra são distintas. Este tipo de amostragem, consiste em fazer uma lista de todas as unidades de observação da população; esta lista é o quadro de amostragem. Em uma amostragem aleatória simples, a unidade de amostragem e a unidade de observação coincidem. Cada unidade recebe um número e uma amostra é selecionada de modo que

cada amostra possível de tamanho n tenha a mesma chance de ser a amostra realmente selecionada [47].

4.2.1.2 Delineamento de amostragem Sistemática

Este delineamento de amostragem é muito utilizado pela sua simplicidade e conveniência, e consistem em elaborar uma lista ordenada de todos os possíveis pontos de amostragem, identificados com um número de 1 até N, depois, divide-se esse conjunto de índices em k grupos, dado por $k = N/n$, onde n é o tamanho de amostra desejado. A seguir, escolhe-se aleatoriamente um número inteiro R de entre 1 e k e seleciona-se dentro do primeiro grupo o elemento com essa posição, que corresponderá ao primeiro elemento da amostra. Os restantes elementos que farão parte da amostra são selecionados a partir da posição do primeiro elemento acima escolhido através de uma sucessão aritmética dada por $R, R+k, R+2k, \ldots, R+(n-1)k$. Em muitas situações o delineamento de amostragem sistemática é estatisticamente eficiente quando comparada a outros delineamento de amostragem [47].

4.2.1.3 Amostragem aleatória Estratificada

De acordo com Cochran [19], este delineamento de amostragem é muito utilizado em pesquisas relacionadas com a pesca e apoiado pela teoria estatística clássica. O delineamento de amostragem aleatória estratificada quando considerada no espaço, visa dividir a área de estudo em pequenas áreas (estratos) de modo que tenham a maior homogeneidade possível em relação à quantidade alvo a considerar, mas que sejam heterogéneas entre si. Para se obter uma amostra estratificada, extrai-se um ponto aleatoriamente de cada estrato de forma independente e, em seguida, agrupa-se as informações para obter estimativas populacionais gerais. Este tipo de delineamento de amostragem assegura que todas as subáreas que compõe a região de estudo sejam amostradas.

4.2.2 Abordagens baseadas em modelos

Na abordagem baseada em modelos, o processo de observação é modelado como um processo estocástico. Este processo estocástico é uma abstração matemática usada para descrever a realidade, onde as probabilidades de ocorrência dos resultados elementares desse processo, os campos de valores não são conhecidos, mas devem ser modeladas [14].

De acordo com Gruijter e Braak [25], em geoestatística, a abordagem baseada em modelos é usada principalmente para predição de médias espaciais. Esta abordagem trata o valor associado a qualquer local como um valor não fixo, mas sim como um valor aleatório. O conjunto de valores associados a todas as realizações possíveis na região de estudo é considerado como apenas uma realização de um processo subjacente, onde algumas características deste processo são assumidas que são conhecidas e essas premissas são formalizadas num modelo geoestatístico, que desempenha o papel que é chamado de modelo de superpopulação na teoria de amostragem.

Na abordagem baseada em modelos, a única fonte de estocasticidade é o processo subjacente postulado, e a inferência baseia-se principalmente no modelo formulado [14].

As abordagens baseadas em modelos e em desenhos amostrais apresentam algumas diferenças, dentre várias destacam-se as seguintes: A abordagem baseada em desenhos, requer menos suposições do que a abordagem baseada em modelos. Na abordagem baseada em desenhos, os locais de amostragem são selecionados por amostragem probabilística e a inferência estatística (por exemplo a estimativa da média espacial) é independente do delineamento de amostragem, enquanto que na abordagem baseada em modelos, não precisa de requisitos para a seleção de locais de amostragem, geralmente são selecionados utilizando a amostragem intencional (direcionada), como por exemplo numa grelha centralizada. [75, 25].

Geralmente, a abordagem baseada em desenhos tem sido a melhor escolha se estivermos interessados na função de distribuição cumulativa espacial para a área como um todo ou para um número restrito de subáreas, e além disso, a validade do resultado realmente importa (validade do resultado é mais importante do que a eficiência), enquanto que a abordagem baseada em modelos é melhor opção se estivermos interessados num mapa que descreva os valores de muitas áreas pequenas, como por exemplo *pixels*, e queremos prever esses valores com maior precisão possível (a eficiência é mais importante que a validade) [14].

Ainda no contexto geoestatístico, um delineamento envolve a seleção e otimização de (maximizar ou minimizar) uma função objetivo adequada [46, 28] que são desenvolvidos com base num determinado critério de delineamento selecionado (como por exemplo a minimização da média ou minimização de variância na estimação) [89].

De acordo com McBratney e Webster [54] e Van Groenigen *et al.* [89], geralmente, a maior parte das abordagens assumem que todos os parâmetros que definem a estrutura de covariância dos dados são conhecidos e o objetivo é de identificar o melhor delineamento que pode minimizar a variância na estimativa. Existem casos em que estas abordagens não são aplicáveis, devido ao conhecimento limitado ou mesmo desconhecimento, sendo necessário recorrer aos dados em análise para a sua estimação da estrutura de covariância dos dados (por exemplo dados do pescado). Todavia, existem outras abordagens alternativas que podem ser usadas que não exigem informações sobre a estrutura de covariância dos dados [74, 46]. Uma das abordagens alternativas foca-se na qualidade da estimativa de semivariograma, que é um elemento principal para análise geoestatística (mais detalhes em 4.2.2.2).

A seguir, descrevem-se alguns critérios de seleção de delineamento de amostragem da abordagem baseada em modelos (que assentam na otimização de uma função objetivo) usados em geoestatística, propostos na literatura.

4.2.2.1 Minimização da média da distância mais curta (MMSD)

Este tipo de critério (MMSD) exige que todos os pontos de amostragem sejam distribuídos regularmente em toda a região de estudo. A distribuição regular pode ser formulada como uma distribuição para qual a média das distâncias entre um ponto escolhido arbitrariamente dentro da região e o seu ponto de amostragem mais próximo é mínimo. Sejam todos os N pontos de uma malha fina como pontos disponíveis para seleção, o critério (MMSD) consiste em minimizar a média das distâncias de todos os pontos da malha até o ponto de amostragem mais próximo:

$$\phi_{MMSD}(S) = \frac{1}{N}\sum_{i}^{N} d(x_i, S), \tag{4.1}$$

onde x é o ponto da grelha e $d(x,S)$ é distância entre o ponto da grelha x e o ponto de amostragem mais próximo no conjunto de amostragem S. O número de pontos da grelha deve ser maior em relação ao número de pontos de amostra para alcançar um valor confiável de ϕ_{MMSD} [44, 76].

4.2.2.2 Distribuição uniforme de pares de pontos para estimativa de variograma (WM)

O critério WM (Warrick e Myers [93]) consiste em otimizar o ajuste da distribuição realizada de pares de pontos para estimar o variograma experimental a uma distribuição ideal definida a priori. Este critério não requer suposições sobre a estrutura de correlação espacial, mas depende apenas das distâncias entre os pontos de amostragem [46, 89, 74, 76], dado por:

$$\phi_{WM}(S) = \sum_{i=1}^{n_c} [a w_i (f_i^* - f_i)^2 + b M_i], \tag{4.2}$$

onde S é um conjunto de pontos amostrais, n_c é o número de classes de distâncias (também chamado de número de classes de atraso), f é a distribuição de pares de pontos realizada, f^* é a distribuição de pares de pontos esperada, a, b *e* w são os coeficientes de ponderação definidos pelo utilizador, e M as outras possíveis objeções responsáveis em classe de distâncias. Por questões de simplificação, geralmente assume-se que o b é zero e a e w são ambos iguais a um. Entretanto, se $b = 0$, a função de minimização é uma simples soma de quadrados do desvio entre a distribuição esperada e a distribuição realizada. Para uma determinada dimensão de amostras (n) e número de classes de distâncias (n_c), o número esperado de pares de pontos para uma distribuição uniforme é dado por

$$f_i^* = \frac{n(n-1)}{2n_c}. \tag{4.3}$$

O terceiro critério que também tem sido usado é a combinação de MMSD e WM (MMDS+WM) (mais detalhes ver [46, 89, 76]).

A seguir vai-se descrever os critérios de amostragem que serão usados neste trabalho para construir os respetivos esquemas de amostragem otimizados.

4.2.3 Abordagem Proposta

A abordagem baseada em desenhos e as abordagens baseadas em modelos antes descritas são adequadas a dados geoestatísticos tradicionais, onde a característica de interesse que se observa é Gaussiana; no entanto neste trabalho, estamos interessados em dados binários, pelo que os critérios anteriores poderão não se conseguir aplicar nestes casos.

A abordagem proposta é considerada uma abordagem adaptativa. De acordo com Chipeta *et al.* [17], delineamentos geoestatísticos adaptativos são aqueles nos quais os locais amostrados são escolhidos em grupos numa sequência temporal, e os locais em qualquer grupo utilizam dados de grupos anteriores para otimizar a recolha de dados para o objetivo do estudo, com base num critério calculado tendo em conta o grupo anterior. O critério do delineamento de amostragem adaptativa assegura que os dados são recolhidos apenas de locais que possam ser úteis para fornecer informações adicionais. As abordagens tradicionais acima referidas não são adaptativas.

Nesta subsecção, será apresentada a abordagem proposta para o delineamento de amostragem, a qual será construída tendo por base a amostragem por modelos com o intuito de obter o critério mais adequado ao problema em análise, e assim contribuindo para a implementação de novos métodos de seleção em dados geoestatísticos binários.

4.2.3.1 Maximização do risco estimado (MaxRSD)

Estamos interessados em construir uma proposta de delineamento amostral geoestatístico para dados binários (modelos com parâmetros desconhecidos), com base num critério que procura amostrar com maior intensidade onde se preconiza uma prevalência mais elevada (adaptativo).

Sendo a estrutura de covariância desconhecida, faz-se primeiramente a estimação de um modelo que melhor se adequa a um conjunto de dados já observado; seguidamente, faz-se a previsão do modelo numa malha fina de pontos que cobre a área em estudo que constituem o conjunto de potenciais pontos a incluir na amostra, obtendo-se as probabilidades estimadas em cada ponto; finalmente, amostram-se destes um número de pontos n, fixo, com probabilidade de seleção proporcional às referidas probabilidades estimadas (risco). Desta forma escolhemos para amostrar os locais de maior risco estimado.

A este critério de seleção de delineamento, chamamos de Maximum risk sampling design (MaxRSD), que depende de n (número de pontos a amostrar) e m (modelo estimado), MaxRSD(n, m), e pode ser obtido como se segue:

1. Considerando $X0 = (x_1, \ldots, x_{n_0})$ pontos iniciais, estima-se o modelo geoestatístico dado pela equação (3.23);

2. Estima-se então o risco nos pontos das potenciais localizações de amostragem numa grelha fina (*mesh*) que cobre a região em estudo $X1 = (x_1, \ldots, x_{n_1})$ (não observados), onde $n < n_1$ e determinam-se as probabilidades de inclusão de cada um desses

pontos, como as probabilidades projetadas nos referidos pontos, padronizadas para somarem 1; e

3. Selecionam-se n pontos de $X1$ de acordo com as probabilidades de inclusão obtidas em 2, obtendo-se uma proposta de delineamento de amostragem.

4.2.3.2 Maximização da variabilidade associada ao risco estimado (MaxVRSD)

Uma alternativa interessante para um delineamento amostral geoestatístico para dados binários (modelos com parâmetros desconhecidos), é usar um critério que procura amostrar com maior intensidade onde a prevalência é estimada com maior incerteza (adaptativo). Sendo a estrutura de covariância desconhecida, faz-se inicialmente estimação de um modelo que melhor se adequa aos dados em análise; depois, faz-se a previsão do modelo numa malha fina de pontos que cobre a área em estudo (que constituem o conjunto de potenciais pontos a incluir na amostra) obtendo-se as probabilidades estimadas em cada ponto e determinam-se os correspondentes erros padrão; finalmente, amostram-se das potenciais escolhas um número de pontos n, fixo, com probabilidade de seleção proporcional aos erros padrão do risco. Desta forma escolhemos para amostrar os locais de maior incerteza no risco estimado.

A este critério de seleção de delineamento, chamamos de Maximum variance risk sampling design (MaxVRSD), que depende de n (número de pontos a amostrar) e de m (modelo a estimar) MaxVRSD(n, m) e pode ser obtido como se segue:

1. Considerando $X0 = (x_1, \ldots, x_{n_0})$ pontos iniciais, estima-se o modelo geoestatístico dado pela equação (3.23);

2. Estima-se então o risco nos pontos das potenciais localizações de amostragem numa grelha fina (*mesh*) que cobre a região em estudo $X1 = (x_1, \ldots, x_{n_1})$ (não observados), onde $n < n_1$ e determinam-se as probabilidades de inclusão de cada um desses pontos com base nos erros padrão do risco projetado nesses pontos, que são padronizados de forma a formarem uma distribuição de probabilidade sobre esses pontos (erro padrão no ponto i/soma de todos os erros padrão nos pontos da grelha); e

3. Selecionam-se n pontos de $X1$ de acordo com as probabilidades de inclusão obtidas em 2, obtendo-se uma proposta de delineamento de amostragem.

A operacionalização destes delineamentos pode depender do seu contexto. Por exemplo, se o contexto for de fiscalização de embarcações no mar, e a amostragem for realizada com recurso a um navio de certa envergadura, poderá ter de se considerar uma política de varrimento da área em que o navio percorre uma certa rota e os pontos de amostragem previstos pelo delineamento são visitados usando barcos ou botes mais pequenos.

4.2.4 Avaliação do delineamento de amostragem

Utilizando os valores estimados dos diferentes parâmetros do modelo adotado no passo 1 e re-estimado com os pontos obtidos no passo 3 das subsecções 4.2.3.1 e 4.2.3.2 é possível avaliar e comparar os resultados em termos do risco estimado e assim identificar o delineamento ótimo.

De forma a se poder avaliar e comparar as propostas de delineamento apresentadas, vai usar-se a raiz do erro quadrático médio (RMSE) de predição, dado por:

$$RMSE = \sqrt{\frac{1}{n}\sum_{\alpha=1}^{n}(z_\alpha^* - z_\alpha)^2}, \tag{4.4}$$

onde n é o número de pontos amostrados, z_α e z_α^* são respetivamente valores preditos no local α, com base no modelo geoestatístico adotado para os dados disponíveis, e estimado no local α com base na proposta de delineamento de amostragem e usando os dados preditos antes referidos. Note-se que os valores preditos amostrados não podem ser usados diretamente para obter os valores estimados, têm de ser convertidos em sucessos ("1") ou insucessos ("0"). Tal é feito comparando o valor predito com uma probabilidade p que idealmente seria 0.5 mas que em estudos de baixa prevalência pode ter de ser diminuída para evitar a questão de poucas observações sucesso ("1") [76].

A proposta de delineamento mais adequada será o que apresentar menor valor de RMSE.

4.3 Construção e análise de delineamentos de amostragem para ações de fiscalização da atividade pesqueira no Algarve

Nesta secção vai-se fazer a aplicação dos critérios descritos em 4.2.3 para seleção de delineamento de amostragem aos dados geoestatísticos binários referentes as infrações pesqueiras, obtidos em ações de fiscalização marítima no Comando de Zona do Sul de Portugal nos anos de 2013 a 2017. Para efeitos de validação das propostas obtidas serão comparadas com o delineamento obtido por amostragem aleatória simples descrita em 4.2.1.1.

4.3.1 Área de estudo

A informação disponível refere-se a dados georeferenciados relativos a ações de fiscalização efetuadas pela Marinha Portuguesa em toda costa Portuguesa, com mais de duzentos e onze mil (211000) pontos fiscalizados em 22 anos (de 1998 a 2019), com uma média de 9620 pontos fiscalizados por ano.

Para este estudo, consideram-se cinco anos (de 2013 a 2017), descritos na secção 3.7 para o comando de zona do Sul (região costeira do Algarve), restringidos ao segundo período do dia compreendido entre as 06:01-12:00 horas, por ser um período com maior

números de infrações. A malha fina considerada para este estudo é a *mesh* definida na figura 3.9.

4.3.2 Resultados e discussão

Com base no modelo dado pela equação (3.23), vai-se considerar nesta aplicação o modelo dado por

$$y_i \sim Bernoulli(\pi_i) \tag{4.5}$$

onde π_i é a probabilidade da presença da referida presumível infração. Então, especifica-se em $logit(\pi_i) = ln\left(\frac{\pi_i}{1-\pi_i}\right)$ um modelo linear incluindo intercepto β_0, e um efeito aleatório espacial $\boldsymbol{u}$

$$logit(\pi_i) = \beta_0 + \boldsymbol{u}, \quad i = 1, \ldots, n \tag{4.6}$$

onde β_0 é o intercepto e $\boldsymbol{u}$ é o campo aleatório. Os valores dos parâmetros considerados para o modelo 4.6 são os estimados de acordo com o que foi apresentado no capítulo 3, utilizando os dados descritos na subsecção 4.3.1.

Na figura 4.1 apresentam-se os valores estimados e preditos para os pontos da grelha considerada com base no modelo adotado em (4.6) bem como os correspondentes erros padrão.

Tabela 4.1: Resumo das distribuições de probabilidade a posteriori para o 2º período dos anos 2013 a 2017.

Parâmetro	Média	SD	Quantil 0.025	Quantil 0.5	Quantil 0.975
β_0	-2.711	0.222	-3.214	-2.688	-2.341
σ_u^2	4.029	1.379	2.020	3.791	7.378
R	4.644	1.279	2.559	4.504	7.546

A seguir, apresentam-se três esquemas de amostragem, delineamento aleatório simples, SSD(n); Maximização do risco estimado, MaxRSD($n, Modelo(4.6)$); e Maximização da variabilidade associada ao risco estimado, MaxVRSD($n, Modelo(4.6)$). Para a avaliação deste estudo, consideramos primeiramente o valor de probabilidade $p = 0.4$ (limite de probabilidade usada para definir os sucessos e insucessos), de acordo com o descrito na subsecção 4.2.4, com três dimensões de amostras diferentes (50, 100 e 200).

As figuras 4.2, 4.4 e 4.6 apresentam os mapas dos três esquemas de amostragem para n=50, 100 e 200 pontos, respetivamente. Fazendo a amostragem com base nos critérios definidos neste estudo, SSD(n), MaxRSD(n, m) e MaxVRSD(n, m), voltou-se a estimar o modelo tendo-se obtido os mapas de valores ajustados ou mapas de risco apresentados pelas figuras 4.3, 4.5 e 4.7, a partir dos quais será possível avaliar a sua qualidade usando o RMSE.

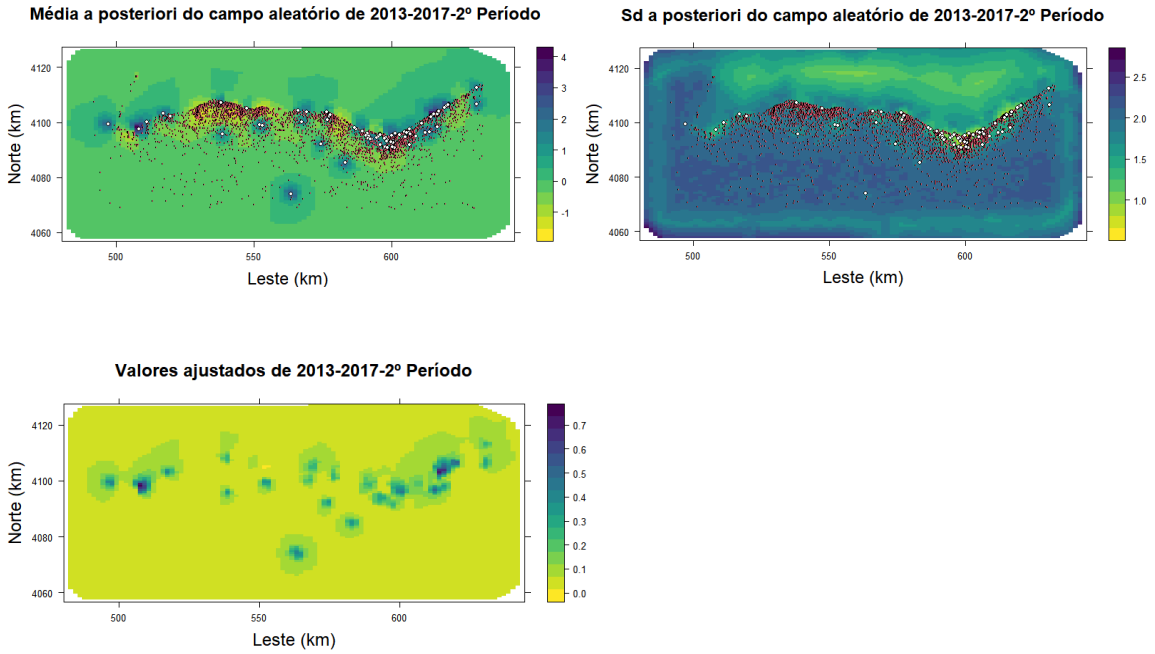

Figura 4.1: Mapa da média (em cima-esquerda) e mapa do erro padrão (em cima-direita) a posteriori do campo aleatório; mapa de risco (em baixo), para os anos de 2013-2017.

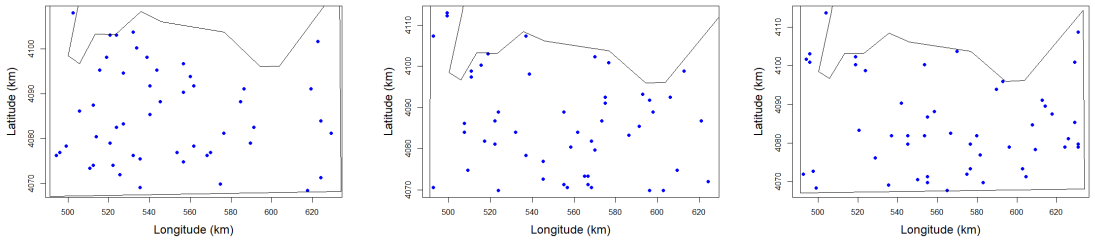

Figura 4.2: Ilustração de distribuição espacial de dados simulados das infrações pesqueiras a partir de três esquemas de amostragem para 50 pontos. Da esquerda para a direita, esquema de amostragem SSD; esquema de amostragem MaxRSD; esquema de amostragem MaxVRSD.

Para cada uma das diferentes propostas de delineamentos de amostragem, procede-se à avaliação do delineamento mais adequado para cada uma das dimensões escolhidas, através da medida RMSE de acordo com secção 4.2.4.

Pode-se verificar na tabela 4.2 que para todos os tamanhos de amostra 50, 100 e 200 pontos o delineamento de amostragem MaxRSD apresenta o melhor valor de RMSE.

Nos mapas dos valores ajustados apresentados pelas figuras 4.3, 4.5 e 4.7, verifica-se que para todos os tamanhos de amostra (50, 100 e 200), o delineamento de amostragem SSD não inclui nenhumas observações sucesso ("1"); isto sucedeu-se também para o delineamento MaxRSD, apenas para $n = 50$, e para o delineamento de amostragem MaxVRSD para $n = 50$ e $n = 100$. Como foi descrito em 4.2.4, vamos baixar o valor da probabilidade

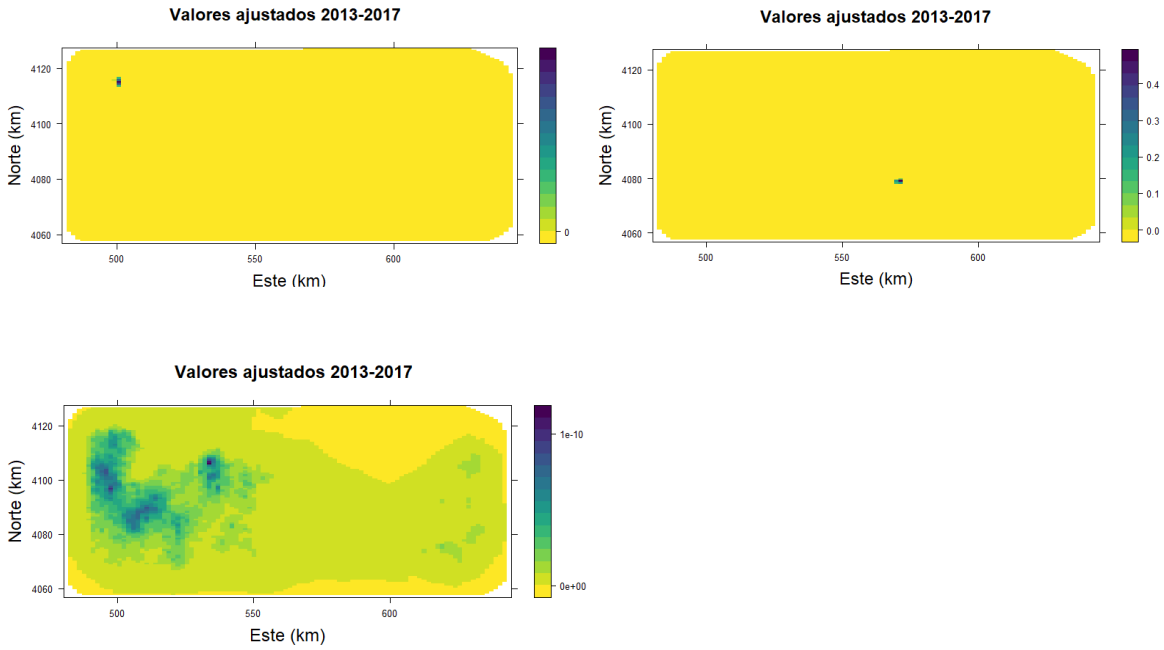

Figura 4.3: Mapas de risco de dados simulados das infrações pesqueiras a partir de três esquemas de amostragem para 50 pontos. Esquema de amostragem SSD (em cima-esquerda); esquema de amostragem MaxRSD (em cima-direita); esquema de amostragem MaxVRSD (em baixo).

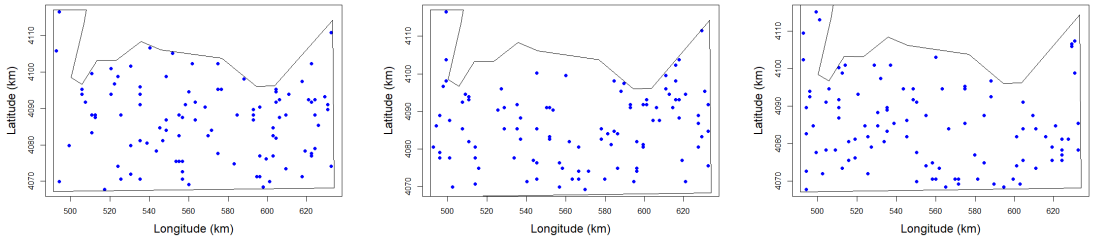

Figura 4.4: Ilustração de distribuição espacial de dados simulados das infrações pesqueiras a partir de três esquemas de amostragem para 100 pontos. Da esquerda para a direita, esquema de amostragem SSD; esquema de amostragem MaxRSD; esquema de amostragem MaxVRSD.

p para $p = 0.2$, então temos os resultados apresentados nas figuras 4.8, 4.10 e 4.12, que ilustram os três esquemas amostrais em análise neste estudo, e as figuras 4.9, 4.11 e 4.13, ilustram os valores ajustados para $n = 50$, $n = 100$ e $n = 200$.

A seguir vai-se proceder a avaliação do delineamento mais adequado para cada uma das dimensões escolhidas, através da medida RMSE descrito na secção 4.2.4.

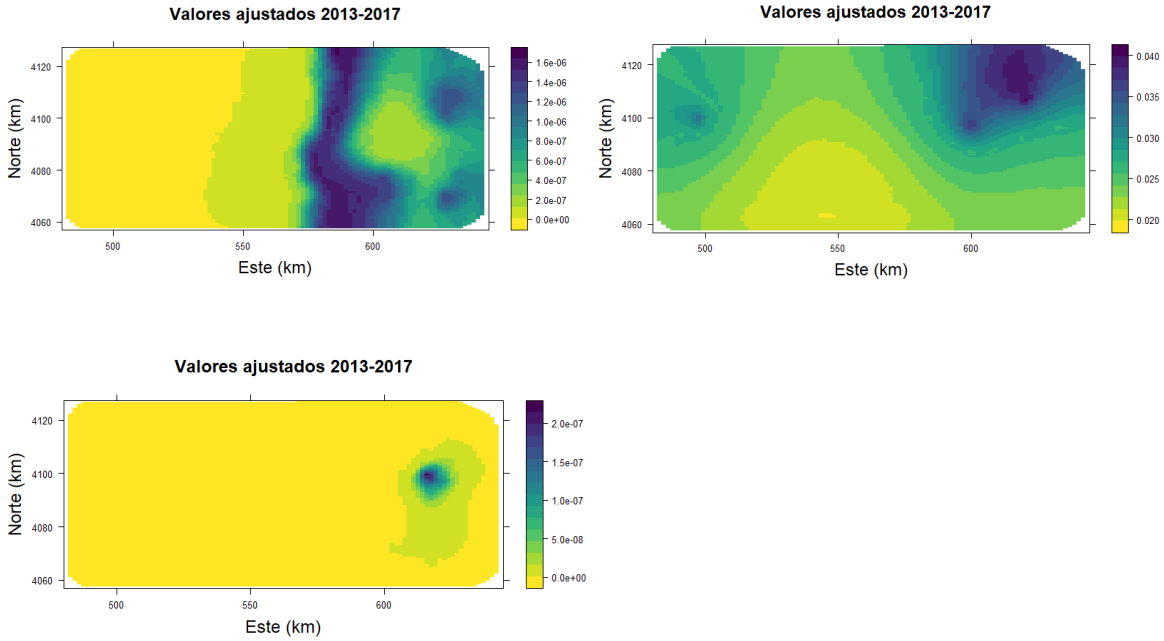

Figura 4.5: Mapas de risco de dados simulados das infrações pesqueiras a partir de três esquemas de amostragem para 100 pontos. Esquema de amostragem SSD (em cima-esquerda); esquema de amostragem MaxRSD (em cima-direita); esquema de amostragem MaxVRSD (em baixo).

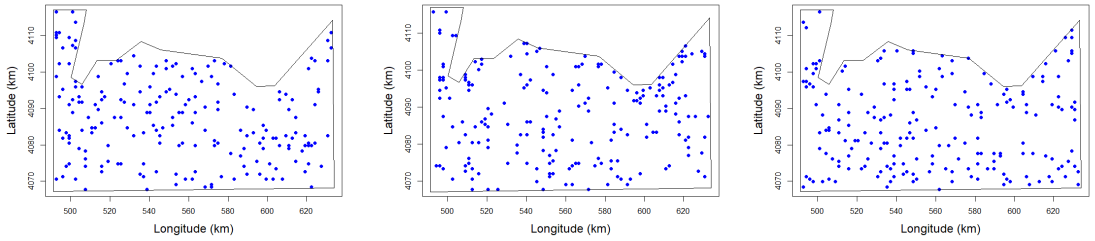

Figura 4.6: Ilustração de distribuição espacial de dados simulados das infrações pesqueiras a partir de três esquemas de amostragem para 200 pontos. Da esquerda para a direita, esquema de amostragem SSD; esquema de amostragem MaxRSD; esquema de amostragem MaxVRSD.

Pode-se verificar na tabela 4.3 que para os tamanhos de amostra 50 e 100 pontos o delineamento de amostragem MaxVRSD apresenta o melhor valor de RMSE, enquanto que para o tamanho de amostra 200 pontos, o delineamento de amostragem MaxRSD apresentou o melhor valor de RMSE.

Neste trabalho, apresentamos delineamentos de amostragem baseados em desenhos e em modelos [67, 14].

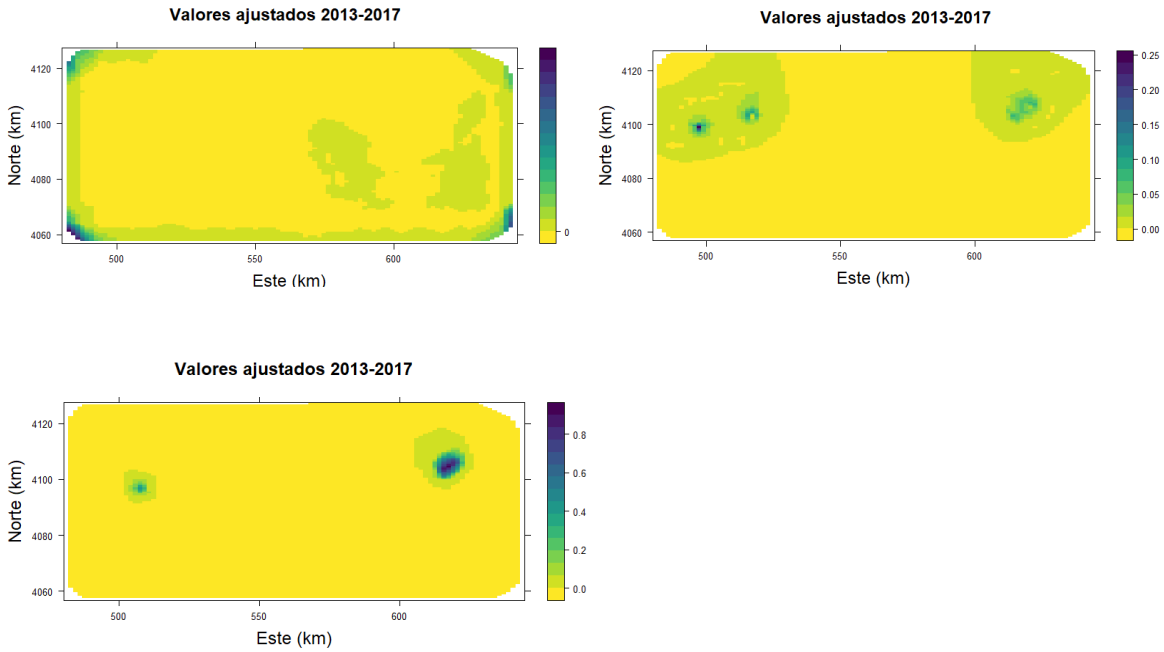

Figura 4.7: Mapas de risco de dados simulados das infrações pesqueiras a partir de três esquemas de amostragem para 200 pontos. Esquema de amostragem SSD (em cima-esquerda); esquema de amostragem MaxRSD (em cima-direita); esquema de amostragem MaxVRSD (em baixo).

Tabela 4.2: Resultados de RMSE para os esquemas de amostragem propostos para $p = 0.4$.

	RMSE		
Tamanho da amostra	SSD	MaxRSD	MaxVRSD
50	0.08222633	**0.08222604**	0.08222984
100	0.08223057	**0.07838906**	0.08223078
200	0.08222861	**0.07299765**	0.08972003

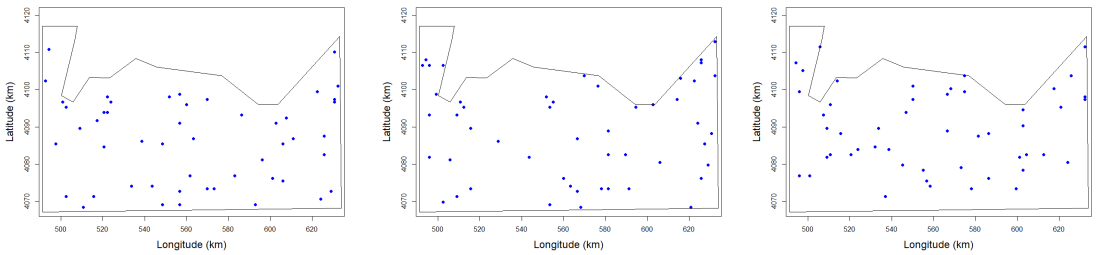

Figura 4.8: Ilustração de distribuição espacial de dados simulados das infrações pesqueiras a partir de três esquemas de amostragem para 50 pontos. Da esquerda para a direita, esquema de amostragem SSD; esquema de amostragem MaxRSD; esquema de amostragem MaxVRSD.

Em geoestatística um dos delineamentos que é frequentemente utilizado para a seleção da amostra é o delineamento de amostragem sistemática [44]. Porém, para este estudo

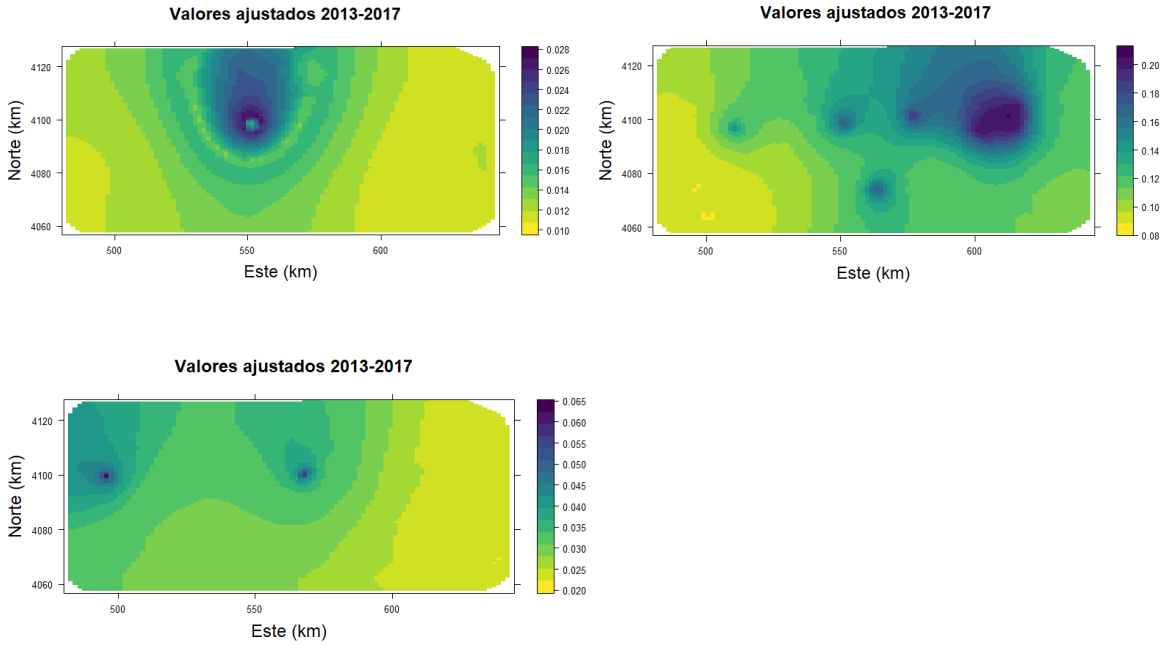

Figura 4.9: Mapas de risco de dados simulados das infrações pesqueiras a partir de três esquemas de amostragem para 50 pontos. Esquema de amostragem SSD (em cima-esquerda); esquema de amostragem MaxRSD (em cima-direita); esquema de amostragem MaxVRSD (em baixo).

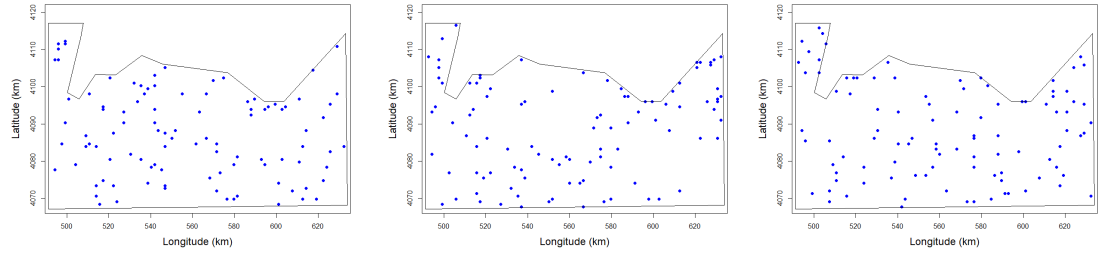

Figura 4.10: Ilustração de distribuição espacial de dados simulados das infrações pesqueiras a partir de três esquemas de amostragem para 100 pontos. Da esquerda para a direita, esquema de amostragem SSD; esquema de amostragem MaxRSD; esquema de amostragem MaxVRSD.

foi experimentado este esquema de amostragem, mas como temos poucos locais de amostragem com probabilidade estimada elevada, este tipo de delineamento de amostragem tem dificuldade em encontrar pontos que sejam informativos em termos da probabilidade da infração, claramente esta não é uma boa estratégia para este problema em concreto.

Inicialmente, foram apresentados resultados do delineamento de amostragem aleatório simples [47] que foram utilizados para comparar aos critérios de delineamento de amostragem propostos, maximização do risco estimado (MaxRSD) e maximização da variabilidade associada ao risco estimado (MaxVRSD) através do RMSE [76] com objetivo

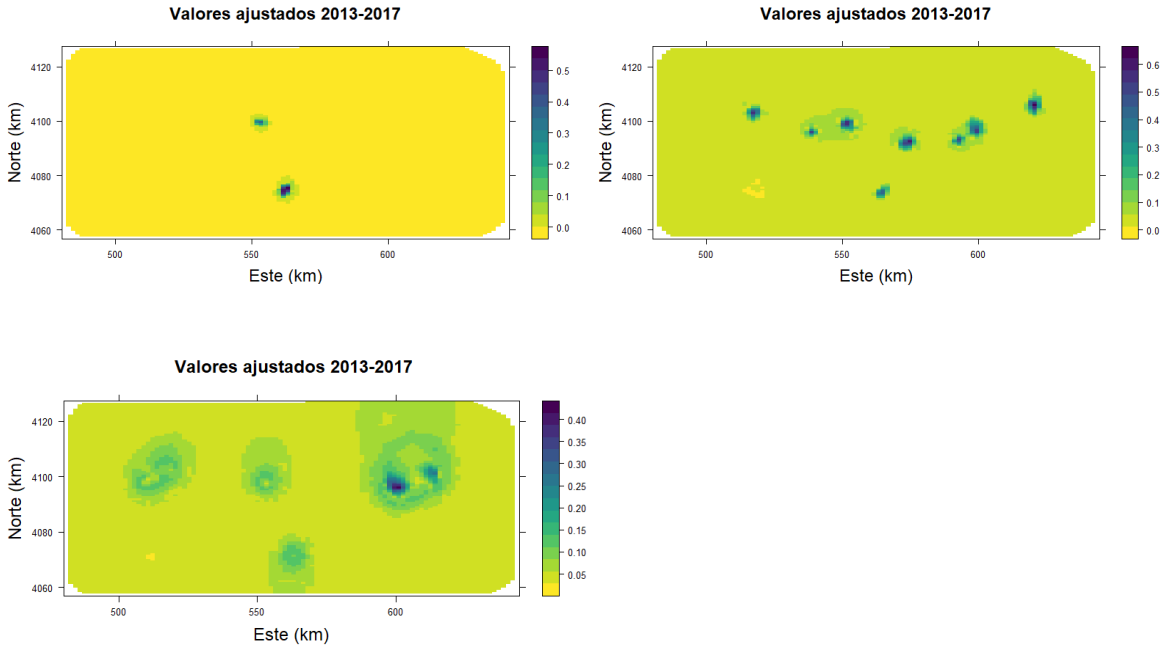

Figura 4.11: Mapas de risco de dados simulados das infrações pesqueiras a partir de três esquemas de amostragem para 100 pontos. Esquema de amostragem SSD (em cima-esquerda); esquema de amostragem MaxRSD (em cima-direita); esquema de amostragem MaxVRSD (em baixo).

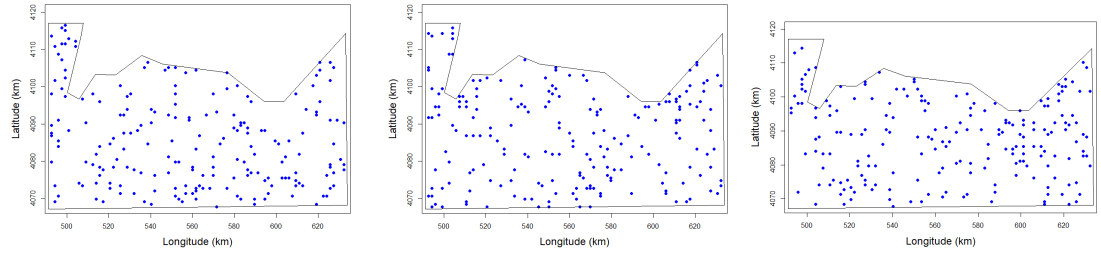

Figura 4.12: Ilustração de distribuição espacial de dados simulados das infrações pesqueiras a partir de três esquemas de amostragem para 200 pontos. Da esquerda para a direita, esquema de amostragem SSD; esquema de amostragem MaxRSD; esquema de amostragem MaxVRSD.

de verificar o melhor delineamento para o problema em análise neste estudo.

Vários critérios de seleção de delineamento de amostragem têm sido amplamente usados em estudos geoestatísticos, tais como MMSD e WM [44, 76, 93, 46]. Estes critérios de seleção são baseados na minimização da média da distância mais curta entre dois pontos e na distribuição uniforme de pares de pontos para estimativa do variograma. Os critérios de seleção de delineamento de amostragem propostos por este estudo são baseados na probabilidade estimada. Estes critérios de seleção de delineamento de amostragem apresentam uma clara vantagem por conseguir amostrar com maior incidência em locais

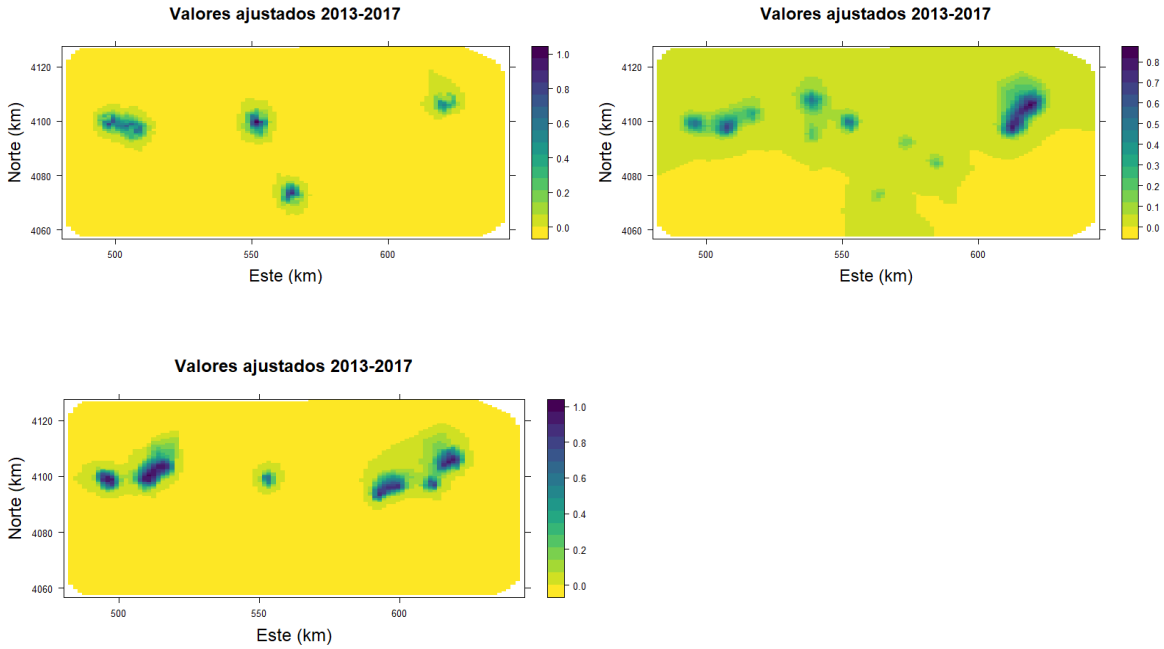

Figura 4.13: Mapas de risco de dados simulados das infrações pesqueiras a partir de três esquemas de amostragem para 200 pontos. Esquema de amostragem SSD (em cima-esquerda); esquema de amostragem MaxRSD (em cima-direita); esquema de amostragem MaxVRSD (em baixo).

Tabela 4.3: Resultados de RMSE para os esquemas de amostragem propostos para $p = 0.2$.

	RMSE		
Tamanho da amostra	SSD	MaxRSD	MaxVRSD
50	0.07671932	0.07276609	**0.06062947**
100	0.08882784	0.059403201	**0.04896903**
200	0.08469035	**0.07071637**	0.07991713

onde a probabilidade de se cometer presumíveis infrações é alta e onde a prevalência é estimada com maior incerteza.

De acordo com os valores de RMSE descritos nas tabelas 4.2 e 4.3, os dois critérios de delineamento de amostragem propostos são os melhores, podendo assim ser uma boa alternativa para o problema em estudo.

4.3.2.1 Operacionalização do esquema de amostragem

A definição das rotas de fiscalização é uma questão muito complexa. Há vários fatores que devem ser levados em consideração quando se está a planear uma ação de fiscalização. Ao se fazer um filtro para uma determinada zona do país, vai-se estar subjacente ao tipo de fiscalização que se efetuam nessa zona, isto é, leva-se em consideração as características intrínsecas das embarcações que circulam nesta zona. Neste caso, os gráficos com as rotas

apresentados nas figuras 4.14 e 4.15, para a região costeira do Algarve é um dos elementos dessas características da zona. Quanto ao fator tempo de navegação disponível para fiscalizar, existe um período máximo que é determinado, não só pelo combustível, mas também pelo mantimento. E o tipo de navio também influencia o período de permanência no mar, onde podemos ter uma lancha ou uma corveta.

A Marinha, na sua forma de atuação comum, tem ao seu cargo o Dispositivo Naval Padrão que é distribuído pelas várias zonas do país de forma que todos os meios estejam empenhados durante todo o ano. No caso do comando de zona do Sul, o requisito são 3 lanchas em permanência. Mas isso não impede que uma corveta ou navio de patrulha oceânica (NPO) seja empenhada no Sul para fiscalização.

Consideremos o seguinte exemplo: quando temos uma amostra de 50 pontos, uma ideia de operacionalização na prática é apresentada na figura 4.14.

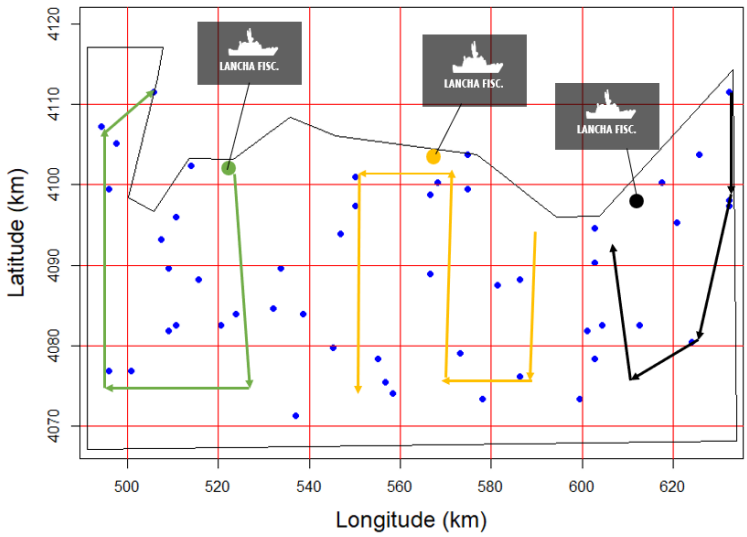

Figura 4.14: Mapas do esquema de amostragem MaxVRSD, para 50 pontos com três rotas de fiscalização propostas.

Na figura 4.14, apresenta três propostas de possíveis rotas de fiscalização feitas por Lanchas alocadas ou empenhadas para cada capitania da região costeira do Algarve. A primeira proposta de rota de fiscalização definida pela linha preta tem início de atividade na zona Este da região costeira do Algarve, concretamente na Vila Real de Santo António, junto a fronteira com a Espanha e com o fim da atividade na região de Olhão; a segunda proposta de rota de fiscalização, definida pela linha amarela, tem inicio de atividade na região costeira de Faro e fim da atividade na região de Portimão; e por fim, a terceira proposta tem início de atividade na região costeira de Alvor e com o fim da atividade na zona Oeste da região costeira do Algarve, concretamente na região de Vila do Bispo.

Na figura 4.15, apresenta-se uma proposta de definição de rotas de fiscalização feitas por três Lanchas e uma Corveta, onde as Lanchas fazem a fiscalização por capitanias nas

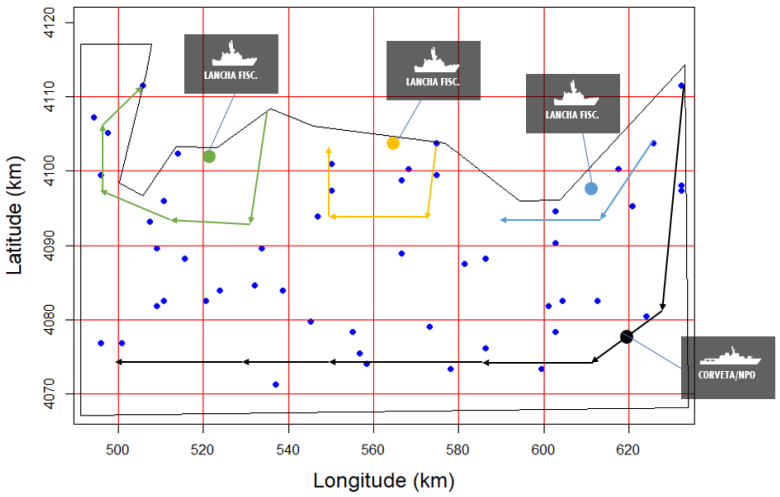

Figura 4.15: Mapas do esquema de amostragem MaxVRSD, para 50 pontos com duas rotas de fiscalização propostas.

zonas próximas a linha da costa e a Corveta faz a fiscalização em zonas mais afastadas a linha da costa. A linha contínua de cor preta representa uma proposta de uma possível rota de fiscalização feita com uma Corveta, com início da atividade na zona Este da região costeira do Algarve, concretamente na Vila Real de Santo António, junto a fronteira com a Espanha e com o fim da atividade na região de Sagres. A segunda rota proposta de fiscalização, é representada pela linha contínua de cor azul, feita por uma Lancha, com início de atividade na região de Tavira e com fim na região Faro. Terceira rota proposta é representada pela linha amarela, com início de atividade na região costeira de Quarteira e com o fim na região Albufeira. Quarta rota proposta de fiscalização tem início em Lagos e fim na zona Oeste da região costeira do Algarve, concretamente na região de Vila do Bispo.

4.4 Conclusões

O delineamento de amostragem é de extrema importância para dar resposta aos objetivos de um determinado estudo, sendo sua resposta limitada, principalmente pelo tipo de delineamento empregue.

Para este estudo foi utilizado a abordagem baseada em modelos por ser a mais adequada em pesquisas desta natureza.

Do conjunto total de dados que engloba os quatro períodos do dia, foram utilizados apenas dados do segundo período, compreendido entre as 06:01-12:00 horas, por ser o período com maior número de infrações.

Construiram-se delineamentos de amostragem, baseado na maximização do risco estimado (MaxRSD) e na maximização da variabilidade associada ao risco estimado (MaxVRSD), que depois foram comparados ao delineamento aleatório simples. Destes, escolheu-se o melhor delineamento para diferentes tamanhos de amostra ($n = 50$, $n = 100$

e $n = 200$), utilizando para tal uma probabilidade de $p = 0.4$ para a transposição da estimativa do risco para valores de uma variável binária. Para todos os tamanhos de amostra o MaxRSD apresentou o melhor valor de RMSE. Verificou-se também que com $p = 0.4$, o delineamento SSD não inclui nenhumas observações sucesso "1"em todos os tamanhos de amostra (baixa prevalência), enquanto que o delineamento MaxRSD, não inclui observações sucesso "1"apenas para o tamanho de amostra $n = 50$, e o delineamento MaxVRSD, não inclui nenhumas observações sucesso "1"nos tamanhos de amostra $n = 50$ e $n = 100$. Portanto, baixou-se o valor da probabilidade para $p = 0.2$, onde verificou-se que para tamanhos de amostra 50 e 100 pontos, o delineamento de amostragem MaxVRSD apresentou o melhor valor de RMSE, enquanto que para o tamanho de amostra de 200 pontos, o delineamento de amostragem MaxVRSD, apresentou o melhor valor de RMSE. Os pontos de amostragem escolhidos neste trabalho são indicadores de uma região onde pode ocorrer uma presumível infração e a necessidade de se prestar atenção na realização da fiscalização.

Construiram-se também os mapas de riscos associados às presumíveis infrações pesqueiras no comando de zona do Sul com os dados simulados, verificou-se que estes mapas confirmam as zonas que apresentam riscos elevados de se cometer presumíveis infrações pesqueira em estudo nesse trabalho.

A aplicação destes delineamentos pressupõe que tenhamos esses delineamentos operacionalizados na prática e essa operacionalização depende de alguns fatores, tais como, o tempo de navegação disponível para a fiscalização que depende do combustível e dos mantimentos; disponibilidade das lanchas e navios para esse tipo de atividades; e as características da zona que se vai fiscalizar.

Os delineamentos de amostragem propostos poderão auxiliar no desenho das possíveis rotas de fiscalização, utilizando os critérios apresentados em 4.2.3. Obtemos assim um subconjunto de pontos, tendo em consideração o ponto de partida de fiscalização e a região de atuação (subconjunto do comando da zona sul neste caso) prevista para esse dia/dias nessa rota. As rotas podem ser atribuídas a uma unidade de pequeno porte (Lanchas) ou uma unidade de grande porte (Corvetas). Depois de se definir a rota de fiscalização, atribui-se a rota mais adequada a uma unidade naval disponível para fiscalizar esta zona otimizando o escalonamento que aumenta a probabilidade de encontrar um número elevado de infratores com menor esforço de recursos.

5

Conclusão e trabalhos futuros

O presente trabalho centrou-se na modelação geoestatística baseada em técnicas de krigagem, modelação de dados geoestatísticos com resposta binária usando a combinação das abordagens INLA-SPDE e no delineamento de amostragem de dados geoestatísticos binários. Aplicou-se a técnica de krigagem ordinária para estimar a quantidade do pescado de lulas, inspecionado em ações de fiscalização da Marinha Portuguesa no ano 2015 na região costeira do Algarve. A metodologia tradicional da geoestatística permitiu o desenvolvimento de um modelo de semivariograma que capturou a correlação espacial, que por sua vez permitiu fazer a krigagem e estimar a quantidade do pescado em pontos de locais onde as quantidades não eram conhecidas. O modelo foi desenvolvido, usando a krigagem ordinária, e através desta técnica gerou-se e apresentou-se o mapa de variâncias. A abordagem da geoestatística com base na krigagem apresenta claras vantagens práticas, ao ser não paramétrica, não necessitando de pressupostos distribucionais sobre o fenómeno espacial. A krigagem ordinária produziu uma superfície predita que capturou qualitativamente a tendência espacial aparente nos dados e foi quase idêntica às previsões obtidas usando o modelo de superfície de tendência linear. Verificou-se através dos mapas que o efeito espacial é alto na zona de Algarve situada entre 550 quilómetros a 592 quilómetros e valores relativamente altos ao largo do quilómetro 620.

Adicionalmente, aplicou-se os modelos espaciais Bayesianos para analisar dados geoestatísticos binários a um conjunto de dados reais de fiscalização marítima da costa Portuguesa para modelar a distribuição espacial das infrações pesqueiras e a previsão do risco associado a estas infrações na costa portuguesa usando a abordagem INLA-SPDE nos anos 2014 e 2015. Produziu-se mapas de médias e erro padrão do efeito espacial subjacente e mapas de riscos associados as presumíveis infrações. Os mapas de média a posteriori do efeito aleatório do campo para os dois anos (2014 e 2015) descreveram a distribuição espacial das infrações em toda a costa Portuguesa, mostrando os locais com maior e com menor efeito aleatório. Os mapas de desvio padrão estimaram a incerteza de previsão para cada ano em estudo. Os mapas de riscos, apresentaram uma distribuição do risco associado as presumíveis infrações de pesca. E verificou-se que o risco associado a essas presumíveis infrações nos dois anos é maior na zona de Lisboa e algumas regiões

da zona Norte de Portugal, mas em 2014 a região de Algarve apresentou um risco elevado. Desta forma, as probabilidades estimadas podem contribuir para definir futuras rotas das ações de fiscalização. As análises foram feitas com recurso ao pacote geoR do *software* R e o pacote R-INLA.

Fez-se também uma análise do comando de zona sul, concretamente na região costeira de Algarve em cinco anos (de 2013 a 2017). Da aplicação feita aos dados de fiscalização marítima, o modelo conseguiu estimar muito bem os dados, mapeando as regiões onde ocorrem as presumíveis infrações obtidas nas ações de fiscalização. Dois modelos foram ajustados, um espacial e outro espaço-temporal [99]. Das análises realizadas pelo critério de comparação DIC, o modelo espacial foi o que melhor se ajusta aos dados de fiscalização marítima ao largo da costa sul de Portugal. Foram construídos mapas de médias onde não foram observadas diferenças significativas durante os 5 anos.

Os mapas de risco apresentaram uma distribuição do risco associado as presumíveis infrações de pesca na região do Algarve. Constatou-se que o risco associado a estas presumíveis infrações nos cinco anos é maior na zona de Sagres e Tavira. Os mapas da média a posteriori e de erro padrão do campo aleatório foram apresentados. Os mapas de efeito aleatório de campo médio a posteriori para os cinco anos (de 2013 a 2017) descreveram a distribuição espacial das infrações na região costeira do Algarve, mostrando os locais com os maiores e os com menores efeitos aleatórios. Os mapas de erro padrão estimaram a incerteza de previsão para os cinco anos em estudo [43, 58] .

Em seguida foram construídos mapas de valores ajustados por períodos do dia que estimaram a probabilidade de encontrar as presumíveis infrações de pesca, onde foram na ordem de 50% no 2º e 3º período do dia e 70% no 1º e 4º período do dia no ano 2013. Construiu-se também mapas de riscos para os quatro períodos do dia para os anos 2014 a 2017 e constatou-se que o risco associado a estas presumíveis infrações é maior no 1º e no 4º período do dia (correspondente ao período noturno), na zona de Sagres e Tavira e, no 2º e 3º período, as mesmas regiões (Sagres e Tavira) apresentaram um risco relativamente alto. As probabilidades estimadas podem contribuir para definir futuras rotas para as ações de fiscalização.

Finalmente, construiram-se delineamentos de amostragem, baseado na maximização do risco estimado (MaxRSD) e na maximização da variabilidade associada ao risco estimado (MaxVRSD), que depois foram comparados ao delineamento aleatório simples. Destes, escolheu-se o melhor delineamento para diferentes tamanhos de amostra, e verificou-se que com valor de probabilidade $p = 0.2$, para tamanhos de amostra 50 e 100 pontos, o delineamento de amostragem MaxVRSD apresentou o melhor valor de RMSE, enquanto que para o tamanho de amostra de 200 pontos, o delineamento de amostragem MaxVRSD, apresentou o melhor valor de RMSE.

Construiram-se também os mapas de riscos associados as presumíveis infrações pesqueiras no comando de zona do Sul com os dados simulados, verificou-se que estes mapas confirmam as zonas que apresentam riscos elevados de se cometer presumíveis infrações pesqueira em estudo nesse trabalho.

A aplicação destes delineamentos pressupõe que tenhamos esses delineamentos operacionalizados na prática e essa operacionalização depende de alguns fatores, tais como, o tempo de navegação disponível para a fiscalização que depende do combustível e dos mantimentos; disponibilidade das lanchas e navios para esse tipo de atividades; e as características da zona que se vai fiscalizar.

Os delineamentos de amostragem propostos poderão auxiliar no desenho das possíveis rotas de fiscalização, utilizando os critérios apresentados em 4.2.3. Obtemos assim um subconjunto de pontos, tendo em consideração o ponto de partida de fiscalização e a região de atuação (subconjunto do comando da zona sul neste caso) prevista para esse dia/dias nessa rota. As rotas podem ser atribuídas a uma unidade de pequeno porte (Lanchas) ou uma unidade de grande porte (Corvetas). Depois de se definir a rota de fiscalização, atribui-se a rota mais adequada a uma unidade naval disponível para fiscalizar esta zona otimizando o escalonamento que aumenta a probabilidade de encontrar um número elevado de infratores com menor esforço de recursos.

Como sugestões para continuidade deste trabalho pretende-se fazer:

- Análise das presumíveis infrações associadas a pesca por comandos de zona continental (Norte, Centro) no âmbito das missões de natureza não militar;
- Mapas de média, erro padrão e mapas de riscos referentes a diferentes períodos horários do dia (diferenciação por quartos de dias, 0h-6h; 6h-12h; 12h-18h; 18h-24h) e suas comparações nos comandos de zona Norte e Centro;
- Análise do padrão da distribuição das presumíveis infrações no tempo, entre 2013 a 2021.
- Análise de outro grupo/tipo de infrações, articulando-se com a futura definição dos percursos por zonas e por grupo de infrações mas com o objetivo de estar disponível a nível nacional.
- Definir rotas de fiscalização tendo em conta os fatores tempo de navegação disponível que depende de combustível e mantimentos; disponibilidades das lanchas e navios para a fiscalização; e as características da zona a ser fiscalizadas.

Bibliografia

[1] A. Agresti. *Categorical data analysis*. Vol. 482. John Wiley & Sons, 2003 (ver pp. 42, 45).

[2] F. M. de Almeida Januário. "Modelos de regressao para dados de contagem e estimaçao da abundância de aves na cidade do Porto". Em: (2012) (ver p. 46).

[3] J. L. S. ANDRIOTTI. "Introdução à geoestatıstica". Em: (1988) (ver pp. 9, 14, 17).

[4] M. Armstrong. *Basic linear geostatistics*. Springer Science & Business Media, 1998 (ver p. 22).

[5] J. P. d. ASSIS, R. P. d. SOUZA e C. T. d. S. DIAS. "Glossário de Estatıstica". Em: *Mossoró/RN, EdUFERSA, 901f* (2019) (ver pp. 5, 9, 14, 23, 44, 48).

[6] H. Bakka et al. "Spatial modeling with R-INLA: A review". Em: *Wiley Interdisciplinary Reviews: Computational Statistics* 10.6 (2018), e1443 (ver p. 57).

[7] S. Banerjee e A. Gelfand. "Prediction, interpolation and regression for spatially misaligned data". Em: *Sankhyā: The Indian Journal of Statistics, Series A* (2002), pp. 227–245 (ver pp. 43, 53).

[8] S. Banerjee, B. P. Carlin e A. E. Gelfand. *Hierarchical modeling and analysis for spatial data*. CRC press, 2014 (ver p. 48).

[9] A. Bardosy. *Introduction to Geostatistics, Institute of Hydraulic Engineering University of Stuttgart, Technical note, 134 sf.* 2002 (ver p. 16).

[10] J. Barry, M. Crowder e P. Diggle. *Parametric estimation of the variogram*. 1997 (ver pp. 23, 24).

[11] L. Berliner. *Hierarchical Bayesian time-series models. InMaximum Entropy and Bayesian Methods (K. Hanson and R. Silver, Eds.)* 1996 (ver p. 50).

[12] M. Blangiardo e M. Cameletti. *Spatial and spatio-temporal Bayesian models with R-INLA*. John Wiley & Sons, 2015 (ver pp. 51–53, 55, 56, 59, 66, 72).

[13] N. E. Breslow e D. G. Clayton. "Approximate inference in generalized linear mixed models". Em: *Journal of the American statistical Association* 88.421 (1993), pp. 9–25 (ver p. 47).

[14] D. Brus e J. De Gruijter. "Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with discussion)". Em: *Geoderma* 80.1-2 (1997), pp. 1–44 (ver pp. 77–79, 87).

[15] M. L. Carvalho e I. Natário. "Análise de dados espaciais". Em: *Sociedade Portuguesa de Estatıstica* (2008) (ver pp. 14, 17, 22, 25).

[16] S. B. Caudill et al. "An advantage of the linear probability model over probit or logit". Em: *Oxford Bulletin of Economics and Statistics* 50.4 (1988), pp. 425–427 (ver p. 45).

[17] M. G. Chipeta, B. Rowlingson e P. J. Diggle. "geosample: an R Package for Geostatistical Sampling Designs". Em: *Under review* (2019) (ver p. 81).

[18] I. Clark. *Practical geostatistics*. Vol. 3. Applied Science Publishers London, 1979 (ver p. 21).

[19] W. G. Cochran. *Sampling techniques*. John Wiley & Sons, 1977 (ver pp. 2, 76–78).

[20] N. Cressie. "Fitting variogram models by weighted least squares". Em: *Journal of the international Association for mathematical Geology* 17.5 (1985), pp. 563–586 (ver p. 23).

[21] N. Cressie. *Statistics for spatial data*. John Wiley & Sons, 1991 (ver p. 6).

[22] N. Cressie. "The origins of kriging". Em: *Mathematical geology* 22.3 (1990), pp. 239–252 (ver p. 32).

[23] N. Cressie e C. K. Wikle. *Statistics for spatio-temporal data*. John Wiley & Sons, 2015 (ver pp. 48–50).

[24] A. Crujeira e D. Alves. *Modelos de Bases de dados*. 2019 (ver p. 48).

[25] J. De Gruijter e C. Ter Braak. "Model-free estimation from spatial samples: a reappraisal of classical sampling theory". Em: *Mathematical geology* 22.4 (1990), pp. 407–415 (ver pp. 78, 79).

[26] V. De Oliveira. "Models for geostatistical binary data: Properties and connections". Em: *The American Statistician* 74.1 (2020), pp. 72–79 (ver pp. 2, 41–43, 47).

[27] P. Diggle e P. Ribeiro. "Model based geostatistics, 14 SINAPE". Em: (2000) (ver pp. 22, 25–27).

[28] P. Diggle e S. Lophaven. "Bayesian geostatistical design". Em: *Scandinavian Journal of Statistics* 33.1 (2006), pp. 53–64 (ver p. 79).

[29] P. J. Diggle e E. Giorgi. "Model-based geostatistics for prevalence mapping in low-resource settings". Em: *Journal of the American Statistical Association* 111.515 (2016), pp. 1096–1120 (ver p. 52).

[30] P. J. Diggle, J. A. Tawn e R. A. Moyeed. "Model-based geostatistics". Em: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 47.3 (1998), pp. 299–350 (ver pp. 2, 43, 47).

[31] P. Diggle e P. Ribeiro. *Model-based Geostatistics*. 2007 (ver pp. 2, 5, 6, 24, 25, 28, 33, 38, 43, 47, 50, 51).

[32] E. Giorgi e P. J. Diggle. "PrevMap: an R package for prevalence mapping". Em: *Journal of Statistical Software* 78 (2017), pp. 1–29 (ver p. 53).

[33] H. Goldstein, W. Browne e J. Rasbash. "Multilevel modelling of medical data". Em: *Statistics in medicine* 21.21 (2002), pp. 3291–3315 (ver p. 50).

[34] C. A. Gotway e W. W. Stroup. "A generalized linear model approach to spatial data analysis and prediction". Em: *Journal of Agricultural, Biological, and Environmental Statistics* (1997), pp. 157–178 (ver p. 43).

[35] G. Guo e H. Zhao. "Multilevel modeling for binary data". Em: *Annual review of sociology* 26.1 (2000), pp. 441–462 (ver p. 48).

[36] Z. Han e V. De Oliveira. "On the correlation structure of Gaussian copula models for geostatistical count data". Em: *Australian & New Zealand Journal of Statistics* 58.1 (2016), pp. 47–69 (ver p. 43).

[37] J. J. Heckman. *Linear probability models of the demand for attributes with an empirical application to estimating the preferences of legislators*. 1996 (ver p. 45).

[38] D. W. Hosmer Jr, S. Lemeshow e R. X. Sturdivant. *Applied logistic regression*. Vol. 398. John Wiley & Sons, 2013 (ver pp. 2, 41).

[39] R. Howarth. "Journel and (Ch. J.) Huijbregts. Mining Geostatistics. London & New York (Academic Press), 1978. x+ 600 pp., 267 figs. Price£ 32· 00." Em: *Mineralogical Magazine* 43.328 (1979), pp. 563–564 (ver pp. 2, 6).

[40] E. Isaaks e R. Srivastava. *An Introduction to Applied Geostatistics, New York: Oxford Univ*. 1989 (ver pp. 27, 32, 42).

[41] P. K. Kitanidis. *Introduction to geostatistics: applications in hydrogeology*. Cambridge university press, 1997 (ver p. 10).

[42] D. G. Kleinbaum et al. *Logistic regression*. Springer, 2002 (ver p. 45).

[43] E. Krainski et al. *Advanced spatial modeling with stochastic partial differential equations using R and INLA*. Chapman e Hall/CRC, 2018 (ver pp. 51, 55–57, 59, 66, 75, 96).

[44] P.-S. Lin e M. K. Clayton. "Analysis of binary spatial data by quasi-likelihood estimating equations". Em: *The Annals of Statistics* 33.2 (2005), pp. 542–555 (ver pp. 43, 77, 80, 88, 90).

[45] F. Lindgren, H. Rue e J. Lindström. "An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach". Em: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73.4 (2011), pp. 423–498 (ver pp. 51, 52, 55–57).

[46] Y. Liu, Y. Chen e J. Cheng. "A comparative study of optimization methods and conventional methods for sampling design in fishery-independent surveys". Em: *ICES Journal of Marine Science* 66.9 (2009), pp. 1873–1882 (ver pp. 76, 77, 79, 80, 90).

[47] S. L. Lohr. *Sampling: design and analysis*. Chapman e Hall/CRC, 2021 (ver pp. 78, 89).

[48] L. Lv, X. Song e W. Sun. "Modify Leave-One-Out Cross Validation by Moving Validation Samples around Random Normal Distributions: Move-One-Away Cross Validation". Em: *Applied Sciences* 10.7 (2020), p. 2448 (ver p. 27).

[49] L. Madsen. "Maximum likelihood estimation of regression parameters with spatially dependent discrete data". Em: *Journal of agricultural, biological, and environmental statistics* 14.4 (2009), pp. 375–391 (ver p. 43).

[50] W. M. Mason, G. Y. Wong e B. Entwisle. "Contextual analysis through the multilevel linear model". Em: *Sociological methodology* 14 (1983), pp. 72–103 (ver p. 48).

[51] G. Matheron. "Random functions and their application in geology". Em: *Geostatistics*. Springer, 1970, pp. 79–87 (ver p. 6).

[52] G. Matheron. "La teorıa de las variables regionalizadas y sus aplicaciones". Em: *Los Cuadernos del Centro de Morfologıa Matemática de Fontainebleau. Fascıculo* 5 (1970), p. 125 (ver pp. 2, 5, 8).

[53] M. Mathur. "Spatial autocorrelation analysis in plant population: An overview". Em: *Journal of Applied and Natural Science* 7.1 (2015), pp. 501–513 (ver p. 14).

[54] A. McBratney, R. Webster e T. Burgess. "The design of optimal sampling schemes for local estimation and mapping of of regionalized variables—I: Theory and method". Em: *Computers & Geosciences* 7.4 (1981), pp. 331–334 (ver p. 79).

[55] C. E. McCulloch e S. R. Searle. *Generalized, linear, and mixed models*. John Wiley & Sons, 2004 (ver pp. 2, 41, 46).

[56] A. M. V. Monteiro et al. "Análise espacial de dados geográficos". Em: *Brasılia: Embrapa* (2004) (ver p. 13).

[57] J.-M. Montero, G. Fernández-Avilés e J. Mateu. *Spatial and spatio-temporal geostatistical modeling and kriging*. Vol. 998. John Wiley & Sons, 2015 (ver pp. 10, 29).

[58] P. Moraga. *Geospatial health data: Modeling and visualization with R-INLA and shiny*. Chapman e Hall/CRC, 2019 (ver pp. 43, 53, 75, 96).

[59] I. Natário. "Métodos Computacionais: INLA, Integrated Nested Laplace Approximation". Em: *Boletim da Sociedade Portuguesa de Estatıstica Outono de* 2013 (2013), pp. 52–56 (ver p. 54).

[60] I. Natário. *Probabilidades e Estatística: Notas produzidas no âmbito da disciplina de probabilidade e Estatística para o curso de engenharia*. Lisboa., 2012 (ver p. 49).

[61] J. A. Nelder e R. W. Wedderburn. "Generalized linear models". Em: *Journal of the Royal Statistical Society: Series A (General)* 135.3 (1972), pp. 370–384 (ver pp. 41, 44).

[62] A. A. Nielsen. "Kriging". Em: (2004) (ver p. 32).

[63] R. A. Olea. "Normalization". Em: *Geostatistics for Engineers and Earth Scientists*. Springer, 1999, pp. 31–38 (ver pp. 6, 19–21, 32).

[64] H. D. Patterson e R. Thompson. "Recovery of inter-block information when block sizes are unequal". Em: *Biometrika* 58.3 (1971), pp. 545–554 (ver p. 26).

[65] E. J. Peberma. *Gstat user's manual, Dep. of Physical Geography*. Utrecht University, Nertherlands, 2014 (ver p. 28).

[66] E. J. Pebesma e C. G. Wesseling. "Gstat: a program for geostatistical modelling, prediction and simulation". Em: *Computers & Geosciences* 24.1 (1998), pp. 17–31 (ver p. 28).

[67] P. Petitgas. "Geostatistics in fisheries survey design and stock assessment: models, variances and applications". Em: *Fish and Fisheries* 2.3 (2001), pp. 231–249 (ver pp. 77, 87).

[68] P. Petitgas et al. "Handbook of geo-statistics in R for fisheries and marine ecology." Em: (2017) (ver pp. 12, 13).

[69] M. J. Pyrcz e C. V. Deutsch. *Geostatistical reservoir modeling*. Oxford university press, 2014 (ver pp. 18, 19).

[70] J. Rivoirard et al. *Geostatistics for estimating fish abundance*. John Wiley & Sons, 2008 (ver pp. 2, 76).

[71] A. D. Rocha et al. "Role of sampling design when predicting spatially dependent ecological data with remote sensing". Em: *IEEE transactions on geoscience and remote sensing* 59.1 (2020), pp. 663–674 (ver p. 77).

[72] Ó. Rodrıguez de Rivera, A. López-Quılez e M. Blangiardo. "Assessing the spatial and spatio-temporal distribution of forest species via Bayesian hierarchical modeling". Em: *Forests* 9.9 (2018), p. 573 (ver pp. 43, 53, 72).

[73] H. Rue, S. Martino e N. Chopin. "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations". Em: *Journal of the royal statistical society: Series b (statistical methodology)* 71.2 (2009), pp. 319–392 (ver pp. 52, 53, 74).

[74] D. Russo. "Design of an optimal sampling network for estimating the variogram". Em: *Soil Science Society of America Journal* 48.4 (1984), pp. 708–716 (ver pp. 77, 79, 80).

[75] C.-E. Särndal et al. "Design-based and model-based inference in survey sampling [with discussion and reply]". Em: *Scandinavian Journal of Statistics* (1978), pp. 27–52 (ver p. 79).

[76] G. C. Simbahan e A. Dobermann. "Sampling optimization based on secondary information and its utilization in soil carbon mapping". Em: *Geoderma* 133.3-4 (2006), pp. 345–362 (ver pp. 77, 80, 83, 89, 90).

[77] P. Simões et al. "A spatial econometric analysis of the calls to the portuguese national health line". Em: *Econometrics* 5.2 (2017), p. 24 (ver pp. 13, 54).

[78] A. Sinclair. *Geostatistical ore reserve estimation| M. David, 1977. Elsevier, Amsterdam, 364 pp., Dfl. 110.00*, 44.95. 1978 (ver p. 2).

[79] A. Soares. *Geoestatıstica para as Ciências da Terra e do Ambiente, 542 IST-Press*. 2014 (ver pp. 8, 9).

[80] A. Soares. *Geostatıstica para as Ciências da Terra e do Ambiente*. 2000 (ver p. 42).

[81] D. J. Spiegelhalter et al. "Bayesian measures of model complexity and fit". Em: *Journal of the royal statistical society: Series b (statistical methodology)* 64.4 (2002), pp. 583–639 (ver p. 66).

[82] A. Stein e C. Ettema. "An overview of spatial sampling procedures and experimental design of spatial studies for ecosystem comparisons". Em: *Agriculture, Ecosystems & Environment* 94.1 (2003), pp. 31–47 (ver pp. 2, 76).

[83] L. Steinbuch, D. J. Brus e G. B. Heuvelink. "Mapping depth to Pleistocene sand with Bayesian generalized linear geostatistical models". Em: *European Journal of Soil Science* (2021) (ver pp. 43, 53).

[84] W. W. Stroup. *Generalized linear mixed models: modern concepts, methods and applications*. CRC press, 2012 (ver p. 46).

[85] J. R. STURARO. "Apostila de geoestatistica básica". Em: *Rio Claro, UNESP, IGCE, 34p* (2015) (ver pp. 17, 21).

[86] W. R. Tobler. *Spectral analysis of spatial series*. Library Photographic Service, U. of California, 1970 (ver p. 13).

[87] P. Tziachris et al. "Spatial modelling and prediction assessment of soil iron using kriging interpolation with pH as auxiliary information". Em: *ISPRS International Journal of Geo-Information* 6.9 (2017), p. 283 (ver p. 28).

[88] F. Valente e M. Mesquita. "Introdução à Geoestatistica". Em: *Instituto Superior de Agronomia, Universidade Técnica de Lisboa, Portugal* (2013) (ver p. 17).

[89] J. W. Van Groenigen, W. Siderius e A. Stein. "Constrained optimisation of soil sampling for minimisation of the kriging variance". Em: *Geoderma* 87.3-4 (1999), pp. 239–259 (ver pp. 77, 79, 80).

[90] J. Vann e D. Guibal. "Beyond Ordinary Kriging–An overview of non-linear estimation". Em: *Proceedings of a one day symposium: Beyond Ordinary Kriging*. 1998 (ver p. 32).

[91] S. Venkatramanan, P. M. Viswanathan e S. Y. Chung. *GIS and geostatistical techniques for groundwater science*. Elsevier, 2019 (ver p. 29).

[92] X. Wang, Y. Yue e J. J. Faraway. *Bayesian regression modeling with INLA*. Chapman e Hall/CRC, 2018 (ver p. 53).

[93] A. Warrick e D. Myers. "Optimization of sampling locations for variogram calculations". Em: *Water Resources Research* 23.3 (1987), pp. 496–500 (ver pp. 80, 90).

[94] S. Watanabe e M. Opper. "Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory." Em: *Journal of machine learning research* 11.12 (2010) (ver p. 66).

[95] R. Webster e M. A. Oliver. *Geostatistics for environmental scientists*. John Wiley & Sons, 2007 (ver pp. 9, 20).

[96] C. K. Wikle, A. Zammit-Mangion e N. Cressie. *Spatio-temporal Statistics with R*. Chapman e Hall/CRC, 2019 (ver p. 13).

[97] J. K. Yamamoto e P. M. B. Landim. *Geoestatıstica: conceitos e aplicações*. Oficina de textos, 2015 (ver pp. 1, 6, 7, 29, 32).

[98] Y. Zhang. "Introduction to Geostatistics—Course Notes". Em: *Dept. of Geology & Geophysics, University of Wyoming* (2011) (ver p. 6).

[99] A. Zuur. "Beginner's Guide to Spatial, Temporal and Spatial-Temporal Ecological Data Analysis with R-Inla: Using Glm and Glmm Volume I". Em: *Hightland Statistics Ltd., Sl OCLC* 973745327 (2017) (ver pp. 61, 68, 75, 96).